Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval

Master's Thesis

Tracy Edinger, ND, MCR
Department of Medical Informatics and Clinical Epidemiology
Oregon Health & Science University

October 2014

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Tracy Edinger

"Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval"

Has been approved

William Hersh, MD

Dina Demner-Fushman, MD, PhD

Aaron Cohen, MD, MS

Steven Bedrick, PhD

TABLE OF CONTENTS

TABLE OF CONTENTS

Acknowledgements

I would like to thank my thesis committee for their invaluable advice and help:

William Hersh, MD
Dina Demner-Fushman, MD, PhD
Aaron Cohen, MD, MS
Steven Bedrick, PhD

I would also like to thank the faculty, staff, and students at OHSU and the NLM who helped me with this work and my informatics training.

Introduction

The use of electronic health records (EHR) in US medical facilities has increased rapidly over recent years. Incorporating EHRs has facilitated clinical practice, resulting in widespread access to patient information and potentially better care. Beyond their primary use for direct clinical care and for billing, EHRs provide a valuable source of data for "secondary uses," including research, quality measures, disease surveillance, and administrative applications.(1) Because of the quantity of data in the EHR of a typical health-care system, secondary use of this data can be of enormous benefit.

Background and significance

EHRs are designed primarily to improve direct patient care and for billing purposes. In the process of recording health information, they become a repository of a vast amount of clinical data. This data can be used for many secondary uses beyond direct clinical care and billing, including research, quality assessment, clinical decision support, disease surveillance and registries, and business applications. All of these uses of EHR data require the ability to find the desired information with a high degree of accuracy and completeness.

In many cases, this means that we need to be able to identify specific patient cohorts. For example, researchers may want to look at records for all patients with a particular disease or all patients who received a specific treatment. For accurate disease surveillance, we must be able to identify patients who had that disease, either by locating specific diagnosis codes or mentions of the disease, or by locating surrogate indicators such as treatments or symptoms. Alternatively, hospital administrators may want to

count the number of patients who received a particular treatment, or track patient volumes in different departments.

The way data is stored in an EHR determines how we need to search for specific patients. Some of the data is stored in structured fields, with data recorded in consistent ways as specific codes, including International Classification of Diseases (ICD) or other codes, or as a finite list of predetermined phrases. Because of the limited content variability in this type of data, a query using structured fields will completely and accurately retrieve all patients matching the criteria. Although searching structured data produces consistent, reliable results, research suggests that searches relying solely on structured data will not retrieve all patients relevant to the topic; retrieval of the full cohort requires a search of clinical text as well.(2,3)

A great deal of EHR data is stored as unstructured text in clinical notes, history and physical exam notes, and reports. This unstructured text can be difficult to access on a large scale. Medical text contains several features that compound this difficulty, including frequent use of abbreviations, lack of standardization of abbreviations, context-dependent differences in word meanings and abbreviations, negation of symptoms or diseases, and documentation of history not directly related to the current visit or even to the current patient.(4)

Several groups have worked on improving the ability to retrieve patient cohorts from EHR data. The Electronic Medical Records and Genomics Network (eMERGE) Consortium has evaluated the ability to retrieve specific patient cohorts from EHRs.(5) In this study, EHR data from five different sites were used to identify patients with at least one of the following diseases: dementia, cataracts, peripheral arterial disease, type 2

diabetes mellitus, and cardiac conduction defects. Patients were identified with a high level of accuracy when the data were stored in structured fields. In some cases, the target information was stored only in the clinical text. The use of natural language processing (NLP) tools increased retrieval significantly—at one site, 129% more cases were identified by including the use of NLP tools than through using structured data and string matching alone. In evaluating results from one site, use of the same terms to mean different things within one document was an issue in correct retrieval of patients; for example, 'potassium' can be a medication or the name of a lab value, and a drug name can be listed as an allergy or as a prescribed medication.(6)

The Informatics for Biology and the Bedside (i2b2) project has sponsored several cohort-identification challenges.(7) In these challenges, participants were given a set of de-identified clinical records and a topic such as obesity or smoking status. Participants developed NLP systems and algorithms to find relevant records, and results were then submitted for evaluation.

Another group that has worked on improving the ability to accurately retrieve patient data from medical text for secondary uses is the Text Retrieval Conference (TREC) of the National Institute of Standards and Technology (NIST). In the TREC 2011 and 2012 Medical Records Track (TRECMed), participants were challenged to find patient cohorts specified by a list of clinical topics in a set of de-identified textual medical records from the University of Pittsburgh.(8) The task of TRECMed was to retrieve patients who might be candidates for participation in research studies described by the topics. The medical records were organized into visits or encounters (e.g., hospitalizations and emergency department visits). Participants developed search systems and algorithms to

retrieve visits relevant to each topic. Retrieved visits were then judged by physicians for relevance to the topic, which in this case meant that the patient was a candidate for a research study on the topic. An analysis of incorrectly retrieved visits identified several key challenges in accurate retrieval of patient cohorts.(4) Factors in retrieving non-relevant visits included terminology similarities, negation of the desired term, and mention of the desired term as a past or future occurrence. Relevant visits were overlooked when the chart notes used different terminology or described rather than named the condition.

Several approaches may be utilized to overcome these issues and facilitate the retrieval process: the query can be constructed to yield a more accurate response, and the original text can be manipulated to make it more searchable. Clinicians are trained to write medical records in a highly structured fashion. Physicians' chart notes are divided into sections that indicate the source and purpose of the information, in a structure referred to as SOAP (Subjective-Objective-Assessment-Plan). Within these sections, a typical chart note for a first encounter with a patient includes the chief complaint, a history of the present illness, a review of systems, past medical history, family history, and social history.

Annotating the text according to these sections would allow the construction of searches targeted to the section most likely to contain relevant information, avoiding or minimizing some of the issues found in previous work. Several tools and strategies exist that segment clinical records.(9–13) Although the effectiveness of each segmentation strategy has been evaluated, no studies could be located that demonstrate whether segmenting improves recall.

Temporality is an information retrieval (IR) issue that is particularly relevant to medical text, which often documents the current illness as well as previous illnesses and potential future complications. Clues to temporality can be found in identifying the section of the medical record: a description of the chief complaint is likely the current issue, whereas a condition listed in the past medical history is something that has resolved or is not the focus of the current visit.

Subject identification can also complicate retrieval of medical information. Chart notes may document illnesses of other family members as well as those of the patient. Identifying who has the disease improves precision by avoiding retrieval based on someone else's disease status. The ability to separate sections of the medical record will facilitate retrieval accuracy by identifying the family history section and allowing that section to be searched only when applicable to the topic.

Clinical text segmentation tools

Several researchers have developed algorithms to segment clinical text. Approaches range from identifying different sections of documents to labeling each sentence type, and clinical text type ranges from ED reports to radiology reports to outpatient chart notes. Following is a brief review of several of these tools.

Apostolova—radiology reports A group of researchers from DePaul University developed an automated tool to segment radiology reports.(9) Their dataset consisted of 215,000 free-text radiology reports generated over a nine-year period. Most reports had been transcribed from dictation, although a small set was transcribed by a speech-

recognition program. Eight sections were identified by manual examination of the data and by consulting professional guidelines.

The tool performs two passes of the data. In the first pass, exact matches to specific combinations of letters and punctuation are identified and marked with the appropriate section heading. For example, the section "Findings" may be indicated by any of the following: *finding, findings, observation, observations, discussion,* or *discussions.* This pass of the tool will identify and label any of the eight sections that are indicated by the pre-specified text.

Because there will occasionally be sections that have not been labeled explicitly, the next pass of the tool looks for patterns that indicate a new section has been started or that the previous section has ended. Examples of these vary and include changes in capitalization, such as all capitals or camel case, presence of white space or extra lines, and the end of the file. This pass also measures the distance of each sentence to the nearest section, and the tool takes this distance into account when determining whether a sentence belongs to the previous section or to a new section.

Performance of this tool varied greatly for the eight document sections. The demographics section was identified with 99% accuracy, whereas the recommendation section was correctly identified only 22% of the time. Average accuracy was 79%.

Cho—medical reports This tool identifies and labels section headings in medical reports.(10) Section indicator test and segmenting rules were identified using a set of training documents. These indicators and rules were then tested on three different

clinical document collections: CT/MR reports, chest radiology reports, and urology reports.

Similar to the Apostolova tool, this tool makes two passes through the data, looking for matching text during the first pass. The list of section headings and matching text was developed manually, with researchers iterating through a subset of the data and using software to insert sections labels and to tag the beginning and end of section headings. The clinical data is modified prior to being searched for sections. All apostrophes, hyphens, asterisks, and parentheses are removed; all letters are converted to upper case; vowels are removed; and letter repeats are omitted to minimize the effect of possible misspellings. Once the heading list was developed, the tool passes through the modified clinical data, searching for any of the tagged text patterns and inserting the appropriate section labels.

The second pass of this tool is designed to identify sections that were not labeled. Sections may be indicated by specific phrases or patterns. For example, a physician's signature on a letter may always be preceded by the phrase "Thank you for referring..." The second pass identifies phrases that recur with specific patterns, which are then used to segment sections. If the expected number of sections is not detected, it is likely that not all sections have been identified. The size of each section is compared to the average size of the section. If the size exceeds the mean by more than three standard deviations, then the tool searches for clues such as white space that might indicate a section break.

Performance of this tool was evaluated for each clinical corpus. Correct identification of all labeled and all unlabeled sections was reported. Overall, this tool performed very well—all sections were correctly identified at least 96% of the time.

Denny—SecTag This tool identifies and labels clinical section headers used to document a history and physical examination.(11) SecTag identifies both explicitly and implicitly labeled section headings in clinical documents. SecTag follows five steps in evaluating chart notes. First, sentences and lists are identified. Intact sentences as well as sentence fragments will be detected. Next, explicit and implicit section headers are located and identified, and sections are labeled. In this step, words are mapped to synonyms and spelling errors are corrected to facilitate section labeling. After that, naive Bayes scoring is used to select the most likely section headers among all those identified. If a section header cannot be identified, the next step uses a calculated score to select the best header. The final step identifies and labels the end of each section.

Section headings were manually evaluated by physicians to assess recall and precision, which were 99.0% and 95.6% for all sections and 98.6% and 96.2% for major section headers.

Mowery—SOAP segments for emergency-department reports This tool identifies and labels the broad categories of the SOAP format (Subjective, Objective, Assessment, Plan) for each sentence in emergency department documents.(12) Labeling sentences this way provides contextual clues to the content of each sentence and facilitates more accurate information retrieval.

This tool uses natural language processing and other text-processing tools to identify textual features such as parts of speech, verb tense, and digit type (date, medication, age,

etc.). A computer algorithm uses these features to assign a SOAP classifier to each sentence.

Two people manually annotated a set of fifty documents, and these annotations were used to score the accuracy of the automated classifier. Accuracy varied, with F_1 scores of 93.9 for subjective, 94.5 for objective, 75.7 for assessment, and 77.0 for plan.

Tepper—Statistical segmentation of clinical documents This tool identifies and labels sections in discharge summaries and radiology reports.(13) Three sets of documents were used in development and validation of the tool: hospital discharge summaries, de-identified research corpus of discharge summaries from three institutions, and a set of hospital radiology reports.

A list of section categories was developed by manually examining the documents. The tool then searches for these categories in the documents, using textual cues such as capitalization patterns, numbers, and blank lines to facilitate the decision of where to place section headings. This is first done on subset of documents to train the tool to find categories. After this training run, the tool is then used on the experimental documents. In some cases, these two sets of documents came from different sources.

Precision for the three datasets varied, although all scores were high. Precision ranged from 82% to 97%. The authors noted worse performance when the data set used for training was different from the dataset being annotated.

Research Question

The goal of the current project is to find out whether searching specific sections of clinical text has an impact on precision. To determine this, text portions of the medical record will be segmented. Queries to retrieve specific patient cohorts will be written with and without utilization of the segmentation. A sample of the retrieved visits will be examined for relevance to the topics, and the effect of segmenting will be evaluated to determine if it improves retrieval accuracy and under what conditions it does so.

Methods

Data This project used a set of de-identified clinical records developed by the Massachusetts Institute of Technology (MIT), Philips Medical Systems, and Beth Israel Deaconess Medical Center.(14) The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) data is a publicly available dataset containing more than 25,000 intensive-care unit (ICU) patients. The MIMIC-II data is stored in a relational database containing structured data and unstructured textual discharge summaries, MD notes, radiology reports, and nursing notes. This project used all four types of text documents for the search corpus. Table 1 shows the number of each document type in the data set.

Table 1. MIMIC-II document types, counts, and patient population.

