Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval

Tracy Edinger, ND, MS[1], Dina Demner-Fushman, MD, PhD[2],
Aaron M. Cohen, MD, MS [1], Steven Bedrick, PhD [1], William Hersh, MD[1]
[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA
[2]National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract

Objective: *Secondary use of electronic health record (EHR) data is enabled by accurate and complete retrieval of the relevant patient cohort, which requires searching both structured and unstructured data. Clinical text poses difficulties to searching, although chart notes incorporate structure that may facilitate accurate retrieval.* ***Methods:*** *We developed rules identifying clinical document sections, which can be indexed in search engines that allow faceted searches, such as Lucene or Essie, an NLM search engine. We developed 22 clinical cohorts and two queries for each cohort, one utilizing section headings and the other searching the whole document. We manually evaluated a subset of retrieved documents to compare query performance.* ***Results:*** *Querying by section had lower recall than whole-document queries (0.83 vs 0.95), higher precision (0.73 vs 0.54), and higher* F_1 *(0.78 vs 0.69).* ***Conclusion****: This evaluation suggests that searching specific sections may improve precision under certain conditions and often with loss of recall.*

Introduction

The use of electronic health records (EHR) in inpatient and outpatient clinical facilities has increased rapidly over recent years. Incorporating EHRs has facilitated clinical practice, resulting in widespread access to patient information and potentially better care. Because of the quantity of data in an EHR and the number of patients in a typical health-care system, using this data for secondary purposes can be of enormous benefit both to the local health care system and nationally.

EHRs are used primarily for direct patient care and for billing purposes, and they are designed to maximize the ability to provide these functions. However, in this process, the EHR becomes a repository of a vast amount of clinical data, both structured and text. This data can be used for many purposes beyond clinical care and billing: research, quality measures, disease surveillance, operational improvement, and administrative applications. All of these uses of EHR data require the ability to find desired information with a high degree of accuracy. 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 the number of procedures performed in a clinic. In some cases, conditions or symptoms may be documented only in clinical notes, and only available through a text search.

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, which are relatively easy to access. This data is recorded in consistent ways as specific codes, including international classification of disease (ICD) or other codes, or as predetermined phrases. Although structured data is relatively easy to search, several studies suggest that a search relying solely on structured data will not retrieve all patients relevant to the topic[1,2,3,4], and other studies suggest that combining structured and unstructured data will improve cohort retrieval[5]. A great deal of EHR data, however, is stored as unstructured text in clinical notes, history and physical exam notes, and reports. This unstructured text is very 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. Because of the difficulty in accurately extracting data from text, most non-research use of EHR data utilizes structured data only. Clinical notes contain highly valuable information not found in structured fields, and they document clinical thinking and decision making. Further, relying solely on structured data for analytic functions results in a greater data-entry burden on providers. Because of these factors, improving retrieval accuracy from text would have great value.

The National Institute of Standards and Technology (NIST) has sponsored challenges to evaluate and improve the ability to accurately retrieve patient data from medical text for secondary uses. In NIST's 2011 and 2012 Text Retrieval Conference Medical Records Track (TRECMed), participants were given a list of clinical topics and a set of de-identified textual medical records. Participants developed search systems and algorithms to retrieve records relevant to each topic. Retrieved visits were then judged for relevance to the topic. An analysis of incorrectly retrieved visits identified several key problems in accurate retrieval of patient cohorts[6]. 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.

The Electronic Medical Records and Genomics Network (eMERGE) Consortium has also evaluated the ability to retrieve specific patient cohorts from EHRs[7]. In this study, EHR data from five different sites were used to identify patients with at least one of five diseases (dementia, cataracts, peripheral arterial disease, type 2 diabetes, and cardiac conduction defects). Patients were identified with a high level of accuracy when the data were stored in structured fields; however, in some cases, target information was stored only in clinical text. The use of natural language processing (NLP) tools increased retrieval significantly—129% more cases were identified by including the use of NLP tools rather 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[8].

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 also 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. Manipulating and searching the text according to these sections would allow the construction of searches targeted to the appropriate section, avoiding or minimizing some of the issues found in previous work. Several tools and strategies[9,10,11,12,13,14] have been documented that segment clinical records. Although the effectiveness of each segmentation strategy has been evaluated, no studies could be located that demonstrate whether segmenting improves the performance of searching clinical text.

