

**Analysis of Search on Clinical Narrative  
within the EHR**

**Karthik Natarajan**

Submitted in partial fulfillment of the  
requirements for the degree  
of Doctor of Philosophy  
under the Executive Committee  
of the Graduate School of Arts and Sciences

**COLUMBIA UNIVERSITY**

2012

©2012

Karthik Natarajan

All Rights Reserved

# ABSTRACT

## Analysis of Search on Clinical Narrative within the EHR

**Karthik Natarajan**

Electronic Health Records (EHRs) are used increasingly in the hospital and outpatient settings, and patients are amassing digitized clinical information. On one hand, aggregating all the patient's clinical information can greatly assist health care workers in making sound decisions. On the other hand, it can result in information overload, making it difficult to browse for information within the health record. Considering the time constraints clinicians face, one way to reduce information overload is through a search utility. However, traditional, free-text search engines within the EHR can potentially miss documents that do not contain the query but that are relevant to the clinical user's search. This dissertation aims at addressing this gap by analyzing within-patient search of the EHR and examining various semantic search approaches on clinical narrative. Our work consists of three studies where clinical users' search needs are examined, traditional string-matching is analyzed, and semantic search approaches on clinical narrative are evaluated.

The first study applied a mixed method approach in order to provide a better understanding of clinical users' search needs within the EHR. It is comprised of a retrospective log analysis of search log files and a survey that was administered to clinical professionals within our institution. The log analysis attempts to categorize how users of a search system query for information, and the survey tries to understand users' search preferences. This study showed that clinical users were very interested in search functionality within the EHR and that various types of users utilize a search utility differently. Overall, most users searched for specific laboratory tests and diseases within the health record.

The last two studies rely on a gold standard, which was developed specifically for this dissertation. The gold standard contained a document collection, a set of queries, and for each document/query pair, a relevance judgment. This gold standard was used to evaluate and compare different search models on clinical narrative. The second study conducted was an error analysis of the traditional, vector-space model search approach. The study examined the false positives and false negatives of this approach and categorized the errors in order to identify gaps that semantic approaches may fill. The last study was a systematic evaluation of five different semantic search approaches. These search methods consisted of distributional semantic approaches and an ontology-based approach. The study identified that a mixed topic modeling and vector-space model approach was the best performing search algorithm on our gold standard.

All of these studies lay the foundation for us to gain a deeper understanding of information retrieval methods within the electronic health record. Ultimately, this will allow health care professionals to easily access pertinent patient information, which could result in better health care delivery.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Understanding the Health Record . . . . .	2
1.2	Retrieval within the Health Record . . . . .	3
1.3	Purpose of Research . . . . .	5
1.4	Research Aims and Associated Research Questions . . . . .	6
1.5	Dissertation Outline . . . . .	7
<b>2</b>	<b>Background</b>	<b>9</b>
2.1	Information Overload . . . . .	9
2.2	Improving Access to Information within the Health Record . . . . .	10
2.3	Information Seeking Models . . . . .	11
2.4	Information Retrieval Approaches . . . . .	12
2.4.1	Keyword Retrieval . . . . .	13
2.4.2	Semantic-based Retrieval . . . . .	14
<b>3</b>	<b>Understanding Users' Search Needs</b>	<b>21</b>
3.1	Log Analysis Study . . . . .	21
3.1.1	Related Work . . . . .	22
3.1.2	Methods . . . . .	24
3.1.3	Results . . . . .	26
3.1.4	Discussion . . . . .	30

3.2	Survey Study . . . . .	34
3.2.1	Methods . . . . .	35
3.2.2	Results . . . . .	35
3.2.3	Discussion . . . . .	36
<b>4</b>	<b>Creation of a Gold Standard for Within-Patient Search</b>	<b>39</b>
4.1	Corpus Selection . . . . .	40
4.2	Query Development . . . . .	41
4.3	Relevance Tagging . . . . .	42
4.4	Results and Discussion . . . . .	43
<b>5</b>	<b>Analysis of VSM on Clinical Narrative</b>	<b>51</b>
5.1	Methods . . . . .	52
5.1.1	Data Pre-Processing . . . . .	52
5.1.2	Search Evaluation . . . . .	54
5.1.3	Error Analysis . . . . .	54
5.2	Results . . . . .	54
5.2.1	Search Results . . . . .	54
5.2.2	Error Analysis Results . . . . .	55
5.3	Discussion . . . . .	59
5.4	Limitations . . . . .	60
<b>6</b>	<b>Evaluation of Semantic Search Approaches on Clinical Narrative</b>	<b>62</b>
6.1	Methods . . . . .	63
6.1.1	Experimental Setup . . . . .	63
6.2	Results . . . . .	68
6.2.1	Overall Results . . . . .	68
6.2.2	Breakdown of Results . . . . .	73
6.3	Discussion . . . . .	80

6.3.1	EHR Implication . . . . .	82
6.3.2	Limitations . . . . .	83
<b>7</b>	<b>Conclusion and Future Work</b>	<b>84</b>
7.1	A Gold Standard Dataset for a Within-Patient EHR Search Evaluation . . . . .	85
7.2	A Within-Patient Deployed Search Engine and its Use . . . . .	86
7.3	Semantic Approach to Within-Patient EHR Search . . . . .	87
7.4	Limitations and Lessons Learned . . . . .	88
7.5	Future Work . . . . .	89
7.5.1	Gold Standard . . . . .	89
7.5.2	Search within the EHR . . . . .	89
7.5.3	Scenario of Use for a Semantic Search Sytem . . . . .	91
	<b>Bibliography</b>	<b>91</b>
<b>A</b>	<b>Appendix</b>	<b>105</b>
A.1	User Survey . . . . .	105
A.2	Annotation Tool . . . . .	116

# List of Figures

2.1	Search process model proposed by Marchionini, 1995 . . . . .	12
2.2	Vector space model example of a query being compared to different documents. (Modified from Dr. Mendonca’s lecture slide) . . . . .	15
2.3	Plate notation of LDA and a description of each variable (Blei, 2003). . . . .	18
2.4	Illustration of how a topic cluster for “stroke” might be created using LDA on clinical documents. . . . .	19
3.1	Results for query “diabetes” on a test patient. . . . .	24
3.2	The various user types of the search utility. . . . .	27
3.3	Number of Unique Queries Per Month . . . . .	28
3.4	The breakdown of survey participants based on their clinical role. . . . .	36
3.5	The average response to the types of information accessed within the health record based on clinical role. . . . .	37
4.1	Document types with most relevant information. . . . .	47
5.1	Creation of the three corpora using the original notes and relevance judgments. . . .	53
6.1	A hierarchical classification of semantic approaches and the specific semantic search methods evaluated in this study. . . . .	64
6.2	11-point precision graph for the top performing search approaches on the lupus cohort. The average precision is calculated for each recall point. . . . .	72



6.3 11-point precision graph for the top performing search approaches on the CHF cohort. The average precision is calculated for each recall point. . . . . 73

# List of Tables

3.1	General Usage Statistics . . . . .	27
3.2	Top 25 queries. . . . .	28
3.3	Percentage of query types based on all unique queries. . . . .	29
3.4	Percentage of query type based on user type. . . . .	29
3.5	Top 5 semantic types of searches. . . . .	30
4.1	This table shows the semantic type of the information need and the number of relevant documents associated to the information need. . . . .	44
4.2	This table shows the semantic type of the information need and the number of relevant documents associated to the information need. . . . .	45
4.3	Top 10 document types in the CHF gold standard corpus. . . . .	46
4.4	Top 10 document types in the lupus gold standard corpus. . . . .	46
4.5	Top ten document types for CHF query types, and the number of times each is found to be relevant. . . . .	48
4.6	Top ten document types for lupus query types, and the number of times each is found to be relevant. . . . .	49
5.1	This table shows the mean average precision scores on the different CHF gold standard corpora. . . . .	55
5.2	This table shows the mean average precision scores on the different lupus gold standard corpora. . . . .	55

5.3	This table is a breakdown of each CHF information need. It shows the semantic type of the information need and the average precision for each corpus. The asterisks indicate cases where the Free-Text average precision is statistically significant over the other searches. . . . .	56
5.4	This table is a breakdown of each lupus information need. It shows the semantic type of the information need and the average precision for each corpus. The asterisks indicate cases where the Free-Text average precision is statistically significant over the other searches ( $p=0.03$ ). . . . .	57
6.1	Top 10 mean average precisions for the LSA, UMLS-QE, and LDA-based KL divergence experiments on the lupus cohort. . . . .	69
6.2	Top 10 mean average precisions for the LDA-based query likelihood and LDA-based query expansion experiments on the lupus cohort. . . . .	70
6.3	Top 10 mean average precisions for the LSA, UMLS-QE, and LDA-based KL divergence experiments on the CHF cohort. . . . .	70
6.4	Top 10 mean average precisions for the LDA-based query likelihood and LDA-based query expansion experiments on the CHF cohort. The asterik indicates a search configuration where its performance was ( $p=0.03$ ) significant over the vector space approach. The standard deviation for 150 topics with 150 terms was 0.161; whereas, the standard deviation for 300 topics with 30 terms was 0.256. . . . .	71
6.5	This table shows the semantic type of the CHF information needs. . . . .	74
6.6	This table shows the semantic type of the lupus information needs. . . . .	75
6.7	This table shows the average percision for the top performing search approaches for each of the CHF information needs. . . . .	76
6.8	This table shows the average percision for the top performing search approaches for each of the lupus information needs. . . . .	77

- 6.9 This table shows the top ten terms of the topic cluster determined to be most similar to each of the lupus information needs for 200 topics. The ‘Topic Label’ column is a manual label identified by a clinical expert after examining the top 20 terms. ‘UNK’ means that a topic label could not be determined from examining the top terms. . . 78
- 6.10 This table shows the top ten terms of the topic cluster determined to be most similar to each of the CHF information needs for 150 topics. The ‘Topic Label’ column is a manual label identified by a clinical expert after examining the top 20 terms. ‘UNK’ means that a topic label could not be determined from examining the top terms. . . 79

# Acknowledgments

There many people I must thank for their guidance and support during this dissertation. First, I must acknowledge my excellent advisor, Noémie Elhadad, who has been very encouraging of my work even during difficult times. I would also like to thank my committee members, Herb Chase, George Hripscak, Stephen Johnson, and Bill Hersh, for all their guidance and advice, especially Herb without whom the creation of the gold standard would not have been possible. Thanks also to the members of Noémie’s research lab, Sharon Gorman and Rimma Pivovarov, whom have been a pleasure to work with.

Thank you to the EzVac team – Melissa Stockwell, Oscar Pena, Ben Dasgupta, David Vawdrey, and Stewin Camargo – for all their support. Working with them these several years has been both educational and fulfilling.

I am indebted to the NLM for funding my research through training grant 5T15-LM007079-18. Thanks also to DBMI for all the support it has provided me through the years. A special thanks to my colleagues and friends who have made these years enjoyable – Daniel Stein, Delano McFarlane, Samuel Brody, Jesse Wrenn, and Adler Perotte.

Lastly, I must thank my family for all their love and support, in particular my father and mother who inspired me to pursue a doctorate and my wife who has had to put up with me during this endeavor.

*To my wife for all her love and support during this journey,  
and to my daughter who has trained me to function with very little sleep*

# Chapter 1

## Introduction

Electronic Health Records (EHRs) are used increasingly in the hospital and outpatient settings, and patients are amassing digitized clinical information [Blumenthal, 2009]. On one hand, aggregating all the patient's clinical information can greatly assist healthcare workers in making sound decisions. On the other hand, it can result in information overload, making it difficult to browse for information within the health record. In a recent qualitative study in Norway, where EHR adoption has reached 95% nationally, researchers observed general practitioners use of EHRs and reported that many of them found it difficult to find information within the system, thereby disincentivizing them from accessing the EHR [Christensen and Grimsmo, 2008]. This was especially true in lengthy patient records, like those of chronically ill patients. Ironically, it is these very patients who require the most care, and the information within these records is especially pertinent to clinical decision-making. This problem of information overload within the EHR is not unique to Norway, and as other countries catch up in EHR adoption, they too will potentially face the same issue.

In fact, this problem of navigating the health record and extracting pertinent clinical facts in the context of a particular complaint or problem is not a new one for experienced clinicians. When clinicians are reviewing a patient's documented medical history in a paper-based charting system, they will generally flip through the pages of the old chart, scanning the notes and study reports for pertinent clinical data. While over time clinicians have become very good at this manual data

review, it is obviously not the most efficient solution. Some low-tech enhancements have evolved in paper charts to optimize this process, such as breaking up the chart into labeled sections and using paper color to signify certain report types (e.g., yellow for urine studies, red for hematological results, etc.). Arguably, even the practice of using the SOAP progress note format evolved at least in part so that specific types of clinical data and narrative were put in specific sections, ensuring a consistent and therefore easily reviewed layout. Though these enhancements have improved navigability, it is still challenging to find information within large health records due to the physical limitations of paper.

With the switch to electronic records, researchers in the informatics community have continued investigating ways to improve access to information. There have been several novel solutions proposed, which range from system enhancements to improved user-interface designs [Tang *et al.*, 1994; Tange *et al.*, 1997; Tange *et al.*, 1998; Zeng *et al.*, 2002; Senathirajah, 2007]. Though these alternative approaches reduce the amount of information presented, they focus primarily on structured data, such as laboratory data, and ignore free-text notes, which contain vital information for clinical decision-making. Therefore there is still much work to be done to reduce information overload in the EHR.

## 1.1 Understanding the Health Record

Since the patient health record is a source for clinical decision-making, it is essential to understand how and why clinicians use it in order to make information more accessible for their review. Nygren and Henriksson conducted the most notable study in 1992 to address this question and inform computer interfaces for EHR systems [Nygren and Henriksson, 1992]. In their study they interviewed seven physicians from various specialties totaling 35 hours. The study showed that clinicians accessed the patient record “to gain an overview of a familiar or new patient, to search for specific details, and to prompt or explore hypotheses” [Nygren and Henriksson, 1992; Nygren *et al.*, 1992; Nygren *et al.*, 1998]. Information overload can stymie clinicians from completing all of these tasks,



especially as more and more information becomes available within the EHR [Blumenthal, 2009; Hripcsak *et al.*, 2011]. However, innovative technological solutions can be leveraged within the EHR to improve access to patient information. The first task from Nygren and Henriksson’s study lends itself well to a summarization tool where clinicians can efficiently familiarize themselves with and gauge the status of a new patient [Wilcox *et al.*, 2005; Van Vleck *et al.*, 2007; Elhadad and McKeown, 2001; McKeown *et al.*, 2001]. The latter two tasks can be achieved through a search utility that can allow clinicians to find useful information buried within clinical notes. For example, if a clinician is trying to remember the last time his patient complained about abdominal pain, the clinician could simply use an EHR search engine, instead of manually flipping through the patient’s record or relying on his memory.

## 1.2 Retrieval within the Health Record

It is hard to find a modern website or application that does not have some type of site indexing and search capabilities; yet, clinical information systems generally lack search functionality. One possible explanation for the lack of search functionality within EHRs is that the system designers made electronic clinical information systems function like the old paper systems, treating it in some ways like a “word-processed paper chart” for documentation in order to lower physician adoption barriers to EHRs [Sujansky, 1998]. From one viewpoint, it makes sense that EHR system designers would want to hold onto paper chart concepts that improve navigation, such as patient folders and chart sections. Furthermore, it is not unreasonable that some would want to defend and preserve these positive aspects of the “old” way of doing things in paper charts, which had spurred much debate [Tange, 1995; Hippisley-cox *et al.*, 2003]. Unfortunately, this way of thinking can lead to missed opportunities in terms of leveraging new digital technologies, such as search. In fact, for this reason there is sparse literature on the design of search tools to help users find clinical information within the EHR. From the few studies that do exist, their systems employ “naive” approaches that focus on string matches within clinical documents and structured data (i.e., lab data, ICD-9 codes, etc.) [Yount *et al.*, 1991; Gregg *et al.*, 2003a; Hanauer, 2006; Schulz *et al.*, 2008a; Seyfried *et al.*,

2009; Zalis and Harris, 2010]. Though the results of these studies showed that clinicians found EHR search functionality useful for both searching within and across patient records, these works did not formally evaluate the use and need of a search engine on clinical narratives within the EHR. Traditional string-matching search engines are ubiquitous and useful for finding information, but their usefulness is dependent on users submitting queries that represent their information need [Gregg *et al.*, 2003a; Hanauer, 2006]. Due to the general nature of the Internet, it is difficult to identify a user’s information need from one single query, which tend to only consist of 1-3 terms, but users adjust to this by spending time refining their queries in order to find the information they are seeking [Chisnell *et al.*, 1995; Jansen *et al.*, 1998; Spink *et al.*, 2001; Jansen and Spink, 2006; Jansen *et al.*, 2007; Scott-Wright *et al.*, 2006]. Like Internet users’ queries, EHR search queries are short; however, in the clinical setting, users do not have time to iteratively refine their queries, making it difficult for a system to retrieve all relevant documents to a user’s information need [Natarajan *et al.*, 2010]. To add to this difficulty, users of various levels of expertise may articulate the same need differently, resulting in different documents being retrieved [Cimino and Shortliffe, 2006; Chase *et al.*, 2009]. For example, a clinician searching for the ejection fraction measurement within a patients record could type “EF,” “ejection fraction,” or “echocardiogram” into a string-matching EHR search engine and get different results. This can be frustrating for a user who is trying to search the electronic record in order to gather information to make an informed diagnosis on a patient. A simple solution to this problem is through synonym expansion, where an ontology, such as SNOMED-CT, is used to add similar terms to a query, thus, resolving the above difference between “EF” and “ejection fraction.” Another form of synonym expansion is when an ontology’s “is-a” hierarchy is traversed. For example, when searching for “congestive heart failure,” the query can be expanded to include child terms (i.e., congestive rheumatic heart failure) or parent terms (i.e., heart disease). The main problem with this type of expansion is determining the correct granular view of the ontology to include [Xu and Croft, 1996; Bhogal *et al.*, 2007; Aronson and Rindfleisch, 1997]. For instance, if a coarse view is included, such as heart disease, the search engine might return many documents related to heart disease, but not relevant to congestive heart failure. Alternatively, if a fine-grained term, such as congestive rheumatic heart failure, is added

to the query, the search engine results may not improve because the query would be too specific. Incorporating these techniques within an EHR search engine can be a first step to improving clinical search results; however, they do not address the above example of typing “echocardiogram” when searching for ejection fraction. A way to assist in bridging this gap between information need and documents returned is through a semantic-based search that is specialized for the clinical domain. In this work, we will address this research gap by understanding users’ search needs and developing methods to improve search functionality within the EHR.

### 1.3 Purpose of Research

There are two types of evaluations for IR systems that attempt to examine this disconnect between information need and resulting documents. The first type is user-oriented (extrinsic) studies, such as examining usability, efficiency, cognitive load, and user satisfaction [Manning *et al.*, 2008; Hersh, 2009]. To our knowledge, there is only one study conducted of this nature on search within the EHR [Tawfik *et al.*, 2011]. In the study, participants were given clinical scenarios where they were asked to search for medication information using their regular EHR as well as a newly developed EHR search system. After completion of their task, researchers examined the study participants’ perceived accuracy (how confident they were with what they found) and measured the reduction in cognitive load. The study showed a reduction in cognitive load through less number of clicks and time spent searching when using the EHR search system. Additionally, participants admitted that finding information within their regular EHR was difficult and that they normally would proceed with a clinical decision despite not having all the relevant information. This corroborates the findings from the Norway study and stresses the need for search functionality within the EHR.

The second type of IR evaluation is system-oriented (intrinsic) studies, such as evaluating system architecture, retrieval methodology and performance [Manning *et al.*, 2008; Hersh, 2009]. A formal evaluation of this type in regards to searching clinical narratives within a patient record is a relatively unexplored area. The only evaluation known to us is that of the TREC Medical Records

Track, which focuses on cohort identification as opposed to retrieval of specific information within the patient record [Voorhees, 2011]. This dissertation implements and examines computational search methods and, thus, will focus predominately on the intrinsic evaluation of semantic search on clinical narrative within the EHR.

This research consists of three studies where clinical users' search needs are examined, traditional string-matching search is analyzed, and various semantic search approaches on clinical narrative are evaluated. All of these studies lay the foundation for us to gain a deeper understanding of information retrieval methods within the electronic health record. The following section will describe the aims and research questions associated to the studies conducted in this dissertation.

## 1.4 Research Aims and Associated Research Questions

In this dissertation, we tackle the following three research aims.

### *Aim 1*

Understand how and what users search for while using a search utility within an EHR. We conduct two complementary studies: a log analysis of search within the EHR and a survey examining users' search preferences.

### Research Questions

1. Are there specific patterns in the types of information being searched over all users (e.g., common data types)?
2. Are there user-specific patterns of information needs?
3. What are users' needs when searching for information within the EHR?

### *Aim 2*

Analyze the performance and investigate the shortcomings of the vector-space model search approach on clinical narrative. We create a gold standard that consists of clinical notes, queries and

relevance judgments. We, then, use this gold standard for our analysis.

### Research Questions

1. Can an error analysis of free-text, within-patient search identify categories of search errors?

### *Aim 3*

Implement and evaluate the use of semantic search approaches to retrieve relevant concepts in the context of searching clinical narratives.

### Research Questions

1. What existing terminologies/ontologies should be incorporated into a clinical search engine?
2. What distributional semantic method should be used to improve semantic search results?
3. What should be the features of the model (i.e., semantic concepts or free-text)?

## 1.5 Dissertation Outline

The next chapter presents relevant background literature for this dissertation from the domains of information retrieval and seeking, biomedical informatics, and computer science. Chapter 3 present two complimentary studies conducted in order to understand clinical users' search needs. The first study is a log analysis on our production search system, and the second is a user survey focused on clinical users' perception of searching within the EHR.

Chapter 4 discusses the creation of a gold standard that uses clinical narratives within the EHR. This gold standard is used to evaluate search methods for the studies conducted in Chapters 5 and 6. Chapter 5 analyzes traditional vector-space model searching on clinical narrative and discusses its drawbacks. Chapter 6 uses the analysis conducted in Chapter 5 to examine different semantic approaches to improving search results over the traditional vector-space model.

Finally, Chapter 7 discusses significant findings, contribution from this work, and their implications on incorporating search functionality within the EHR. The chapter ends with limitations and future

directions that this work can be expanded upon to study within-patient search in the EHR.

## Chapter 2

# Background

In this chapter, I discuss background knowledge on the need for search within the clinical environment, foundational concepts to information retrieval, and algorithms pertinent to this dissertation.

### 2.1 Information Overload

Information overload is a problem faced in many fields. It occurs when an individual receives too much information, which hinders his/her decision-making [Bawden and Robinson, 2008; Eppler and Mengis, 2004]. According to many business studies, information overload happens when the amount of information exceeds an individual's processing capacity for a given time-constrained task. When faced with this problem, individuals conduct limited search strategies and arbitrarily analyze information (ignoring potentially important information), which all lead to suboptimal decisions [Bawden and Robinson, 2008; Eppler and Mengis, 2004]. This concern of information overload leading to suboptimal decisions in the business world is also voiced in the medical domain. In a 1997 article, Dr. Weed argues that when clinicians are "faced with information overload, [they] fall back on 'clinical judgment,' that is a global, intuitive assessment of findings rather than organized investigation and explicit linkage of each finding in the patient to the relevant diagnostic or management options in the medical literature" [Weed, 1997]. This "clinical judgment," he states,

is proven to be “inferior to judgments based on thorough analysis of specific data.” The problem of information overload faced with managing medical literature can also be considered in regards to EHRs, as witnessed in other studies [Christensen and Grimsmo, 2008; Tawfik *et al.*, 2011]. An intelligent search engine could potentially aid in alleviating this problem and improve access to information.

## 2.2 Improving Access to Information within the Health Record

The medical community has been investigating how to address the issue of information overload within the health record since its origin. In order to improve access to information, they have attempted to orient the health record in multiple ways: time-oriented, source-oriented, and problem-oriented [Zeng *et al.*, 2002; Tange, 1996; Weed, 1968].

The health record originated as a journal for clinicians, so information about a patient was captured in a chronological order (time-oriented). However, as the health record began to be used for other purposes and multiple individuals were accessing it, the layout of the record was organized based on the different departments where the information originated (source-oriented). Finally, in the mid-sixties, Dr. Weed suggested that the record should be organized around a patient’s problems (problem-oriented). Each organizational style has its benefits of accessing information; however, once an orientation has been created for the paper record it becomes too time-consuming and difficult to re-orient it into another layout.

As the patient record moves to an electronic format, there have been many novel solutions proposed within the informatics field to improve access to information and thus reduce information overload within the EHR [Tang *et al.*, 1994; Tange *et al.*, 1997; Senathirajah and Bakken, 2007; Tange, 1996; Zeng and Cimino, 1999; Zeng *et al.*, 2002]. These studies focused on improving the EHR design either by the user-interface or by how information was organized. Zeng *et al.* worked on displaying patient information based on clinical concepts. She created a system, QCIS (Query Clinical Information System), that allowed users to access laboratory information through multiple



views [Zeng and Cimino, 1999; Zeng *et al.*, 2002]. In a lab experiment, she showed that QCIS allowed users to quickly find information on a patient and that it reduced users' information overload through the use of concept-oriented views of the patient's health record.