Document Type	Number of Documents	Patient Age
Nursing Note	420,057	All patients
Radiology Report	356,917	All patients
Discharge Summary	48,044	All patients
MD Note	176	Neonates only

Term Mapping to the Unified Medical Language System (UMLS) The search engine used for this project (described below) employs the Unified Medical Language System (UMLS), which is a set of tools developed by the National Library of Medicine (NLM) to facilitate development of computer applications capable of interpreting biomedical language.(15) The UMLS contains three components: the UMLS Metathesaurus, the UMLS semantic network, and the SPECIALIST lexicon and lexical tools.(16) The first of these, the UMLS Metathesaurus, "represents biomedical concepts derived from a variety of controlled vocabularies, classifications, and other biomedical terminologies" in a hierarchical structure depicting linkages between concepts and terms.(17) At the highest level are concepts, each representing a distinct biomedical topic or disease.(18) Each concept is identified by a number, called a concept unique identifier, or CUI. The various terms that refer to a concept are each identified by a term unique identifier (LUI), each of which is linked to the main concept. Variations of terms, differing by word endings or order or words, are identified by string unique identifiers (SUI), and the source vocabulary for each string is identified by an atomic unit identifier (AUI). The linkages mapped in the UMLS allow association of dissimilar phrases that refer to the same concept. The UMLS SPECIALIST Lexicon maps terms to spelling variations, abbreviations, and acronyms.(19)

Search Engine For this project, queries were run using the Essie search engine, which was originally developed by the National Library of Medicine to search the clinical-research registry ClinicalTrials.gov (20) and later adapted to facilitate the ability of non-clinicians to search medical literature.(21) Essie uses the UMLS Metathesaurus and

SPECIALIST Lexicon to map terms, allowing comparison of different but equivalent terms. For example, the phrase 'heart attack' will be mapped to 'myocardial infarction,' which is the same concept phrased in a way more likely to occur in medical literature. Because Essie uses the UMLS to map synonyms, equivalent terms may be found without being listed explicitly, facilitating queries by eliminating the need to identify all possible synonyms for the desired search term.

When a query is run, Essie returns a ranked, scored list of documents retrieved for that query.(21) Scores indicate the probability that documents are relevant to the query and range between zero and one. Calculation of scores takes into account the document section where the term was found, whether the document contains the original search term or a synonym instead, and the combination of search terms found. All document sections identified in this project were given an equal weight of 0.60. If a search for one term found the term in one section of the document, that document was estimated to have a 60% probability of relevance. If terms are found in multiple sections, the weights of the sections are combined to yield the overall probability of relevance. Likewise, the scores for presence of multiple search terms are calculated using the section weights and combined to find the overall probability.

Essie's performance has been validated through use in a number of TREC evaluations.(21) Essie achieved the highest score of 25 groups and 49 runs participating in the 2003 TREC Genomics track, and Essie's interactive run had the second highest score of 92 runs in 2006. Essie achieved the highest performance in TRECMed 2012.(8)

Segmentation Researchers at the NLM segmented a set of clinical documents for use in TRECMed. Segmentation was accomplished with a programming script that inserted XML tags at section headings. The section headings from that project were used as a starting point to segment the current set of documents, and the same script was used to insert the section headers into the MIMIC-II data.

A subset of each type of document in the MIMIC-II data was manually examined to identify the most common section headings. All spelling and punctuation variations were recorded for each heading. A text file was then created containing rules for identifying section headings and the corresponding heading tags for the new sections. The rules written for the TRECMed project were adjusted for the MIMIC-II data section headings and added to the new rules. Table 2 shows examples of the headings inserted for the indicator text listed. Appendix A contains a full list of section headings, and Appendix B lists the entire set of rules used for this project.

Table 2. Examples of headings and associated indicator text found in the documents.

Inserted Heading	Indicator Text
AdmissionDiagnosis	admitting diagnoses: diagnosis on admission: pre-op diagnosis
DCInstructions	discharge planning and instructions: follow up instructions: follow up instructions are as follows-
LabRadResults	cta chest: important diagnostics and labs: radiographs-

Documents were searched for exact matches to the indicator text variations listed in the text file, and the appropriate heading tags were inserted at those locations. Two tags were inserted for each heading, one signaling the start of the section and the other signaling the end. When indicator text signifying a new heading was located, an opening

tag was inserted for that heading, and an ending tag for the previous heading (if any) was inserted just before it. The heading for the last section was closed when the end of the document was reached. Section tags were set up in XML format; for example, <AdmissionDiagnosis> indicates the start of the section and </AdmissionDiagnosis> signals the end of the section. The original documents each contain a single block of text surrounded by opening and closing tags:

<text>
DATE: [3305-8-7**] 1:51 PM**
CHEST (PORTABLE AP)
Reason: CHECK ETT TUBE PLACEMENT
?PNA, CHF
REASON FOR THIS EXAMINATION:
CHECK ETT TUBE PLACEMENT
?PNA
CHF
UNDERLYING MEDICAL CONDITION:
85 y/o male s/p acute mi and catheterization now
in ccu with cardiogenic shock.
FINAL REPORT
CLINICAL INDICATION: Assess endotracheal tube placement in patient with congestive heart failure.
Comparison is made to previous study of one day earlier. An endotracheal tube is present, in satisfactory position. A Swan-Ganz catheter terminates in the proximal left pulmonary artery and has been withdrawn in the interval. An intraaortic balloon pump terminates about 3.3 cm below the superior aspect of the aortic knob, and a nasogastric tube terminates in the region of the gastroduodenal junction.
Cardiac and mediastinal contour are stable in the interval and pulmonary vascularity is within normal limits for technique. There has been improvement in the left retrocardiac opacity and there remains a patchy right basilar opacification which is slightly increased. A small amount of fluid is seen in the minor fissure.
IMPRESSION:
1) Lines and tubes in satisfactory position, as detailed above, with no evidence of pneumothorax.
2) Improved left retrocardiac opacity and worsened right lower lobe opacity likely due to atelectasis.
</text>

After segmenting this text, the document is broken into blocks, with the preamble, allergies, addendum, and discharge disposition separated by XML tags:

<text>
<preamble>**DATE: [**3305-8-7**] 1:51 PM**
CHEST (PORTABLE AP)</preamble>
<indication>**Reason: CHECK ETT TUBE PLACEMENT**
?PNA, CHF
REASON FOR THIS EXAMINATION:
CHECK ETT TUBE PLACEMENT
?PNA
CHF</indication>
<condition>**UNDERLYING MEDICAL CONDITION:**

**85 y/o male s/p acute mi and catheterization now
in ccu with cardiogenic shock.**</condition>
<procedure_details>**FINAL REPORT
CLINICAL INDICATION: Assess endotracheal tube placement in patient with congestive heart failure.
Comparison is made to previous study of one day earlier. An endotracheal tube is present, in satisfactory position. A Swan-Ganz catheter terminates in the proximal left pulmonary artery and has been withdrawn in the interval. An intraaortic balloon pump terminates about 3.3 cm below the superior aspect of the aortic knob, and a nasogastric tube terminates in the region of the gastroduodenal junction.
Cardiac and mediastinal contour are stable in the interval and pulmonary vascularity is within normal limits for technique. There has been improvement in the left retrocardiac opacity and there remains a patchy right basilar opacification which is slightly increased.
A small amount of fluid is seen in the minor fissure.**</procedure_details>
<study_impression>**IMPRESSION:
1) Lines and tubes in satisfactory position, as detailed above, with no evidence of pneumothorax.
2) Improved left retrocardiac opacity and worsened right lower lobe opacity likely due to atelectasis.**</study_impression>
</text>

The segmented documents were indexed in Essie, allowing searches to target specific sections of the documents. Search queries utilized the XML tags to locate text in specific sections of the documents.

Queries A set of clinical topics to be retrieved from the text were developed from topics in TRECMed 2012.(8) This set contained fifty clinical topics drawn from the Institute of Medicine's clinical comparative effectiveness priorities (16 topics), meaningful use clinical quality measures (12 topics), and the OHSUMED literature retrieval test collection (22 topics). Because the original query topics were developed for a wider range of patients, not all topics are relevant to the ICU documents in the MIMIC-II data. A subset of 22 topics was used, modified as necessary to fit the current clinic population (see Table 3 on next page for a list of all topics).

Queries were developed in an iterative fashion, refining the search details to maximize the number of relevant visits returned without using the sections. After an

initial query was run against the data, the text of a subset of the returned visits was examined to determine if any details needed to be added to or removed from the query.

Table 3. Description of query topics.

Topic Number	Topic Description
136	Patients with dental caries
137	Patients with inflammatory disorders receiving TNF-inhibitor treatments
140	Patients who developed disseminated intravascular coagulation in the hospital
141	Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes
143	Patients who had a carotid endarterectomy during this admission
144	Patients with diabetes mellitus who also have thrombocytosis
146	Patients treated for the post-partum problems depression, hypercoagulability or cardiomyopathy
147	Patients with left lower quadrant abdominal pain
149	Patients with delirium, hypertension, and tachycardia
150	Patients who have cerebral palsy and depression
153	Patients admitted to the hospital with end-stage chronic disease who are discharged on hospice care
158	Patients with esophageal cancer who develop pericardial effusion
160	Patients with low back pain who had imaging studies
165	Patients who have gluten intolerance or celiac disease
167	Patients with HIV/AIDS who develop pancytopenia
171	Patients with thyrotoxicosis treated with beta-blockers
173	Patients who received pneumonia vaccination during this admission
174	Patients with ventilator-associated pneumonia
179	Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression
183	Patients with acute vision loss
184	Patients with colon cancer receiving chemotherapy
185	Patients who develop thrombocytopenia in pregnancy

After the base query was complete, the segmented query was developed from the base. Specific sections were used as appropriate. For example, if a topic specified patients with diabetes mellitus, the segmented query would search sections where the

patient's diagnosis would be found but not the family history, where non-relevant references were likely.

For all topics, both queries search for terms that are as similar as possible, to avoid introducing variation based on term differences. Appendix C contains the final queries used for each topic. Below are examples of two clinical topics and the two queries for each topic:

> Topic: Patients with esophageal cancer who develop pericardial effusion
> Base Query: *esophageal cancer AND "pericardial effusion"*
> Segmented Query: *esophageal cancer AND (AREA[FinalDiagnosis] "pericardial effusion" OR AREA[Course] "pericardial effusion" OR AREA[LabRadResults] "pericardial effusion" OR AREA[AssessmentAndPlan] "pericardial effusion")*
>
> Topic: Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes
> Base Query: *alzheimers AND EXPAND[concept] (bed sore OR pressure ulcer) AND (NOT home OR facility OR "nursing home" OR "extended care" OR "assisted living") AND NOT expired*
> Segmented Query: *alzheimers AND EXPAND[concept] (bed sore OR pressure ulcer) AND AREA[DCDisposition] (NOT home OR facility OR nursing OR extended) AND AREA[DCDisposition] NOT expired*

Query Analysis After the queries were developed, they were run on the data corpus, and the retrieved visits were recorded. A subset of retrieved documents was examined to understand the effect of segmenting on retrieval performance. Several sets of documents were examined. The ten highest ranked documents retrieved only by the base query were manually reviewed and evaluated. The ten highest ranked documents retrieved only by the segmented query were also evaluated. In some cases, fewer than ten documents were retrieved by one query only. In those cases, all documents retrieved by only one query were evaluated.

To compare results when both queries retrieved the same document, the difference in score assigned to those documents by each query was used to decide which to examine. The difference was calculated by subtracting the score assigned by the segmented query

from the score assigned by the base query. When this value was highly positive, the base query had assigned a much greater score than the segmented query, indicating that the base query had a much higher probability of relevance to the topic. When this value was highly negative, the segmented query had assigned a much greater score than the base query, indicating that the segmented query had a much higher probability of relevance to the topic. The ten documents with the largest positive difference in score and the ten documents with the largest negative difference in score were also examined. Because the queries retrieved different numbers of documents, and because some documents received the same score from both queries, the actual number analyzed varied widely. Table 4 (next page) shows the topics and the number of documents analyzed for each topic.

To evaluate the results, several observations were recorded for all documents retrieved for each topic. These observations indicated whether the documents were relevant to the topic, whether use of the segmented query improved performance, reason for success or failure of the segmented query, and reason for retrieving non-relevant documents.

First, relevance to the topic was assessed and recorded. In the few cases where relevance could not be determined with certainty, the document was noted to be possibly relevant. Some non-relevant documents were relevant to parts of the topic, and these documents were noted to be relevant to some but not all of the topic components. A list of relevance categories is below:

Relevance to the query
- Relevant
- Possibly relevant
- Relevant to some, but not all, query components
- Not relevant

Table 4. Query topics, total number of documents analyzed, and breakdown of retrieval by query type.