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. 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 accuracy of recall by avoiding retrieval based on someone else's disease status. The ability to separate sections of the medical record may facilitate retrieval accuracy by identifying the family history section and allowing that section to be searched only when applicable.

The hypothesis of this study is that searching for patient cohorts by looking in specific, relevant sections of clinical text will improve retrieval accuracy.

Methods

Data This project used de-identified clinical records developed by the Massachusetts Institute of Technology (MIT), Philips Medical Systems, and Beth Israel Deaconess Medical Center[15]. The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) data is a publicly available dataset containing more than 30,000 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.

Search Engine Queries for this project were run using the Essie 4 search engine developed by the National Library of Medicine[16]. Essie 4 maps terms to the UMLS and allows differential weighting of terms to alter the order of documents retrieved. Mapping terms to the UMLS allows comparison of different but equivalent terms; for example, 'myocardial infarction' and 'heart attack' refer to the same concept using different words. Because Essie 4 maps to the UMLS, equivalent terms are found without being listed explicitly, allowing the queries to focus on detecting the difference in retrieval when using sections, rather than including an exhaustive list of all possible synonyms.

When a query is run, Essie 4 returns a list of documents retrieved for that query. Because of term weighting, each document is assigned a rank that indicates the relative likelihood that it is relevant to the query. Weights are given values greater than 0 and less than 1.

Segmentation To segment the documents, we examined a subset of each type of document to identify the most common section headings, recording all terminology, spelling, and punctuation variations. We then created a text file containing these heading variations and the corresponding XML heading tags for the new sections. Table 1 shows examples of the headings inserted for the indicator text listed.

Table 1. Examples of indicator text in the documents and inserted headings.

Inserted Heading	Indicator Text
AssessmentAndPlan	disposition/plan treatment/plan: overall assessment and plan:
Course	clinical course in the emergency department: ed course: institution course:
DCDisposition	transferred to: dispo:
LabRadResults	cta chest: important diagnostics and labs: radiographs-
FinalDiagnosis	diagnoses at the time of discharge: final discharge diagnosis: diagnosis on transfer:

The documents were searched for exact matches to the heading variations listed in the text file, and the appropriate heading tags were inserted at those locations. Two tags were inserted for each heading, one at the start of the section and the other at the end. When a new heading was located, an opening tag was inserted for that heading, and an ending tag for the previous heading was inserted just before it. Tags were set up in XML format; for example, <AdmissionDiagnosis> at the start of the section and </AdmissionDiagnosis> at the end of the section.