Though Zeng's study on concept-orient views showed improved access to structured information, it did not examine accessing information within unstructured clinical narratives, which is more difficult. A search engine, however, could easily improve access to information within unstructured notes, and help clinical users accomplish two of the tasks that the health record is often used for: searching for specific details and exploring hypotheses [Nygren *et al.*, 1998].

## 2.3 Information Seeking Models

In order to develop a good search system within the EHR, the theoretical information seeking process must be understood. There have been many studies in the information science domain that examined users' search behavior within an information retrieval (IR) system in order to improve the search experience. One early cognitive model was Belkin's Anomalous States of Knowledge (ASK) framework, which was developed to improve the design of IR systems [Belkin, 1980]. He stated that the fundamental problem with IR systems was that they were designed for best match, thus, assuming that the user knew exactly how to describe his/her information need or ASK. The difficulty of articulating one's ASK is what Belkin coins as "non-specifiability of need," which is the cause of the disconnect between a user's ASK and the search result returned. For example, a clinician-in-training (novice) might search for "creatinine of 1.6" when they are actually looking for kidney failure [Chase *et al.*, 2009]. For this reason, Belkin suggests that IR systems be designed to aid the user in articulating his/her ASK through an interactive search process.

More than a decade later Marchionini developed a very general, process-oriented IR search model for all electronic environments [Marchionini, 1995]. The model consists of eight states – recognize and accept an information problem, define and understand the problem, choose a search system, formulate a query, execute search, examine results, extract information, reflect/iterate/stop. Mar-

Marchionini’s model captures the iterative nature of seeking for information. It allows the user to loop-back to previous states and reformulate new search strategies. A diagram of the model is depicted in Figure 2.1.

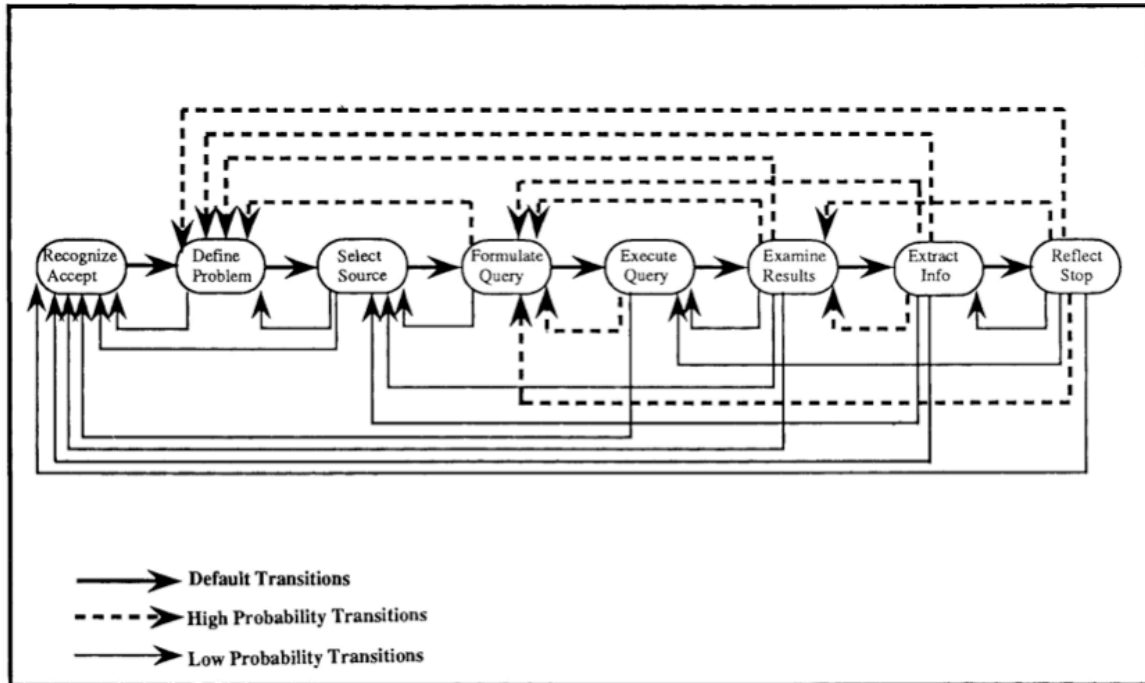


Figure 2.1: Search process model proposed by Marchionini, 1995

There have been other models that have been developed to study how individuals seek information, but Marchionini’s general, process-oriented model best captures the retrieval tasks within the electronic health record assuming there is only one source, the EHR [Hung *et al.*, 2008; Case, 2002]. For retrieval tasks within the EHR, users would select the patient record as a source and iteratively formulate their query in order to find the information they are seeking.

## 2.4 Information Retrieval Approaches

There are two types of IR systems - free-text systems and semantic-based systems. Free-text search systems find documents that contain query terms entered by the user. Semantic-based systems go beyond string matching of terms and retrieve documents that match the “meaning” of the query. In

both cases, the metrics used to evaluate an IR system, depending on the corpus size, are precision and recall. Precision is the percentage of retrieved relevant documents out of the total number of retrieved documents. Recall is the percentage of relevant documents out of the total number of documents in the corpus [Manning *et al.*, 2008; Hersh, 2009]. In order to use these metrics, a set of test queries are developed and relevant documents within the corpus are tagged accordingly. Depending on the intended use of the IR system, one metric may be more important than the other. For example, when casually searching the Web, a user is more concerned that all the documents returned are relevant (i.e., high precision) and not that all the relevant documents were retrieved. The opposite would be true for a researcher searching PubMed; he/she is more interested that the IR system retrieves all relevant documents (i.e., high recall). Another metric that has become standard in the IR community is the mean average precision (MAP). It is the *mean* of all the average precisions over all queries submitted to an IR system. In other words, for each information need, a set of precision values are calculated for each relevant document retrieved in the ranked result set. These values are then averaged for each information need, and the mean of all the average precision scores is equal to MAP. The MAP score is approximately the area under the average precision-recall curve for a set of information needs. This single value has shown to be a sound metric for evaluating the overall performance of a system [Manning *et al.*, 2008]. Besides these and other quantitative metrics, IR systems' are also qualitatively evaluated for their usefulness through user satisfaction studies [Manning *et al.*, 2008; Hersh, 2009].

### 2.4.1 Keyword Retrieval

Keyword retrieval systems are based on the vector-space model, which had its beginnings in the late 1950's [Salton *et al.*, 1975]. The vector-space model retrieves documents relevant to a query by representing documents and queries as vectors. First, a term-document frequency matrix is constructed, where each row represents a unique term in the collection of documents and each column represents an individual document in the collection. The cells within the matrix represent the frequency a term,  $t$ , occurs in a specific document,  $d$ , and is denoted as  $tf_{t,d}$  [Manning *et al.*,

2008].

The  $t$ - $d$  frequency matrix is biased by terms that occur frequently across many documents. In order to remove this bias and be able to discriminate between documents more effectively, a new weighting is calculated for each term within a document. In order to calculate this weight,  $\omega$ , the inverse document frequency,  $idf$ , is calculated for each term,  $t$ , and multiplied to the term frequency for each document,  $tf_{t,d}$ .

$$idf_t = \log \frac{N}{df_t},$$

where  $df_t$  is the number of documents that contain the term  $t$  and  $N$  is the total number of documents in the collection.

$$\omega_{t,d} = tf_{t,d} \times idf_t$$

These weight calculations are done also for a query, so documents and queries can be represented as a vector of these weights. By representing documents and queries as vectors in a vector space, the cosine similarity between the documents and the query can be calculated. Thus, the documents with a similarity score closest to zero are the ones most similar to the query. Figure 2.2 is a graphical example of a query,  $q$ , compared to the documents in a collection.

## 2.4.2 Semantic-based Retrieval

There are two levels to incorporate semantics into search. The first is through simple synonym expansion [Manning *et al.*, 2008; Bhogal *et al.*, 2007]. The second level is via semantic relations. This is a more sophisticated method that requires documents to be parsed by a natural language processor. In this case, relevant documents are retrieved based on terms it contains which are semantically associated with the query terms. There are two types of approaches to incorporate semantic relations into an IR system – human-curated knowledge representations (i.e., MeSH) and

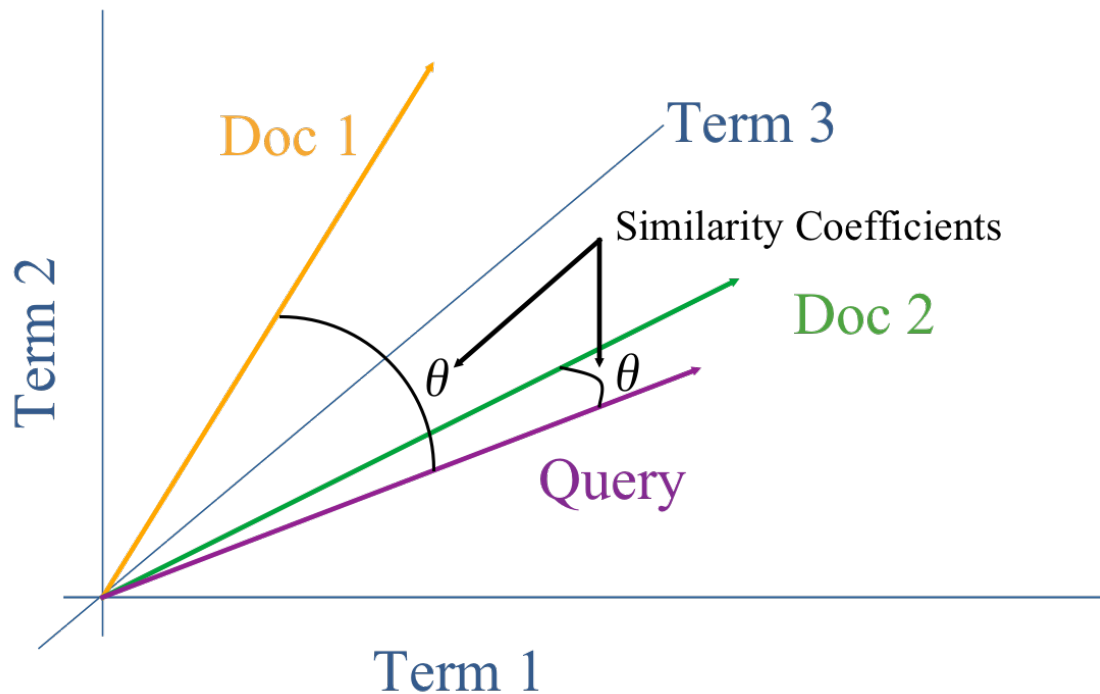


Figure 2.2: Vector space model example of a query being compared to different documents. (Modified from Dr. Mendonca’s lecture slide)

distributional semantics. Human-curated knowledge representations, such as ontologies, are useful in incorporating conceptual relations in order to improve search results. However, the maintenance of such knowledge structures is both time-consuming and expensive. There has been much work in the past few decades focused on distributional semantic approaches in order to extract semantics from text. These approaches examine large document collections and attempt to cluster similar terms within the corpus [Mendonça and Cimino, 2000; Cohen and Widdows, 2009].

Topic modeling has become a popular distributional semantic approach to categorizing the content of documents and has shown to improve IR results [Park and Ramamohanarao, 2009; Wei and Croft, ; Wei and Croft, 2006; Deerwester *et al.*, 1990; Hofmann, 1999] . The general idea behind topic modeling approaches is to find underlying themes covered in a collection of documents. These underlying themes will be referred to as *latent topics*, and the collection of documents will be referred to as *corpus*. Each latent topic consists of a collection of terms found within the corpus. By representing documents as a mixture of these topics, “noisy,” irrelevant terms are removed,

thus, reducing the search space.

### 2.4.2.1 Latent Semantic Analysis

The first successful attempt at finding latent topics was latent semantic analysis (LSA) [Deerwester *et al.*, 1990]. It was a geometric solution, which was used to illustrate how to improve information retrieval results [Deerwester *et al.*, 1990]. LSA achieved this by reducing a term-document matrix to reveal “latent semantics” in documents. For example, relevant documents that do not contain query terms, “human computer interaction,” would also be retrieved without the need of synonym expansion [Deerwester *et al.*, 1990]. LSA accomplishes this matrix reduction through a linear algebra method called singular value decomposition (SVD) [Deerwester *et al.*, 1990; Eldén, 2007; Strang, 1998]. SVD splits a matrix,  $X$ , into three components matrices –  $U S V$  – as seen below.

$$X = USV^T$$

$U$  and  $V$  are orthogonal matrices and  $S$  is a diagonal matrix containing the eigenvalues of  $X$  in descending order. SVD is used in many applications such as image compression and information retrieval. Once the matrix is decomposed, a low-rank approximation of the original matrix can be made. This is useful both for computation and for saving space. A reasonable rank  $K$  is selected to represent the original matrix,  $X$ , such that  $K < \text{Rank}(X)$ . By reducing the matrix by  $K$  topics, “noisy” terms are removed within a document and a more accurate representation of the corpus is created for better retrieval.

$$\begin{bmatrix} X \\ t \times d \end{bmatrix} = \begin{bmatrix} U \\ t \times k \end{bmatrix} \begin{bmatrix} S \\ k \times k \end{bmatrix} \begin{bmatrix} V^T \\ k \times d \end{bmatrix}$$

The conventional way of determining a suitable  $K$  is by examining matrix  $S$ 's Frobenius Norm (or Euclidean Norm),  $SF$ , which is the square root of the sum of the absolute squares of all the matrix elements. Essentially, the differences between  $SF$  and various Frobenius Norms of reduced

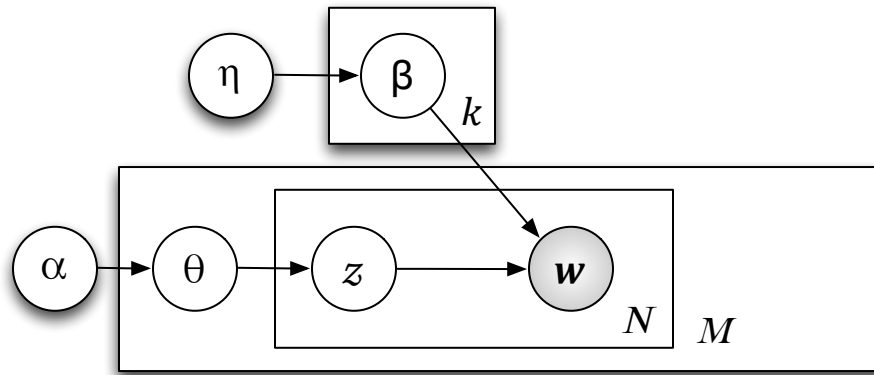
S are calculated, and a reasonable minimum is selected. This method is used in determining the appropriate rank-approximation for image compression. However, when dealing with text in LSA, there is no gap in the sequence of singular values, so finding the minimum Frobenius is not useful [Eldén, 2007]. In order to find an optimal  $k$  with LSA, there have been several approaches, but the most popular requires manual review of sharp drops in a scree plot. A scree plot is a plot of the eigenvalues of a matrix in descending order (the values in  $S$ ). The rank where there is the greatest drop is usually labeled as the optimal  $k$  [Bradford, 2008]. This solution can be useful, but requires manual review. There have been other automated approaches from calculating the entropy of singular values to testing the results of an information retrieval system based on LSA [Bradford, 2008; Zhu and Ghodsi, 2006].

The way retrieval is handled with LSA is similar to the vector space model, except it uses the reduced matrix,  $X$ , as the matrix of weights when calculating similarity between the query and document.

#### 2.4.2.2 Latent Dirichlet Allocation

Though LSA showed retrieval improvement, according to Hoffman, LSA lacked the “statistical foundation” that probabilistic approaches, such as probabilistic latent semantic analysis (pLSA), offered [Hofmann, 1999]. He also showed that pLSA outperformed LSA in information retrieval tasks. In recent years, latent Dirichlet allocation (LDA), a popular probabilistic topic modeling method, has attracted much attention in many research areas including the IR community [Wei and Croft, 2006; Zhai, 2008; Park and Ramamohanarao, 2009; Griffiths and Steyvers, 2004; Elhadad and Gabay, 2010; Blei *et al.*, 2003; Cohen and Widdows, 2009]. Research has shown that incorporating LDA can be promising for improving IR results. LDA is a generative model that improves on pLSA’s approach. It treats each document as a mixture of topics like pLSA, but the topic mixture is a Dirichlet distribution across all documents in the corpus as opposed to each individual document. This prevents LDA from overfitting to the corpus and allows it to infer topic distributions on new documents not found in the training corpus [Blei *et al.*, 2003]. This latter point is what makes

LDA especially useful for information retrieval because it allows for new documents to be easily added to an index for searching. Figure 2.3 is a graphical depiction, called plate notation, of how LDA models a document collection. Figure 2.4 illustrates how terms from documents might be clustered to create a topic that would be labelled as “stroke.”



#### Variable Definitions

$\alpha, \eta$  = Hyper-parameters

$k$  = Number of topics

$M$  = Number of documents in the corpus  $[1..j]$

$N$  = Number of words in the document  $[1..i]$

$\beta$  = Matrix of probability distribution of topics for each word in the vocabulary ( $k \times V$  matrix)

$\theta$  = Matrix of probability distribution of topics for each document ( $k \times M$  matrix)

$z$  = The latent topic of  $i^{th}$  word in the  $j^{th}$  document

$w$  = The observed word in document

Figure 2.3: Plate notation of LDA and a description of each variable (Blei, 2003).

LDA can be incorporated into retrieval in different ways. There is the language model approach and the traditional vector-space model approach using query expansion. A simple method to determine similarity between a query and document under a language model framework is to calculate the Kullback-Leibler (KL) divergence of a query and a document since both are represented as a distribution of topics under LDA [Zhai, 2008]. This similarity method, however, has not yielded useful results for information retrieval purposes [Wei and Croft, 2006; Zhai, 2008]. The accepted approach to incorporate a language model for retrieval is the query likelihood model [Zhai and Lafferty, 2001; Liu and Croft, 2004; Wei and Croft, 2006]. With the query likelihood model, a



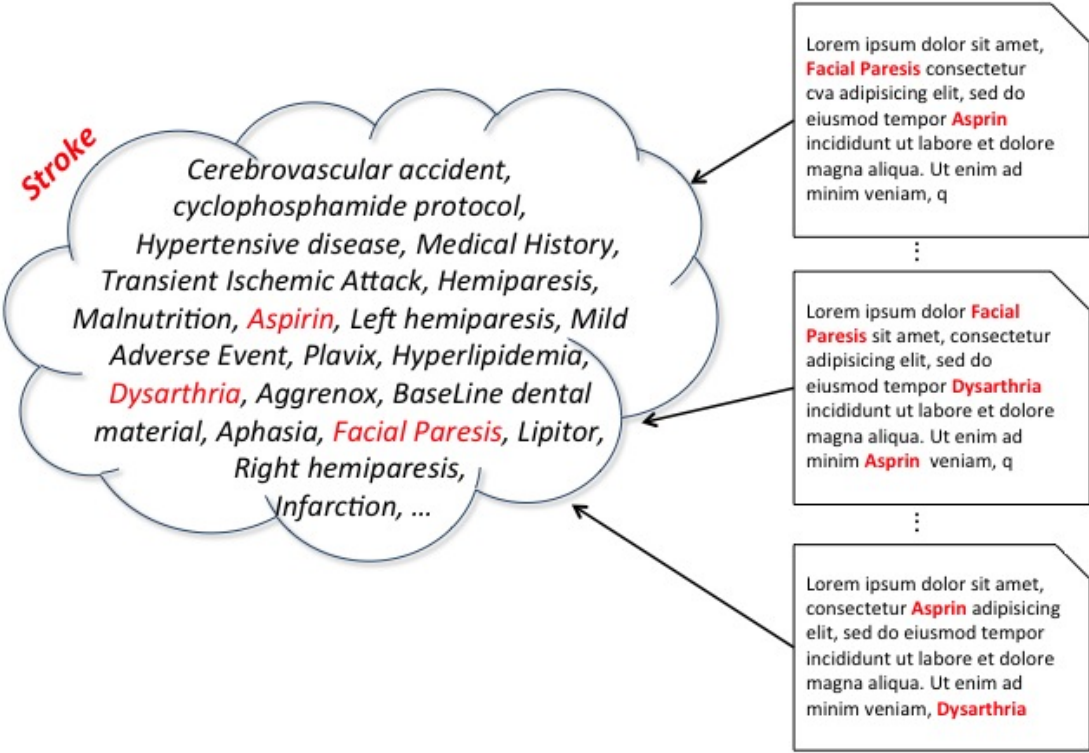


Figure 2.4: Illustration of how a topic cluster for “stroke” might be created using LDA on clinical documents.

document  $D$  is scored based on its probability of generating query  $Q$  after being trained on a corpus,

$$P(Q|D) = \prod_{q \in Q} P(q|D)$$

where  $q$  is a query term within  $Q$  and  $P(q|D)$  is the maximum likelihood estimate of document  $D$  containing term  $q$ . In order to properly incorporate topic models into a language model, smoothing techniques have been studied and applied to the language models [Wei and Croft, 2006; Zhai and Lafferty, 2001; Azzopardi *et al.*, ]. Using these technique, Xing and Croft demonstrated improvements in retrieval on TREC data using the query likelihood model by incorporating LDA through a linear combination [Wei and Croft, 2006],

$$P(q|D) = \lambda P_{ML}(q|D) + (1 - \lambda) P_{lda}(q|D)$$

where  $P_{ML}(q|D)$  is the maximum likelihood of term  $q$  within the document and smoothed based on the collection of documents.  $P_{lda}(q|D)$  is the probability of a term  $q$  appearing in document  $D$  under the LDA model.

As mentioned earlier, LDA can be used as a thesaurus for query expansion within the vector-space model approach. In this scenario, a topic distribution is inferred for a query. Since each topic is a cluster of terms, the topic most relevant to a query can be used to supply terms for query expansion [Park and Ramamohanarao, 2009; Yi, 2009]. Determining the number of terms to include, as well as their weighting, for any query expansion is an area of much research [Bhogal *et al.*, 2007], and thus, an in-depth analysis of query expansion techniques is outside the scope of this dissertation.

## Chapter 3

# Understanding Users' Search Needs

This chapter will cover two studies that were conducted to address Aim 1. The first study is a retrospective examination of search log files, which has been published [Natarajan *et al.*, 2010]. It attempts to categorize how users of a search system query for information. The second study is a survey that was administered to clinical professionals within our institution in order to understand their search preferences. The combined studies will provide a better understanding of clinical users' search needs within the EHR.

### 3.1 Log Analysis Study

Our institution has a Web-based clinical information system, WebCIS, that acts as a portal to all clinical narrative documents and laboratory test results within our clinical data repository [Hripcsak *et al.*, 1999]. It is used regularly during clinical workflow for accessing clinical information; however, it lacks search functionality. The absence of an EHR search feature and the relative dearth of literature on the subject inspired us to build and study a search utility. We designed and implemented a simple keyword search utility called CISearch, which is integrated within WebCIS.

The topic of search within the EHR has many unexplored research questions. In this retrospective

study, we attempt to answer one of the fundamental questions in order to guide future research: what are the characteristics of users' searches within the EHR? We hypothesize that general Web search classification schemas can be leveraged to categorize EHR-based queries and that these queries can be mapped further to medical semantic types derived from the Unified Medical Language System (UMLS).

### 3.1.1 Related Work

In order to improve search utilities and the search experience of any system, understanding users' search intent is essential. Although the medical informatics field has studied search and clinician information needs, the research has focused on accessing medical reference information, which is different from EHR-based search [Hersh, 2009; Osheroff *et al.*, 1991; Currie *et al.*, 2003; Smith, 1996; Allen *et al.*, 2003]. From a different perspective, investigators in the computer science and information science fields have examined search on a broad scale. Broder was the first to categorize and study why people searched the Web [Broder, 2002]. He determined three broad search categories: *navigational*, *informational*, and *transactional*. Navigational searches are searches that involve a user seeking a specific site (e.g., searching for the International Journal of Medical Informatics homepage). Informational searches are searches that involve a user seeking information on a topic (e.g., searching 'what is biomedical informatics'). Transactional searches are searches that involve a user seeking a site to perform another transaction (e.g., searching for a research paper in PubMed). Other search taxonomies have had essentially the same three high-level categories [Li *et al.*, 2005; Rose and Levinson, 2004]. Li and colleagues analyzed intranet queries in a more domain-specific setting than Broder. Their high-level classification followed Broder's scheme, and they expanded the analysis to include domain-specific sub-categories of search types. The categories were derived in an iterative process by manually examining the intranet queries. Li's intranet search study suggests that medical searches within EHRs, which are also domain specific, could be categorized into Broder's three search categories.

There are many ways to capture users' information needs in order to understand search intent.