Number of Documents Analyzed				Topic	Topic Description
Total	Retrieval Query				
	Base Only	Both Queries	Segmented Only		
40	10	20	10	160	Patients with low back pain who had imaging studies
40	14	26	0	183	Patients with acute vision loss
40	16	24	0	136	Patients with dental caries
40	20	20	0	171	Patients with thyrotoxicosis treated with beta-blockers
40	3	37	0	147	Patients with left lower quadrant abdominal pain
38	18	20	0	146	Patients treated for the post-partum problems depression, hypercoagulability or cardiomyopathy
30	10	20	0	140	Patients who developed disseminated intravascular coagulation in the hospital
30	10	20	0	165	Patients who have gluten intolerance or celiac disease
30	10	16	0	184	Patients with colon cancer receiving chemotherapy
29	10	19	0	174	Patients with ventilator-associated pneumonia
22	12	10	0	149	Patients with delirium, hypertension, and tachycardia
20	10	10	0	137	Patients with inflammatory disorders receiving TNF-inhibitor treatments
20	10	10	0	143	Patients who had a carotid endarterectomy during this admission
20	10	10	0	144	Patients with diabetes mellitus who also have thrombocytosis
20	10	10	0	153	Patients admitted to the hospital with end-stage chronic disease who are discharged on hospice care
20	10	9	0	158	Patients with esophageal cancer who develop pericardial effusion
20	10	10	0	167	Patients with HIV/AIDS who develop pancytopenia
20	10	10	0	173	Patients who received pneumonia vaccination during this admission
16	0	10	6	141	Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes
15	10	5	0	185	Patients who develop thrombocytopenia in pregnancy
14	7	7	0	150	Patients who have cerebral palsy and depression
10	0	0	10	179	Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

To evaluate performance of the segmented queries, document retrieval was compared across several factors: document relevance, which query retrieved the document, and the relative difference in scores assigned by the two query types. A set of codes was developed to incorporate these factors; the codes are shown in table 5.

Table 5. Codes indicating relative success/failure of the segmented compared to the base query.

Code	Document Relevance	Retrieved by Query	Rank Difference
-2	Relevant	Base only	N/A
-2	Non-relevant	Segmented only	N/A
-1	Relevant	Both	Segmented query incorrectly assigned lower rank
-1	Non-relevant	Both	Segmented query incorrectly assigned higher rank
0	Any	Both	Same
1	Relevant	Both	Segmented query correctly assigned higher rank
1	Non-relevant	Both	Segmented query correctly assigned lower rank
2	Relevant	Segmented only	N/A
2	Non-relevant	Base only	N/A

Another code indicated the reason for success or failure of the segmented queries. In some cases, no reason could be identified for success or failure, and this was noted for that topic. Reasons for failure included the following:

Query error—the query did not look in the section documenting the presence or absence of the issue, or the query searched for the wrong thing
Segmenting error—the text was not segmented properly
Chart error—search condition was documented in the wrong section of the chart

Reasons for success of the segmented queries included the following:

Targeted searching of a specific section rather than the whole document avoided retrieval of a non-relevant document
Targeted searching of a specific section allowed retrieval of a relevant document

In many cases, non-relevant documents were retrieved. These documents were examined to identify and record the reason for retrieval. Many of these reflected those found in previous research (4), although several additional reasons were identified. Table 6 describes the high-level categories and descriptions for retrieval of non-relevant documents.

Table 6. Reasons for retrieval of non-relevant documents.

Reason Category	Description
Negation	Search condition was documented as denied, ruled out, or resolved
Temporality (past)	Search condition was a past or preexisting condition or medication
Temporality (past)	Past condition or medication was mentioned because of relevance to current condition
Temporality (future)	Search condition was a future condition or a condition to be evaluated, or was care planned but not carried out yet
Temporality (temporary treatment)	Search medication was given during the current admission but not prescribed ongoing, as specified by the query
Non-relevant reference	Non-relevant reference was made to the search condition; this included references to the condition in people other than the patient
Non-relevant reference	Search condition was mentioned as part of the differential diagnosis but not definitively diagnosed
Non-relevant reference	Search procedure was started but not completed (topic specified patients with completed procedures)
Terminology differences	Clinical text used different terminology than the query
Synonymy	Search engine used incorrect synonymy when mapping terms
Query failure	Poorly written query

Codes were developed in an iterative fashion, with new codes added as necessary to capture new reasons.

Results

A total of 574 documents were examined; this set of documents represented those with the highest difference in scores between the base and segmented queries and those retrieved by one set of queries only. Of those, 344 were relevant to the corresponding retrieval topic, 18 were possibly relevant, and 230 were not relevant. Of the relevant documents, 77 were retrieved by the base query only, 20 were retrieved by the segmented query only, and 247 were retrieved by both queries. Of the non-relevant documents, 143 were retrieved by the base query only, six non-relevant documents were retrieved by the segmented query only, and 81 were retrieved by both queries. Table 7 shows retrieval counts for both sets of queries; documents with only possible relevance are treated as non-relevant documents in this table.

Table 7. Retrieval counts of relevant and non-relevant documents by query type.

	Base Query Only N(% of Total)	Both Queries N(% of Total)	Segmented Query Only N(% of Total)	Total
Relevant Documents	77 (22%)	247 (72%)	20 (6%)	344
Non-relevant Documents	143 (62%)	81 (35%)	6 (3%)	230
Total	220 (38%)	328 (57%)	26 (5%)	574

In most cases, the base query retrieved more documents than the segmented query. The segmented query returned 78% of the relevant documents examined, and the base query returned 94%. Of the non-relevant documents examined, the segmented query returned 38%, and the base query returned 97%. Table 8 shows recall, precision, and F_1 measures for both query types.

Table 8. Recall, precision, and F_1 measures for both query types.

	Recall	Precision	F_1 Measure
Base Queries	0.95	0.54	0.69
Segmented Queries	0.83	0.73	0.78

Retrieval of relevant documents Of the relevant documents returned by segmented queries, 78% had higher scores using the segmented queries as compared to the base queries, and 20% had lower scores as compared to the base queries. Twenty relevant documents were returned only by the segmented queries and 66 only by the base queries.

In most cases where the segmented query did not return a relevant document, the query failed to look in the section containing the desired information and thus did not identify the document. However, in eleven documents, the section specified by the query was not identified during the segmentation process because it did not have an explicit heading. In these cases, the previous section was labeled, and the specified section was set apart by a blank line but no explicit heading. Because these sections did not have an

explicit heading, they were included in the previous sections, so a search for information in these unlabeled sections did not retrieve any documents.

Retrieval of non-relevant documents Of the non-relevant documents retrieved by the segmented queries, 31% were assigned higher scores by the segmented query than by the base query, and 60% were assigned a lower score. Six non-relevant documents were returned only by the segmented queries and 134 by the base queries only. Reasons for returning non-relevant documents included conditions that were denied or ruled out, past conditions or medications, future or possible conditions, irrelevant references to conditions (for example, in the family history, or one word with multiple meanings), and, in one case, a procedure that was aborted prior to completion. Table 9 lists the reasons for retrieving non-relevant documents and the frequency for each reason.

Table 9. Reasons for retrieval of non-relevant documents.

Failure Reason	**Frequency**
No failure	328
Non-relevant reference: Non-relevant reference was made to the search condition; this included references to the condition in people other than the patient	85
Temporality (past): Search condition was a past or preexisting condition or medication	58
Non-relevant reference: Search condition was mentioned as part of the differential diagnosis but not definitively diagnosed	24
Negation: Search condition was documented as denied, ruled out, or resolved	22
Synonymy: Search engine used incorrect synonymy when mapping terms	20
Temporality (future): Search condition was a future condition or a condition to be evaluated, or was care planned but not carried out yet	17
Query failure: Poorly written query	11
Temporality (past): Past condition or medication was mentioned because of relevance to current condition	5
Terminology differences: Clinical text used different terminology than the query	2
Temporality (temporary treatment): Search medication was given during the current admission but not prescribed ongoing, as specified by the query	1
Non-relevant reference: Search procedure was started but not completed (topic specified patients with completed procedures)	1

Matthews Correlation Coefficient Matthews correlation coefficient (MCC) was used to evaluate performance of the segmented queries. MCC is used in machine learning as a measure of the quality of binary classifications. It was chosen for this project because it yields reliable results with small samples and can measure both increases and decreases in performance.

To calculate MCC, retrieval results were first classified according to the score differential between the segmented and base queries. For each topic, documents were grouped into true and false positives and negatives. One axis indicated relevance to the topic, with documents categorized as relevant or non-relevant. Documents that were only partially or possibly relevant were included with non-relevant documents. The other axis reflected whether the score based on the segmented query was higher than the score from the base query. Table 10, below, illustrates the classification of documents according to these criteria.

Table 10. Classification of retrieved documents based on relevance to topic and relative scores from segmented and base queries.

	Segmented score higher than base	**Segmented score lower than base**
Document relevant	True Positive	False Negative
Document not relevant	False Positive	True Negative

In some cases, only one type of query retrieved a document. For example, the base but not the segmented query might have retrieved a document. In this case, the score assigned by the base query was used, and a score of zero was used for the segmented query. Because of this, the MCC values do not reflect a difference between documents that were retrieved and documents that were not retrieved.

Possible values for MCC range from -1 to 1. If the segmented queries overall decrease performance, MCC will be less than zero. A score of -1 indicates that the segmented query yields only false positive and false negative results. If the segmented query increases performance, MCC will be greater than zero. An MCC of 1 indicates that the segmented query yielded only true positive results.

MCC was calculated for each topic as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}}$$

MCC could not be calculated for six topics because the sum of true negative and false negative or the sum of true positive and false positive was zero, resulting in a denominator of zero.

The average of all MCC values was 0.422, which is significant at $p<0.01$ ($p(|T|>|t|)=0.003$). Although there was great variability in the values of MCC, the segmented queries for eight topics performed very well, showing statistical improvement over the base queries using Fisher's exact test for significance. The segmented queries for *143: Patients who had a carotid endarterectomy during this admission* and *144: Patients with diabetes mellitus who also have thrombocytosis* performed very well with MCCs of 0.905 ($p<0.001$). The segmented query for *185: Patients who develop thrombocytopenia in pregnancy* had an MCC of 0.853 ($p<0.001$). Other high performers were *183: Patients with acute vision loss* (MCC=0.756, $p<0.001$), *173: Patients who received pneumonia vaccination during this admission* (MCC=0.734, $p<0.01$), *165: Patients who have gluten intolerance or celiac disease* (MCC=0.666, $p<0.001$), *137: Patients with inflammatory disorders receiving TNF-inhibitor treatments* (MCC=0.503,

p<0.05), and *184: Patients with colon cancer receiving chemotherapy* (MCC=0.400, p<0.05). MCC scores for each topic are listed in Table 11, next page.

Overall, eight topics showed statistically significant improvement using the segmented queries, and no topics had statistically significant decreased performance. MCC values were charted as shown in Figure 1.

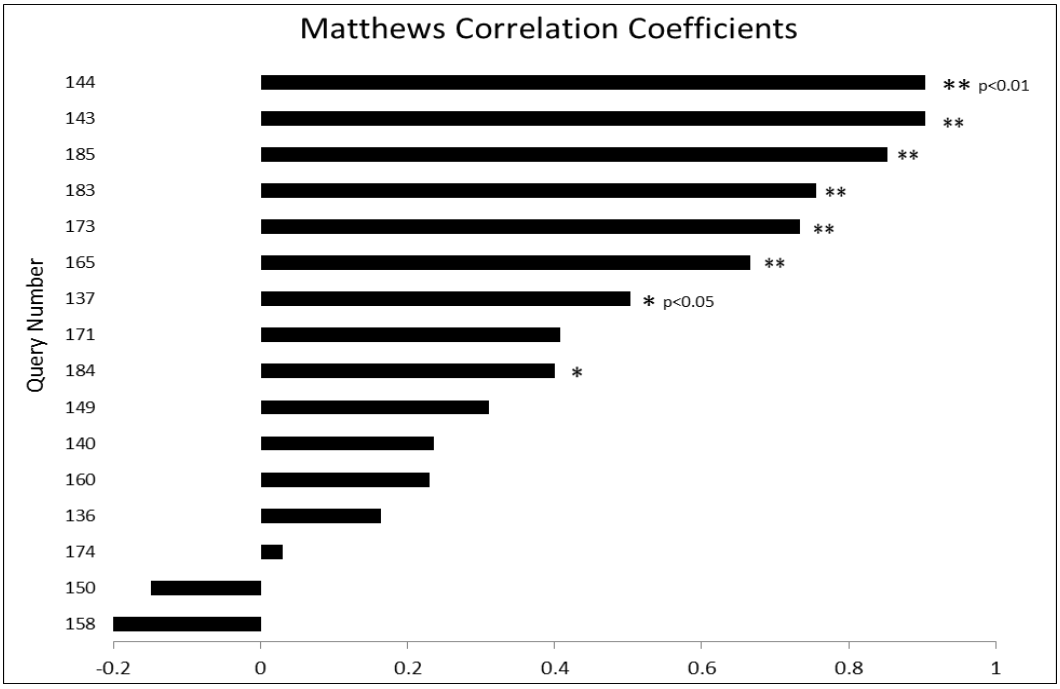

Figure 1. Matthews Correlation Coefficients for topics. * p<0.05, ** p<0.01

Table 11. Matthews correlation coefficients for each topic and Fisher's exact test for significance. * p<0.05, ** p<0.01