Prior to having the section tags inserted, the original documents contained a single block of text surrounded by opening and closing tags, as shown in this example:

```
<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, indication, condition, procedure details, and study impression 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 4. 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, a project which used a set of de-identified clinical records for hospital and emergency-department patients. This set contains 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)[17]. Because the original query topics were developed for a wider range of patients, not all topics were relevant to ICU patients. A subset of 22 topics was used, modified as necessary to fit the current clinic population.

Queries were developed in an iterative fashion, refining the search details to maximize the number of relevant visits returned. An initial query was run against the data. The text of a subset of the returned visits was examined to determine if any new terms or operators needed to be added to the query, or if any query components needed to be removed. Although information from the returned visits was used to guide query development, details that appeared to be site-specific were not used. Once a set of queries was developed, they were run on the data, and the retrieved visits were recorded. The queries were then revised to utilize the sections, and visits retrieved with the revised queries were recorded. 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"*
> Query by sections: *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*
> Query by sections: *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 A subset of retrieved documents was examined to understand the effect of segmenting on retrieval performance. Three sets of documents were examined; two of these sets were the ten highest ranked documents retrieved only with the base query or only with the query by sections. The third set compared results when both queries retrieved the same document; the difference in rank assigned to those documents by each query was used to decide which to examine. The difference was calculated by subtracting the rank assigned by the query by sections from the rank assigned by the base query. When this value was highly positive, the base query had assigned a much greater rank than the query by sections; ie, the document was placed much lower in the list of ranked retrieved results. When this value was highly negative, the query by section had assigned a much greater rank than the base query. The ten documents with the largest positive difference in rank and the ten documents with the largest negative difference in rank were examined. Because the queries retrieved different numbers of documents, the actual number analyzed for each topic was usually lower than 40. Table 2 below shows the topics and the number of documents analyzed for each topic.

Table 2. Query topics and number of documents analyzed.

Number of Documents Analyzed	Query Topic
40	Patients with dental caries
40	Patients with thyrotoxicosis treated with beta-blockers
40	Patients with acute vision loss
40	Patients with left lower quadrant abdominal pain
40	Patients with low back pain who had imaging studies
38	Patients treated for the post-partum problems depression, hypercoagulability or cardiomyopathy
30	Patients who developed disseminated intravascular coagulation in the hospital
30	Patients who have gluten intolerance or celiac disease
30	Patients with colon cancer receiving chemotherapy
29	Patients with ventilator-associated pneumonia
22	Patients with delirium, hypertension, and tachycardia
20	Patients admitted to the hospital with end-stage chronic disease who are discharged on hospice care
20	Patients who had a carotid endarterectomy during this admission
20	Patients who received pneumonia vaccination during this admission
20	Patients with HIV/AIDS who develop pancytopenia
20	Patients with diabetes mellitus who also have thrombocytosis

20	Patients with esophageal cancer who develop pericardial effusion
20	Patients with inflammatory disorders receiving TNF-inhibitor treatments
16	Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes
15	Patients who develop thrombocytopenia in pregnancy
14	Patients who have cerebral palsy and depression
10	Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression

For each document, several observations were recorded. Retrieved documents were evaluated for relevance to the topic. Non-relevant documents were examined to determine the reason for retrieval; specifically, the documents were examined to identify occurrences of the search terms. We assessed the reason for performance differences between the two queries, and the reason for performance difference when querying by sections. Codes to reflect these assessments were developed in an iterative fashion, with new codes added as necessary to capture new reasons.

Results

A total of 574 documents were examined. Of those, 344 were relevant to the given topic. Both queries retrieved 247 of these documents, querying by sections retrieved an additional 20 documents, and querying the whole document retrieved an additional 77 documents. The remaining 230 non-relevant documents included 146 that were not at all relevant, 53 that were relevant to portions of the topic but not the entire topic, and 18 that were possibly relevant. Eighty-one non-relevant documents were retrieved by both queries, querying by sections retrieved an additional six documents, and querying the whole document retrieved an additional 143 documents. Figure 1 illustrates the comparative retrieval rates for queries of specific sections versus querying the whole document.

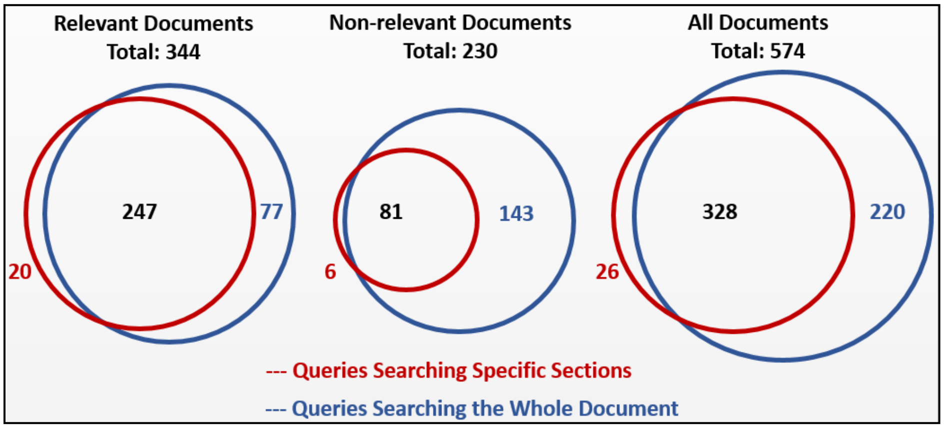

Figure 1. Retrieval rate comparisons between searches of document sections and searching whole documents. Text and circles in red (left) represent the results of querying by sections, and text and circles in blue (right) represent the results of querying whole documents.