Research methods, such as surveys, interviews, and focus groups, provide a deep understanding of the subjects' behaviors and needs. Another method, the analysis of transaction logs, provides an unobtrusive way to capture user behavior. Transaction logs are files that contain records of the interactions between a system and its users. The methodology of analyzing these transaction logs in order to investigate research questions is called transaction log analysis (TLA) [Jansen, 2006]. TLA has been employed in studies across many domains in order to understand users' behavior when interacting with a system [Chisnell *et al.*, 1995; Beitzel *et al.*, 2004; Chen and Cimino, 2003; Jansen, 2006; Joachims, 2002; Kelly and Teevan, 2003; Klink, 2004; Mat-Hassan and Levene, 2005; Murata and Saito, 2006; Scott-Wright *et al.*, 2006; Silverstein *et al.*, 1999; Fox *et al.*, 2005; Teevan *et al.*, 2007]. These studies range from examining general usage to examining implicit features such as clickthrough data to improve search. TLA has been utilized previously at our institution to study clinician information needs within the clinical information system [Chen and Cimino, 2003]. The study found that laboratory and radiology reports were the most accessed. Our study followed the steps of TLA to understand clinician searches because of its unobtrusiveness in collecting data and its ability to examine the behavior of all search users.

#### 3.1.1.1 CISearch

CISearch is a general search utility, which searches free-text clinical reports within a patient's electronic health record. Unlike most search engines, which display search results based on relevancy, CISearch displays results in reverse chronological order. In its current version, CISearch indexes all free-text, clinical documents (e.g., radiology reports, discharge summaries, and nursing notes). It does not search structured, coded data that can be represented numerically, such as laboratory results (e.g., CHEM-7 test), because accessing such information within our EHR is relatively user-friendly and efficient. CISearch was integrated into WebCIS in July 2008. The search box was placed at the top of the main, left navigation area within WebCIS so that it was easy to access. In order to reduce the barrier of implementation, we customized a widely used, open-source search engine, Lucene, to index and search clinical notes within a particular patient [Apa, 2007]. Lucene

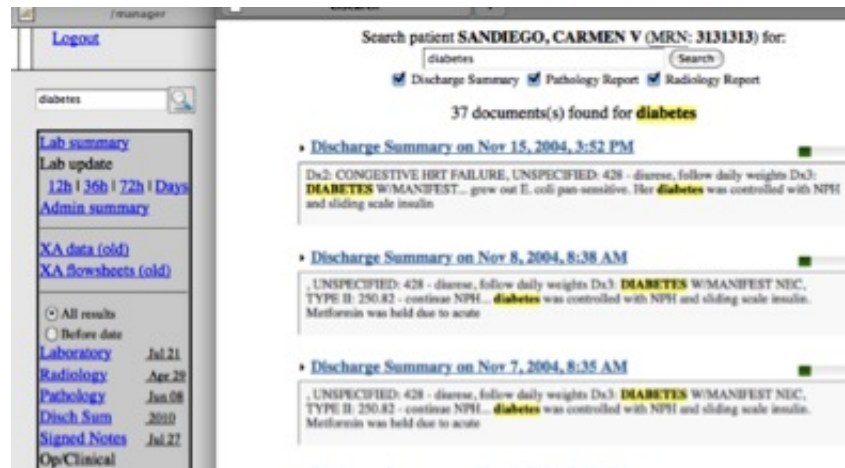


Figure 3.1: Results for query “diabetes” on a test patient.

is based on the vector-space model and has several built-in features. Features utilized in CISearch were in-memory indexing, advanced query grammar, stop-word removal, text snippets, and results highlighting. At the time this study was performed, CISearch indexed and searched discharge summaries, radiology reports, and pathology reports only. It was decided that an in-memory search was acceptable for the initial version of CISearch because it did not require the creation and maintenance of a database of indexed documents and because the initial search document space was small. Figure 3.1. shows the results for 'diabetes' on a fictitious test patient.

## 3.1.2 Methods

### 3.1.2.1 Data Collection

WebCIS log files were collected for 6 months (from July 14th to December 31, 2008). The files contained all CISearch transactions within WebCIS. There were two types of CISearch log entries: query and clickthrough. We define *query* as the entire string that a user enters and define *query term* as the individual strings separated by whitespace that comprise a query. The query entry contained a timestamp, the user identifier and its IP address, the patient health record number for the patient currently viewed, the document types that were selected to be searched, the search query, the number of documents retrieved from the search, the total number of documents in the

patient record, and the document retrieval time. The clickthrough entry is similar to the query view. It contained the document selected, the document's relevancy score, and the document's rank in the result set.

### 3.1.2.2 Pre-Processing

Once the data was collected, the log files were cleaned before analysis. First, the log files were filtered to remove entries of hospital information-technology employees and system developers. Then the log files were de-identified by replacing Medical Record Numbers (MRN) and user ids with unique numbers. Finally, the query and clickthrough log entries were extracted and inserted into respective database tables.

### 3.1.2.3 Analysis

The analysis of the queries was carried out using Broder's categories (navigational, transactional, and informational). Two investigators manually categorized all the unique queries and inter-annotator agreement was analyzed. For example, a query containing a patient MRN was labeled as a navigational search because it was most likely that the user was trying to switch patients rather than searching for the MRN within the current patient's health record. Queries that represented an action were labeled as a transactional search. For instance, the query 'add note' most likely referred to the user's intent to create a new note as opposed to searching for those words within the health record. All other queries were labeled as informational searches.

During the analysis it became apparent that informational searches were most frequently performed. Considering the large proportion of informational searches and our future goal of extracting pertinent information from the health record, we further categorized informational searches. Three physicians categorized a random sample of informational searches with semantic information. The reviewers were given overlapping data sets so that two clinicians categorized each query. In order to reduce the burden of categorizing the queries, an abbreviated list of UMLS semantic types was

provided to the clinicians. The abbreviated list was created by iteratively filtering and clustering UMLS concepts with similar meaning from a clinical perspective. For example, the UMLS semantic type 'Chemical Viewed Structurally' and its children were removed since, in a clinical context, it is more likely that clinicians refer to chemicals functionally (e.g., how a patient is reacting to an antibiotic) as opposed to structurally (e.g., the molecular structure of the chemical compound); likewise, the semantic types 'Body Part' and 'Body Location' were merged. The reviewers practiced categorizing on a standard set of 20 queries with an investigator before proceeding with their individual sets of 119 queries. The inter-annotator agreement between the three physicians was analyzed as well.

### 3.1.3 Results

#### 3.1.3.1 General Usage Results

There were a total of 436 unique users of CISearch in the first six months of its deployment within WebCIS. This represents roughly 5.3% of WebCIS users (the total number of WebCIS users was estimated to be the average number of active users within a month for the past year, which is approximately 8,200). Figure 3.2 shows the breakdown based on user types of the 436 unique users.

A total of 6,117 search log lines were analyzed. We removed highly repetitive queries from our data set. In particular, three users conducted approximately 2,200 searches with variations of the same query containing a specific drug and medical condition over several hundred patient records. All three individuals were researchers. These queries were considered outliers because of their high frequency and were excluded from the analysis. The users that only submitted these outlier queries were also removed from the analysis. Table 3.1 lists the general usage statistics of CISearch. A unique query was defined as a distinct string, ignoring case as well as leading and ending whitespaces. A view occurred when a user clicked on one of the documents returned by a query. Figure 3.3 describes the monthly usage of the search utility. Of the 980 unique queries,



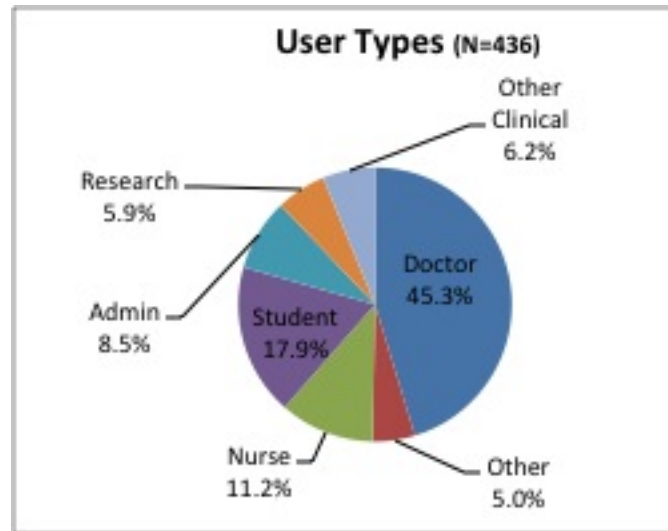


Figure 3.2: The various user types of the search utility.

Number of Queries After the Removal of Outlier Queries	2,207
Number of Unique Queries After the Removal of Outlier Queries	980
Average Number of Terms Per Query	1.2
Number of Clickthroughs (Total Number of Queries=4,427)	618 (13.9%)
Number of Unique Users (Approx # of active users =8,200)	436 (5.3%)

Table 3.1: General Usage Statistics

only four utilized the built-in features in the query language of Lucene, such as quotes around queries (i.e., “chest tube”). Additionally, out of the 980 queries, 148 were abbreviations (e.g., “chf” for congestive heart failure) and 78 were part of a word (e.g., “tach” for either tachycardia or tachyarrhythmia).

Table 3.2 shows the top 25 unique searches. The most frequent query, ‘class,’ was used predominately in conjunction with “nyha”, “III”, and “IV.” In this case, we suspect that users were searching for mentions of New York Heart Association Classifications within the health records.

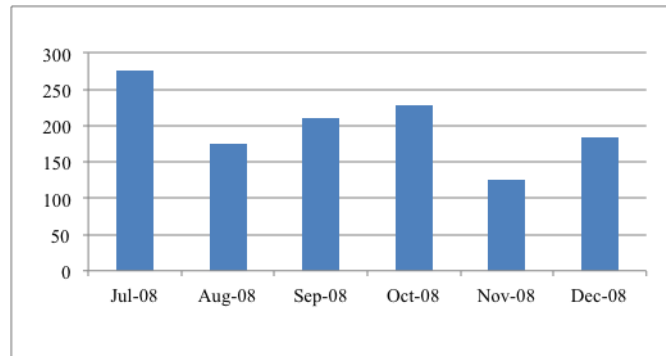


Figure 3.3: Number of Unique Queries Per Month

Query	Frequency	Percentage (N=2,207)
class	217	9.8%
nyha	99	4.5%
hodgkins	64	2.9%
iii	52	2.4%
iv	50	2.3%
nephrogenic	39	1.8%
hysterectomy	33	1.5%
cva	24	1.1%
ef	23	1.0%
hf	19	0.9%
fibronectin	19	0.9%
cmv	17	0.8%
chf	16	0.7%
embol	15	0.7%
sirolimus	13	0.6%
hiv	13	0.6%
pericardiocentesis	13	0.6%
renal	13	0.6%
subdural	12	0.5%
lvad	12	0.5%
dsum	12	0.5%
placenta	12	0.5%
accreta	12	0.5%
mri	11	0.5%

Table 3.2: Top 25 queries.

Query Type	Percentage (N=980)
Informational	85.1%
Navigational	14.5%
Transactional	0.4%

Table 3.3: Percentage of query types based on all unique queries.

	Informational	Transactional	Navigational
Research (N=684)	95.3%	0.0%	4.7%
Doctor (N=649)	91.8%	0.0%	8.2%
Student (N=331)	87.6%	0.0%	12.4%
Other Clinical (N=65)	72.3%	0.0%	28.7%
Nurse (N=102)	70.6%	0.0%	29.4%
Admin (N=63)	52.4%	3.2%	44.4%
Other (N=313)	94.9%	0.6%	4.5%

Table 3.4: Percentage of query type based on user type.

### 3.1.3.2 Analysis Results

#### *High-Level Classification of Search Queries*

980 unique queries were categorized as informational (e.g., “chf”), navigational (e.g., medical record number), or transactional (e.g., “add drug”). The inter-annotator agreement was a Kappa of 0.93 [Cohen, 1960]. A large majority of the queries were categorized as informational. Table 3.3 shows the breakdown of the queries and Table 3.4 shows what types of queries each user type performed.

#### *Semantic Classification of Informational Search Queries*

357 unique queries were categorized, each by two clinicians. There were a total of three reviewers. Each reviewer categorized 238 queries of which 119 queries overlapped with one other reviewer. The inter-annotator agreement was a Kappa of 0.56 (the reviewers agreed on 161 queries and disagreed on 74). Among the disagreements, 52 were due to ambiguities with the semantic type “Laboratory or Test Results” (e.g., immunoglobulin can be classified as a laboratory test or a biologically active substance), and 16 were due to ambiguities with “Disease or Syndrome” (e.g., atrial fibrillation can be classified as a finding or a disease). 122 of the total 357 queries were left uncategorized by either one or both reviewers due to uncertainty in the query. Some of these uncategorized queries were

<b>Semantic Type</b>	<b>Percentage (N = 161)</b>
Laboratory or Test Result	29.2%
Disease or Syndrome	21.7%
Body Part, Organ, or Organ Component	8.1%
Pharmacologic Substance	7.5%
Diagnostic Procedure	6.2%

Table 3.5: Top 5 semantic types of searches.

ambiguous abbreviations, part of a word, first names, or simply not medical terms.

The informational searches that the reviewers agreed upon showed that a majority of searches were about “Laboratory or Test Results” and “Disease or Syndromes.” Table 3.5 lists the top five semantic types of searches and their percentage from the total number of queries where the reviewers agreed with one another.

### 3.1.4 Discussion

The analysis of search logs yielded several design implications for future versions of CISearch, and possibly for others who wish to integrate search into their EHR.

#### 3.1.4.1 Adoption

It is premature to perform a formal study of EHR-based search adoption in our institution because the search utility is at the beginning stages of development and new functionalities are being identified. Furthermore, we have not performed any marketing or training to WebCIS users. Yet, the consistent search usage from month-to-month of the system suggests the potential usefulness of such a tool. A multi-year evaluation of the use and adoption of the system could address the usefulness of the system [Cimino, 2006].

### 3.1.4.2 User Type and High-Level Query Classification

Overall, users show a strong bias toward informational searches. When stratified by user types, however, different user behaviors emerge. All clinical users (e.g., doctors, nurses, and students) who provide direct care to patients tend to perform more informational searches (with doctors at 91.8%). Administrative staff's queries are evenly balanced between navigational and informational searches, confirming that their information needs differ from clinical users. Finally, researchers exhibit different behavior, with hardly any navigational searches (95.3% informational and 4.7% navigational). Contrary to clinical users, researchers approach the EHR as an interface tool for cohort selection, explaining the negligible number of navigational searches. The unanticipated use of the system to frequently search the same set of terms across multiple patients suggests that cross-patient search functionality would be useful for research purposes. There have been systems and studies designed to examine the use of cross-patient searches for cohort eligibility through the use of EHRs [De Bruijn *et al.*, 2008; Pakhomov *et al.*, 2007; Schulz *et al.*, 2008b; Gregg *et al.*, 2003b; Hanauer, 2006]. Though the cross-patient searches imply the need for such a system, our objective is to understand how clinicians search within an individual patient's record in order to extract pertinent information at the point of care.

### 3.1.4.3 EHR Implications

Overall, the CISearch queries adhered to the broad Web search categories (transactional, navigational, and informational). While most of the searches were deemed informational, we did note occurrences of transactional and navigational queries. The finding of transactional searches was unexpected considering WebCIS is predominately a read-only system and is not an order-entry system; however, there were a few transactional searches such as "add note" or "add drug". This may be a consequence of having multiple clinical information systems with overlapping functionality. The majority of the navigational searches were patient lookups (i.e., MRN or patient names). This might indicate the need for a more efficient way of switching patients within the EHR.

The semantic type “Laboratory or Test Result” was the most frequent informational search. This supports Chen’s findings that the most frequented section within WebCIS was laboratory results [Chen and Cimino, 2003]. WebCIS is efficient at displaying laboratory results in a structured format, so it would be interesting to understand why users are looking for them with free-text queries. We could hypothesize that viewing laboratory results mentioned within clinical narratives conveys more relevant and contextualized information than merely viewing the raw data in the structured section of the EHR. This could be assessed through user interviews or surveys, and may inform better display of laboratory results. As more usage data is collected, patient displays and navigation within the EHR can be further tailored for individual users, potentially improving users’ ability to access patient information.

#### 3.1.4.4 Concept-Based Searching

Our original plan to improve search was to map queries to UMLS concepts within machine processed notes. When entering a query, a user would be prompted to select the semantic type that best represents the query. However, we found that mapping query terms to the UMLS is inherently ambiguous because of its multi-hierarchical structure. Table 3.5 suggests that semantic types could be leveraged to inform preference rules for disambiguating UMLS concepts during the retrieval process. For example, a preference rule that favors laboratory test/procedure types over biological substance types would classify the query “fibrinogen” as a laboratory procedure and then search for that concept within the machine processed notes.

On the other hand, the large presence of queries with abbreviations and incomplete words, which do not map to the UMLS, suggests that indexing and searching based on UMLS concepts cannot be the sole solution. Rather, a combination of free-form text and concept-based search is needed. This finding is supported by Nadkarni et al.’s study that determined that both free-text and concept-based indexing was needed for concept-based searching of clinical notes [Nadkarni *et al.*, 2001].

#### 3.1.4.5 Semantic Categorization

The low agreement between reviewers was predominately due to the ambiguous nature of the classification discussed in the Limitations section below. For example, two reviewers classified the query “amylase” as a laboratory test and a biologically active substance. The reviewers commented that a query could easily be placed in either category.

#### 3.1.4.6 Limitations

Log analysis is an efficient, unobtrusive way to obtain information about a user's actions; however, it does not give insight into the user's underlying motivations or background for performing a search [Jansen, 2006]. While it provides an abundant and rich source of data, TLA cannot be solely used to model a user's information seeking behavior [Jansen, 2006]. There have been many studies examining log files to determine features that represent a successful search. One such feature is examining users' search sessions. In the clinical domain, it would entail examining a user's entire EHR session, which would include the user's activity before and after conducting a search. This additional information would aid in better inferring users' search behavior and search needs. Another feature that has proven to be representative of whether a search result is relevant is clickthrough data [Thorsten, 2002]. It is mainly used for ranking search results, and it is effective because the search results contain query-based snippets, allowing a user to determine whether a document is relevant or not before clicking on it [Tombros and Sanderson, 1998]. Though clickthrough analysis is useful in determining a document's relevancy, it is also limited because it does not account for documents that the user deems relevant based on the snippet. Thus, to truly understand what users are searching and the usefulness of a search utility, log analysis must be supplemented with observational and survey studies.

Another limitation in the study concerns the semantic categorization of queries. Besides the inherent ambiguity of labeling with the UMLS, it was difficult to disambiguate the queries because the reviewers were not provided the context of the queries (e.g., the query was performed while in the

laboratory section of WebCIS), resulting in the low kappa score. This manual process is also not scalable, a limitation which other search log studies face [Broder, 2002; Li *et al.*, 2005; Rose and Levinson, 2004]. The only solution to this problem would be a semi-manual approach whereby a trained classifier program would categorize a random sample of queries, and then these categorized queries would be manually reviewed. This review process would occur rarely.

Finally, there were two limitations related to CISearch itself. First, no marketing or formal training on CISearch was provided to the users within the institution, which may explain the relatively low adoption rate of 5.3%, the infrequent use of the advanced search features, and the low number of clickthroughs. Second, the search functionality was narrow in scope. For its initial release, CISearch only searched discharge summaries, radiology reports and pathology reports (in our current implementation, all note types are supported). As such, it is possible that the users were biased in their queries, based on the behavior of the search engine. However, when examining the search logs, it became evident that users were unaware of the limited search space and the internal workings of the search engine (this phenomenon also relates to the absence of user training), and instead queried the search functionality in a genuine fashion (as evidenced by the high frequency of navigational and transactional queries, typically not supported by domain-specific search engines). Overall, our approach to understanding information needs of clinicians in the context of EHR search follows an iterative process traditionally employed in software engineering, where a prototype with limited functionality is deployed in order to capture user needs in a real-world setting rather than in a laboratory setting. Once user needs are understood better, the prototype can be refined.

## 3.2 Survey Study

As mentioned above, log analysis alone cannot fully capture user needs [Jansen, 2006]. For this reason, an electronic survey was developed to gather users' search preferences. By combining the results of the survey and the log analysis, a better understanding of users' search needs can be obtained in order to design an effective EHR search utility.



### 3.2.1 Methods

The electronic survey was administered to healthcare providers at our institution. The survey was divided into four parts: *EHR usage*, *general search engine usage*, *searching within the EHR*, and *ideal clinical search*.

The first section, *EHR usage*, collected the user's clinical role in order to match the survey results to the log analysis study. It also gathered users' general experience with an EHR. The second section, *general search engine usage*, attempted to assess users' familiarity of Web search engines, specifically the use of advanced features such as wildcard or boolean retrieval. This information would allow us to understand how sophisticated the users were in regards to searching. The third section, *searching within the EHR*, gathered users' knowledge and experience of the current EHR search engine at our institution. The final section, *ideal clinical search*, surveyed clinical users' expectation of a within-patient EHR search engine. This section examined users' opinion on semantic based searching as well as visualization of search results. The complete survey can be viewed in Appendix A.

### 3.2.2 Results

Survey participants were recruited via targeted email blasts and word of mouth. 122 participants completed the survey with a response rate of approximately 25%. They were composed of nurses, attendings, residents, and medical students. The participants mainly practiced in the inpatient setting, except for pediatric attendings, and were all very familiar with the EHR systems at our institution. Figure 3.4 shows the breakdown of the participants' clinical role. All groups accessed test results and medications on a daily basis. The frequency of which different clinical professionals access various types of information within the EHR is shown in Figure 3.5.

All the survey participants reported using search engines in their daily lives and feeling very confident in their ability to use such systems. When prompted about their preferences towards an ideal EHR search utility, most participants were interested in an intelligent search system that possessed

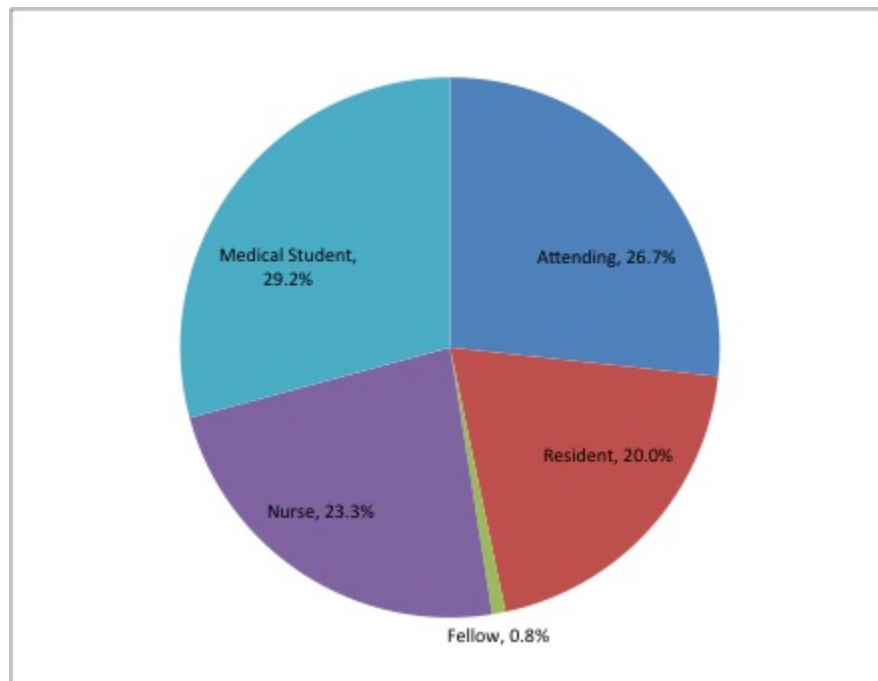


Figure 3.4: The breakdown of survey participants based on their clinical role.

a flexible user interface that would allow them to visualize searched patient data. They felt that a search system would assist them in familiarizing themselves with a patient (66%). The participants also acknowledged that a search system that would allow them to search for information across patients on their patient list (75%) would be very useful. In fact, many were surprised to find that a search utility existed within our institution's EHR and agreed that such a tool would be a useful feature.

### 3.2.3 Discussion

Useful search implications can be gleaned from this survey, namely the interest in an EHR search utility. These results also confirm our previous log analysis study that EHR users predominately seek information regarding laboratory results.