Query	True Positive	False Positive	True Negative	False Negative	MCC	Fisher's exact
143 Patients who had a carotid endarterectomy during this admission	9	1	10	0	0.905**	0.0000595
144 Patients with diabetes mellitus who also have thrombocytosis	9	1	10	0	0.905**	0.0000595
185 Patients who develop thrombocytopenia in pregnancy	4	1	10	0	0.853**	0.00366
183 Patients with acute vision loss	10	0	16	4	0.756**	0.0000333
173 Patients who received pneumonia vaccination during this admission	10	0	7	3	0.734**	0.00155
165 Patients who have gluten intolerance or celiac disease	9	1	16	4	0.666**	0.000405
137 Patients with inflammatory disorders receiving TNF-inhibitor treatments	8	2	7	3	0.503*	0.0322
171 Patients with thyrotoxicosis treated with beta-blockers	8	2	6	4	0.408	0.0750
184 Patients with colon cancer receiving chemotherapy	10	0	10	15	0.400*	0.0178
149 Patients with delirium, hypertension, and tachycardia	8	2	6	6	0.311	0.130
140 Patients who developed disseminated intravascular coagulation in the hospital	7	3	11	9	0.236	0.139
160 Patients with low back pain who had imaging studies	20	0	2	18	0.229	0.244
136 Patients with dental caries	7	3	6	7	0.164	0.252
174 Patients with ventilator-associated pneumonia	7	2	5	15	0.0300	0.358
150 Patients who have cerebral palsy and depression	4	3	2	5	-0.149	0.367
158 Patients with esophageal cancer who develop pericardial effusion	5	5	3	7	-0.204	0.240
179 Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression	10	0	0	0		
141 Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes	6	10	0	0		
146 Patients treated for the post-partum problems depression, hypercoagulability or cardiomyopathy	0	0	26	12		
167 Patients with HIV/AIDS who develop pancytopenia	0	0	19	1		
153 Patients admitted to the hospital with end-stage chronic disease who are discharged on hospice care	0	0	7	13		
147 Patients with left lower quadrant abdominal pain	0	0	1	2		

Precision, Recall, and Fall-Out Analysis also included calculation of precision, recall, and fall-out for each query for each topic. Precision is the proportion of retrieved relevant documents to all retrieved documents, recall is the proportion of retrieved relevant documents to all relevant documents (true positive rate), and fall-out is the proportion of retrieved non-relevant documents to all non-relevant documents (false positive rate). Whereas the MCC calculations looked at score differences without distinguishing between retrieved and non-retrieved documents, these calculations do make that distinction, allowing comparison based on whether the query retrieved a document or not.

As shown in Figure 2, precision for the segmented queries ranged from 0.1 to 1, and precision for the base queries ranged from 0.05 to 0.93. For all but three topics, precision for the segmented queries exceeded that of the base queries. In other words, more of the documents retrieved by the segmented queries were relevant. This result concurs with the MCC results, where all but three topics were improved by using the segmented queries and the three topics where performance decreased with the segmented queries did not show statistical significance.

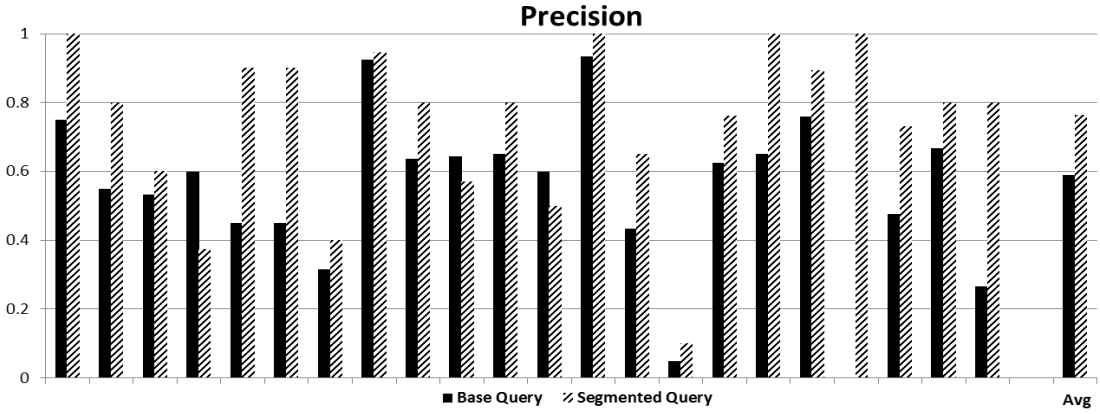

Figure 2. Precision for all topics, and average precision over all topics.

As shown in Figure 3, recall for the segmented queries ranged from 0.42 to 1, and recall for the segmented queries ranged from 0 to 1. Recall for the base queries was higher than that of the segmented queries for thirteen topics, the same for seven, and lower for two. This means that, overall, the base queries retrieved more relevant documents than the segmented queries.

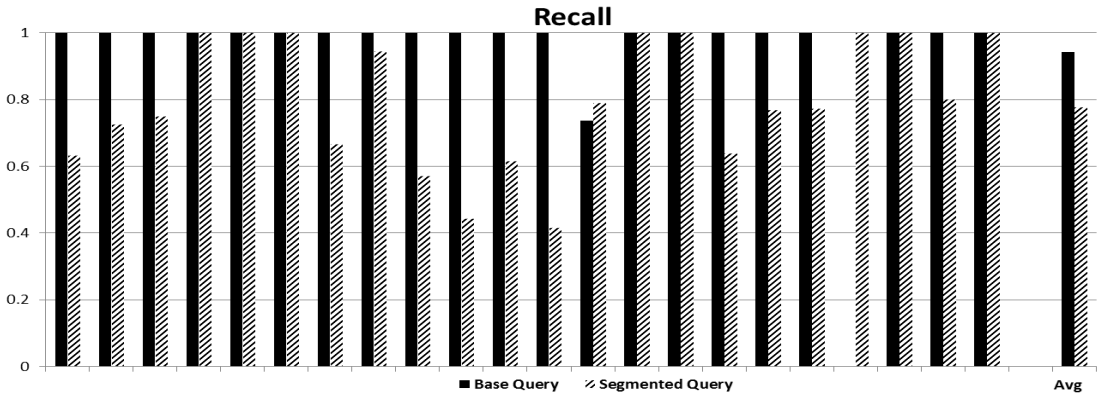

Figure 3. Recall for all topics, and average recall over all topics.

As shown in Figure 4, fall-out for the segmented queries ranged from 0 to 1, and fall-out for the base queries ranged from 0.4 to 1. Fall-out for the base queries was higher for twenty topics, and lower for only one topic, indicating that the base queries retrieved far more non-relevant documents than the segmented queries.

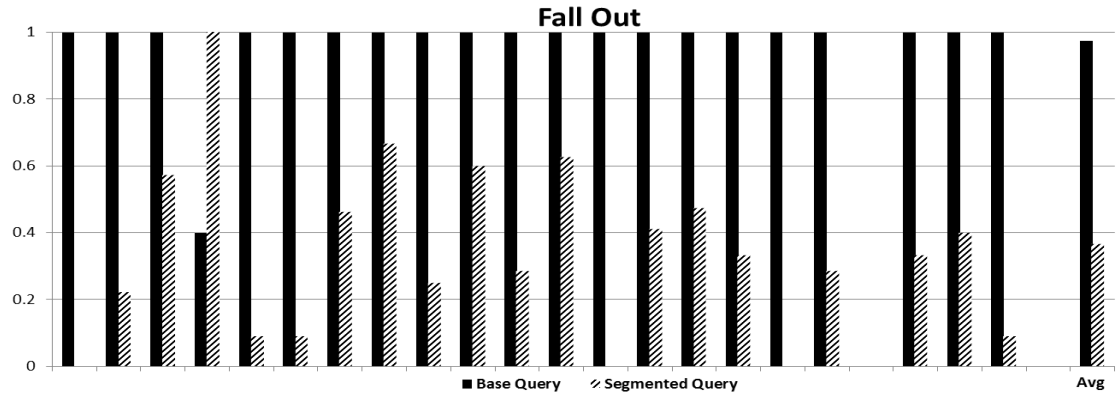

Figure 4. Fall out for all topics, and average fall out over all topics.

The F_1 measure reflects the accuracy of a test. It is calculated as 2 * precision * recall / (precision + recall). For most topics, the F_1 measures for the segmented and base queries were fairly close. In a few cases, one set of queries outperformed the others. The graph below illustrates the F_1 measures for both sets of queries.

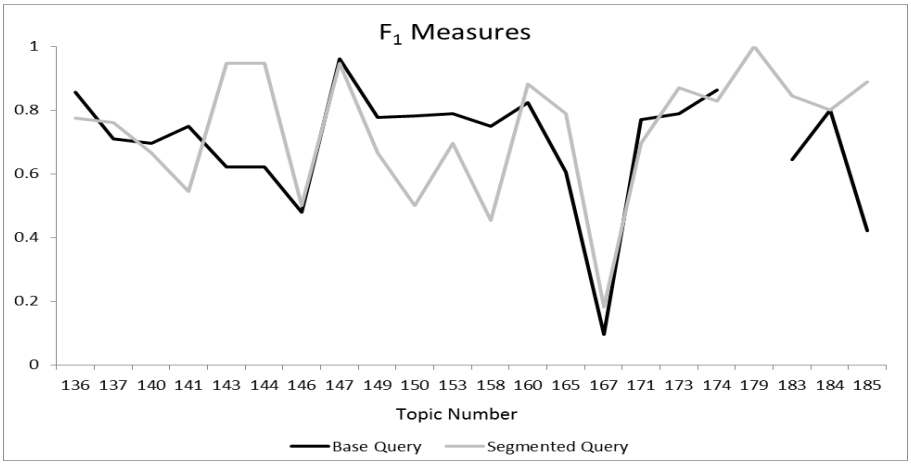

Figure 5. F_1 measures for all topics.

Discussion

In the documents examined, searching for information in specific sections of the document resulted in improved accuracy of retrieved documents, mostly by improving the precision of the query. Overall, the segmented queries retrieved only about half the number of non-relevant documents than the base queries. However, they also retrieved only about 80% of the number of relevant documents. This suggests that using a segmented query will not retrieve as large a set of relevant documents, but it is much more likely to retrieve a concentrated set of relevant documents and to avoid retrieving many non-relevant documents.

For several topics, the denominator for the MCC calculation was zero and thus could not be calculated. One of these was the topic *Patients taking atypical antipsychotics*

without a diagnosis of schizophrenia or bipolar depression. The base query looked for documents that did not contain *schizophrenia* or *bipolar depression*. The segmented query looked for documents that did not contain these conditions in sections containing the patient diagnosis but did not restrict documents that contained these conditions in non-patient fields such as family and social history. Both queries for this topic returned more than 4500 documents; all but ten of these had the same score.

The ten with differences were returned by the segmented query only and were the ten documents evaluated for this query. All ten documents were relevant to the topic. Each of these documents contained references to schizophrenia and bipolar disorder in the social history or family history sections, where mentions were made of these conditions in family members. Because the segmented query avoided looking in these sections, it was able to retrieve these relevant documents. The base query, on the other hand, specified that documents retrieved not contain any references to these conditions and thus incorrectly eliminated these relevant documents from those retrieved.

The performance differential for these queries points to an issue in document retrieval. Because documents are evaluated for retrieval at the document level, including a specification of *NOT*, as in this case, will result in the failure to retrieve a document containing the specified phrase. For topics like this, it is desirable to be able to ignore some references but not others. Searching specific sections of the document is one way to get around this issue.

Expanding the segmented queries to look at more sections for the desired information will result in retrieval of more documents, both relevant and non-relevant. One way to improve retrieval of relevant documents is to ensure greater accuracy of

labeling section headers.

Although these methods may improve the chance of retrieving relevant documents, it is important to keep in mind that the documents themselves are not perfect. Clinical text is a tool used to communicate medical information, and is often created in high-stress situations. The current examination found several cases where information was documented in non-typical sections; for example, one document noted patient's Alzheimer's disease in the social history when describing the living condition. Additionally, the sections themselves may vary slightly between institutions or clinicians. The current data set contained problem lists in some documents, while other documents listed ongoing problems only in the past medical history.

Another factor that affects retrieval accuracy is the nature of medical text. As noted in previous works, retrieving information from medical text is complicated by the use of negation, temporality, and context-dependent meaning of words.(4) Looking in specific sections alleviates some of these issues, especially temporality, but not all of them. However, constructing queries to search specific sections adds to the complexity of writing queries. As shown in Appendix C, most of the segmented queries are much longer than the base queries because of the need to duplicate terms in each specified section of the search.