To assess overall query performance, several statistics were calculated. Recall is a measure of the proportion of relevant documents returned by a search; a high recall indicates that most of the relevant documents were retrieved. Precision indicates what proportion of retrieved documents are relevant; high precision indicates that most of the retrieved documents are relevant. The F_1 measure reflects both precision and recall to give an overall picture of how well the queries performed. As shown in Table 3, queries of specific sections have lower recall than queries across the whole document; however, these queries also have higher precision and overall F_1 measures.

Table 3. Recall, precision, and F_1 measures for both query types.

	Recall	**Precision**	**F_1 Measure**
Whole-document Queries	0.95	0.54	0.69
Section Queries	0.83	0.73	0.78

Examination of the documents revealed that not all sections were identified correctly. This was not surprising, given our method of section identification. Because sections were ended only when new ones were identified, unidentified sections were erroneously included in the previous section. Our identification rules specified text patterns that signified new sections, so sections that were implicitly identified by extra spaces or changes in wording or content were overlooked. In all, we identified 44 documents in which relevant sections were not correctly tagged, factoring into incorrect retrieval. Eleven of these were relevant documents not retrieved when querying by section; accurate segmenting would have enabled retrieval of this set.

Of the relevant documents, 77 were retrieved by the base query only, and 20 were retrieved only by the query by section. The remaining 247 were retrieved by both queries. Querying by section returned 78% of the relevant documents examined, and the base query returned 94%. Designing queries to examine only specific sections of the text led to overlooking a number of relevant documents: fifty-two relevant documents were not retrieved because of overly restrictive queries of specific sections of the document.

Looking at the 20 documents retrieved when querying by section revealed two topics these queries handled especially well. One topic was "Patients with low-back pain who had imaging studies." The query for this topic included the word *lumbar* as a synonym for *low back*. Adding this term resulted in retrieving a number of documents in which patients had lumbar punctures, so the base query was adapted to eliminate documents with any mention of that procedure. Querying by section eliminated documents that mentioned lumbar puncture only in the lab and radiology results, allowing this query to place additional relevant documents higher in the ranked list. The other topic handled well by querying by section was "Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression." The base query looked for documents without these diagnoses, leading to omission of documents in which non-relevant mention was made in the family or social history. The other query was able to specify the absence of these in sections of the chart pertaining to the patient, while ignoring sections mentioning diagnoses of relatives or other family members.

Queries of specific sections returned only 38% of the non-relevant documents examined, whereas the base queries returned more than 97%. Only six non-relevant documents were retrieved only by querying specific sections, and 143 were retrieved by whole-document queries only. The remaining 81 were retrieved by both queries. Reasons for returning non-relevant documents included mention of conditions that had been denied or ruled out, past conditions or medications, future or possible conditions, non-relevant 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. Because of their ability to search in specified sections of the chart, these queries were able to avoid retrieval of many non-relevant documents.

Matthews correlation coefficient (MCC) was used to evaluate performance of individual 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 as true and false positives and negatives, indicating relevance to the topic and score differential. Table 4, below, illustrates the classification of documents according to these criteria.

Table 4. Classification of retrieved documents based on relevance to topic and relative scores.

	Querying by section has higher score than querying whole document	**Querying by section has lower score than querying whole document**
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, a document may be retrieved by the base query but not the query of specific sections. In this case, the score assigned by the base query was used, and a score of zero was used for the section 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 querying by sections produces an overall performance decrease, MCC will be less than zero. A score of -1 indicates that querying by sections yields only false positive and false negative results. If querying by sections increases performance, MCC will be greater than zero. An MCC of 1 indicates that querying by sections yielded only true positive results. MCC was calculated for each topic as follows (scores listed in Table 5):

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

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 $p<0.01$. Although there was great variability in the values of MCC, the queries using sections for eight topics performed very well, showing statistical improvement over the base queries using Fisher's exact test for significance, and no topics had statistically significant decreased performance. The queries of sections for *Patients who had a carotid endarterectomy during this admission* and *Patients with diabetes mellitus who also have thrombocytosis* performed very well with MCCs of 0.905 ($p<0.001$). The sectioned query for *Patients who develop thrombocytopenia in pregnancy* had an MCC of 0.853 ($p<0.001$). Other high performers were *Patients with acute vision loss* (MCC=0.756, $p<0.001$), *Patients who received pneumonia vaccination during this admission* (MCC=0.734, $p<0.01$), and *Patients who have gluten intolerance or celiac disease* (MCC=0.666, $p<0.001$).