When comparing the different clinical users, no statistical differences were found between their responses; whereas, the log analysis study indicated that clinical user types varied in their search

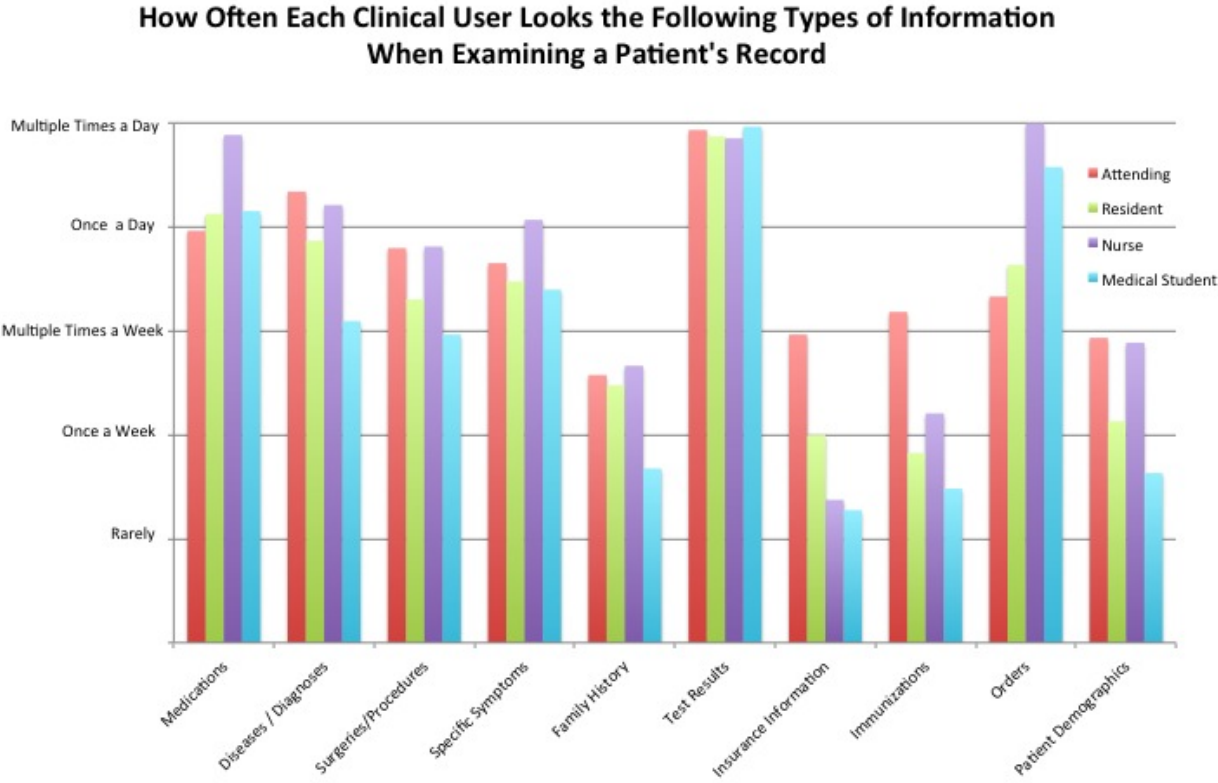


Figure 3.5: The average response to the types of information accessed within the health record based on clinical role.

behavior. These two findings could mean that clinical users may differ in how they search, but they generally search for the same types of information, implying the need for an intelligent search utility that is personalized based on user type.

There were some limitations to this study, mainly the low number of participants. It was difficult to recruit individuals because there was no remuneration for completing the survey. This made it difficult to conduct a proper comparative analysis between the different clinical user types. After further review of the survey responses, it was also evident that some questions needed further refinement, namely the question that asked participants to indicate which clinical note contained the most useful information when caring for a patient. Since no clinical context was provide with the questions, those who filled out the survey found the question difficult to answer and either skipped it or listed all the notes in the "Other" answer option. Despite the vagueness of the question, most individuals selected discharge summaries as the most useful source of information, which corroborates other anecdotal experiences.

## Chapter 4

# Creation of a Gold Standard for Within-Patient Search

In order to intrinsically evaluate search within the EHR, a gold standard must be created. A gold standard for search contains a document collection, a set of queries, and for each document/query pair a relevance judgment. Equipped with such a gold standard, different search models can be evaluated and compared against each other.

When creating such a resource, there are a few trade-offs to consider. Ideally, a gold standard contains a large number of documents and a diverse set of information needs in order to get a fair assessment of a search engine's performance. Creating such large resources, however, is a costly endeavor, requiring many judges to carefully examine every document in a collection and agree upon their relevance for each information need. In creating our gold standard, sets of homogenous real patient records were selected based on a particular disease diagnose, as opposed to a random selection of patient reports. This enabled us to develop meaningful, clinically relevant information needs for each disease. At the same time, it ensured that, for each information need, there was a useful amount of relevant documents in the patient records. We will refer to these sets of homogeneous patients as cohorts even though our intention is not to conduct a cohort analysis

or evaluate search specifically for a particular disease.

## 4.1 Corpus Selection

Two cohorts based on different diseases were selected for generating the gold standard – lupus and congestive heart failure (CHF). Lupus is a specific autoimmune disease that causes chronic inflammation. CHF is a broad disease that pertains to any condition where the heart is unable to properly pump blood to the rest of the body. Both diseases were chosen because of their complex set of signs and symptoms, which make them difficult to diagnose. CHF, for instance, is a broad disorder with vague symptoms that make it difficult to diagnosis in its early stage. Likewise, lupus is often misdiagnosed, and often times the diagnosis is missed for years. As such, from an information retrieval perspective, both conditions are interesting diseases; however, our search evaluation is not intended to be disease specific.

Ten patients (eight diseased patients and two control patients) from each cohort were selected from our clinical data warehouse (CDW). Our selection criteria was any patient with an ICD-9 code for either CHF or lupus since 2007 and with at least one visit at one of Columbia University’s clinic, Associates in Internal Medicine (AIM), in 2010. By adding the AIM-visit criteria, we improved our chances that the patient’s primary care physician was located at our institution, thus, ensuring the retrieval of more complete patient records. Two comparison patients were randomly selected using the same criteria with the exception that neither had a diagnosis for CHF nor lupus. Both these control patients were included in the lupus and CHF cohort to test for false negatives when evaluating search. A clinical faculty member reviewed all patient records to ensure that the document collection contained enough clinically useful information. All notes without any clinical information (such as notes documenting patient no-shows or telephone calls among providers) were removed from the collection. At the end of the selection process, each patient chart consisted of approximately 30 clinical notes, which roughly totaled to 300 notes per cohort. This collection of notes for the two cohorts will be referred to as our CHF and lupus document collections.

Once the final set of patients was identified, all the patients' clinical notes were retrieved from the CDW and processed using a NLP framework developed in our research lab [Gorman and Elhadad, 2011]. After the processing step, the notes were represented in XML format, providing structure (i.e., sections, paragraphs, list items, and sentences) to the free text notes as well as identifying concepts within the UMLS.

## 4.2 Query Development

In developing a gold standard, a set of queries based on information needs must be generated to run against the search engine to evaluate search performance. Usually, the information needs are articulated in the form of a question, which is the case for our gold standard. A set of questions for each cohort was developed with assistance of a clinical faculty member. All the questions were based on both information needs that might arise in a clinical setting as well as findings from our log analysis and survey study [Natarajan *et al.*, 2010]. As mentioned in the previous chapter, we analyzed search logs from a free-text search engine deployed in our institution's EHR. We found that clinical users' queries could be mapped to UMLS semantic types, such as laboratory results, medications, diseases, and diagnostic procedures. The information needs developed covered a good mix of these semantic types in order to model topics that were actually searched in a live system. Additionally, the questions were constructed as yes/no questions, as opposed to questions with open-ended or specific answers, which are more for question answering systems. The questions went through several iterations for clinical validity before being finalized. Then, each question was translated to a query that could be submitted to a search engine. The query terms used were extracted from the developed questions. The only changes made were morphological (i.e., "fever" was added to the query that contained "fevers"). The queries were a mix of terms that could and could not be mapped to a knowledge-based concept (i.e., CUI within the UMLS) in order to assess a search engine's ability to handle free-text and concept terms. Since the focus of this dissertation is on search method and not on query translation, concept mapping was done manually, where CUIs were selected based on their relevance to the questions and their presence in the corpus.

### 4.3 Relevance Tagging

In order to evaluate the performance of search engines, the number of relevant documents must be known for a given information need; therefore, a medical expert manually tagged each clinical note as relevant or non-relevant at the paragraph level for each information need in the test corpus. Fourth-year medical students were chosen as medical experts because their medical knowledge was deemed rich enough to accurately carry out the annotation task as well as for practical reasons.

Four medical students were recruited to annotate the test corpus. Two were randomly selected to annotate the lupus patients, and the other two were assigned to the CHF patients. Each annotator went through a 15-minute training session. During this session, the purpose of their task and the use of the annotation tool were explained. At the end of the training, the medical students annotated three training notes – outside of the test corpus – in order to familiarize themselves with the information needs they were assigned. The annotators tagged the paragraphs within each document that were relevant to an information need. They were instructed to tag paragraphs as relevant even when the answer to the information need was implicit. For instance, a paragraph mentioning that the patient is on Tikosyn, an anti-arrhythmia medication would be marked as relevant for the information need, “does this patient have an arrhythmia?” Since the gold standard was developed for information retrieval as opposed to information extraction, the annotators were instructed to tag negated terms as relevant, such as “no MI,” since the purpose of the retrieval task was to present all relevant information to the users so that they can make an informed decision. After the medical students finished the annotations, as a consensus step, a resident resolved all disagreements, finalizing the gold standard. All relevance tagging was performed using a web-based annotation tool that I developed. More details on the annotation tool can be found in the Appendix.



## 4.4 Results and Discussion

The completed gold standard consisted of two corpora and contained 15 information needs per cohort. The combined corpus contained 638 documents, or 20,643 paragraphs overall. There were 321 clinical notes in the lupus cohort and 319 notes in the CHF cohort. 2,114 paragraphs, from 299 documents, were judged relevant. There were many cases where the relevant text did not contain terms within the information need. “There is trivial aortic regurgitation. The mitral leaflets are mildly thickened. Trace mitral regurgitation is seen,” is an example of non-explicit text marked as relevant for the information need on valvular disease. Another example of this can be seen with medication type information needs – “resume prednisone 5 mg po daily, continue HCQ,” which is pertinent to the information needs on antimalarial and corticosteroid medications. Tables 4.1 and 4.2 present a summary of the gold standard at the document level for each cohort. There were a total of 110 different document types in the CHF corpus and 119 in lupus. A document type is defined by the title of a clinical note. Tables 4.3 and 4.4 show the top ten document types found in each cohort and the frequency in which they occur. Figure 4.1 also graphically displays the the overlap of the top document types between each cohort. To better understand which document types contained information for the different information needs, the queries were grouped by query type. The query type was a manual mapping of the information need to a UMLS semantic type. Tables 4.5 and 4.6 show the top ten document types that are relevant for each query type.

CHF Information Needs	Query Type	No. Relevant Docs
Is there documentation that suggests the severity of the patient's heart failure?	Disease	29
Is there documentation that suggests the patient had a stroke or had signs of cerebral ischemia?	Disease	32
Is there documentation that the patient has valvular disease?	Disease	18
Has the patient been admitted for CHF?	Disease	13
Has the patient ever had a MI?	Disease	55
Is there documentation that suggests that the patient has been diagnosed with CHF?	Disease	34
Does this patient have a history of smoking?	Finding	40
Is there documentation that suggests lower extremity edema?	Finding	25
Have results of an echocardiogram been documented?	Finding	61
Has this patient experienced shortness of breath?	Finding	41
Does this patient have an arrhythmia?	Finding	44
Is or has this patient been on any anticoagulant medication?	Medication	17
Is or has this patient been on diuretics?	Medication	50
Is or has the patient been on statins?	Medication	53
Is or has this patient been on beta-blockers?	Medication	63

Table 4.1: This table shows the semantic type of the information need and the number of relevant documents associated to the information need.

<b>Lupus Information Needs</b>	<b>Query Type</b>	<b>No. Relevant Docs</b>
Has the inflammation rate in the body been measured?	Diagnostic Procedure	38
Has the patient had a biopsy done on any organs?	Diagnostic Procedure	44
Has an antinuclear antibody test been done on this patient?	Diagnostic Procedure	59
Has the patient been anemic?	Disease	26
Is there any documentation that suggests when this patient was diagnosed with lupus?	Disease	43
Does this patient have drug-induced lupus?	Disease	3
Has the patient had unexplained fevers?	Finding	19
Has the patient experienced hair loss?	Finding	38
Has the patient had kidney involvement?	Finding	21
Has the patient experienced joint pain or joint swelling?	Finding	85
Has the patient had skin rash?	Finding	57
Has the patient experienced chest-pain?	Finding	26
Is or has the patient been on any corticosteroids or immunosuppressive medications?	Medication	32
Is or has the patient been on anti-malarial medications?	Medication	46
Is or has the patient been on nonsteroidal anti-inflammatory drugs?	Medication	57

Table 4.2: This table shows the semantic type of the information need and the number of relevant documents associated to the information need.

<b>Document Type</b>	<b>No. Docs</b>
12-Lead Electrocardiogram	39
Clinical Note	33
Ambulatory Internal Medicine Structured Note	19
Primary Provider Clinic Note	18
Signout	14
X-Ray of Chest, 2 Views	9
Follow-up	8
Transthoracic Echocardiography	7
Exercise Thallium Test	6
Ambulatory AIM Care Triage Telephone Triage Form	6

Table 4.3: Top 10 document types in the CHF gold standard corpus.

<b>Document Type</b>	<b>No. Docs</b>
Clinical Note	30
Surgical Pathology Event	28
12-Lead Electrocardiogram	16
Rheumatology	12
Follow-up	10
Primary Provider Clinic Note	10
Ob/Gyn Ultrasound	8
Ambulatory Internal Medicine Structured Note	8
Event Note	8
X-Ray, Other and Unspecified	7

Table 4.4: Top 10 document types in the lupus gold standard corpus.

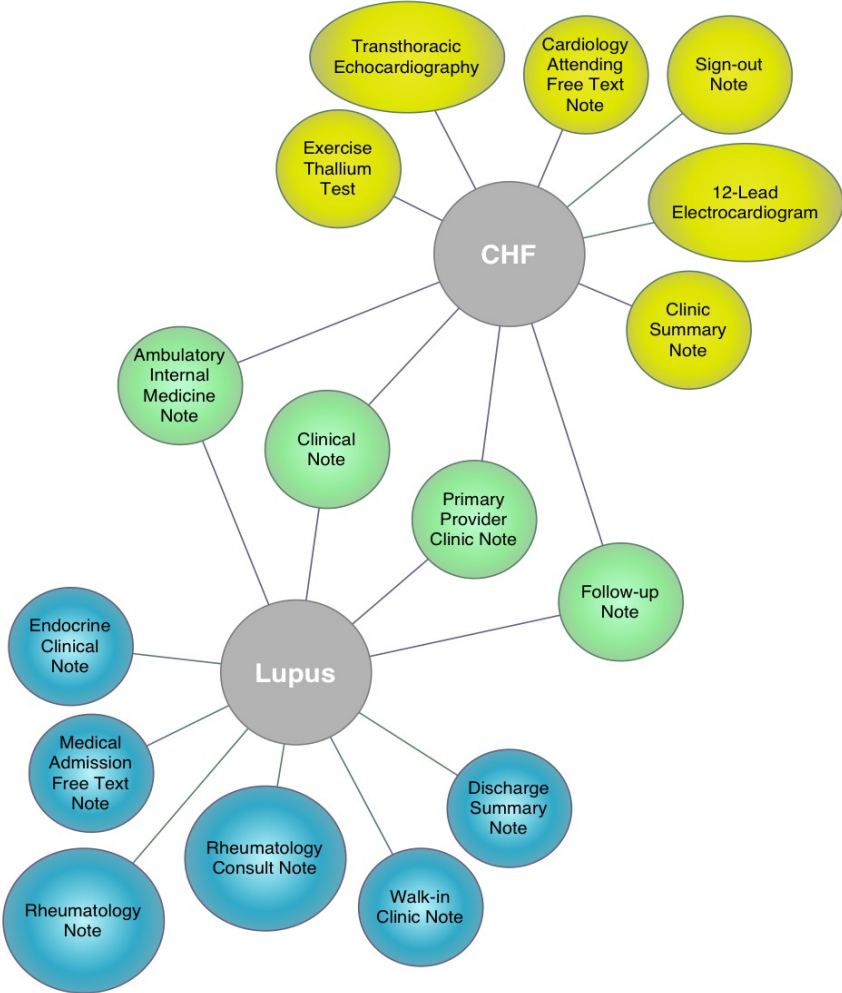


Figure 4.1: Document types with most relevant information.

Document Type	CHF Query Types		
	Disease	Finding	Medication
12-Lead Electrocardiogram	10	24	
Adult Echocardiographic Report	6		
Amb Internal Medicine Structured Note	23	16	31
Ambulatory Nutrition Reassessment			6
Cardiology Attending Free Text Note	6	9	6
Clinic Summary	8	9	
Clinical Note	26	26	26
Exercise Thallium Test		7	11
Follow-up	12	18	8
Myocardial Perfusion SPECT			6
Primary Provider Clinic Note	10	24	21
Signout	11	15	13
Transthoracic Echocardiography	8	7	
Walk-in Clinic Note			6

Table 4.5: Top ten document types for CHF query types, and the number of times each is found to be relevant.

Document Type	Lupus Query Types			
	Diagnostic Procedure	Disease	Finding	Medication
Amb Rheumatology Note	11	7	15	11
Amb Care TriageTelephone Triage Form				3
Amb Primary ProviderStructured Note	12	8	13	13
Amb Walk-In Clinic Structured Note	3	2		5
Amb Internal MedicineStructured Note	15	6	21	10
Clinic Summary				4
Clinical Note	18	9	33	31
Discharge Summary Note		3	5	
Endocrine Clinical Note			8	
Follow Up Note	10	5	10	10
Medicine AdmissionFree Text Note	4	2	10	
Primary Provider Clinic Note	15	10	27	13
Rheumatology	28	8	30	10
Surgical Pathology Event	7			

Table 4.6: Top ten document types for lupus query types, and the number of times each is found to be relevant.

The document type with the most relevant information was “Clinical Note,” which was also one of the most prevalent documents in both corpora. This document type is a generic name used for all clinical notes that did not contain a title within our institution’s EHR system. Except for “Diagnostic Procedure” information needs, all other questions seemed to be mostly answered in Medicine related notes (i.e., Primary Provider Notes and Ambulatory Internal Medicine Note). This could be due to the nature of the specific information need or due to the fact that Medicine

physicians tend to rely more on documentation and thus document more than other specialties. This gold standard with the location of relevant information could be used to inform different search strategies based on a query type. For example, a search engine could weigh certain document types (i.e., primary provider notes) higher than others (i.e., progress notes) for medication type queries, and it could also give a higher weight to different sections of a document as well (i.e., problem list vs family history).

To our knowledge, this is one of the first attempts to develop a gold standard for within-patient search in the EHR. It lays the foundation for formal IR evaluation within the EHR domain. The annotations were conducted on each document at the paragraph level. By including paragraph level annotations, relevance based on location within a document could be used as an evaluation criteria. The different types of information needs, along with their relevance judgments, are shown to belong to a representative set of note types, providing a realistic and rich dataset for information retrieval and other informatics research.



## Chapter 5

# Analysis of VSM on Clinical Narrative

As mentioned in the introduction chapter, most search engines deployed within EHRs rely on the traditional, vector-space model approach to unstructured, free-text notes. A shortcoming of such search engines is the inability to handle synonymy, which is very common in clinical notes. For instance, the concept of atrial fibrillation can be conveyed in a note as “atrial fibrillation”, “afib”, “a-fib”, or “AF.” For such a case, a search engine should take into account all these variations in order to retrieve all relevant documents; however, today’s conventional search engines place this burden upon the users, making search cumbersome. One way to abstract away from search at the string level towards a semantic search is to identify and index the concepts mentioned in the clinical notes and then query the concepts as opposed to the strings. Concepts from a terminology, like SNOMED-CT or more generally the UMLS, can be the unit of processing for the search. Though studies have examined EHR search queries, there has yet to be research that examines how search performs on clinical text in order to improve baseline search systems [Natarajan *et al.*, 2010; Yang *et al.*, 2011]. This study hopes to address this gap.

In this chapter, I will describe an error analysis study that addresses aim 2 of this dissertation. In this study, our primary research goal is to classify errors in clinical search results, which can later be used to improve search performance. A secondary research goal is to examine whether there is any advantage in indexing and querying at the level of semantic concepts rather than strings.

## 5.1 Methods

To answer our research questions, we used the gold standard discussed in the previous chapter. Three search models using the vector-space model method for document retrieval were evaluated. Each search model varied in its representation of the patient notes – one consisted of the original free text, one consisted of all UMLS terminologies (except for GO vocabularies) concepts identified in the collection, and one consisted of identified concepts found in SNOMED-CT and RxNorm terminologies only. In addition to evaluating the search performance, I manually examined the false positives and false negatives of the free-text search results. This in-depth error analysis was not conducted on the other search models because such an analysis is more of an evaluation of the NLP system than the search engine.

### 5.1.1 Data Pre-Processing

As mentioned in the previous chapter, all notes in the gold standard were processed by an NLP system. The system outputted an XML document for each note analyzed, which contained both the original text of the note as well as the meta-data identified. From these processed notes, three gold standard corpora were created for evaluating a vector-space model search: the original free-text notes, semantically processed notes, and a semantically filtered set of notes. These corpora will be referred to as *Free-Text*, *All CUI*, and *Filtered CUI*, respectively. The purpose of the three data sets was to observe how search performs on semantically tagged text versus free-text. The Free-Text data set contained the original clinical notes and, therefore, was not processed by our NLP system. The other two corpora contained files with only the CUIs identified in the processed notes. The Filtered CUI corpus was a subset of the All CUI corpus and contained only clinically oriented SNOMED-CT and RxNorm CUIs. The filenames and relevance judgments remained intact for these two corpora. Figure 5.1 depicts the creation of the three corpora.

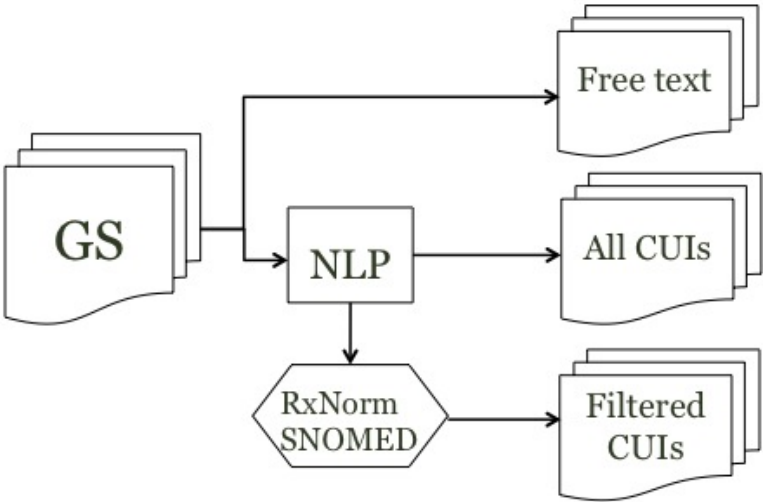


Figure 5.1: Creation of the three corpora using the original notes and relevance judgments.

### 5.1.2 Search Evaluation

An open-source search engine named Lucene was implemented for the search experiments on all three corpora. Lucene is a robust, Java-based text search API that is based on the vector-space model and is implemented in many applications on the Internet [Apa, 2007].

For the search experiments conducted, each information need was searched under the three corpora – Free-Text, All CUI, and Filtered CUI. MAP scores were calculated over all the information needs for the three corpora, and average precisions scores were calculated for each information need.

### 5.1.3 Error Analysis

In addition to calculating the IR metrics, an error analysis was conducted on the search results from Free-Text gold standard corpus . For each information need, the non-relevant documents retrieved (false-positive) and the relevant documents that were not retrieved (false-negative) were manually examined. Each error was recorded for all the information needs. After this step, the errors were grouped in an iterative process until all the errors were categorized into error types.

## 5.2 Results

### 5.2.1 Search Results

Three search experiments (one for each corpus) were conducted for each cohort. Tables 5.1 and 5.2 show the mean average precision for each cohort for the three corpora. Tables 5.3 and 5.4 show a breakdown of each information need and the average precision on each corpus. Overall searching on the Free-Text corpus outperformed the search on the CUI corpora, but there were only two instances where the difference between average precisions was statistically significant, which are marked by asterisks. For the first instance, the semantic abstraction was too high. For example, the query was “kidney involvement”, but the CUI-based search retrieved non-relevant documents containing the term “renal” since there is no CUI for “kidney involvement”. In the other cases,

the search experiment on the Free-Text corpus performed better because of ontological resolution issues that were present in CUI-based search. For example, in the Free-text search, the query “joint pain” retrieved relevant documents referring to “knee pain” because the search matched the query term “pain”. In the CUI-based search, these documents were not retrieved because they were too specific. Likewise, the CUI search failed to retrieve many relevant documents for the query “anemia” because documents containing “normocytic anemia” were tagged with a granular CUI as opposed to the general CUI for anemia.

<b>CHF Corpus</b>	<b>Mean Average Precision</b>
Free-Text	0.323
All CUI	0.271
Filter CUI	0.260

Table 5.1: This table shows the mean average precision scores on the different CHF gold standard corpora.

<b>Lupus Corpus</b>	<b>Mean Average Precision</b>
Free-Text	0.511
All CUI	0.420
Filter CUI	0.404

Table 5.2: This table shows the mean average precision scores on the different lupus gold standard corpora.