Limitations

There were several limitations to this study. One person did all of the topic development, query development, relevance judging, and coding of the data. It would have been preferable to split these tasks among different people to reduce the likelihood

of introducing bias in the results. Additionally, the sample size was very small. Typically an information-retrieval task would include several hundred documents for each topic, ensuring that results are more generalizable.

The method of defining sections produced several inaccuracies in segmenting the documents, leading to decreased performance of the segmented queries. After reviewing the results, it was apparent that many of the queries themselves could have been better developed to take advantage of the document sections. In addition, for this project, the segmented queries were developed to reflect the base queries as much as possible. Many would be improved by developing them without reflecting the base queries, allowing them to fully utilize the characteristics of the segmented documents.

Future Work

Future work should take several approaches. First, because of inaccuracies in labeling section headings in the current project, a validated sectioning tool should be used. The tool chosen should have the ability to identify section changes that are not explicitly identified by headings. Second, some of the common retrieval issues in medical text, such as negation and temporality, should be identified using a published tool. After taking these steps, a set of queries can be developed and run, and a quantitative analysis of a large sample of the results can be done to provide greater insight into the effectiveness and limitations of segmenting documents on retrieval accuracy.

Future work should also take into account a document format found in some healthcare communications. Situation Background Assessment Recommendation

(SBAR) is a structure used to communicate patient status between non-physician providers. This format bears a great deal of resemblance to SOAP; however, the initials S and A in SBAR have very different meanings from the same initials in SOAP. Mislabeling them would lead to misinterpretation of the content. The sectioning tool used in future work should be able to distinguish these two note formats.

Improving the ability to search clinical text will have great benefit for secondary use of EHR data. Increasing attention is being paid to improving clinical quality, improving population health, utilizing EHR data for research, and other secondary uses. All of these applications depend on the ability to retrieve complete and accurate patient cohorts. As suggested by numerous studies, complete cohorts can rarely be retrieved without utilizing clinical text, which is difficult to search accurately. The findings of the current study suggest that segmenting clinical documents greatly improves retrieval accuracy. Because of this, continuing this line of investigation will ultimately improve the ability to utilize EHR data for secondary use.

References

1. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc JAMIA. 2007 Feb;14(1):1–9.

2. Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol. 2012;8(12):e1002823.

3. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc JAMIA. 2014 Mar 1;21(2):221–30.

4. Edinger T, Cohen AM, Bedrick S, Ambert K, Hersh W. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2012;2012:180–8.

5. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011 Apr 20;3(79):79re1.

6. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc JAMIA. 2010 Feb;17(1):19–24.

7. i2b2: Informatics for Integrating Biology & the Bedside [Internet]. [cited 2014 May 18]. Available from: https://www.i2b2.org/index.html

8. Voorhees E, Hersh W. Overview of the TREC 2012 Medical Records Track. Twenty-First Text Retr Conf TREC 2012 Proc [Internet]. 2012; Available from: http://trec.nist.gov/pubs/trec21/t21.proceedings.html

9. Apostolova E, Channin DS, Demner-Fushman D, Furst J, Lytinen S, Raicu D. Automatic segmentation of clinical texts. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Conf. 2009;2009:5905–8.

10. Cho PS, Taira RK, Kangarloo H. Automatic section segmentation of medical reports. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2003;155–9.

11. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc JAMIA. 2009 Dec;16(6):806–15.

12. Mowery D, Wiebe J, Visweswaran S, Harkema H, Chapman WW. Building an automated SOAP classifier for emergency department reports. J Biomed Inform. 2012 Feb;45(1):71–81.

13. Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical Section Segmentation in Free-Text Clinical Records. LREC [Internet]. 2012 [cited 2012 Oct 8]. Available from: http://staff.washington.edu/melihay/publications/LREC_2012.pdf

14. Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol. 2002;29:641–4.

15. The Unified Medical Language System (UMLS) [Internet]. [cited 2014 Mar 23]. Available from: http://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html

16. Bodenreider O, Willis J, Hole W. The Unified Medical Language System: What is it and how to use it? [Internet]. MEDINFO; 2004 Sep 8; San Francisco, CA. Available from: http://www.nlm.nih.gov/research/umls/presentations/2004-medinfo_tut.pdf

17. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993 Apr;81(2):217–22.

18. Hersh W. Information Retrieval: A Health and Biomedical Perspective. Springer; 2008. 496 p.

19. Fact Sheet SPECIALIST Lexicon [Internet]. [cited 2014 Mar 23]. Available from: https://www.nlm.nih.gov/pubs/factsheets/umlslex.html

20. McCray AT, Ide NC. Design and implementation of a national clinical trials registry. J Am Med Inform Assoc JAMIA. 2000 Jun;7(3):313–23.

21. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc JAMIA. 2007 Jun;14(3):253–63.

Appendix A. Section Headings and Weights

Weight	Section Name
0.60	Addendum
0.60	AdmissionDiagnosis
0.60	Allergies
0.60	AssessmentAndPlan
0.70	Category
0.25	ChartTime
0.60	ChiefComplaint
0.60	CodeStatus
0.60	Comments
0.60	Complications
0.60	Condition
0.60	Course
0.60	DCCondition
0.60	DCDisposition
0.60	DCInstructions
0.60	DCMeds
0.60	Environment
0.60	FamilyHistory
0.60	FinalDiagnosis
0.60	FluidsLytesNutr
0.60	HistoryOfPresentIllness
0.60	HomeMeds
0.75	HospitalAdmissionId
0.60	HospitalMeds
0.60	Indication
0.60	LabRadResults
0.60	Objective
0.60	PastMedicalHistory
0.60	PhysicalExam
0.60	Preamble
0.60	ProblemList
0.60	ProcedureDetails
0.60	ProcedureType
0.60	RHCM
0.60	ROS
0.60	SocialHistory
0.60	StudyImpression
0.99	SubjectId
0.60	Subjective
0.60	SubjectiveAndObjective
0.65	Title

Appendix B. Section Heading Rules

Each line contains one rule with sections separated by '|' character. The first section indicates the document type to be searched: DSC is the discharge summary, MDN MD notes, NUR nursing notes, RAD radiology reports, and ALL searches all document types. PreMarker, the next section, indicates text or characters that must precede the indicator text, and NegPreMarker indicates text that must not be present preceding the indicator text. Tag indicated the section heading to be inserted when the indicator text is found. PostMarkers are text or characters that must follow the indicator text. GtOrLt and Percent together indicate which portion of the document the indicator text must occur in; for example, <50 indicates that the text must be in the first half of the document. The last field on each line is for comments.

```
#DocumentType|PreMarker|NegPreMarker|Tag|Indicator|PostMarkers|GtOrLt|Percent|Comment
DSC|^||addendum|addend|um:#a:|||
ALL|||admission_diagnosis|admission diagnoses|$#:# are:# at this time:# at this time are:#sat this time:|||
ALL|^||admission_diagnosis|admission diagnosis|$#:#(es)$#(es):|||
ALL|^||admission_diagnosis|admitting diagnoses|$#:|||
ALL|^||admission_diagnosis|admitting diagnosis|$#:#(es)$#(es):|||
ALL|^||admission_diagnosis|diagnoses on admission|:|||
ALL|^||admission_diagnosis|diagnosis on admission|:|||
ALL|^||admission_diagnosis|pre op diagnosis|$#:|||
ALL|^||admission_diagnosis|preop diagnosis|$#:|||
ALL|^||admission_diagnosis|pre-op diagnosis|$#:|||
ALL|^||admission_diagnosis|preoperative diagnoses|$#:|||
ALL|^||admission_diagnosis|preoperative diagnosis|$#:#(es)$#(es):|||
ALL|||admission_diagnosis|preoperative impression|:|||
ALL|^||admission_diagnosis|primary diagnoses on admission|:|||
ALL|^||admission_diagnosis|reasons for admission|:|||
ALL|^||admission_diagnosis|referring diagnosis|$#:|||
ALL|||allergies|allergic to||||
ALL|||allergies|allergies|$#:#-# to# are to|||
ALL|||allergies|allergies and reactions|:|||
ALL|||allergies|allergies to medications|:#-#are|||
ALL|||allergies|allergy/reaction profile|:#-#to#are to|||
ALL|||allergies|allergy/reaction profile (all)|:#-#to#are to|||
ALL|||allergies|medicine allergies|:#-#to#are to|||
ALL|||allergies|penicillin allergy||||
ALL|has an #has ||allergies|allergy to||||
MDN|^||assessment_and_plan|a|:#-# -|||
NUR|^||assessment_and_plan|a|:#-# -|||
ALL|^||assessment_and_plan|a/p|:#-# -|||
ALL|||assessment_and_plan|assessment|:#-|||
ALL|^||assessment_and_plan|assessment and plan|$#:#-|||
ALL|||assessment_and_plan|assessment and recommendations|:|||
ALL|^||assessment_and_plan|assessment/plan|$#:|||
DSC|||assessment_and_plan|clinical impression|$#:#-#s$#s:#s-|||
MDN|||assessment_and_plan|clinical impression|$#:#-#s$#s:#s-|||
NUR|||assessment_and_plan|clinical impression|$#:#-#s$#s:#s-|||
ALL|||assessment_and_plan|disposition/plan|:|||
MDN|^||assessment_and_plan|imp|/#:|||
DSC|^||assessment_and_plan|impression|$#:#-#s$#s:#s-|||
MDN|^||assessment_and_plan|impression|$#:#-#s$#s:#s-|||
```

NUR|^||assessment_and_plan|impression|$#:#-#s$#s:#s-|||
ALL|^||assessment_and_plan|impression and plan|$#:#-|||
ALL|||assessment_and_plan|impression and recommendations|:|||
ALL|||assessment_and_plan|impression/plan|:|||
ALL|^||assessment_and_plan|"in summary, this is a"||||
MDN|^||assessment_and_plan|p|:#-# -|||
NUR|^||assessment_and_plan|p|:#-# -|||
ALL|||assessment_and_plan|plan/followup|:#-|||
ALL|^||assessment_and_plan|recommendations|:#-#are as follows|||
ALL|||assessment_and_plan|treatment/plan|:|||
DSC|||chief_complaint|chief complaint|$# is# of#:#-# -|||
ALL|^||chief_complaint|reason for presentation|$# is#:#-# -|||
ALL|^||code_status|code status|$#:# is# was#-|||
ALL|\d+[.:]||code_status|code status||||
RAD|||comments|comment|:|||
ALL|^||complications|complications|$#:#-|||
ALL|||condition|admitting condition|:#-# is# was|||
ALL|||condition|clinical condition at present|:#-# is# was|||
DSC|||condition|condition|:#-# is# was|||
MDN|||condition|condition|:#-# is# was|||
NUR|||condition|condition|:#-# is# was|||
RAD|||condition|underlying medical condition|:#-# is# was|||
ALL|||condition|patient condition|:#-# is# was|||
ALL|||course|assessment and ed course|:#-|||
DSC|||course|brief hospital course|:#-# -|||
ALL|||course|clinical course|:|||
ALL|^||course|clinical course in emergency department|:|||
ALL|^||course|clinical course in the emergency department|:|||
ALL|^||course|course|:# in th ward:# in the ward:# in the icu:# in the emergency department:# of treatment:|||
ALL|||course|death summary|:|||
ALL|||course|discharge summary|:|||
ALL|||course|ed course|:#-# and plan:|||
ALL|||course|ed course and medical decision making|:#-|||
ALL|||course|ed course and medical decision-making|:#-|||
ALL|||course|emergency department course|:#-|||
ALL|||course|emergency department course/medical decision making|:#-|||
ALL|||course|emergency room course|:|||
ALL|||course|hospital course|$#:#-# (by systems):# and labs:# and plan:# and plan-# and treatment:|||
ALL|^||course|hospital presentation|:|||
ALL|||course|hospitalization course|:|||
ALL|||course|hospitalization course and treatment|:|||
ALL|||course|institution course|:|||
DSC|||course|intensive care unit course|:#-# -|||
ALL|||course|medical decision making|:#-|||
ALL|||course|medical decision making in the emergency room|:#-|||
ALL|^||course|medical decision-making|:#-|||
ALL|||course|medical management|:|||
ALL|^||course|medical-decision making|:#-|||
ALL|||course|other interventions|:|||
ALL|^||course|patient status|$|||
ALL|||course|perioperative complications|:|||
ALL|^||course|review of admission|:|||
ALL|^||course|summary|:|||
ALL|||course|summary of hospitalization|:|||
ALL|||course|transfer summary|:|||
ALL|^||course|treatment|:|||
ALL|||dc_condition|condition at the time of discharge|:#-# is# was|||