### 5.2.2 Error Analysis Results

The false positive and false negative Free-Text search results were manually reviewed for both cohorts. False positive results – non-relevant, retrieved documents – were due to query terms

CHF Information Needs	Query Type	Average Precision		
		Free-Text	All CUI	Filtered CUI
Is there documentation that suggests the severity of the patient's heart failure?	Disease	0.213	0.352	0.354
Is there documentation that suggests the patient had a stroke or had signs of cerebral ischemia?	Disease	0.144	0.38	0.377
Is there documentation that the patient has valvular disease?	Disease	0.523*	0*	0*
Has the patient been admitted for CHF?	Disease	0.692	0.658	0.643
Has the patient ever had a MI?	Disease	0.4	0.288	0.35
Is there documentation that suggests that the patient has been diagnosed with CHF?	Disease	0.889	0.873	0.869
Does this patient have a history of smoking?	Finding	0.043	0.036	0.008
Is there documentation that suggests lower extremity edema?	Finding	0.331	0.06	0.06
Have results of an echocardiogram been documented?	Finding	0.206	0.288	0.082
Has this patient experienced shortness of breath?	Finding	0.608	0.614	0.64
Does this patient have an arrhythmia?	Finding	0.259	0.023	0.023
Is or has this patient been on any anticoagulant medication?	Medication	0.059	0.059	0.059
Is or has this patient been on diuretics?	Medication	0.04	0.02	0.02
Is or has the patient been on statins?	Medication	0.189	0.189	0.189
Is or has this patient been on beta-blockers?	Medication	0.25	0.222	0.222

Table 5.3: This table is a breakdown of each CHF information need. It shows the semantic type of the information need and the average precision for each corpus. The asterisks indicate cases where the Free-Text average precision is statistically significant over the other searches.

Lupus Information Needs	Query Type	Average Precision		
		Free-Text	All CUI	Filtered CUI
Has the inflammation rate in the body been measured?	Diagnostic Procedure	0.039	0.005	0.005
Has the patient had a biopsy done on any organs?	Diagnostic Procedure	0.268	0.216	0.218
Has an antinuclear antibody test been done on this patient?	Diagnostic Procedure	0.803	0.683	0.695
Has the patient been anemic?	Disease	0.538	0.231	0.231
Is there any documentation that suggests when this patient was diagnosed with lupus?	Disease	0.549	0.599	0.582
Does this patient have drug-induced lupus?	Disease	1	1	1
Has the patient had unexplained fevers?	Finding	0.654	0.445	0.461
Has the patient experienced hair loss?	Finding	0.58	0.825	0.84
Has the patient had kidney involvement?	Finding	0.716	0.631	0.402
Has the patient experienced joint pain or joint swelling?	Finding	0.735*	0.300*	0.291*
Has the patient had skin rash?	Finding	0.825	0.602	0.586
Has the patient experienced chest-pain?	Finding	0.785	0.529	0.523
Is or has the patient been on any corticosteroids or immunosuppressive medications?	Medication	0	0	0
Is or has the patient been on anti-malarial medications?	Medication	0	0	0
Is or has the patient been on nonsteroidal anti-inflammatory drugs?	Medication	0.175	0.228	0.228

Table 5.4: This table is a breakdown of each lupus information need. It shows the semantic type of the information need and the average precision for each corpus. The asterisks indicate cases where the Free-Text average precision is statistically significant over the other searches ( $p=0.03$ ).

found in many non-relevant documents. They ranged from broad query terms (i.e., ‘heart,’ ‘lupus,’ and ‘kidney’), which, given the corpus, were frequent terms, to terms that were part of a medical phrase (i.e., “respiratory rate,” “alpha blocker,” and “pain rating scale”). These cases comprised of approximately 96% of the errors, and the other 4% were due to structured template notes containing the query term (i.e., “Admitted:,” “Skin:,” and “Smoking History:”).

The false negative results – relevant, not retrieved documents – were categorized into three error types – abbreviations, synonymy, and implicit reference. Abbreviation errors represented about 12% of the false negatives and were primarily found in disease and findings type queries. ‘NSTEMI’ for non-ST-segment elevation myocardial infarction, ‘SLE’ for systemic lupus erythematosus, and ‘TTE’ for transthoracic echocardiogram are a few examples. Most errors were due to synonymy and implicit reference. The most prevalent error, synonymy, was broken into two types – simple synonyms and ontological similarity – and represented approximately 11% and 51% of all false negatives, respectively. Simple synonyms errors, such as alopecia, dyspnea, febrile, and arthritis, were mostly found in findings type queries; whereas, ontological similarity errors were predominantly found in medication type queries. Ontological similarity refers to words that are related via the is-a relationship. For instance, Lipitor is a statin and prednisone is a corticosteroid. All medication queries had very low precision because relevant notes contained specific drug names as opposed to the drug class – in some cases it was the brand name and in others it was the generic name of the drug. Finally, implicit references comprised 26% of all the false negatives. These were cases where medical knowledge was needed to access the relevance of the note and where clinical training was needed. Phrases, such as “slight left facial droop,” which suggests that the patient had a stroke, and “decreased hematocrit from 40-22,” which indicates that the patient is anemic, are such examples.



### 5.3 Discussion

In many cases, vector-space model search does not perform well, regardless of the corpus used. The search performance is query dependent, even within query types. This variability within query type is not easily observed when examining the MAP score since it is an aggregate value over all information needs. Thus, when intrinsically evaluating a search engine on clinical records, there might be no other way than to conduct an evaluation using a much larger corpus and query set, where there are several information needs per query type.

Even though searching on the Free-Text corpus outperformed search on the CUI-based corpora, the fact that most non-retrieved, relevant documents were due to implicit references suggests a need for a concept-based search. In many cases, the CUI search failed due to parsing and ontological granularity issues. For example, there were cases where phrases were not parsed properly, so the proper CUI was not identified (i.e., “significant valvular disease” was parsed as “significant valvular” and “disease”). Also, there were cases where the concept mapper identified specific CUIs. For instance, the CUI used for the query “arrhythmia” was for “cardiac arrhythmia” (C0003811), but the relevant documents contained CUIs for “sinus arrhythmia”. Clearly, the performance of a CUI-based search is very dependent on the parser and concept mapper used on the clinical notes, and therefore, it should be taken into consideration when integrating semantics into an EHR search.

These evaluation findings have many design implications for future concept-based search functionality within the EHR. Even though CUIs provide a semantic abstraction that intuitively is favorable, it is clear that any concept-based search must be a hybrid system that incorporates free-text. This finding corroborates the results of other search studies [Nadkarni *et al.*, 2001; Natarajan *et al.*, 2010]. The fact that the search on the All CUI corpus slightly outperformed the search on the Filtered CUI corpus reinforces the idea that there is no one terminology that contains all clinically relevant concepts. Thus, a concept-based search should consider including all CUIs when searching. In order to improve a concept-based search’s performance, there must be a way to integrate concept resolution along the is-a relationship tree into the search process. This way CUI terms such

as “sinus arrhythmia” could be retrieved for queries about arrhythmia. Additionally, since a quarter of all non-retrieved, relevant documents contained implicit references to the information need, a concept-based search must inject more sophisticated semantic approaches. This would require utilizing other relationships between concepts either through an ontology or through statistical correlation between concepts, neither of which is a trivial task.

The challenge with ontologies is that there is no one ontology that fully encompasses all clinical data with all possible relationships. Even within the UMLS, which is composed of other ontologies, the abbreviation “cp” is not mapped to “chest pain,” which was a common abbreviation within our gold standard corpus. Data-driven approaches could address some of these shortcomings, but the nature of clinical data can pose particular challenges. Since data-driven methods are dependent on the underlying clinical data, there are several things to take into consideration when employing these methods. First, clinical data is not uniform, meaning there are few data points when a patient is healthy, so modeling non-sick patients is difficult. Second, clinical notes have many abbreviations and other shorthand notations that make it difficult to process and thus difficult to aggregate data for processing. Last, with electronic notes becoming more prevalent, copy and paste has become an issue [Wrenn *et al.*, ]. Redundant information can skew results for data-driven methods by artificially making a data point (i.e., CUI) more important. In some cases, incorrect information is propagated. There have been attempts to assess redundancy within text and its implications on data-driven methods, but this is still an area of open research [Wrenn *et al.*, ; Zhang *et al.*, 2011; Elhadad *et al.*, 2012] .

## 5.4 Limitations

As mentioned earlier, the small corpus and query sets were a limitation to this study, as well as the use of only two cohorts. In order to fully evaluate search in EHRs, a larger corpus and query set must be used, such as the ones found in the TREC competitions. Besides the limitation of the gold standard, the search evaluation is sensitive to query translation. Translation from information need

to query was carried out manually in the study. This has its advantages since the study's focus was on the search process not on query translation. We tried to be consistent in our translation process, but one could argue that we would need to add more morphological variants into the free-text queries. We attempted to balance the length of the queries with the number of variants included in the query, as the query length has an impact on the similarity metric. Also, we tried to replicate the way users would enter a query, where the query contained a few terms and not many query term variants [Natarajan *et al.*, 2010].

## Chapter 6

# Evaluation of Semantic Search Approaches on Clinical Narrative

Improving the discovery of relevant documents through the use of semantic structures has been an area of focus in IR research for several decades. Work on incorporating semantics has ranged from human-curated knowledge bases to data-driven approaches [Manning *et al.*, 2008]. Human-curated knowledge representations, such as ontologies, are useful for incorporating high-fidelity conceptual relations in order to improve search results. Such semantic structures are found in many specialty search systems [Ide *et al.*, 2007; Ope, 2008; Pub, ; Zeng and Cimino, 2001; Baud *et al.*, 2001]. A popular implementation of such an approach can be found in the clinical domain – PubMed. It is an ontology-driven IR system for finding scientific literature in the biomedical domain [Ide *et al.*, 2007]. The difficulty with such a system is the maintenance of a knowledge base, since it is both time-consuming and expensive. In the past few decades, there has been a focus on statistical approaches, such as probabilistic topic modeling, that are generally more computationally intensive, but do not suffer the same disadvantages of an ontology-based search. These data-driven approaches are designed to identify semantics within a collection of documents. Although topic modeling for IR is continually being researched, there has not been much work in the clinical domain, specifically on clinical narrative.

In this chapter, I will outline a study that addresses aim 3 of this dissertation. The study evaluates various semantic search methods on clinical notes within the EHR and compares them to a baseline vector-space model approach.

## 6.1 Methods

As mentioned in the background chapter, there are two approaches to incorporating semantics into search – a knowledge-base and data-driven approaches. Within data-driven approaches there are two general methods used – geometric and probabilistic. Figure 6.1 present the semantic approaches surveyed in this evaluation. They include: *vector-space model* (VSM), *latent semantic analysis* (LSA), *UMLS-based query expansion* (UMLS-QE), *KL-divergence retrieval model* (KL), *LDA-based query likelihood model* (LDA-QL), and *LDA-based query expansion* (LDA-QE). As in the previous vector-space error analysis study, this study evaluates IR performance on both free-text and processed notes, and as such, the evaluation uses the same three gold standard corpora: the original free-text notes, semantically processed notes, and a semantically filtered set of notes. The framework developed for the previous study was extended to incorporate the multiple semantic search approaches in this study.

### 6.1.1 Experimental Setup

In this section, I will describe the experimental setup for each search evaluation. Each search algorithm was evaluated on the three different representations of the gold standard. The average precision values were calculated for each of the 15 information needs, as well as the mean average precision (MAP) for each search evaluation [Manning *et al.*, 2008; Hersh, 2009].

#### 6.1.1.1 LSA Setup

***Additional Processing on Data.*** As mentioned in the background chapter, LSA performs a singular value decomposition of the term-document matrix. For efficiency purposes, this process

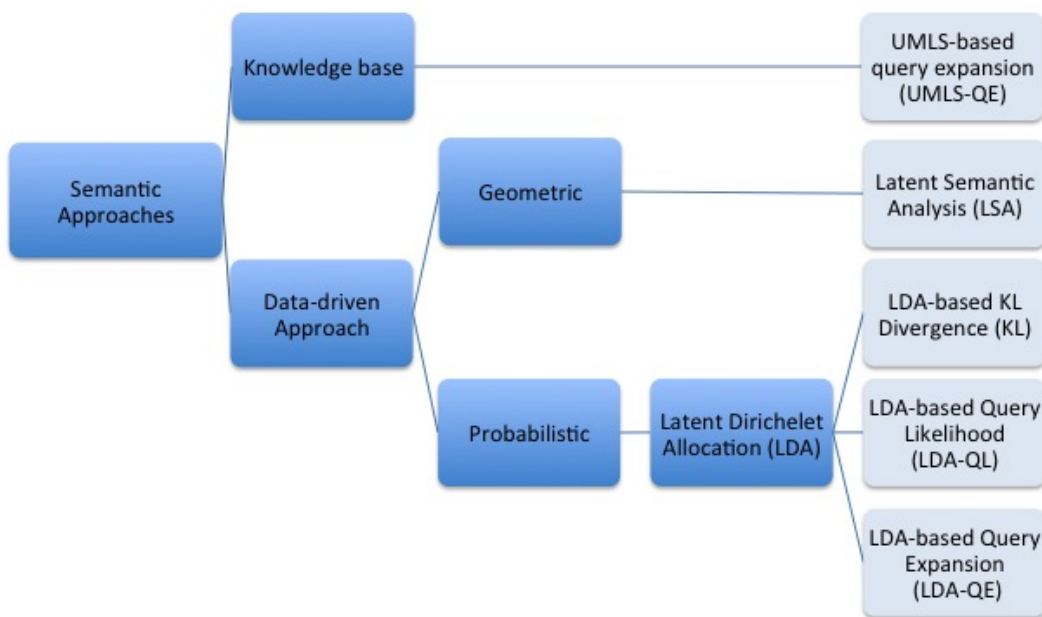


Figure 6.1: A hierarchical classification of semantic approaches and the specific semantic search methods evaluated in this study.

was performed offline and the reduced matrices for different  $K$  were stored for the evaluation. This matrix reduction was performed on the three different gold standard corpora.

**Queries.** Just as the gold standard documents were created for the three sets, the queries were replicated to represent each data set (Free-text, All CUIs, and Filtered CUIs). These were the same queries used in the vector-space model study conducted in the previous chapter. I will refer to the collection of queries from the different data sets as the *query set*.

**Resources.** The SVD was performed using the LingPipe API [Alias-i, 2008]. All other calculations used in the LSA-base search were performed using native JAVA functions.

#### 6.1.1.2 UMLS-QE Setup

**Additional Processing on Data.** There was no additional processing required for this evaluation. The experiment was similar to VSM except for the query set submitted to the search engine.

**Queries.** The query set created in this evaluation used the UMLS to expand queries from the original *query set* mentioned above. For both the All CUI and Filtered CUI corpora, each CUI was traversed one level above and below along the “is-a” relationship in order to include the direct parents and children. Additionally, the CUIs were filtered based on semantic type; only CUIs matching the semantic types of the information need were included as an expansion term. This strategy is similar to the methodology employed by Liu, which showed improved retrieval performance [Liu and Chu, 2007]. The Free-text corpus queries were constructed from the expanded All CUI query set. The CUIs from this query set were used to retrieve the string representation of each of the concepts and were added as expansion terms to the Free-text queries.

**Resources.** The query expansion terms were identified using a modified version of UMLS-Query [Shah and Musen, 2008].

### 6.1.1.3 KL Setup

**Additional Processing on Data.** For each of the three gold standard corpora, an LDA model was trained on a larger set of lupus and CHF patients, which did not contain patients from the gold standard. The pool of patients were selected from the CDW with a visit between 1/1/2007 and 10/1/2010. For the lupus cohort, all the notes from 1,631 patients were used for training the model. There were a total of 232,779 clinical notes. The number of tokens (i.e., words or CUIs) for each of the free-text, semantically processed, and semantically filtered corpora were 137,563, 44,220, and 27,753, respectively. The CHF cohort was too large to train all the patients on our computing environment, so two levels of filtering were performed – note-level and patient-level. First, note types that were deemed to not have clinically relevant information were removed (i.e., patient no show notes, social work notes, phone call notes, miscellaneous notes). Second, pediatric patients were removed, and then the remaining patients were randomly sampled based on the time span in years between the oldest and newest note within a health record. The time spans ranged from 0 to 20 years. Half the patients for each time span year were randomly selected, resulting in a total of 7,669 patients. There were a total of 1,273,492 clinical notes that were used to train the CHF models. There were 320,682 words in the free-text corpus, 48,725 CUIs in the semantically processed corpus, and 32,412 CUIs in the semantically filtered corpus. All trainings for various numbers of topics were performed offline on the three different corpora. From this point, I will refer to these trained models as the *LDA models*.

After the LDA models were created, they were used to infer topic distributions for the documents in the appropriate gold standard corpus. For example, the LDA models that were trained on the semantically filtered set of notes were used to infer topic distributions on all the similar gold standard corpora of semantically filtered notes. The inferred topic distributions for the gold standard documents were used for evaluating similarity between them and the query using KL divergence as the similarity measure, which was discussed in the background chapter.

**Queries.** The query set was the same as the one used in the LSA experiments. As with the gold standard documents, a topic distribution was inferred for each query in the query set. All



inferencing was done offline for efficiency purposes.

**Resources.** The MALLET toolkit was used to train the LDA models and was used to calculate KL divergence between topic distributions [McCallum, 2002]. Other LDA parameters were assigned to the default values set in MALLET for all the model trainings.

#### 6.1.1.4 LDA-QL Setup

**Additional Processing on Data.** The LDA models created and the inferred topic distribution of the gold standard documents were the same as discussed above in the KL experiment.

**Queries.** The query set was the same as the one used in the LSA experiments. The probability of the query terms given the topic was calculated using the term-topic matrix created by MALLET.

**Resources.** LDA training and inferencing were performed using the MALLET toolkit, as well as any matrix calculations required for the query likelihood model.

#### 6.1.1.5 LDA-QE Setup

**Additional Processing on Data.** The LDA models created were the same as discussed above in the KL experiment. Topic distributions for the gold standard documents were not calculated for this search method.

**Queries.** The query set was the same as the one used in the LSA experiments. A topic distribution was inferred for each query in the query set. These topic distributions were the same as the ones created in the KL experiment. The highest ranked topic for each query was selected for query expansion. Each topic is a cluster of terms, where the terms are ordered based on how strongly associated they are to that topic. Using the appropriate topic clusters, each term was added in order to the original query.

**Resources.** LDA training and inferencing were performed using the MALLET toolkit. Additionally, Lucene was used to perform the retrieval on the gold standard corpora [Apa, 2007].

## 6.2 Results

### 6.2.1 Overall Results

There were a total of 1,854 semantic search evaluations conducted (927 per cohort). Tables 6.1- 6.4 present the top 10 MAP scores for each of the search methods on the two cohorts. Figures 6.3 and 6.2 are 11-point average precision graphs for a select number of top search methods for each cohort. For the 11-point average precision graphs, the average precision over all the information needs is calculated at eleven recall levels (0.0, 0.1, 0.2, . . . , 1.0) and plotted on a graph [Manning *et al.*, 2008].

The vector-space model (VSM) performed better than both the LSA and LDA-based KL methods. The query likelihood approach performed better than the VSM only on the CHF cohort. UMLS-QE performed very well on both cohorts on the Free-text corpus, but had mixed results on the other corpora. The LDA-QE method performed the best compared to all other search methods on the two cohorts; however, many expansion terms were required to achieve this high performance – 150 terms for CHF and 70 for lupus. The LDA-QE approach with 150 topics and 150 expansion terms on the CHF corpus was the only search configuration that was statistically significant ( $p=0.03$ ) compared to VSM on the CHF Free-Text corpus, which was the highest performing VSM approach.

<b>LSA</b>	<b>MAP</b>	<b>UMLS-QE</b>	<b>MAP</b>	<b>KL</b>	<b>MAP</b>
vsm_freertext	0.511	umls_freertext	0.545	vsm_freertext	0.511
lsa_freertextnumFactors260	0.439	vsm_freertext	0.511	vsm_allCui	0.420
lsa_freertextnumFactors300	0.429	umls_allCui	0.435	vsm_filteredCui	0.404
lsa_freertextnumFactors320	0.423	umls_filteredCui	0.423	kl_allCui_topics50	0.093
lsa_freertextnumFactors240	0.422	vsm_allCui	0.420	kl_allCui_topics100	0.088
lsa_freertextnumFactors180	0.421	vsm_filteredCui	0.404	kl_freertext_topics50	0.088
vsm_allCui	0.420			kl_freertext_topics100	0.086
lsa_allCuiumFactors320	0.414			kl_freertext_topics200	0.084
lsa_freertextnumFactors220	0.413			kl_allCui_topics150	0.083
lsa_freertextnumFactors280	0.411			kl_freertext_topics250	0.083

Table 6.1: Top 10 mean average precisions for the LSA, UMLS-QE, and LDA-based KL divergence experiments on the lupus cohort.

<b>LDA-QL</b>	<b>MAP</b>	<b>LDA-QE</b>	<b>MAP</b>
vsm_freertext	0.511	qe_allCui_topics200_terms70	0.569
ql_filteredCui_topics50_lambda_0.0	0.485	qe_allCui_topics200_terms140	0.569
ql_filteredCui_topics50_lambda_0.1	0.467	qe_allCui_topics200_terms110	0.568
ql_filteredCui_topics100_lambda_0.0	0.466	qe_allCui_topics200_terms90	0.567
ql_filteredCui_topics500_lambda_0.0	0.465	qe_allCui_topics200_terms120	0.567
ql_filteredCui_topics250_lambda_0.0	0.459	qe_allCui_topics200_terms80	0.567
ql_filteredCui_topics300_lambda_0.0	0.458	qe_allCui_topics200_terms50	0.566
ql_filteredCui_topics200_lambda_0.0	0.457	qe_allCui_topics200_terms130	0.566
ql_filteredCui_topics50_lambda_0.2	0.449	qe_allCui_topics200_terms60	0.566
ql_filteredCui_topics100_lambda_0.1	0.447	qe_allCui_topisc200_terms150	0.566

Table 6.2: Top 10 mean average precisions for the LDA-based query likelihood and LDA-based query expansion experiments on the lupus cohort.

<b>LSA</b>	<b>MAP</b>	<b>UMLS-QE</b>	<b>MAP</b>	<b>KL</b>	<b>MAP</b>
lsa_freertext_numFactors280	0.367	umls_freertext	0.454	vsm_freertext	0.323
lsa_freertext_numFactors240	0.366	vsm_freertext	0.323	vsm_allCui	0.271
lsa_freertext_numFactors320	0.363	vsm_allCui	0.271	vsm_filteredCui	0.260
lsa_freertext_numFactors260	0.363	umls_allCui	0.270	kl_allCui_topics50	0.101
lsa_freertext_numFactors250	0.360	vsm_filteredCui	0.260	kl_freertext_topics100	0.096
lsa_freertext_numFactors140	0.353	umls_filteredCui	0.246	kl_allCui_topics100	0.095
lsa_freertext_numFactors160	0.352			kl_freertext_topics350	0.094
lsa_freertext_numFactors220	0.352			kl_freertext_topics150	0.094
lsa_freertext_numFactors180	0.351			kl_freertext_topics200	0.094
lsa_freertext_numFactors200	0.343			kl_freertext_topics50	0.093

Table 6.3: Top 10 mean average precisions for the LSA, UMLS-QE, and LDA-based KL divergence experiments on the CHF cohort.

<b>LDA-QL</b>	<b>MAP</b>	<b>LDA-QE</b>	<b>MAP</b>
ql_freertext_topics200_lambda.0.0	0.460	qe_allCui_topics150_terms150*	0.500
ql_freertext_topics500_lambda.0.0	0.452	qe_allCui_topics300_terms30	0.500
ql_freertext_topics250_lambda.0.0	0.449	qe_allCui_topics150_terms130	0.497
ql_freertext_topics150_lambda.0.0	0.448	qe_allCui_topics150_terms100	0.496
ql_freertext_topics100_lambda.0.0	0.447	qe_allCui_topics150_terms140	0.495
ql_freertext_topics350_lambda.0.0	0.441	qe_allCui_topics150_terms90	0.492
ql_freertext_topics400_lambda.0.0	0.440	qe_allCui_topics150_terms120	0.492
ql_freertext_topics450_lambda.0.0	0.433	qe_allCui_topics150_terms80	0.492
ql_freertext_topics300_lambda.0.0	0.428	qe_allCui_topics150_terms60	0.491
ql_filteredCui_topics200_lambda.0.0	0.425	qe_allCui_topics450_terms80	0.488

Table 6.4: Top 10 mean average precisions for the LDA-based query likelihood and LDA-based query expansion experiments on the CHF cohort. The asterik indicates a search configuration where its performance was ( $p=0.03$ ) significant over the vector space approach. The standard deviation for 150 topics with 150 terms was 0.161; whereas, the standard deviation for 300 topics with 30 terms was 0.256.