ALL|||dc_condition|condition on discharge|:#-# is# was|||
ALL|||dc_condition|condition on transfer|:#-# is# was|||
ALL|||dc_condition|condition upon discharge|:#-# is# was|||
DSC|^||dc_condition|discharge condition|:#-# is# was|||
ALL|||dc_condition|discharge status|:#-# is# was|||
DSC|||dc_disposition|discharge disposition|:#-# -|||
DSC|||dc_disposition|dispo|$#:#-# to#sition$#sition:#sition-#sition to|||
MDN|^||dc_disposition|dispo|$#:#-# to#sition$#sition:#sition-#sition to|||
DSC|||dc_disposition|transferred to|$#:#-# -# care of|||
DSC|||dc_instructions|activities|:||| DSC|||dc_instructions|activity|:|||
DSC|||dc_instructions|diet|:|||
ALL|||dc_instructions|discharge information|:# are as follows#-|||
ALL|||dc_instructions|discharge instructions|:# are as follows|||
ALL|||dc_instructions|discharge orders|:|||
ALL|^||dc_instructions|discharge plan|$#:# are as follows|||
ALL|^||dc_instructions|discharge planning and instructions|:|||
ALL|||dc_instructions|follow up|:# are as follows#-|||
ALL|||dc_instructions|follow up instructions|:# are as follows#-|||
ALL|||dc_instructions|follow up with||||
ALL|||dc_instructions|followup|:# are as follows#-|||
ALL|||dc_instructions|follow-up|:# are as follows#-|||
ALL|||dc_instructions|followup instructions|:#-# - # are|||
ALL|||dc_instructions|follow-up instructions|:|||
ALL|||dc_instructions|home instructions|:|||
ALL|||dc_instructions|instructions|:# are as follows#-|||
ALL|^||dc_instructions|patient instructions|$|||
ALL|||dc_instructions|to have followup in||||
DSC|||dc_instructions|transfer instructions|:# are as follows#-|||
ALL|||dc_meds|discharge medications|:# are|||
DSC|||dc_meds|discharge meds|:|||
DSC|||dc_meds|medications|:|>|51|
DSC|||dc_meds|medications at discharge|:|||
DSC|||dc_meds|medications at tansfer|:|||
DSC|||dc_meds|medications at time of discharge|:|||
DSC|||dc_meds|medications on discharge|:|>|50|Added >50 b/c of dschg rpt that noted meds on discharge from referring institution
DSC|||dc_meds|medications on transfer|:|||
DSC|||dc_meds|medications upon discharge|:|||
DSC|||dc_meds|transfer medications|:#-# are# include|||
MDN|^||environment|env't|:|||
DSC|^||family_history|family history|$#:#-# of# to# is# shows|||
ALL|^||family_history|family hx|:|||
ALL|^||family_history|fhx|:|||
ALL|^||final_diagnosis|cause of death|:|||
DSC|||final_diagnosis|clinical impression|:#-|||
ALL|^||final_diagnosis|diagnoses|$#:#-# at discharge# on transfer# at transfer# upon discharge|||
DSC|^||final_diagnosis|diagnoses at time of death|:|||
ALL|^||final_diagnosis|diagnosis|$#:#-# is# was#(es):#(es)$#(es)-|||
DSC|^||final_diagnosis|diagnosis at time of death|:|||
ALL|^||final_diagnosis|discharge diagnoses|$#:#-|||
DSC|^||final_diagnosis|discharge diagnosis|$#:#-#(es):#(es)-#(es)$|||
DSC|||final_diagnosis|discharge impression|:#-|||
DSC|^||final_diagnosis|discharge/transfer diagnosis|:#(es):|||
DSC|||final_diagnosis|discharged diagnosis|:|||
ALL|||final_diagnosis|final diagnoses|:#-# at the time of discharge:|||
ALL|||final_diagnosis|final diagnosis|:#-#(es):#(es)-|||
ALL|||final_diagnosis|final discharge diagnoses|:|||

ALL|||final_diagnosis|final discharge diagnosis|:|||
ALL|||final_diagnosis|final impressions|:#-|||
DSC|||final_diagnosis|pertinent secondary|:|||
DSC|^||final_diagnosis|post op diagnosis|$#:|||
DSC|^||final_diagnosis|post-op diagnosis|$#:|||
DSC|^||final_diagnosis|postoperative diagnoses|$#:|||
DSC|^||final_diagnosis|postoperative diagnosis|$#:#(es)$#(es):|||
DSC|||final_diagnosis|postoperative impression|:#s:|||
DSC|||final_diagnosis|primary diagnosis|:#(es):|||
DSC|||final_diagnosis|primary diagnosis(es)|:#-|||
DSC|^||final_diagnosis|principal transfer diagnosis|:#(es):|||
DSC|||final_diagnosis|resident impression|:#-|||
DSC|^||final_diagnosis|secondary diagnoses|$#:#(es)$#(es):|||
DSC|||final_diagnosis|tcu diagnosis(es)|:|||
DSC|^||final_diagnosis|transfer diagnoses|:# are:|||
DSC|||final_diagnosis|transfer diagnosis|:#(es):#(es)-|||
MDN|^||fluids_electrolytes_nutrition|fen|:|||
ALL|||history_of_present_illness|admission history and summary|:|||
DSC|||history_of_present_illness|background history|:|||
ALL|^||history_of_present_illness|brief clinical history|$#:|||
DSC|||history_of_present_illness|brief history|:|||
ALL|||history_of_present_illness|brief history and physical|:|||
ALL|||history_of_present_illness|brief hpi|:|||
DSC|^||history_of_present_illness|brief presentation|:|||
ALL|^||history_of_present_illness|clinical history|$#:|||
DSC|^||history_of_present_illness|clinical presentation|:|||
ALL|^||history_of_present_illness|history|:|||
ALL|||history_of_present_illness|history and physical exam|:|||
ALL|^||history_of_present_illness|history of present illness|$#:#-|||
DSC|^||history_of_present_illness|history of presentation|:|||
ALL|||history_of_present_illness|history of presenting disease|:#-|||
ALL|||history_of_present_illness|history of presenting illness|:|||
ALL|^||history_of_present_illness|hpi|:#-|||
ALL|^||history_of_present_illness|hx|:|||
ALL|||history_of_present_illness|institution admission|:|||
ALL|||history_of_present_illness|patient history|:|||
DSC|^||history_of_present_illness|presentation|:|||
ALL|||history_of_present_illness|presenting history|:|||
ALL|^||history_of_present_illness|reason for admission|$#:|||
DSC|||history_of_present_illness|reason for consult|:|||
DSC|||history_of_present_illness|reason for consultation|:|||
ALL|^||history_of_present_illness|reason for hospitalization|:|||
ALL|||home_meds|current medications|:#-# -|||
ALL|||home_meds|home medications|:#-# -# are|<|70|
ALL|||home_meds|home meds|:#-# -|<|70|
DSC|||home_meds|medications|:|<|70|
DSC|||home_meds|medications include||<|70|
ALL|^||home_meds|medications on admission|:|||
ALL|^||home_meds|medications prior to admission|:|||
ALL|||hospital_medications|medication administration orders|$|||
DSC|||hospital_meds|medications at time of dictation|:|||
RAD|^||indication|indication|:|||Often followed by findings without their own separate heading--findings will be included in indication section
RAD|^||indication|reason|:|||
RAD|^||indication|reason for this examination:||||
ALL|||lab_rad_results|additional imaging|:|||
DSC|||lab_rad_results|admission labs|:|||

MDN|^||lab_rad_results|bili|:#-# -|||
ALL|^||lab_rad_results|chest x-ray|:|||
DSC|||lab_rad_results|ct head|:|||
DSC|||lab_rad_results|ct of the abdomen with iv contrast|:|||
DSC|||lab_rad_results|ct of the pelvis with iv contrast|:|||
DSC|||lab_rad_results|cta chest|:|||
ALL|||lab_rad_results|current imaging|:|||
DSC|^||lab_rad_results|cxr|:|||
ALL|||lab_rad_results|diagnostic data|:|||
DSC|||lab_rad_results|diagnostic imaging|:|||
DSC|||lab_rad_results|diagnostic studies|:|||
DSC|||lab_rad_results|diagnostics|:|||
DSC|^||lab_rad_results|ekg|:|||
ALL|^||lab_rad_results|general results|$|||
ALL|^||lab_rad_results|imaging|:|||
DSC|^||lab_rad_results|important diagnostics|:|||
DSC|||lab_rad_results|important diagnostics and labs|:|||
ALL|^||lab_rad_results|lab values|:|||
ALL|||lab_rad_results|lab/imaging|:|||
ALL|||lab_rad_results|laboratories|:#-|||
ALL|^||lab_rad_results|laboratories and diagnostics|:|||
DSC|||lab_rad_results|laboratories prior to transfer|:|||
ALL|^||lab_rad_results|laboratories/diagnostics|:|||
DSC|||lab_rad_results|laboratory|:#-|||
DSC|||lab_rad_results|laboratory and diagnostic data|:#-|||
ALL|^||lab_rad_results|laboratory and diagnostics|:|||
DSC|||lab_rad_results|laboratory and diagnostics|:#-|||
ALL|^||lab_rad_results|laboratory data|$#:#-|||
DSC|||lab_rad_results|laboratory data|$#:#-|||
ALL|||lab_rad_results|laboratory data and imaging|:|||
ALL|||lab_rad_results|laboratory data/diagnostic data|:|||
ALL|||lab_rad_results|laboratory data/imaging|:|||
ALL|||lab_rad_results|laboratory evaluation|:|||
ALL|||lab_rad_results|laboratory exam|:#s:|||
ALL|||lab_rad_results|laboratory findings|:|||
ALL|||lab_rad_results|laboratory results|:#-|||
ALL|||lab_rad_results|laboratory studies|:#-|||
ALL|||lab_rad_results|laboratory studies and imaging|:|||
ALL|||lab_rad_results|laboratory studies/imaging|:|||
ALL|||lab_rad_results|laboratory studies/other imaging|:|||
ALL|||lab_rad_results|laboratory testing and imaging|:|||
ALL|||lab_rad_results|laboratory tests|:|||
ALL|||lab_rad_results|laboratory values|:|||
ALL|||lab_rad_results|"laboratory, radiographic and other diagnostic study findings"|:#-|||
ALL|||lab_rad_results|laboratory/diagnostic data|:#-|||
ALL|||lab_rad_results|laboratory/imaging|:|||
DSC|^||lab_rad_results|laboratory/radiologic data|:|||
ALL|^||lab_rad_results|labs|$#:#-|||
ALL|^||lab_rad_results|labs and diagnostics|:|||
ALL|||lab_rad_results|labs/imaging|:|||
ALL|||lab_rad_results|pertinent lab data and radiology studies|:|||
ALL|||lab_rad_results|pertinent laboratory data|:|||
DSC|^||lab_rad_results|pertinent results|:|||
ALL|||lab_rad_results|principal procedures and test|:|||
ALL|||lab_rad_results|procedures/imaging|:|||
ALL|||lab_rad_results|radiographs|:#-|||
ALL|||lab_rad_results|radiological data|:|||

ALL|^||lab_rad_results|review of diagnostics|:|||
ALL|||lab_rad_results|studies|:#-|||
ALL|||lab_rad_results|test results|:|||
ALL|||lab_rad_results|tests/imaging|:|||
ALL|||lab_rad_results|transfer labs|:|||
ALL|^||lab_rad_results|x-ray|:|||
MDN|^||objective|o|:|||
ALL|^||past_medical_history|immunization history|:|||
ALL|^||past_medical_history|immunizations|:|||
ALL|^||past_medical_history|medical history|$#:#-|||
DSC|^||past_medical_history|past gynecologic history|:|||
DSC|||past_medical_history|past medical history|$#:#-|||
ALL|||past_medical_history|past medical/surgical history|$#:#-|||
DSC|||past_medical_history|past obstetric history|:|||
DSC|||past_medical_history|past surgical history|$#:#-|||
ALL|||past_medical_history|pmh|:#-#x$#x:#x-|||
ALL|||past_medical_history|prior medical history|:|||
ALL|||past_medical_history|psh|:#-|||
ALL|||past_medical_history|pshx|$#:#-|||
MDN|^||physical_exam|cvr/resp|:#-# -|||
DSC|||physical_exam|discharge physical|:|||
ALL|^||physical_exam|general exam|$#:#-|||
DSC|^||physical_exam|initial physical|:|||
DSC|^||physical_exam|on physical exam|","|||
DSC|^||physical_exam|pe|:|||
MDN|^||physical_exam|pe|:#x|||
ALL|||physical_exam|physical evaluation|:|||
ALL|^||physical_exam|physical exam|$#:#-#ination:#ination-#ination$|||
DSC|||physical_exam|physical exam is as follows|:|||
ALL|^||physical_exam|physical examination on admission|$#:#-|||
ALL|||physical_exam|vitals|:|||
ALL|^||problem_list|active problem list|$#:|||
ALL|||problem_list|problem list|:#s:|||
ALL|^||problem_list|problem list|$|||
ALL|^||problem_list|problems|$#:#-|||
RAD|^||procedure_details|final report|$#:#-# (revised)$|||
RAD|||procedure_details|findings|:#-|||
RAD|^||procedure_details|findings|$|||
DSC|^||procedure_details|procedure|:|||
RAD|^||procedure_details|procedure and findings|$#:|||
RAD|||procedure_details|technique|:|||
DSC|||procedure_type|major surgical or invasive procedure|:|||
MDN|^||rhcm|rhcm|:|||
ALL|^||ros|review of systems|$#:#-# was significant for# was positive for|||
ALL|||ros|ros|:|||
ALL|^||social_history|shx|:|||
ALL|||social_history|social history|$#:#-|||
ALL|^||social_history|social history and habits|:|||
ALL|^||social_history|social hx|:|||
ALL|||social_history|tobacco usage|:|||
RAD|^||study_impression|conclusion|$#:#-#s$#s:#s-|||
RAD|||study_impression|final result|:#-|||
RAD|^||study_impression|impression|:|||
RAD|^||study_impression|interpretation|$#:#-|||
MDN|^||subjective|s|:|||
NUR|^||subjective_and_objective|s/o|:|||