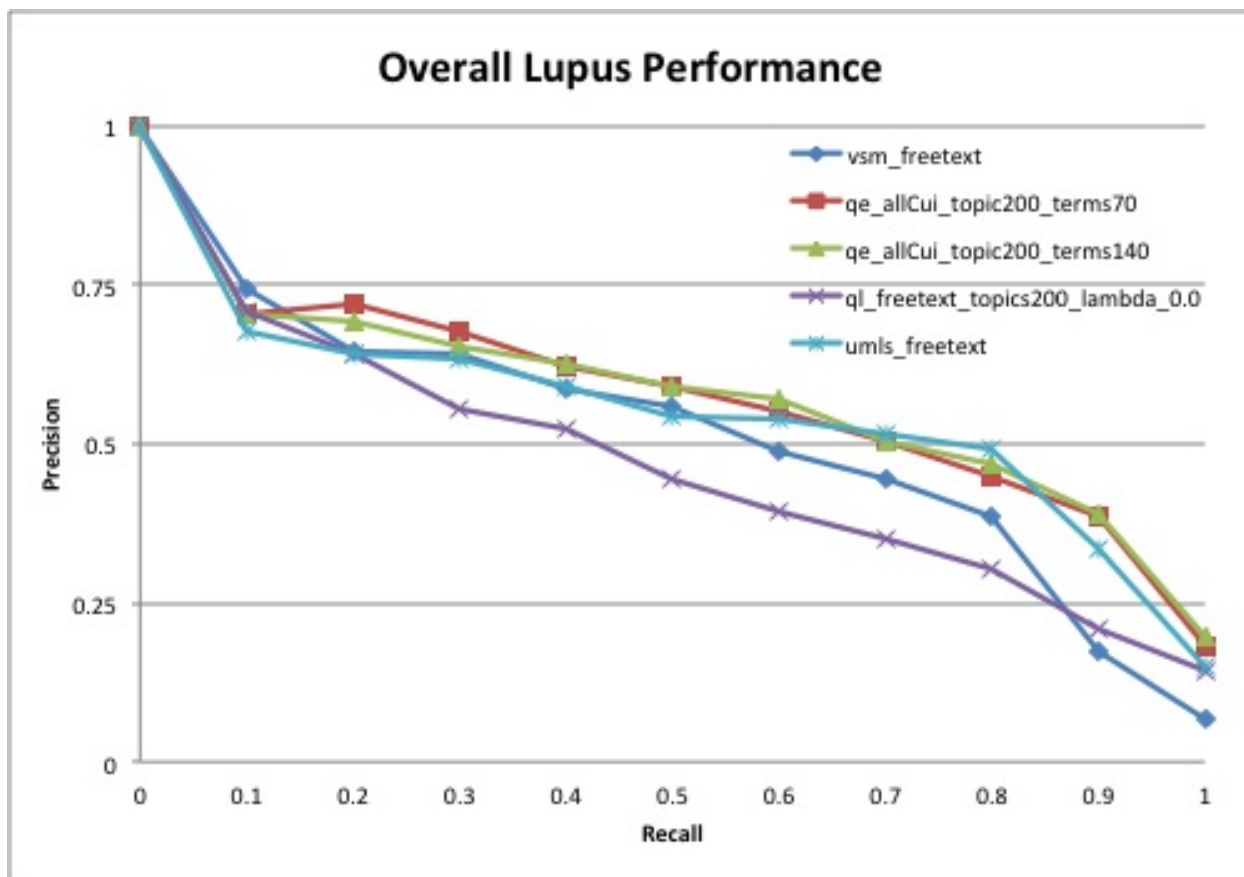


Figure 6.2: 11-point precision graph for the top performing search approaches on the lupus cohort. The average precision is calculated for each recall point.

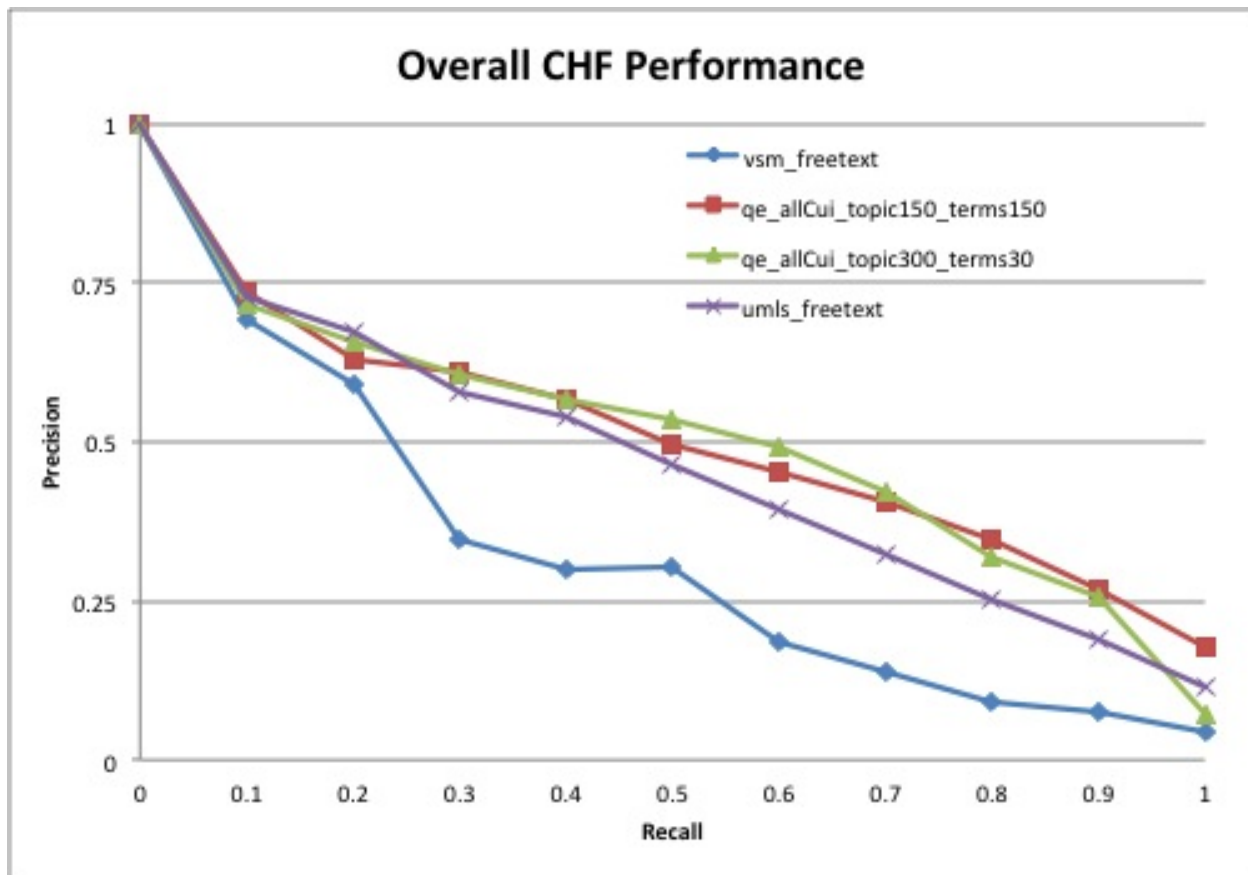


Figure 6.3: 11-point precision graph for the top performing search approaches on the CHF cohort. The average precision is calculated for each recall point.

### 6.2.2 Breakdown of Results

Tables 6.7 and 6.8 are a breakdown of the average precisions for each information need for the top performing query likelihood and LDA-based query expansion approaches. These approaches considerably outperformed VSM when handling medication related information needs. Overall, these search configurations achieved a recall of 1 for most of the information needs, meaning all relevant documents were retrieved; however, many non-relevant document were retrieved as well, resulting in lower precision. In order to further understand the performance of these approaches, the topic clusters created by LDA were examined. Tables 6.9 and 6.10 present the top ten terms of the topics most related to each information need. Tables 6.5 and 6.6 are presented to remind

the readers of the information needs for each cohort and to use them as a reference for the other tables.

Query No.	CHF Information Needs	Query Type
1	Is there documentation that suggests the severity of the patient's heart failure?	Disease
2	Has the patient ever had a MI?	Disease
3	Has the patient been admitted for CHF?	Disease
4	Is there documentation that suggests that the patient has been diagnosed with CHF?	Disease
5	Is there documentation that the patient has valvular disease?	Disease
6	Is there documentation that suggests the patient had a stroke or had signs of cerebral ischemia?	Disease
7	Have results of an echocardiogram been documented?	Finding
8	Is there documentation that suggests lower extremity edema?	Finding
9	Does this patient have a history of smoking?	Finding
10	Has this patient experienced shortness of breath?	Finding
11	Does this patient have an arrhythmia?	Finding
12	Is or has the patient been on statins?	Medication
13	Is or has this patient been on any anticoagulant medication?	Medication
14	Is or has this patient been on diuretics?	Medication
15	Is or has this patient been on beta-blockers?	Medication

Table 6.5: This table shows the semantic type of the CHF information needs.



Query No.	Lupus Information Needs	Query Type
1	Has an antinuclear antibody test been done on this patient?	Diagnostic Procedure
2	Has the inflammation rate in the body been measured?	Diagnostic Procedure
3	Has the patient had a biopsy done on any organs?	Diagnostic Procedure
4	Has the patient been anemic?	Disease
5	Is there any documentation that suggests when this patient was diagnosed with lupus?	Disease
6	Does this patient have drug-induced lupus?	Disease
7	Has the patient had kidney involvement?	Finding
8	Has the patient experienced joint pain or joint swelling?	Finding
9	Has the patient experienced hair loss?	Finding
10	Has the patient had unexplained fevers?	Finding
11	Has the patient had skin rash?	Finding
12	Has the patient experienced chest-pain?	Finding
13	Is or has the patient been on nonsteroidal anti-inflammatory drugs?	Medication
14	Is or has the patient been on anti-malarial medications?	Medication
15	Is or has the patient been on any corticosteroids or immunosuppressive medications?	Medication

Table 6.6: This table shows the semantic type of the lupus information needs.

Query Type	CHF Query No.	VSM FreeText	UMLS QE FreeText	QL FreeText 200 Topics & $\lambda=0.0$	QE AllCui 150 Topics & 150 Terms	QE AllCui 300 Topics & 30 Terms
Disease	1	0.213 $\pm$ 0.23	0.437 $\pm$ 0.26	0.342 $\pm$ 0.20	0.533 $\pm$ 0.233	<b>0.584 <math>\pm</math> 0.27</b>
	2	0.400 $\pm$ 0.42	0.547 $\pm$ 0.25	0.532 $\pm$ 0.13	0.595 $\pm$ 0.24	<b>0.662 <math>\pm</math> 0.26</b>
	3	<b>0.692 <math>\pm</math> 0.30</b>	0.511 $\pm$ 0.36	0.587 $\pm$ 0.29	0.532 $\pm$ 0.35	0.634 $\pm$ 0.33
	4	<b>0.889 <math>\pm</math> 0.16</b>	0.624 $\pm$ 0.33	0.574 $\pm$ 0.35	0.648 $\pm$ 0.25	0.641 $\pm$ 0.24
	5	<b>0.523 <math>\pm</math> 0.50</b>	0.386 $\pm$ 0.14	0.385 $\pm$ 0.19	0.385 $\pm$ 0.15	0.061 $\pm$ 0.04
	6	0.144 $\pm$ 0.24	0.485 $\pm$ 0.22	0.523 $\pm$ 0.34	<b>0.589 <math>\pm</math> 0.18</b>	0.587 $\pm$ 0.19
Finding	7	0.206 $\pm$ 0.38	0.727 $\pm$ 0.30	0.786 $\pm$ 0.24	0.782 $\pm$ 0.19	<b>0.869 <math>\pm</math> 0.20</b>
	8	0.331 $\pm$ 0.26	0.271 $\pm$ 0.11	0.232 $\pm$ 0.23	<b>0.343 <math>\pm</math> 0.15</b>	0.103 $\pm$ 0.08
	9	0.043 $\pm$ 0.13	<b>0.532 <math>\pm</math> 0.14</b>	0.321 $\pm$ 0.08	0.351 $\pm$ 0.09	0.333 $\pm$ 0.12
	10	<b>0.608 <math>\pm</math> 0.36</b>	0.577 $\pm$ 0.25	0.498 $\pm$ 0.27	0.431 $\pm$ 0.15	0.604 $\pm$ 0.26
	11	0.259 $\pm$ 0.41	0.259 $\pm$ 0.15	<b>0.531 <math>\pm</math> 0.16</b>	0.209 $\pm$ 0.11	0.291 $\pm$ 0.11
Medication	12	0.189 $\pm$ 0.40	0.461 $\pm$ 0.34	0.413 $\pm$ 0.14	0.690 $\pm$ 0.16	<b>0.801 <math>\pm</math> 0.16</b>
	13	0.059 $\pm$ 0.24	0.148 $\pm$ 0.24	0.111 $\pm$ 0.23	<b>0.303 <math>\pm</math> 0.26</b>	0.190 $\pm$ 0.23
	14	0.040 $\pm$ 0.20	0.265 $\pm$ 0.24	0.477 $\pm$ 0.24	0.474 $\pm$ 0.23	<b>0.587 <math>\pm</math> 0.25</b>
	15	0.250 $\pm$ 0.43	0.585 $\pm$ 0.31	0.597 $\pm$ 0.20	<b>0.639 <math>\pm</math> 0.30</b>	0.558 $\pm$ 0.36

Table 6.7: This table shows the average precision for the top performing search approaches for each of the CHF information needs.

Query Type	Lupus Query No.	VSM FreeText	UMLS QE FreeText	QL FilteredCui 50 Topics & $\lambda=0.0$	QE AllCui 200 Topics & 70 Terms	QE AllCui 200 Topics & 140 Terms
Diagnostic Procedure	1	0.803 $\pm$ 0.30	0.541 $\pm$ 0.20	0.751 $\pm$ 0.14	0.865 $\pm$ 0.18	<b>0.918</b> $\pm$ 0.13
	2	0.039 $\pm$ 0.07	0.261 $\pm$ 0.07	0.290 $\pm$ 0.08	<b>0.393</b> $\pm$ <b>0.16</b>	0.362 $\pm$ 0.19
	3	0.268 $\pm$ 0.38	<b>0.610</b> $\pm$ <b>0.19</b>	0.254 $\pm$ 0.14	0.453 $\pm$ 0.19	0.417 $\pm$ 0.19
Disease	4	0.538 $\pm$ 0.51	0.596 $\pm$ 0.34	0.373 $\pm$ 0.27	0.608 $\pm$ 0.26	<b>0.639</b> $\pm$ <b>0.23</b>
	5	0.549 $\pm$ 0.21	0.767 $\pm$ 0.20	0.619 $\pm$ 0.11	0.812 $\pm$ 0.15	<b>0.839</b> $\pm$ <b>0.18</b>
	6	<b>1</b> $\pm$ 0.0	0.421 $\pm$ 0.08	0.019 $\pm$ 0.01	0.082 $\pm$ 0.04	0.117 $\pm$ 0.07
Finding	7	0.716 $\pm$ 0.33	0.606 $\pm$ 0.26	0.218 $\pm$ 0.06	0.793 $\pm$ 0.27	<b>0.839</b> $\pm$ <b>0.28</b>
	8	0.735 $\pm$ 0.28	0.772 $\pm$ 0.19	0.758 $\pm$ 0.24	<b>0.809</b> $\pm$ <b>0.27</b>	0.799 $\pm$ 0.24
	9	0.578 $\pm$ 0.40	<b>0.929</b> $\pm$ <b>0.13</b>	0.561 $\pm$ 0.17	0.518 $\pm$ 0.15	0.509 $\pm$ 0.11
	10	<b>0.654</b> $\pm$ <b>0.28</b>	0.208 $\pm$ 0.04	0.324 $\pm$ 0.24	0.295 $\pm$ 0.16	0.290 $\pm$ 0.17
	11	0.825 $\pm$ 0.25	<b>0.850</b> $\pm$ <b>0.14</b>	0.801 $\pm$ 0.12	0.827 $\pm$ 0.10	0.818 $\pm$ 0.10
	12	0.785 $\pm$ 0.24	<b>0.785</b> $\pm$ <b>0.17</b>	0.372 $\pm$ 0.26	0.234 $\pm$ 0.13	0.214 $\pm$ 0.11
Medication	13	0.175 $\pm$ 0.38	0.453 $\pm$ 0.26	0.505 $\pm$ 0.07	<b>0.593</b> $\pm$ 0.18	0.557 $\pm$ 0.14
	14	0 $\pm$ 0	0.161 $\pm$ 0.18	<b>0.662</b> $\pm$ <b>0.21</b>	0.644 $\pm$ 0.12	0.644 $\pm$ 0.14
	15	0 $\pm$ 0	0.218 $\pm$ 0.05	<b>0.774</b> $\pm$ <b>0.13</b>	0.606 $\pm$ 0.18	0.568 $\pm$ 0.20

Table 6.8: This table shows the average precision for the top performing search approaches for each of the lupus information needs.

Query Type	Lupus No.	Topic Label	Top Ten Terms
Diagnostic	1	Labs	Negative; Antibodies; AB hearing assessment list; AB - Zebrafish; AB Term Type; Of Each; Antibodies; Antinuclear; Antinuclear Antibody Assay; nanolitre; Netherlands
	2	Biopsy or Pathology Related	Biopsy; Biopsy Domain; Specimen Source Codes - Biopsy; Consent Type - biopsy; biopsy characteristics; Box Dosing Unit; Increase; What subject filter - Status; Focal; Status
3			
Disease	4	Blood Labs	Erythrocyte sedimentation rate measurement; Extended Rotated Sidebent; erythrocyte sedimentation rate result; Electron Spin Resonance Spectroscopy; Iron measurement; Iron; Dietary Iron; Genus Anemia; Anemia; C-reactive protein
	5	Lupus Related	Lupus Erythematosus, Systemic; Lupus Erythematosus, Discoid; Lupus Erythematosus; Lupus Vulgaris; Flare; Rheumatologist; Systemic; Plaquenil; Rheumatology specialty; Prednisone
	6	Seizure Related	Seizures; Seizure Adverse Event; Keppra; Electroencephalography; Electroencephalographic Patterns; EEG - Technician, Other - NUCCProviderCodes; Specialist / Technologist - EEG; Epilepsy; BID protein; Twice a day
7			
Finding	12	Follow-up Mgmt	Follow-Up; Follow-Up Report; Follow-up status; Physicians; Appointments; Instructions; Week; Question (inquiry); Decision; Body Fluid Discharge
	10	GI related	Diarrhea; Diarrhea Adverse Event; Nausea; Nausea Adverse Event; Vomiting; Fever; Vomiting Adverse Event; Abdominal Pain; day; Symptoms
	8	Lupus Symptoms and Treatment	Exanthema; Prednisone; Arthralgia; Plaquenil; week; Rheumatologist; month; Rheumatology specialty; Alopecia; Synovitis
	9		
11			
15			
Medication	13	UNK	Cerebral Palsy; CP protocol; centipoise; Propionibacterium acnes; Cleft Palate; cyclophosphamide/prednisone; ABD tumor staging notation; CDISC SDTM Not Done Terminology; Naturopathic Doctor; Norrie disease
	14	UNK	Sierra Leone; Encephalitis, St. Louis; Lupus Erythematosus, Systemic; Prednisone; Plaquenil; Hypertensive disease; Rheum; Flare; Nephritis; Woman

Table 6.9: This table shows the top ten terms of the topic cluster determined to be most similar to each of the lupus information needs for 200 topics. The ‘Topic Label’ column is a manual label identified by a clinical expert after examining the top 20 terms. ‘UNK’ means that a topic label could not be determined from examining the top terms.

Query Type	CHF No.	Topic Label	Top Ten Terms
Finding	7	Echo Report Terms	Mild Adverse Event; Normal assessment finding; Skin appearance normal (finding); Moderate Adverse Event; physiological aspects; Mitral Valve Insufficiency; Pericardial effusion; Pericardial effusion body substance; Tricuspid Valve Insufficiency; Entire right ventricle
	8	LEE & DVT Sx	Edema; Edema:Finding:Point in time:Patient:Ordinal; Lower Extremity; Leg; Lewis Blood-Group System; Entire lower limb; Entire lower leg, from knee to ankle; Cellulitis; Swelling; Deep Vein Thrombosis
	9	Cholesterol Related	Serum LDL cholesterol measurement; Low-Density Lipoproteins; Hypertensive disease; Lipids; Serum HDL cholesterol measurement; Serum total cholesterol measurement; High Density Lipoproteins; BID protein; 4-azido-7-phenylpyrazolo-(1,5a)-1,3,5-triazine; Lipitor
	10	Pulmonary and TB Related	Coughing; Cough Adverse Event; Fever; Dyspnea; Peptide Nucleic Acids; Pneumonia; Lung; Tuberculosis; Purified Protein Derivative of Tuberculin; tributyrin
	11	Cardiac Vascular Related	Bruit; Heart; Aspirin; Myocardial Infarction; Dipyridamole; Angioplasty, Transluminal, Percutaneous Coronary; History of myocardial infarction within 6 mo; Cardiac Catheterization Procedures; Patient; Cardiology studies
Disease	1 3 4	CHF Related	Congestive heart failure; Lasix; BID protein; Coreg Butoxamine HCl; Coreg; Diuresis; Dyspnea; Body weight:Mass:Point in time:Patient:Quantitative; Edema:Finding:Point in time:Patient:Ordinal; Edema
	5	Heart Valve Related	doxorubicin/vincristine protocol; Surgical repair; aVR; Anteroventral Thalamic Nucleus; Heart; Operative Surgical Procedures; anatomic valve; Adverse reactions; Surgical Replantation; methotrexate/vinblastine protocol
	6	Cerebrovascular Related	Cerebrovascular accident; cyclophosphamide/doxorubicin/vincristine protocol; Hypertensive disease; Congestive heart failure; Medical History; hypercholesterolemia; Hypercholesterolemia result; Weakness; Asthenia; HTN Adverse Event
	2	MI related	Anterior descending branch of left coronary artery; LEUKOCYTE ADHESION DEFICIENCY, TYPE I; Myocardial Infarction; History of myocardial infarction within 6 mo; Aspirin; Plavix; Chest Pain; Troponin; Electrocardiography; Heart
Medication	12	Medical History Related	History of previous events; Medical History; Hypertension Adverse Event; Blood pressure finding; Blood pressure determination; Admission activity; Emergency Situation; Act Code - emergency; Blood Pressure; past medical history
	14	UNK	Blood Clot; Milrinone; Tachycardia, Ventricular; Tidal Volume; Amiodarone; Male gender; Primary idiopathic dilated cardiomyopathy; Thrombus; I Blood-Group System; 3',5'-dichloromethotrexate
	15	Cardiac Related Test	Normal assessment finding; Skin appearance normal (finding); Scanning; Stress; Exercise Pain Management; Adenosine; Tomography, Emission-Computed, Single-Photon; Perfusion (procedure); Radionuclide Imaging; THYROID HORMONE PLASMA MEMBRANE TRANSPORT DEFECT

Table 6.10: This table shows the top ten terms of the topic cluster determined to be most similar to each of the CHF information needs for 150 topics. The ‘Topic Label’ column is a manual label identified by a clinical expert after examining the top 20 terms. ‘UNK’ means that a topic label could not be determined from examining the top terms.

### 6.3 Discussion

In this comparative analysis, we found that not all semantic-based approaches outperformed the traditional, vector-space model approach in retrieving relevant documents. LSA performed best on the freetext corpus; however, it did not perform better than VSM. LSA does well at finding latent topics, but, traditionally, it has had mixed IR results in TREC [Atreya and Elkan, 2011; Dumais, 1995]. Our ontology-based query expansion approach showed improved performance over VSM, which is consistent with most research findings for medical literature retrieval [Hersh *et al.*, 2000; Yang and Chute, 1994; Aronson *et al.*, 1994; Aronson and Rindflesch, 1997; Liu and Chu, 2007]. UMLS-QE identified synonyms that VSM missed, such as “bx” for biopsy or “hx” for history.

With the LDA-based topic modeling approaches, KL was the only one that did not outperform VSM on any of the three gold standard corpora for each cohort. Its low performance is consistent with other evaluations of incorporating topic modeling into information retrieval [Zhai, 2008]. The other two LDA-based approaches performed well on all the corpora. The majority of top performing query likelihood configurations for both cohorts were when  $\lambda=0.0$ , meaning that the retrieval was solely based on the LDA model and not at all on the corpus level language model. This is contrary to other evaluations where  $\lambda$  was found to be closer to 0.7 [Yi, 2009; Wei and Croft, 2006; Heinz, 2007]. A possible explanation for this difference could be due to the domain of the corpus. The other evaluations were using TREC data sets, which are composed of documents on various subject areas. This makes LDA topic clusters “too coarse”; whereas, our evaluation was on a specialized set of documents from two disease groups, making the LDA topic clusters more apt than the corpus-level language model. These specialized topic clusters allowed the LDA-QE approach to have the best overall performance, specifically on medication related questions. For medication related questions, most of the time LDA-QE was able to match the information need to a topic cluster that contained appropriate medications. For example, lupus information need #14, which examined if a patient was on anti-malarial medications, was mapped to the topic cluster that contained anti-malarial drugs - hydroxychloroquine (generic name) and Plaquenil (brand name). The reason UMLS-QE did not perform as well on medication queries was because the UMLS did not contain specific drug

names along its “is-a” relationships, such as Plaquenil.

The other aspect where LDA topic clusters improved retrieval was where medical knowledge was required. In some cases, the LDA algorithm was able to group concepts that require medical knowledge to associate them. For example, one of the topic clusters was able to associate the CUI for “facial droop” to “stroke,” which allowed LDA-QE to retrieve a document that most other search methods missed. This reveals the potential that LDA topic modeling has on the medical domain. Using SNOMED-CT the shortest path between the two concepts using the “is-a” link was seven:

C0038454 (Cerebrovascular accident, NOS) → C0007820 (Disorder Cerebrovascular) →  
C0042373 (Diseases, Vascular) → C0425654 (Blood vessel finding) →  
C0424722 (General finding of soft tissue) → C0427052 (Finding of power of skeletal muscle) →  
C0151786 (Weakness, Muscle) → C0427055 (Paresis, Facial)

Examining other relationship types within SNOMED-CT, the path between the two concepts was also long and unclear, making it difficult for any general ontology algorithm to find the connection between the two concepts.