Appendix C. Queries

141: Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes	
Baseline Query	alzheimers AND EXPAND[concept] (bed sore OR pressure ulcer) AND (NOT home OR facility OR "nursing home" OR "extended care" OR "assisted living") AND NOT expired
Sectioned Query	alzheimers AND EXPAND[concept] (bed sore OR pressure ulcer) AND AREA[DCDisposition] (NOT home OR facility OR nursing OR extended) AND AREA[DCDisposition] NOT expired
143: Patients who had a carotid endarterectomy during this admission	
Baseline Query	carotid endarterectomy
Sectioned Query	AREA[ProcedureType] carotid endarterectomy OR (AREA[FinalDiagnosis] carotid endarterectomy AND NOT (AREA[FinalDiagnosis] ("status post carotid endarterectomy" OR "s/p carotid endarterectomy" OR "status post left carotid endarterectomy" OR "s/p left carotid endarterectomy" OR "status post right carotid endarterectomy" OR "s/p right carotid endarterectomy" OR "status post bilateral carotid endarterectomy" OR "status post bilateral carotid endarterectomies" OR "s/p carotid endarterectomy" OR "past carotid endarterectomy" OR "s/p l carotid endarterectomy" OR "s/p r carotid endarterectomy" OR "s/p bilateral carotid endarterectomy" OR "s/p bilateral carotid endarterectomies")))
147: Patients with left lower quadrant abdominal pain	
Baseline Query	EXPAND[concept] left lower quadrant abdominal pain
Sectioned Query	EXPAND[concept] left lower quadrant abdominal pain AND NOT AREA[PastMedicalHistory] left lower quadrant abdominal pain AND NOT AREA[FamilyHistory] left lower quadrant abdominal pain
149: Patients with delirium, hypertension, and tachycardia	
Baseline Query	delirium AND hypertension AND tachycardia
Sectioned Query	(AREA[Addendum] delirium OR AREA[AdmissionDiagnosis] delirium OR AREA[AssessmentAndPlan] delirium OR AREA[ChiefComplaint] delirium OR AREA[Complications] delirium OR AREA[Condition] delirium OR AREA[Course] delirium OR AREA[DCDisposition] delirium OR AREA[DCInstructions] delirium OR AREA[FinalDiagnosis] delirium OR AREA[HistoryOfPresentIllness] delirium OR AREA[Indication] delirium OR AREA[LabRadResults] delirium OR AREA[Objective] delirium OR AREA[PhysicalExam] delirium OR AREA[Preamble] delirium OR AREA[ProblemList] delirium OR AREA[ROS] delirium OR AREA[Subjective] delirium OR AREA[SubjectiveAndObjective] delirium) AND (AREA[Addendum] hypertension OR AREA[AdmissionDiagnosis] hypertension OR AREA[AssessmentAndPlan] hypertension OR AREA[ChiefComplaint] hypertension OR AREA[Complications] hypertension OR AREA[Condition] hypertension OR AREA[Course] hypertension OR AREA[DCDisposition] hypertension OR AREA[DCInstructions] hypertension OR AREA[FinalDiagnosis] hypertension OR AREA[HistoryOfPresentIllness] hypertension OR AREA[Indication] hypertension OR AREA[LabRadResults] hypertension OR AREA[Objective] hypertension OR AREA[PastMedicalHistory] hypertension OR AREA[PhysicalExam] hypertension OR AREA[Preamble] hypertension OR AREA[ProblemList] hypertension OR AREA[ROS] hypertension OR AREA[Subjective] hypertension OR AREA[SubjectiveAndObjective] hypertension) AND tachycardia

Appendix C. Queries, continued

153: Patients admitted to the hospital with end-stage chronic disease who are discharged on hospice care	
Baseline Query	(TERM_EXPAND end-stage OR TERM_EXPAND terminal NOT[FUZZY] "terminal ileum") AND hospice
Sectioned Query	(TERM_EXPAND end-stage OR TERM_EXPAND terminal NOT[FUZZY] "terminal ileum") AND hospice AND (AREA[DCDisposition] hospice OR AREA[DCInstructions] hospice OR AREA[DCCondition] hospice)
165: Patients who have gluten intolerance or celiac disease	
Baseline Query	celiac disease AND NOT celiac artery
Sectioned Query	(AREA[Addendum] celiac disease OR AREA[AdmissionDiagnosis] celiac disease OR AREA[AssessmentAndPlan] celiac disease OR AREA[ChiefComplaint] celiac disease OR AREA[Course] celiac disease OR AREA[Condition] celiac disease OR AREA[DCCondition] celiac disease OR AREA[DCDisposition] celiac disease OR AREA[DCInstructions] celiac disease OR AREA[FinalDiagnosis] celiac disease OR AREA[HistoryOfPresentIllness] celiac disease OR AREA[HomeMeds] celiac disease OR AREA[HospitalMeds] celiac disease OR AREA[Indication] celiac disease OR AREA[PastMedicalHistory] celiac disease OR AREA[PhysicalExam] celiac disease OR AREA[ProblemList] celiac disease OR AREA[Preamble] celiac disease OR AREA[ROS] celiac disease OR AREA[Subjective] celiac disease OR AREA[SubjectiveAndObjective] celiac disease) AND_NOT celiac artery
165: Patients with HIV/AIDS who develop pancytopenia	
Baseline Query	pancytopenia AND (aids OR hiv)
Sectioned Query	pancytopenia AND (AREA[Addendum] (aids OR hiv) OR AREA[AdmissionDiagnosis] (aids OR hiv) OR AREA[AssessmentAndPlan] (aids OR hiv) OR AREA[ChiefComplaint] (aids OR hiv) OR AREA[Course] (aids OR hiv) OR AREA[Condition] (aids OR hiv) OR AREA[DCCondition] (aids OR hiv) OR AREA[DCDisposition] (aids OR hiv) OR AREA[DCInstructions] (aids OR hiv) OR AREA[FinalDiagnosis] (aids OR hiv) OR AREA[HistoryOfPresentIllness] (aids OR hiv) OR AREA[HomeMeds] (aids OR hiv) OR AREA[HospitalMeds] (aids OR hiv) OR AREA[Indication] (aids OR hiv) OR AREA[PastMedicalHistory] (aids OR hiv) OR AREA[PhysicalExam] (aids OR hiv) OR AREA[ProblemList] (aids OR hiv) OR AREA[Preamble] (aids OR hiv) OR AREA[ROS] (aids OR hiv) OR AREA[Subjective] (aids OR hiv) OR AREA[SubjectiveAndObjective] (aids OR hiv))
173: Patients who recieved pneumonia vaccination during this admission	
Baseline Query	pneumovax OR ("pneumonia vaccine" AND_NOT "no pneumonia vaccine")
Sectioned Query	AREA[Course] (pneumovax OR ("pneumonia vaccine" AND_NOT "no pneumonia vaccine")) OR AREA[ProcedureType] (pneumovax OR ("pneumonia vaccine" AND_NOT "no pneumonia vaccine"))

Appendix C. Queries, continued.

174: Patients with ventilator-associated pneumonia	
Baseline Query	ventilator-associated pneumonia OR "vap"
Sectioned Query	AREA[Course]("ventilator-associated pneumonia" OR "vap") OR AREA[LabRadResults] ("ventilator-associated pneumonia" OR "vap") OR AREA[AssessmentAndPlan] ("ventilator-associated pneumonia" OR "vap") OR AREA[Condition] ("ventilator-associated pneumonia" OR "vap") OR AREA[Addendum] ("ventilator-associated pneumonia" OR "vap") OR AREA[FinalDiagnosis] ("ventilator-associated pneumonia" OR "vap") OR AREA[Indication] ("ventilator-associated pneumonia" OR "vap")
179: Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression	
Baseline Query	(Amisulpride OR Aripiprazole OR Asenapine OR Blonanserin OR Clotiapine OR Clozapine OR Iloperidone OR Lurasidone OR Mosapramine OR Olanzapine OR Paliperidone OR Perospirone OR Quetiapine OR Remoxipride OR Risperidone OR Sertindole OR Sulpiride OR Ziprasidone OR Zotepine) AND_NOT (schizophrenia OR bipolar)
Sectioned Query	(Amisulpride OR Aripiprazole OR Asenapine OR Blonanserin OR Clotiapine OR Clozapine OR Iloperidone OR Lurasidone OR Mosapramine OR Olanzapine OR Paliperidone OR Perospirone OR Quetiapine OR Remoxipride OR Risperidone OR Sertindole OR Sulpiride OR Ziprasidone OR Zotepine) AND_NOT (AREA[Addendum] (schizophrenia OR bipolar) OR AREA[AdmissionDiagnosis] (schizophrenia OR bipolar) OR AREA[AssessmentAndPlan] (schizophrenia OR bipolar) OR AREA[ChiefComplaint] (schizophrenia OR bipolar) OR AREA[Course] (schizophrenia OR bipolar) OR AREA[Condition] (schizophrenia OR bipolar) OR AREA[DCCondition] (schizophrenia OR bipolar) OR AREA[DCDisposition] (schizophrenia OR bipolar) OR AREA[DCInstructions] (schizophrenia OR bipolar) OR AREA[FinalDiagnosis] (schizophrenia OR bipolar) OR AREA[HistoryOfPresentIllness] (schizophrenia OR bipolar) OR AREA[HomeMeds] (schizophrenia OR bipolar) OR AREA[HospitalMeds] (schizophrenia OR bipolar) OR AREA[Indication] (schizophrenia OR bipolar) OR AREA[PastMedicalHistory] (schizophrenia OR bipolar) OR AREA[PhysicalExam] (schizophrenia OR bipolar) OR AREA[ProblemList] (schizophrenia OR bipolar) OR AREA[Preamble] (schizophrenia OR bipolar) OR AREA[ROS] (schizophrenia OR bipolar) OR AREA[Subjective] (schizophrenia OR bipolar) OR AREA[SubjectiveAndObjective] (schizophrenia OR bipolar))
183: Patients with acute vision loss	
Baseline Query	EXPAND[concept] vision loss OR EXPAND[concept] blindness
Sectioned Query	AREA[HistoryOfPresentIllness] (EXPAND[concept] vision loss OR EXPAND[concept] blindness) OR AREA[AdmissionDiagnosis] (EXPAND[concept] vision loss OR EXPAND[concept] blindness) OR AREA[Condition] (EXPAND[concept] vision loss OR EXPAND[concept] blindness) OR AREA[ChiefComplaint] (EXPAND[concept] vision loss OR EXPAND[concept] blindness) OR AREA[Course] (EXPAND[concept] vision loss OR EXPAND[concept] blindness)

Appendix C. Queries, continued.