Though the topic clusters created allowed LDA-QE to outperform other search methods, LDA-QE required a high number of expansion terms from the topic clusters to achieve this performance. This was due to the concept mapping technique applied on the processed clinical notes. As mentioned in the vector-space error analysis study, the Filtered CUI and All CUI corpora were created by extracting the CUIs from the processed clinical notes; the program that mapped text to the UMLS did not disambiguate between matched concepts. For example, lupus information need #13’s topic cluster contained “Cerebral Palsy,” “cleft palate,” “CP protocol,” and “centipoise,” all of which were CUI matches for the abbreviation “CP” found in the original notes. In this case, the mapping program did not properly identify the abbreviation for “chest pain” and including four unrelated terms that make the topic cluster unclear. The hypothesis that the lack of term disambiguation was the cause of the high number of terms needed for LDA-QE to perform well was confirmed by examining LDA-QE on the Free-text corpus. On this corpus, the best performing LDA-QE

approach, which also outperformed VSM, only required five expansion terms.

Besides the semantic approach, the corpus utilized affected search performance. As found in the previous study, the LDA-based search method performed better on the All CUI corpus over the Filtered CUI corpus, which reinforces the idea that there is no one terminology that contains all clinically relevant concepts. Thus, a semantic-based search should consider incorporating multiple terminologies into the retrieval process.

### 6.3.1 EHR Implication

Besides being the best performing search, the benefit of the LDA-QE approach over the other semantic approaches is that it is easier to operationalize. The LDA created topic clusters are used as a thesaurus; therefore, the terms in the topic clusters can be indexed separately from the document index and used during query time [Park and Ramamohanarao, 2009; Park and Ramamohanarao, 2008]. Having a separate index allows for LDA models to be refreshed without affecting the original term-document index. Other semantic approaches either require re-indexing when new documents are added (i.e., LSA), or are more time and disk storage intensive when having to retrain topic models on a corpus (i.e., LDA-QL).

Depending on the retrieval task, how the data is stratified and sampled requires much consideration when training LDA models. Training models on homogenous data sets create more granular, cohesive clusters as opposed to heterogeneous data sets. A general purpose model trained from heterogenous data could be used to select or classify a patient, such as in cohort identification. With disease-specific models, an LDA-QE based search system could appropriately expand the user's query based on the patient's disorder. The correct method to stratify and sample patient data, especially when dealing with patient's with comorbidities, is an open area of research and outside the scope of this work.



### 6.3.2 Limitations

A limitation of our evaluation was due to our gold standard. Search performance is query dependent, even within query types. Thus, when intrinsically evaluating a search engine on clinical records, to better generalize the results of a search evaluation a larger query set and corpus must be utilized.

Another limitation was with the concept mapping program used in our evaluation. The mapping technique did not disambiguate terms, adding “noise” to the topic clusters. As a result of the the noisy topic clusters, the LDA-QE approach required a high number of query expansion terms, and since concept terms existed in groups, relevant documents were artificially ranked higher in the All CUI corpus versus the Filtered CUI corpus. One possibility of handling the “noisy” topic clusters that was not evaluated in this study is re-weighting the query terms using an ontology so that semantically relevant terms are weighted higher than extraneous ones. Though it was found that the performance of any semantic-based search is very dependent on the concept mapper used on the clinical notes, we are only concerned with the optimal configuration for a semantic-based search for this evaluation. For operationalizing a semantic search in the future, one should take into careful consideration the NLP system employed.

## Chapter 7

# Conclusion and Future Work

Clinicians, both in the inpatient and ambulatory setting, are faced with an ever-growing amount of digitized clinical data that they must manage. Navigating through this clinical data can be cumbersome and potentially detrimental to patient care. Search functionality within the EHR is one potential solution to this problem.

This dissertation aimed at understanding search within the EHR and at evaluating the value of a semantic approach for an EHR-based search engine. This research primarily focused on a system-oriented evaluation of various search methods on clinical narrative. To assess the performance of the semantic search methods, a gold standard was created using clinically derived information needs and real patient notes. The semantic approaches evaluated were a geometric reduction approach called latent semantic analysis, an ontological approach that used the UMLS, and various topic modeling approaches that used latent Dirichlet allocation to cluster topics. All semantic methods were compared against a baseline vector-space model approach. Our evaluation demonstrated that a hybrid topic modeling and vector space model approach outperforms other all semantic search methods.

The studies in this dissertation make primary contributions to the field of clinical informatics and secondary contributions to the general information retrieval community through the potential use

of the gold standard. We discuss each aspect next.

## 7.1 A Gold Standard Dataset for a Within-Patient EHR Search Evaluation

To our knowledge, our gold standard is the first attempt to develop a data set for retrieval within a patient's electronic health record. In 2011, TREC for the first time included a medical records track, which focused on cohort identification for comparative effectiveness studies [Voorhees, 2011]. This is a very different task than the one addressed in this dissertation – identifying specific information within the health record. The annotations for our gold standard were done at the paragraph and document level; whereas, the TREC data were carried out at the patient visit level. By including paragraph level annotations, relevance based on location within a document could be used as an evaluation criteria.

Developing a gold standard is a time and resource intensive task. The development of this data set took the majority of the time spent on this dissertation. It entailed: acquiring proper access and identifying clinically rich patient records, creating an application for annotating the clinical notes, recruiting and training participants, and validating the annotations. Acquiring access to clinical data requires much scrutiny due to privacy issues, and as such, identifying and subsequently retrieving the clinical notes was the most challenging step in the gold standard creation. To hasten the process of acquiring data, we investigated alternative data sources that were publicly available and de-identified. Currently, there are two repositories online – the MIMIC II data set from Harvard and the other from the University of Pittsburgh's Biomedical Language Understanding Lab (BLULab) [Saeed *et al.*, 2011]. For security purposes, both data sets require signed data use agreements. The MIMIC II data is predominately ICU data and thus did not contain a diverse set of document types for our purposes. Even though the BLULab data is used by the TREC medical track for its evaluation, we were unable to obtain a copy of the clinical notes for our research. After multiple attempts to acquire the BLULab data, we resolved to creating our gold standard from

clinical notes found at our institution. We believe that the resulting gold standard developed in this dissertation is a realistic and rich dataset for information retrieval and as well as for other informatics research. We hope to make this data set publicly available to the informatics and general information retrieval communities in order to bolster IR research within the domain of the EHR.

Since there is no other way than the creation of a gold standard to intrinsically evaluate search, a more ambitious work would be to expand this gold standard to include a larger sample of random patients and information needs. This would be a large endeavor, but would result in a data set that can be used to make general claims on search performance. For example, we can evaluate the need to train special LDA models for particular diseases or to train one general model for all patients.

## 7.2 A Within-Patient Deployed Search Engine and its Use

The research conducted in this dissertation provides several application-oriented results towards the field of clinical informatics. First, our implementation of a real search engine within the EHR was one of the first documented retrieval tools within the EHR, and as such makes it a novel endeavor. Second, the knowledge acquired from studying users' search queries provided us a model of who uses such a system and what information they seek. It was discovered that clinical users predominately search for labs and diseases within clinical narrative. Surprisingly, it was discovered that non-clinical users, such as billing coders, also accessed the system to find information within the record. The last major application-oriented contribution was derived from our error analysis work, which categorized the shortcomings of a traditional, string-matching search algorithm. The main categories where retrieval errors arise were simple synonyms, ontological similarity (i.e., nodes along the "is-a" relationship), and implicit references (i.e., "slight left facial droop" suggesting a stroke).

The needs assessment studies revealed that clinical users strongly support the idea of search functionality within the EHR. The studies also showed that though users differed in their clinical role,

they all predominantly performed informational searches related to laboratory results and specific diseases, suggesting that search engine results could be specially tailored for these types of queries. For example, when searching for specific lab results, the results pages could not only contain documentation of the labs, but a graph of the lab values for the past 48 hrs. Besides seeking specific information, healthcare professionals also used the EHR search engine as a navigational aid within the electronic health record (i.e., navigating to a patient's labs or switching patients), suggesting the need for shortcuts within the EHR.

### 7.3 Semantic Approach to Within-Patient EHR Search

It was shown that search users tend to only enter one to two query terms when seeking information [Natarajan *et al.*, 2010]. This makes it challenging for a simple VSM approach to capture all relevant documents for a user's information need, which is essential for clinicians in order to make informed patient care decisions. A semantic search has the potential to address this gap.

The main contribution of this dissertation is the systematic assessment of various semantic search approaches on clinical narratives within a patient's health record. This evaluation demonstrated that a hybrid topic modeling and VSM approach improves search performance over other semantic search approaches. The benefit of this approach is that it can be easily incorporated into a traditional VSM search engine as a thesaurus lookup. Therefore, there is minimal custom modification needed to integrate it into an open-source VSM search engine, such as Lucene. The main modification required is for query translation. The proper mapping of a query to a topic or semantic type can be difficult and is therefore beyond the scope of this dissertation. However, from our log analysis study, we do know that a number of user queries were part of words (i.e., "hyper"), implying the need for a free-text module during query translation to be included in any semantic search approach.

## 7.4 Limitations and Lessons Learned

There are a few limitations in this dissertation that must be restated regarding the needs assessment, gold standard, and semantic search evaluation. As mentioned in the log analysis study, logs are only a proxy to understanding a user's information need and do not provide a complete picture. Therefore, a user survey was conducted. Our initial intention was to survey individuals at search time to better understand their information needs. However, this was not feasible due to logistic issues of administrating a survey within a production EHR system. Thus, we employed an electronic survey outside of the EHR that targeted all clinical users, regardless of their use of our search engine within the EHR - CISearch. Since these users were not familiar with our search engine, it was difficult to understand users' search information needs. Even with the help of key clinical users, recruiting individuals to participate in the survey was very challenging since we did not provide any remuneration. In any future survey studies, I have learned to include incentives to recruit participants.

As mentioned above, creating a gold standard is difficult task, so in order to make the task more manageable, our gold standard was manually created using two disease cohorts. Though the data set was representative of the two diseases and their information needs, its size is small compared to data sets for general IR evaluations such as TREC, which contains millions of documents and hundreds of information needs depending on the track. However, TREC medical records track only contained 35 information needs, which is comparable in size to the 30 that we developed. The main drawback of the size of our gold standard is that generalizing results from our studies beyond these diseases is not possible, making it difficult to state that one search approach is the best for all EHR systems.

Finally, the nature of the intrinsic evaluation conducted in the last study focused on semantic search approaches that improve content retrieval. It does not provide a complete picture of what search method is best for searching within a patient's health record from a user's perspective. To access the utility of a search engine, a task based evaluation, which examines the system's usability and

users' satisfaction, must be conducted [Hersh, 2009; Manning *et al.*, 2008]. As stated earlier, the objective of this dissertation was to evaluate different algorithmic search methods, and as such, a user-oriented evaluation was outside the scope of this work.

## 7.5 Future Work

### 7.5.1 Gold Standard

Creating a gold standard is a time and resource intensive task and using our gold standard only for this dissertation would be a loss to the general IR domain. One of the goals post-dissertation is to share the gold standard with the informatics community in order to cultivate the growing interest in IR work on clinical narrative. All the notes in the gold standard have been processed using a modified de-identification script provided by Harvard [Douglass *et al.*, 2004]. Afterwards, I manually examined every note and removed identifiable information missed by the script. The next steps that are underway for the gold standard are that two faculty members will review the de-identified notes sequentially making sure that no identifiable information is missed, and then the data set will be sent to New York-Presbyterian's security and quality committees for final approval. Even though the clinical notes are de-identified, security and privacy are a concern. Thus, the data cannot be anonymously downloaded and should require data use agreements (DUA). Hosting and the administrative overhead to maintain the data set is still an issue being examined. One possibility is to partner with other public data groups such as MIMIC and provide them the data to administer.

### 7.5.2 Search within the EHR

This work has shown that a hybrid topic modeling and VSM approach outperforms other semantic search methods when dealing with clinical narrative. There are many directions that this work can be expanded. An immediate possibility is to operationalize a semantic search. Since the prototype

version of CISearch was integrated into WebCIS, other clinical systems have rolled into production, reducing the usage on WebCIS. I am currently working with one of the new systems to implement search and have the support of the systems development team. The main benefit from this support is that we are able to index documents within the database, making retrieval time significantly faster compared to the in-memory search implemented in CISearch. Once this system goes live, we will have a live testbed to explore other search related research questions. Concurrently, we are conducting a multi-institutional search log analysis in order to expand and validate our initial log analysis findings.

In the clinical setting, it is important that clinical users have all the relevant information on a patient in order to make sound medical decisions. Our evaluation has shown that some semantic approaches both outperform VSM in terms of precision and also in recall. However, given the time constraints that clinicians face, it is essential that they are not burdened with reviewing non-relevant documents. Thus, high precision and recall are equally important and therefore an area where future work is planned. We will examine other methods to improve upon LDA-QE, such as incorporating relevance feedback and ontological knowledge.

One main research topic that was outside the scope of this dissertation is the user-oriented evaluation of an EHR search tool. By recruiting active users of the search tool, studies can be designed to examine how to best display search results based on the user's information need (i.e., display lab results differently from disease type questions, order results based on relevance, personalize results display, etc.). Another unanswered usability question to examine is how to relate results from a semantic search engine? With traditional VSM search engines, the convention of displaying search results is to highlight query terms that were found in relevant documents; however, with semantic results there can be relevant documents that do not contain any query terms, resulting in a cognitive disconnect between what the user types and what the system returns. Finally, a semantic search engine could provide a summary of the patient record based on a topic of interest (i.e., the clinician's query). For instance, this would allow a clinician to assess the status of a patient's disease, such as chronic kidney disease, by displaying all relevant information in reverse



chronological order as well as graphically displaying the patient's GRF, creatinine, and BUN values for the past 48-hrs.

### 7.5.3 Scenario of Use for a Semantic Search System

This section describes our vision of a fully functional semantic search system within the EHR through a use case scenario. Ultimately, from this work we are closer to an EHR search engine where all relevant information is presented to a clinical user at a timely fashion. For example, when a 50 year old male heart patient comes to his local emergency room complaining of chest pain, the physician on-call must get an accurate overview of the patient before treating him. Traditionally, the physician would get a quick medical history from the patient and scan the last few notes in the patient's record. With an ideal search utility, the system would conduct different search strategies based on the user's query. The physician could type "chest pain" into the system and get a query-based summary view of the patient's record. The system would know the patient was a cardiac patient based on clinical notes from previous visits and expand the query to contain cardiac information related to the semantic type of "chest pain." The resulting summary view would be multimodal, including snippets of narrative from all relevant clinical notes, a list of pertinent prescribed medications such as Tikosyn, a graphical chart of all recorded INR values, and a timeline of all cardiac related hospitalizations. The physician also would be able to drill down on any note or lab value to get more detailed information, such as a viewing an EKG report and its corresponding wave recording. From this flexible, high-level view of the patient's chart, the physician would quickly notice that the patient underwent a mitral valve replacement five years prior, which the patient failed to mention. Such missed information would be vital in treating a patient, and such a semantic search system could better aid clinical providers in their assessment of a patient and, ultimately, improve their clinical decision-making. An intelligent semantic system, such as this, is the vision of our work in this domain, and this dissertation sets the foundational efforts towards reaching this goal.

# Bibliography

Alias-i. LingPipe 4.1.0, 2008.

Mureen Allen, Leanne M Currie, Mark Graham, Suzanne Bakken, Vimla L Patel, and James J Cimino. The classification of clinicians' information needs while using a clinical information system. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 26–30, January 2003.

Apache Lucene - Query Parser Syntax, 2007.

A R Aronson and T C Rindflesch. Query expansion using the UMLS Metathesaurus. *Proceedings : a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, pages 485–9, January 1997.

Alan R. Aronson, Thomas C. Rindesch, and Allen C. Browne. Exploiting a Large Thesaurus for Information Retrieval. 1994.

Avinash Atreya and Charles Elkan. Latent semantic indexing (LSI) fails for TREC collections. *SIGKDD Explor. Newsl.*, 12(2):5–10, March 2011.

L. Azzopardi, M. Girolami, and C.J. van Rijsbergen. Topic based language models for ad hoc information retrieval. *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, pages 3281–3286.

Robert H Baud, Christian Lovis, Patrick Ruch, and Anne-marie Rassinoux. Conceptual Search in Electronic Patient Record. *Medical Informatics*, pages 156–160, 2001.

- David Bawden and Lyn Robinson. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2):180–191, 2008.
- S.M. M Beitzel, E.C. C Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–328. ACM New York, NY, USA, 2004.
- N.J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.
- J Bhogal, a Macfarlane, and P Smith. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886, 2007.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, May 2003.
- David Blumenthal. Stimulating the adoption of health information technology. *The New England journal of medicine*, 360(15):1477–9, April 2009.
- R B Bradford. An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications. *Performance Evaluation*, pages 153–162, 2008.
- Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
- Donald Owen Case. *Looking for information : a survey of research on information seeking, needs, and behavior*. Academic Press, Amsterdam ; New York, 2002.
- Herbert S Chase, David R Kaufman, Stephen B Johnson, and Eneida a Mendonca. Voice capture of medical residents’ clinical information needs during an inpatient rotation. *Journal of the American Medical Informatics Association : JAMIA*, 16(3):387–94, 2009.
- Elizabeth S Chen and James J Cimino. Automated discovery of patient-specific clinician information needs using clinical information system log files. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 145–149, 2003.

- C Chisnell, K Dunn, and D F Sittig. Determining educational needs for the biomedical library customer: an analysis of end-user searching in MEDLINE. *Medinfo. MEDINFO*, 8 Pt 2:1423–7, January 1995.
- Tom Christensen and Anders Grimsmo. Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP’s use of electronic patient records. *BMC medical informatics and decision making*, 8:12, January 2008.
- James J. Cimino and Edward H. Shortliffe. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Springer-Verlag, New York, New York, USA, 3rd edition, 2006.
- J J Cimino. Use, usability, usefulness, and impact of an infobutton manager. *AMIA Annu Symp Proc*, pages 151–155, 2006.
- Trevor Cohen and Dominic Widdows. Empirical distributional semantics: methods and biomedical applications. *Journal of biomedical informatics*, 42(2):390–405, 2009.
- Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- L M Currie, M Graham, M Allen, S Bakken, V Patel, and J J Cimino. Clinical information needs in context: an observational study of clinicians while using a clinical information system. *AMIA Annu Symp Proc*, pages 190–194, 2003.
- Berry De Bruijn, Simona Carini, Svetlana Kiritchenko, Joel Martin, and Ida Sim. Automated information extraction of key trial design elements from clinical trial publications. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 141–145, 2008.
- S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- M Douglass, GD Clifford, A Reisner, G.B. Moody, and R.G. Mark. Computer-assisted de-

- identification of free text in the MIMIC II database. In *Computers in Cardiology, 2004*, pages 341–344. IEEE, 2004.
- Susan T. Dumais. Latent Semantic Indexing (LSI): TREC-3 Report. In *Overview of the Third Text REtrieval Conference*, pages 219–230, 1995.
- Lars Eldén. *Matrix methods in data mining and pattern recognition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2007.
- Michael Elhadad and David Gabay. Automatic Evaluation of Search Ontologies in the Entertainment Domain using Text Classification. *Evaluation*, 2010.
- Noemie Elhadad and Kathleen R Mckeown. Towards generating patient specific summaries of medical articles. *Proc. of NAACL Workshop on Automatic Summarization*, 2001.
- Noémie Elhadad, Michael Elhadad, and Raphael Cohen. Redundancy in Electronic Health Record Corpora: Analysis, Impact on Text Mining Performance and Mitigation Strategies. 2012.
- Martin Eppler and Jeanne Mengis. The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society*, 20(5):325–344, 2004.
- Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, 2005.
- Sharon R. Lipsky Gorman and Noemie Elhadad. ClinNote and HealthTermFinder: A pipeline for processing clinical notes. Technical report, Columbia University, 2011.
- William Gregg, Jim Jirjis, Nancy M Lorenzi, and Dario Giuse. StarTracker: an integrated, web-based clinical search engine. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, (1):855, January 2003.
- William Gregg, Jim Jirjis, Nancy M Lorenzi, and Dario Giuse. StarTracker: an integrated, web-based clinical search engine. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, page 855, 2003.

- Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl:5228–35, April 2004.
- David A Hanauer. EMERSE: The Electronic Medical Record Search Engine. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 331(7531):941, January 2006.
- Daniel Heinz. *Comparison of Language Models for Information Retrieval*. PhD thesis, Carnegie Mellon University, 2007.
- W Hersh, S Price, and L Donohoe. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 344–8, January 2000.
- William R Hersh. *Information retrieval : a health and biomedical perspective*. Springer, New York, NY, 3rd edition, 2009.
- Julia Hippisley-cox, Mike Pringle, Ruth Cater, Alison Wynn, Vicky Hammersley, Carol Coupland, Rhydian Hapgood, Peter Horsfield, Sheila Teasdale, and Christine Johnson. The electronic patient record in primary care: regression or progression? A cross sectional study. *BMJ*, 326:1439–1443, 2003.
- Thomas Hofmann. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, pages 50–57, 1999.
- G Hripcsak, J J Cimino, and S Sengupta. WebCIS: large scale deployment of a Web-based clinical information system. *Proc AMIA Symp*, pages 804–808, 1999.
- George Hripcsak, David K Vawdrey, Matthew R Fred, and Susan B Bostwick. Use of electronic clinical documentation: time spent and team interactions. *Journal of the American Medical Informatics Association : JAMIA*, 18(2):112–7, March 2011.
- Peter W Hung, Stephen B Johnson, David R Kaufman, and Eneida A Mendonça. A multi-

- level model of information seeking in the clinical domain. *Journal of biomedical informatics*, 41(2):357–70, April 2008.
- Nicholas C Ide, Russell F Loane, and Dina Demner-Fushman. Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association : JAMIA*, 14(3):253–63, 2007.
- BJ Jansen and A Spink. How are we searching the World Wide Web? A comparison of nine search engine. *Information Processing and Management*, pages 92–109, 2006.
- B.J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the Web. In *ACM SIGIR Forum*, volume 32, pages 5–17. ACM New York, NY, USA, 1998.
- B.J. Jansen, D.L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, pages 1149–1150. ACM New York, NY, USA, 2007.
- Bernard J. Jansen. Search log analysis: What it is, what’s been done, how to do it. *Library & Information Science Research*, 28(3):407–432, 2006.
- T Joachims. Evaluating retrieval performance using clickthrough data. In *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, pages 12–15, 2002.
- D Kelly and J Teevan. Implicit feedback for inferring user preference: a bibliography. *ACM SIGIR Forum*, 37:18–28, 2003.
- S Klink. Improving document transformation techniques with collaborative learned term-based concepts. *Lecture notes in computer science*, pages 281–305, 2004.
- Hang Li, Yunbo Cao, Jun Xu, Yunhua Hu, Shenjie Li, and Dmitriy Meyerzon. A new approach to intranet search based on information extraction, 2005.
- Zhenyu Liu and Wesley W Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202, January 2007.

- Xiaoyong Liu and W. Bruce Croft. Cluster-based retrieval using language models. *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 186, 2004.
- CD Manning, P Raghavan, and H Schtze. *Introduction to information retrieval*. Number c. Cambridge University Press New York, NY, USA, 2008.
- Gary Marchionini. *Information seeking in electronic environments*. Cambridge University Press, New York, 1995.
- Mazlita Mat-Hassan and Mark Levene. Associating search and navigation behavior through log analysis. *Journal of the American Society for Information Science and Technology*, 56:913–934, 2005.
- Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit, 2002.
- Kathleen R. McKeown, Judith L. Klavans, André Kushniruk, Vimla Patel, Simone Teufel, Shih-Fu Chang, James Cimino, Steven Feiner, Carol Friedman, Luis Gravano, Vasileios Hatzivasiloglou, Steven Johnson, and Desmond A. Jordan. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries - JCDL '01*, pages 331–340, New York, New York, USA, 2001. ACM Press.
- E a Mendonça and J J Cimino. Automated knowledge extraction from MEDLINE citations. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 575–9, January 2000.
- Tsuyoshi Murata and Kota Saito. Extracting Users' Interests from Web Log Data. *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, pages 343–346, 2006.
- P Nadkarni, R Chen, and C Brandt. UMLS concept indexing for production databases: a feasibility study. *Journal of the American Medical Informatics Association : JAMIA*, 8:80–91, 2001.
- Karthik Natarajan, Daniel Stein, Samat Jain, and Noémie Elhadad. An analysis of clinical queries



- in an electronic health record search utility. *International journal of medical informatics*, 79(7):515–22, July 2010.
- E Nygren and P Henriksson. Reading the medical record. I. Analysis of physicians’ ways of reading the medical record. *Computer methods and programs in biomedicine*, 39(1-2):1–12, 1992.
- E Nygren, M Johnson, and P Henriksson. Reading the medical record. II. Design of a human-computer interface for basic reading of computerized medical records. *Computer methods and programs in biomedicine*, 39(1-2):13–25, 1992.
- E Nygren, J C Wyatt, and P Wright. Helping clinicians to find data and avoid delays. *Lancet*, 352(9138):1462–6, October 1998.
- OpenCalais, 2008.
- J A Osheroff, D E Forsythe, B G Buchanan, R A Bankowitz, B H Blumenfeld, and R A Miller. Physicians’ information needs: analysis of questions posed during clinical teaching. *Ann Intern Med*, 114(7):576–581, 1991.
- Serguei Pakhomov, Susan a Weston, Steven J Jacobsen, Christopher G Chute, Ryan Meverden, and Véronique L Roger. Electronic medical records for clinical research: application to the identification of heart failure. *The American journal of managed care*, 13:281–288, 2007.
- Laurence a. F. Park and Kotagiri Ramamohanarao. Efficient storage and retrieval of probabilistic latent semantic information for information retrieval. *The VLDB Journal*, 18(1):141–155, February 2008.
- Laurence A F Park and Kotagiri Ramamohanarao. The Sensitivity of Latent Dirichlet Allocation for Information Retrieval. In *ECML PKDD*, pages 176–188, 2009.
- PubMed.
- Daniel E Rose and Danny Levinson. Understanding user goals in web search, 2004.
- Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter

- Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960, May 2011.
- G Salton, A Wong, and C S Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 1975.
- Stefan Schulz, Philipp Daumke, Pascal Fischer, and Marcel Lucas Müller. Evaluation of a document search engine in a clinical department system. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 647–51, January 2008.
- Stefan Schulz, Philipp Daumke, Pascal Fischer, and Marcel Lucas Müller. Evaluation of a document search engine in a clinical department system. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 647–651, 2008.
- Alicia Scott-Wright, Jonathan Crowell, Qing Zeng, David Bates, and Robert Greenes. Analysis of information needs of users of MEDLINEplus, 2002 - 2003. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 699–703, January 2006.
- Y Senathirajah and S Bakken. Development of user-configurable information source pages for medical information retrieval. *AMIA Annu Symp Proc*, page 1109, 2007.
- Yalini Senathirajah. Development of User-configurable Information Source Pages for Medical Information Retrieval 1. *Journal of the Medical Library Association*, pages 2007–2007, 2007.
- Lisa Seyfried, David a Hanauer, Donald Nease, Rashad Albeiruti, Janet Kavanagh, and Helen C Kales. Enhanced identification of eligibility for depression research using an electronic medical record search engine. *International journal of medical informatics*, pages 1–6, June 2009.
- Nigam Shah and Mark Musen. UMLS-Query: A Perl Module for Querying the UMLS. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, (Table 1):652–6, January 2008.
- C Silverstein, H Marais, M Henzinger, and M. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 1999.