184: Patients with colon cancer receiving chemotherapy	
Baseline Query	EXPAND[concept] colon cancer AND chemotherapy
Sectioned Query	chemotherapy AND (AREA[AdmissionDiagnosis]EXPAND[concept] colon cancer OR AREA[AssessmentAndPlan]EXPAND[concept] colon cancer OR AREA[ChiefComplaint]EXPAND[concept] colon cancer OR AREA[Condition]EXPAND[concept] colon cancer OR AREA[Course]EXPAND[concept] colon cancer OR AREA[DCCondition]EXPAND[concept] colon cancer OR AREA[DCDisposition]EXPAND[concept] colon cancer OR AREA[DCInstructions]EXPAND[concept] colon cancer OR AREA[FinalDiagnosis]EXPAND[concept] colon cancer OR AREA[Indication]EXPAND[concept] colon cancer OR AREA[LabRadResults]EXPAND[concept] colon cancer OR AREA[ProblemList]EXPAND[concept] colon cancer OR AREA[ProcedureDetails]EXPAND[concept] colon cancer OR AREA[StudyImpression]EXPAND[concept] colon cancer)
136: Patients with dental caries	
Baseline Query	caries AND NOT ("no evidence of caries" OR "caries absent" OR "no caries identified" OR "- caries" OR "without caries" OR "no dental caries" OR "no definite caries")
Sectioned Query	NOT ("no evidence of caries" OR "caries absent" OR "no caries identified" OR "- caries" OR "without caries" OR "no dental caries" OR "no definite caries") AND (AREA[StudyImpression] caries OR AREA[FinalDiagnosis] caries OR AREA[PhysicalExam] caries OR AREA[Addendum] caries OR AREA[AdmissionDiagnosis] caries OR AREA[AssessmentAndPlan] caries OR AREA[ChiefComplaint] caries OR AREA[Comments] caries OR AREA[Condition] caries OR AREA[Course] caries OR AREA[DCCondition] caries OR AREA[DCDisposition] caries OR AREA[DCInstructions] caries OR AREA[HistoryOfPresentIllness] caries OR AREA[LabRadResults] caries OR AREA[Objective] caries OR AREA[Preamble] caries OR AREA[ProblemList] caries OR AREA[ROS] caries OR AREA[Subjective] caries OR AREA[SubjectiveAndObjective] caries)
137: Patients with inflammatory disorders receiving TNF-inhibitor treatments	
Baseline Query	(tnf-inhibitor OR enbrel OR remicade OR humira OR trental OR adalimubab OR infliximab OR cerolizumab OR cimzia OR golimumab OR simponi OR etancercept) AND NOT cancer AND_NOT mbs
Sectioned Query	(AREA[DCCondition] (etancercept OR simponi OR golimumab OR cimzia OR cerolizumab OR infliximab OR tnf-inhibitor OR enbrel OR remicade OR humira) OR AREA[DCDisposition] (etancercept OR simponi OR golimumab OR cimzia OR cerolizumab OR infliximab OR tnf-inhibitor OR enbrel OR remicade OR humira) OR AREA[DCInstructions] (etancercept OR simponi OR golimumab OR cimzia OR cerolizumab OR infliximab OR tnf-inhibitor OR enbrel OR remicade OR humira) OR AREA[DCMeds] (etancercept OR simponi OR golimumab OR cimzia OR cerolizumab OR infliximab OR tnf-inhibitor OR enbrel OR remicade OR humira) OR AREA[HomeMeds] (etancercept OR simponi OR golimumab OR cimzia OR cerolizumab OR infliximab OR tnf-inhibitor OR enbrel OR remicade OR humira) OR AREA[HospitalMeds] (etancercept OR simponi OR golimumab OR cimzia OR cerolizumab OR infliximab OR tnf-inhibitor OR enbrel OR remicade OR humira)) AND NOT (cancer OR mbs)

Appendix C. Queries, continued.

140: Patients who developed disseminated intravascular coagulation in the hospital	
Baseline Query	dic
Sectioned Query	AREA[Addendum] dic OR AREA[AdmissionDiagnosis] dic OR AREA[AssessmentAndPlan] dic OR AREA[ChiefComplaint] dic OR AREA[Complications] dic OR AREA[Condition] dic OR AREA[Course] dic OR AREA[FinalDiagnosis] dic OR AREA[Indication] dic OR AREA[LabRadResults] dic OR AREA[Objective] dic OR AREA[ProblemList] dic OR AREA[ROS] dic
144: Patients with diabetes mellitus who also have thrombocytosis	
Baseline Query	thrombocytosis AND diabetes
Sectioned Query	thrombocytosis AND (AREA[Addendum] diabetes OR AREA[AdmissionDiagnosis] WEIGHT[0.90] diabetes OR AREA[AssessmentAndPlan] diabetes OR AREA[ChiefComplaint] diabetes OR AREA[Comments] diabetes OR AREA[Condition] diabetes OR AREA[Course] diabetes OR AREA[DCCondition] diabetes OR AREA[DCDisposition] diabetes OR AREA[DCInstructions] diabetes OR AREA[FinalDiagnosis] diabetes OR AREA[HistoryOfPresentIllness] diabetes OR AREA[Indication] diabetes OR AREA[LabRadResults] diabetes OR AREA[Objective] diabetes OR AREA[PastMedicalHistory] diabetes OR AREA[PhysicalExam] diabetes OR AREA[Preamble] diabetes OR AREA[ProblemList] diabetes OR AREA[ROS] diabetes)
146: Patients treated for the post-partum problems depression, hypercoagulability or cardiomyopathy	
Baseline Query	post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy
Sectioned Query	(AREA[Addendum] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[AdmissionDiagnosis] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[AssessmentAndPlan] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[ChiefComplaint] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[Comments] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[Complications] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[Condition] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[Course] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[DCCondition] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[DCDisposition] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[DCInstructions] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[FinalDiagnosis] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[HistoryOfPresentIllness] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[Indication] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[Objective] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[Preamble] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[ProblemList] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy) OR AREA[ROS] (post-partum depression OR post-partum hypercoagulability OR post-partum cardiomyopathy)) AND (cardiomyopathy OR hypercoagulability OR CONCEPT_EXPAND depression)

Appendix C. Queries, continued.

150: Patients who have cerebral palsy and depression	
Baseline Query	EXPAND[concept] cerebral palsy AND EXPAND[concept] depression
Sectioned Query	(AREA[Addendum] EXPAND[concept] cerebral palsy OR AREA[AdmissionDiagnosis] EXPAND[concept] cerebral palsy OR AREA[AssessmentAndPlan] EXPAND[concept] cerebral palsy OR AREA[ChiefComplaint] EXPAND[concept] cerebral palsy OR AREA[Comments] EXPAND[concept] cerebral palsy OR AREA[Complications] EXPAND[concept] cerebral palsy OR AREA[Condition] EXPAND[concept] cerebral palsy OR AREA[Course] EXPAND[concept] cerebral palsy OR AREA[DCCondition] EXPAND[concept] cerebral palsy OR AREA[DCDisposition] EXPAND[concept] cerebral palsy OR AREA[DCInstructions] EXPAND[concept] cerebral palsy OR AREA[FinalDiagnosis] EXPAND[concept] cerebral palsy OR AREA[HistoryOfPresentIllness] EXPAND[concept] cerebral palsy OR AREA[Indication] EXPAND[concept] cerebral palsy OR AREA[Objective] EXPAND[concept] cerebral palsy OR AREA[PastMedicalHistory] EXPAND[concept] cerebral palsy OR AREA[PhysicalExam] EXPAND[concept] cerebral palsy OR AREA[Preamble] EXPAND[concept] cerebral palsy OR AREA[ProblemList] EXPAND[concept] cerebral palsy OR AREA[ROS] EXPAND[concept] cerebral palsy OR AREA[Subjective] EXPAND[concept] cerebral palsy OR AREA[SubjectiveAndObjective] EXPAND[concept] cerebral palsy) AND (AREA[Addendum] EXPAND[concept] depression OR AREA[AdmissionDiagnosis] EXPAND[concept] depression OR AREA[AssessmentAndPlan] EXPAND[concept] depression OR AREA[ChiefComplaint] EXPAND[concept] depression OR AREA[Comments] EXPAND[concept] depression OR AREA[Complications] EXPAND[concept] depression OR AREA[Condition] EXPAND[concept] depression OR AREA[Course] EXPAND[concept] depression OR AREA[DCCondition] EXPAND[concept] depression OR AREA[DCDisposition] EXPAND[concept] depression OR AREA[DCInstructions] EXPAND[concept] depression OR AREA[FinalDiagnosis] EXPAND[concept] depression OR AREA[HistoryOfPresentIllness] EXPAND[concept] depression OR AREA[Indication] EXPAND[concept] depression OR AREA[Objective] EXPAND[concept] depression OR AREA[Preamble] EXPAND[concept] depression OR AREA[ProblemList] EXPAND[concept] depression OR AREA[Subjective] EXPAND[concept] depression OR AREA[SubjectiveAndObjective] EXPAND[concept] depression)
158: Patients with esophageal cancer who develop pericardial effusion	
Baseline Query	esophageal cancer AND "pericardial effusion"
Sectioned Query	esophageal cancer AND (AREA[FinalDiagnosis] "pericardial effusion" OR AREA[Course] "pericardial effusion" OR AREA[LabRadResults] "pericardial effusion" OR AREA[AssessmentAndPlan] "pericardial effusion" OR AREA[Course] "pericardial effusion")

Appendix C. Queries, continued.

185: Patients who develop thrombocytopenia in pregnancy	
Baseline Query	thrombocytopenia AND pregnant
Sectioned Query	(AREA[Addendum] pregnant OR AREA[AdmissionDiagnosis] pregnant OR AREA[AssessmentAndPlan] pregnant OR AREA[ChiefComplaint] pregnant OR AREA[Comments] pregnant OR AREA[Condition] pregnant OR AREA[Course] pregnant OR AREA[DCCondition] pregnant OR AREA[DCDisposition] pregnant OR AREA[DCInstructions] pregnant OR AREA[FinalDiagnosis] WEIGHT[0.90] pregnant AREA[HistoryOfPresentIllness] pregnant OR AREA[Indication] pregnant OR AREA[LabRadResults] pregnant OR AREA[Objective] pregnant OR AREA[PhysicalExam] pregnant OR AREA[ProblemList] pregnant OR AREA[ProcedureDetails] pregnant OR AREA[ROS] pregnant OR AREA[StudyImpression] pregnant OR AREA[Subjective] pregnant OR AREA[SubjectiveAndObjective] pregnant) AND (AREA[Addendum] thrombocytopenia OR AREA[AdmissionDiagnosis] thrombocytopenia OR AREA[AssessmentAndPlan] thrombocytopenia OR AREA[ChiefComplaint] thrombocytopenia OR AREA[Comments] thrombocytopenia OR AREA[Condition] thrombocytopenia OR AREA[Course] thrombocytopenia OR AREA[DCCondition] thrombocytopenia OR AREA[DCDisposition] thrombocytopenia OR AREA[DCInstructions] thrombocytopenia OR AREA[FinalDiagnosis] thrombocytopenia AREA[HistoryOfPresentIllness] thrombocytopenia OR AREA[Indication] thrombocytopenia OR AREA[LabRadResults] thrombocytopenia OR AREA[Objective] thrombocytopenia OR AREA[PhysicalExam] WEIGHT[0.90] thrombocytopenia OR AREA[ProblemList] thrombocytopenia OR AREA[ROS] thrombocytopenia OR AREA[SubjectiveAndObjective] thrombocytopenia)
160: Patients with low back pain who had imaging studies	
Baseline Query	(MRI OR "CT scan") AND ("Low Back Pain" OR ((lumbar OR lspine OR "l1" OR "l2" OR "l3" OR "l4" OR "l5") AND "back pain") AND NOT "lumbar puncture")
Sectioned Query	((AREA[ProcedureDetails](MRI OR "CT scan") AND AREA[ProcedureDetails] (lumbar OR lspine OR "l1" OR "l2" OR "l3" OR "l4" OR "l5")) OR (AREA[LabRadResults](MRI OR "CT scan") AND AREA[LabRadResults] (lumbar OR lspine OR "l1" OR "l2" OR "l3" OR "l4" OR "l5"))) AND ("Low Back Pain" OR ((lumbar OR lspine OR "l1" OR "l2" OR "l3" OR "l4" OR "l5") AND "back pain" AND NOT AREA[LabRadResults] "lumbar puncture"))

Appendix D. Script to combine search results

```
#!/usr/bin/env python
#
# Tracy Edinger
# Thesis
# 12/28/13
#
# This program creates a dictionary from data in a csv file
#

import string

# get file names
querynum = raw_input("Please enter the query number: ")
expansion = raw_input("Please enter the expansion type: ")
basename = "".join(['q',querynum,'-base-',expansion,'-list-csv'])
segname = "".join(['q',querynum,'-seg-',expansion,'-list-csv'])
outfile = "".join(['q',querynum,'-',expansion,'-py-out'])

# Read in data from files
basedict = {}
segdict = {}
subjdict = {}
admdict = {}

basefile = open(basename,'r')
basefile.readline()
linecount = 0

for line in basefile:
  data = line.split(",")
  if ((linecount > 4) and (len(data) >3)):
    docnum = data[2]
    basedict[data[2]] = data[1]
    segdict[data[2]] = '"-1"'
    subjdict[data[2]] = data[4]
    admdict[data[2]] = data[5]
  linecount = linecount + 1

basefile.close()

# Read in data from seg file
segfile = open(segname,'r')
```

```
segfile.readline()
linecount = 0

for line in segfile:
  data = line.split(",")
  if ((linecount > 4) and (len(data) > 3)):
    if data[2] in segdict:
      del segdict[data[2]]
    segdict[data[2]] = data[1]
    if not (data[2] in basedict):
      basedict[data[2]] = '"-1"'
      subjdict[data[2]] = data[4]
      admdict[data[2]] = data[5]
  linecount = linecount + 1
segfile.close()
```