- R Smith. What clinical information do doctors need? *BMJ*, 313(7064):1062–1068, 1996.
- A. Spink, D. Wolfram, M.B.J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.
- Gilbert Strang. *Introduction to linear algebra*. Wellesley-Cambridge Press, Wellesley, MA, 2nd edition, 1998.
- W V Sujansky. The benefits and challenges of an electronic medical record: much more than a "word-processed" patient chart. *The Western journal of medicine*, 169(3):176–83, September 1998.
- P C Tang, J Annevelink, H J Suermondt, and C Y Young. Semantic integration of information in a physician's workstation. *International journal of bio-medical computing*, 35(1):47–60, February 1994.
- H J Tange, V A Dreessen, A Hasman, and H H Donkers. An experimental electronic medical-record system with multiple views on medical narratives. *Computer methods and programs in biomedicine*, 54(3):157–72, November 1997.
- H J Tange, H C Schouten, A D Kester, and A Hasman. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 5(6):571–82, 1998.
- H J Tange. The paper-based patient record: is it really so bad? *Computer methods and programs in biomedicine*, 48(1-2):127–31, 1995.
- H Tange. How to approach the structuring of the medical record? Towards a model for flexible access to free text medical data. *International journal of bio-medical computing*, 42(1-2):27–34, July 1996.
- Andrew A Tawfik, Karl M Kochendorfer, Dinara Saparova, Said Al Ghenaimi, and Joi L Moore. Using semantic search to reduce cognitive load in an electronic health record. In *2011 IEEE*

- 13th International Conference on e-Health Networking, Applications and Services*, pages 181–184. IEEE, June 2011.
- Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A S Potts. Information Re-Retrieval: Repeat Queries in Yahoos Logs. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 151, 2007.
- Joachims Thorsten. Optimizing search engines using clickthrough data, 2002.
- Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pages 2–10, 1998.
- Tielman T Van Vleck, Daniel M Stein, Peter D Stetson, and Stephen B Johnson. Assessing data relevance for automated generation of a clinical summary. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 761–5, January 2007.
- Ellen (NIST) Voorhees. TREC Medical Records Track, 2011.
- L L Weed. Medical records that guide and teach. *The New England journal of medicine*, 278(11):593–600, March 1968.
- L L Weed. New connections between medical knowledge and patient care. In *BMJ (Clinical research ed.)*, volume 315, pages 231–5, July 1997.
- Xing Wei and W Bruce Croft. Investigating Retrieval Performance with Manually-Built Topic Models. *Information Retrieval*, pages 333–349.
- Xing Wei and W Bruce Croft. LDA-Based Document Models for Ad-hoc Retrieval. In *SIGIR*, Seattle, WA USA, 2006. ACM.
- Adam Wilcox, Spencer S Jones, David A Dorr, Wayne Cannon, Laurie Burns, P T Ms, Kelli Radican, Kent Christensen, Cherie Bruncker, Ann Larsen, Scott P Narus, Sidney N Thornton, and Paul D Clayton. Use and Impact of a Computer-Generated Patient Summary Worksheet

- for Primary Care Intermountain Health Care , Salt Lake City , UT Department of Medical Informatics and Clinical Epidemiology ,. In *AMIA 2005 Symposium*, pages 824–828, 2005.
- Jesse O Wrenn, Daniel M Stein, Suzanne Bakken, and Peter D Stetson. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association : JAMIA*, 17(1):49–53.
- Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '96*, pages 4–11, 1996.
- Yiming Yang and Christopher G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems*, 12(3):252–277, 1994.
- Lei Yang, Qiaozhu Mei, Kai Zheng, and David A Hanauer. Query log analysis of an electronic health record search engine. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:915–24, January 2011.
- Xing Yi. A comparative study of utilizing topic models for information retrieval. *Advances in Information Retrieval*, pages 29–41, 2009.
- RJ Yount, JK Vries, and CD Council. The Medical Archival System: an information retrieval system based on distributed parallel. *Information Processing and Management: an*, 27(4):379–389, 1991.
- Michael Zalis and Mitchell Harris. Advanced search of the electronic medical record: augmenting safety and efficiency in radiology. *Journal of the American College of Radiology : JACR*, 7(8):625–33, August 2010.
- Q Zeng and J J Cimino. Evaluation of a system to identify relevant patient information and its impact on clinical information retrieval. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 642–6, January 1999.

- Q Zeng and J J Cimino. A knowledge-based, concept-oriented view generation system for clinical data. *Journal of biomedical informatics*, 34(2):112–28, 2001.
- Qing Zeng, James J Cimino, and Kelly H Zou. Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. *Journal of the American Medical Informatics Association : JAMIA*, 9(3):294–305, 2002.
- Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, pages 334–342, New York, New York, USA, 2001. ACM Press.
- ChengXiang Zhai. *Statistical Language Models for Information Retrieval*, volume 1. Morgan &cLaypool publishers, January 2008.
- Rui Zhang, Serguei Pakhomov, Bridget T McInnes, and Genevieve B Melton. Evaluating measures of redundancy in clinical texts. In *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, volume 2011, pages 1612–20, January 2011.
- M Zhu and a Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

## Appendix A

# Appendix

### A.1 User Survey

Below are the screen shots from the online survey conducted to understand clinical users' search preferences.

[Exit this survey](#)

## Understanding user needs for an EHR-based search utility

### Introduction

1 / 6	
-------	--

Thank you for your participation in this anonymous survey. Your responses will help researchers understand how people utilize an electronic health record (EHR) search engine and will inform future enhancements to our institution's search utility.

If you would like to know more about this study or have any question, please contact Karthik Natarajan ([kan7003@dbmi.columbia.edu](mailto:kan7003@dbmi.columbia.edu)). This survey was reviewed and approved by the Columbia University Institutional Review Board (IRB-AAAD5377). To learn more about protection of human subjects at Columbia University, please go to: <http://www.cumc.columbia.edu/dept/irb/info.html>

[Next](#)



[Exit this survey](#)

Understanding user needs for an EHR-based search utility

**General EHR Usage**

2 / 6

In this section, we are interested in your experience and background in using an EHR system.

**1. Please specify the primary role in which you use an EHR:**

- Attending
- Resident
- Fellow
- Nurse
- Researcher
- Medical Student
- Other Healthcare Professional
- Quality Management
- Risk Management
- Administration (billing, chart analyst, etc.)

Other (please specify)

**2. How long have you used WebCIS?**

- Less than 6 months
- 6 - 12 months
- 13 months - 3 years
- More than 3 years

**3. Besides lab results, which document do you find to contain the most useful information when caring for a patient:**

- Nursing Flowsheet
- Discharge Summary
- Radiology Report
- Admission Note
- Pathology Report
- Cardiology Report
- Consult Note



Other (please specify)

[Prev](#) [Next](#)



[Exit this survey](#)

Understanding user needs for an EHR-based search utility

**Search Engine Use**

3 / 6

**7. How often do you use search engines (i.e., Google, Bing, PubMed) in your everyday life?**

- Daily
- Weekly
- Rarely
- Never

If you do use search engine(s), which one?

**8. When using a search engine, I feel confident in my ability to:**

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree	N/A
Enter effective search keywords	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Use advanced search features (i.e., quotation marks, limits in PubMed, etc)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Efficiently refine search until I find what I'm looking for	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Prev](#) [Next](#)



**Exit this survey**

Understanding user needs for an EHR-based search utility

**Searching the EHR**

4 / 6	
-------	--

We installed a search engine called CISearch within WebCIS in July 2008 in hopes to improve users' access to clinical information within the EHR. In its initial version it searches discharge summaries, radiology reports, and pathology reports. In this section, we hope to understand users' perception of the search utility in order to improve its functionality.

**9. How often do you use the search utility within WebCIS?**

- Daily
- Weekly
- Rarely
- Never
- I did not know there was a search utility

**10. If you use the search utility within WebCIS, how satisfied are you with the results you get?**

- Very satisfied
- Somewhat satisfied
- Somewhat dissatisfied
- Very dissatisfied
- N/A

If you answered "Somewhat dissatisfied" or "Very dissatisfied," please describe what could be improved.

**11. If you used the search utility within WebCIS, what is your level of agreement with the following statements?**

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree	N/A
I find it hard to use the search utility because of its user interface	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The search utility helps me answer my clinical questions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

If you answered "Disagree" or "Strongly Disagree," please describe what could be improved.

**12. Do you have any additional feedback that you would like to provide about the current version of the WebCIS search utility? Positive and negative comments are welcome and appreciated!**

Prev

Next



**Exit this survey**

Understanding user needs for an EHR-based search utility

**Ideal Search Utility**

5 / 6

In this section, we would appreciate your input on how an ideal search utility should function in order to improve our current electronic health record (EHR) search utility.

**13. Please rate your level of agreement with the following:**

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree	N/A
Searching across patients on your patient list would be useful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search results ordered with most recent documents first would be useful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search results displayed based on relevance would be useful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A search utility would be useful for familiarizing myself with a patient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A search utility that allows me to enter complex queries would be useful (i.e., "mri AND brain")	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A search utility that also returned relevant documents that did NOT contain the query terms would be useful. (i.e., if you searched for "CHF", documents that contained "myocardial infarction" would also be returned.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**14. Which of the following would you like to see in an EHR-based search utility regarding searched lab results? (You can select multiple choices).**

- Most recent lab result is displayed as a numerical value
- Lab is displayed in a graph over time
- Lab is displayed in a graph along with other similar lab results
- Lab is highlighted in documents they are mentioned

Please specify other ways you would like to see lab results displayed in search results.

[Prev](#) [Next](#)





**Exit this survey**

Understanding user needs for an EHR-based search utility

**Thank you!**

6 / 6	
-------	--

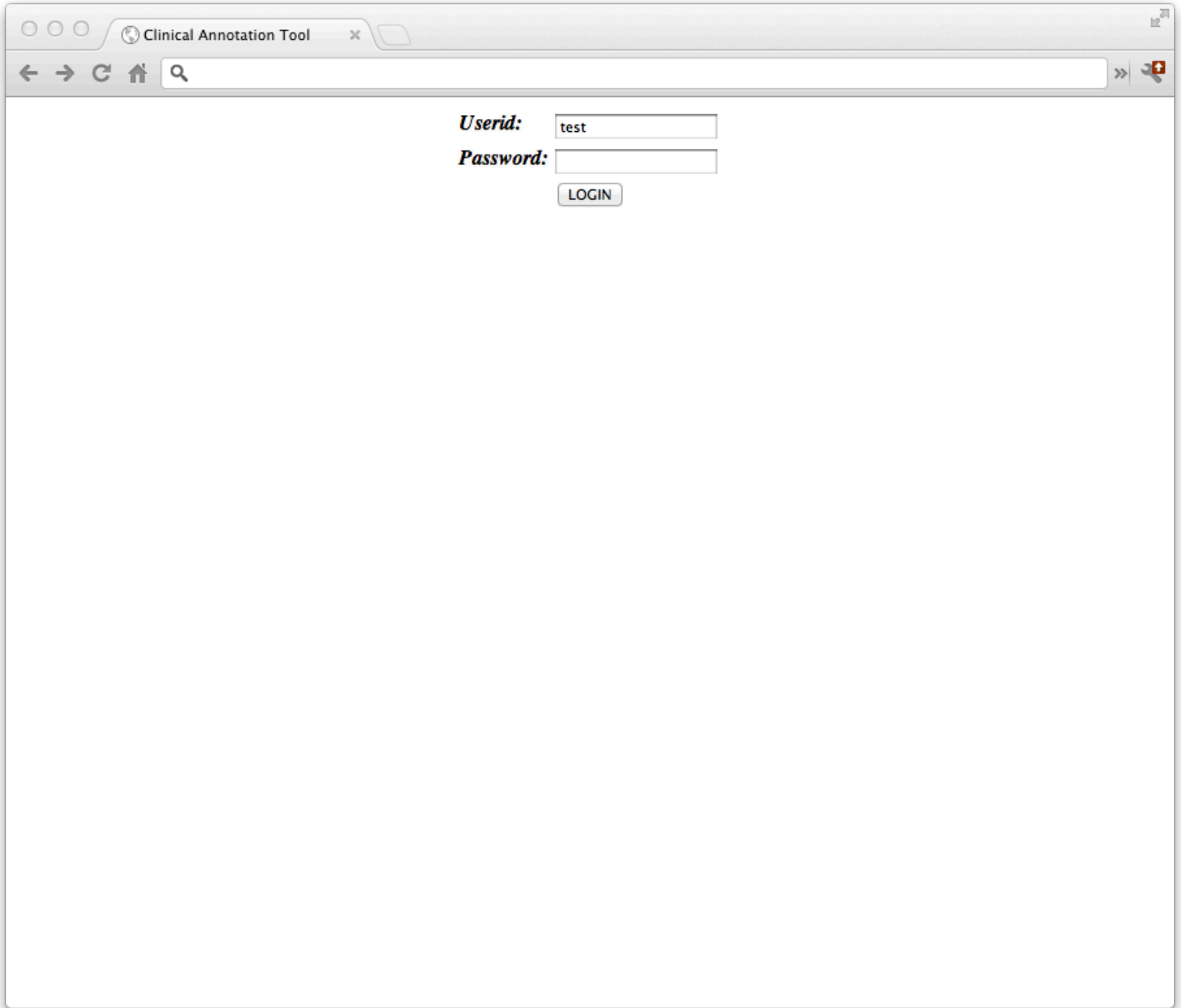
Thank you for participating in this survey. Your answers are very valuable to us and will help improve the future versions of the search utility within WebCIS.

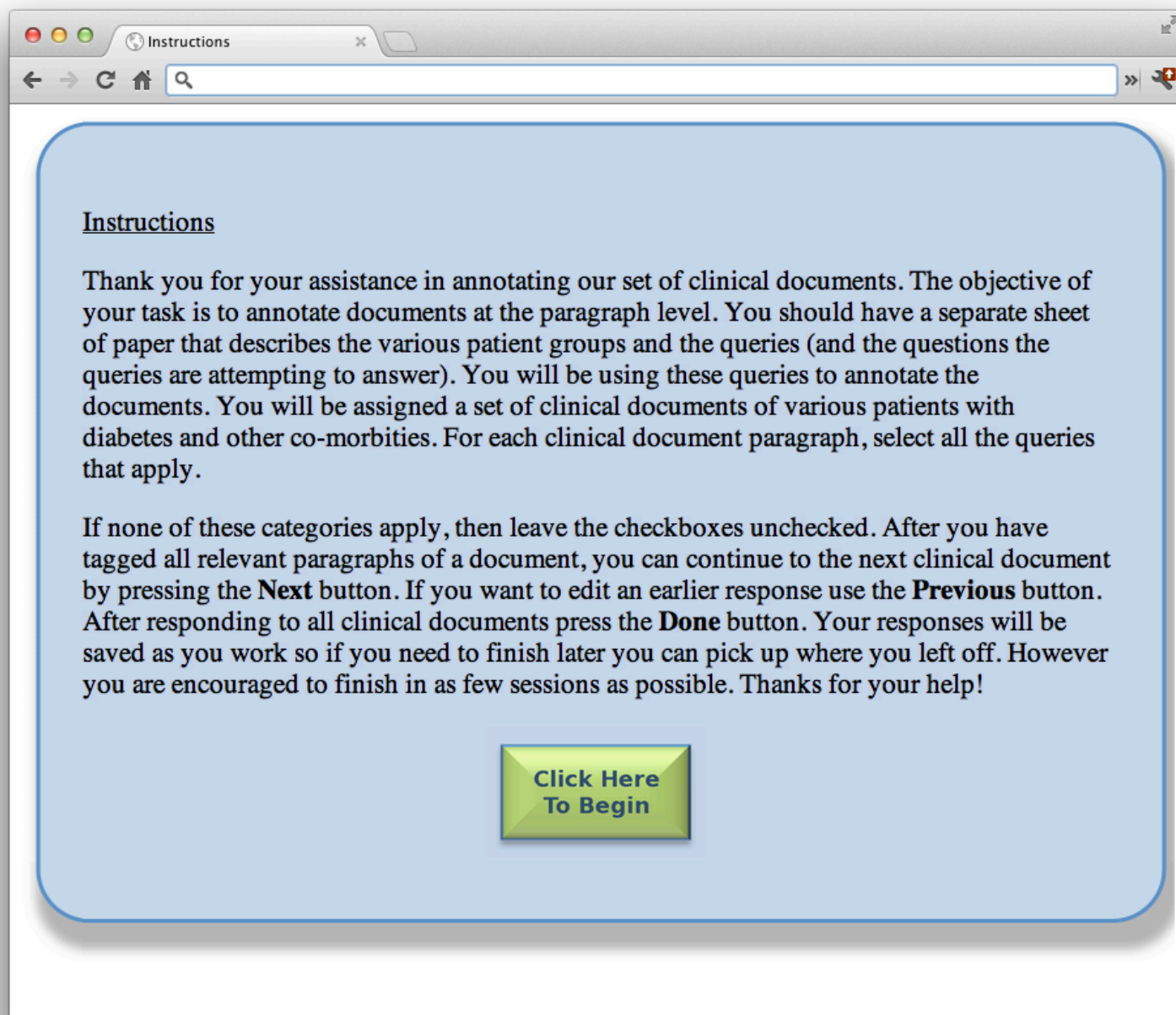
Prev Done

## A.2 Annotation Tool

Before the gold standard was created, a web-based annotation tool was specifically developed for the relevance tagging task. The application was written in PHP and the annotations were persisted to a MySQL database. Since the annotators were examining real patient notes, the web application ran on a secure server and was only accessible to a limited number of computers based on IP addresses. Besides the network-based security, application-level authentication was incorporated into the annotation tool; each annotator was assigned a username and password to access the application.

When the annotator successfully logged into the system, he/she was allowed to select a patient record to annotate. Once a patient record was selected, the annotator was shown in chronological order each document in the record. The application used XSLT to render the XML document that was created in the preprocessing step mentioned in the corpus selection section above. The application displayed the document in paragraph chunks. Above each paragraph there were checkboxes that represented each of the information needs being annotated so that the annotator could easily tag relevant paragraphs. Though this display was easy for selecting relevant documents, it was distracting for a user to read and understand the entire note, so a link to the plain text version of the clinical note was provided to the annotator. When the annotators navigated to the next document, their paragraph-level relevance tags were persisted to the database. Screen shots of the annotation application can be seen below.





The image shows a screenshot of a web browser window. The browser's address bar contains the word "Instructions". The main content area has a light blue background with rounded corners. At the top left of this area is the heading "Instructions". Below the heading is a paragraph of text explaining the task: "Thank you for your assistance in annotating our set of clinical documents. The objective of your task is to annotate documents at the paragraph level. You should have a separate sheet of paper that describes the various patient groups and the queries (and the questions the queries are attempting to answer). You will be using these queries to annotate the documents. You will be assigned a set of clinical documents of various patients with diabetes and other co-morbidities. For each clinical document paragraph, select all the queries that apply." This is followed by a second paragraph: "If none of these categories apply, then leave the checkboxes unchecked. After you have tagged all relevant paragraphs of a document, you can continue to the next clinical document by pressing the **Next** button. If you want to edit an earlier response use the **Previous** button. After responding to all clinical documents press the **Done** button. Your responses will be saved as you work so if you need to finish later you can pick up where you left off. However you are encouraged to finish in as few sessions as possible. Thanks for your help!" At the bottom center of the blue area is a green button with a 3D effect and the text "Click Here To Begin".

Instructions

Thank you for your assistance in annotating our set of clinical documents. The objective of your task is to annotate documents at the paragraph level. You should have a separate sheet of paper that describes the various patient groups and the queries (and the questions the queries are attempting to answer). You will be using these queries to annotate the documents. You will be assigned a set of clinical documents of various patients with diabetes and other co-morbidities. For each clinical document paragraph, select all the queries that apply.

If none of these categories apply, then leave the checkboxes unchecked. After you have tagged all relevant paragraphs of a document, you can continue to the next clinical document by pressing the **Next** button. If you want to edit an earlier response use the **Previous** button. After responding to all clinical documents press the **Done** button. Your responses will be saved as you work so if you need to finish later you can pick up where you left off. However you are encouraged to finish in as few sessions as possible. Thanks for your help!

[Click Here To Begin](#)

	Patient ID	Cohort	Total # of Docs	# of Docs Tagged
<input type="radio"/>	39	Congestive Heart Failure	31	31
<input type="radio"/>	40	Congestive Heart Failure	36	36
<input type="radio"/>	41	Congestive Heart Failure	30	30
<input type="radio"/>	42	Congestive Heart Failure	32	32
<input type="radio"/>	43	Congestive Heart Failure	35	35
<input type="radio"/>	44	Congestive Heart Failure	33	33
<input type="radio"/>	45	Congestive Heart Failure	30	30
<input type="radio"/>	46	Congestive Heart Failure	31	31
<input checked="" type="radio"/>	47	CHF TRAINING	2	1
<input type="radio"/>	49	NON-COHORT	34	34
<input type="radio"/>	51	NON-COHORT	27	27
<input type="radio"/>	52	LUPUS	36	36
<input type="radio"/>	53	LUPUS	35	35
<input type="radio"/>	54	LUPUS	42	42
<input type="radio"/>	55	LUPUS	33	33
<input type="radio"/>	56	LUPUS	30	30
<input type="radio"/>	57	LUPUS	26	26
<input type="radio"/>	58	LUPUS	31	31

Clinical Note

When you are finished click the "NEXT" button below. This will save your tags and then send you to the next note.

Document Name: [Nephrology HD^CVVH at 13/23/1809 11:50](#) 1 out of 2 documents.

---

SEEN AND EXAMINED ON DIALYSIS

kidney  joint  hair loss  fever  rash  chest pain  ANA  inflammation  pericarditis  biopsy  anemia  NSAID  anti-malaria  steroid   
 lupus  drug-induced

55 yo male with CKD 2/2 to HPT (baseline Cr 3.8 in 3/09) with acute kidney failure likely due to sepsis.

kidney  joint  hair loss  fever  rash  chest pain  ANA  inflammation  pericarditis  biopsy  anemia  NSAID  anti-malaria  steroid   
 lupus  drug-induced

Other medical problems include: stroke in 2004 with residual (L) hemiparesis, COPD & NSCL CA of lung S/P LUL lobectomy.

kidney  joint  hair loss  fever  rash  chest pain  ANA  inflammation  pericarditis  biopsy  anemia  NSAID  anti-malaria  steroid   
 lupus  drug-induced

During sepsis, Pt was in shock on pressors and CVVHD. Continues to have fevers (TM 101.1 9/6) and he is still on broad spectrum antibiotics.

kidney  joint  hair loss  fever  rash  chest pain  ANA  inflammation  pericarditis  biopsy  anemia  NSAID  anti-malaria  steroid   
 lupus  drug-induced

S/Pp fistulogram by IR for marked LUE swelling. Found to have central venous stenosis and 2 sites of stenosis within graft. IR attempted angioplasty on graft that was unsuccessful..

kidney  joint  hair loss  fever  rash  chest pain  ANA  inflammation  pericarditis  biopsy  anemia  NSAID  anti-malaria  steroid   
 lupus  drug-induced