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

Discussion

In the documents examined, searching for information in specific sections of the document did provide some improvement. Overall, queries of specific sections retrieved only about half the number of non-relevant documents as compared to the base queries. However, they also retrieved only about 80% of the relevant documents. This suggests that querying specific sections will not retrieve as large a set of relevant documents, but it is more likely to avoid retrieving non-relevant documents.

Expanding the queries to look at more sections for the desired information would result in retrieval of more documents, both relevant and non-relevant. Because Essie returns documents that match the search criteria, queries developed for this project were designed to retrieve documents that were most likely to be relevant. The queries utilizing the section headings were written to look only in the relevant section, resulting in two different queries for each topic. Some NLP search engines will identify each relevant item of information as well as the document section in which it was found. While the methodology in the present project may reflect real-world practice, using an NLP tool like this would allow the use of one set of queries and an evaluation that could focus more closely on the use of looking in specific sections. Improving retrieval of relevant documents could also be achieved through greater accuracy in labeling section headers. In the sample of documents validated for this project, better section detection may have allowed retrieval of 13 additional relevant documents, about 16% of the overlooked relevant documents.

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 the patient's Alzheimer's disease only 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.

For some topics, the query must look for documents that do not reference a specific condition or medication, as in the topic "Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression." Searching for documents that do not reference these eliminates those that mention the conditions in other people. Being able to search specific sections improves recall in this situation by avoiding sections likely to contain false positives.

Future work should take several approaches. First, a formal evaluation of the section identification in these documents would provide a baseline to help interpret the results of the query evaluation. An alternative approach would be the use of a validated sectioning algorithm, which may provide greater accuracy in labeling section headings. The tool chosen should have the ability to identify section changes that are not explicitly identified by headings. Using regular expressions or other pattern matching algorithms would improve section detection and accuracy, which would improve overall performance.

Second, some of the common retrieval issues in medical text, such as negation, should be identified using a published tool. Next, a set of queries can be run, and a quantitative analysis of the results can be done to provide greater insight into the effectiveness of segmenting documents on retrieval.

Another avenue of future work is to develop algorithms that combine the two search types to predict document relevance. Because of the differences in recall and precision between the two types of searches, knowing which documents are retrieved by both searches provides a clue to predict relevance. Future work in this area would develop prediction models based on retrieval results.

The queries used in this project contained multiple components. Breaking down the cohort criteria into individual concepts would allow a more focused evaluation. For example, in the search for patients who have cerebral palsy and depression, one set of queries could look for cerebral palsy and another set for depression. This would allow an assessment of which individual concepts may be more accurately identified in specific sections of the document, and which can be accurately identified when searching the whole document.

Because of the high value of clinical text and the documentation burden on clinicians, pursuing further work in this area would provide great value to providers, healthcare organizations, and patients. Identifying and searching specific sections of clinical text significantly improves precision with only modest decreases in recall, resulting in greater usability of this data source.

References

1. 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. 2014 Mar 1;21(2):221–30.
2. Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol. 2012;8(12):e1002823.
3 Friedlin J, Overhage M, Al-Haddad MA, et al. Comparing methods for identifying pancreatic cancer patients using electronic data sources. AMIA Annu Symp Proc. 2010;237-241.
4. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc. 2016;23:e20-e27.
5. Kocbek S, Cavedon L, Martinez D, et al. Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources. J Biomed Inform. 2016 Oct 11;64:158-167.
6. 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. 2012:180–8.
7. 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.
8. 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. 2010 Jan 1;17(1):19–24.
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. 2003;155–9.
11. Denny JC, Miller RA, Johnson KB, Spickard A 3rd. Development and evaluation of a clinical note section header terminology. AMIA Annu Symp Proc. 2008;156–60.
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. Ganeson KA, Subotin M. A general supervised approach for segmentation of clinical texts. IEEE Intl Conf on BigData. 2014.
14. Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical section segmentation in free-text clinical records. Proceeding of the International Conference on Language Resources and Evaluation (LREC); Istanbul. May, 2012.
15. 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.
16. Ide NC, Loane RF, Demner-Fushman D. Essie: A Concept-based Search Engine for Structured Biomedical Text. J Am Med Inform Assoc. 2007 Feb 28;14(3):253–63.
17. Voorhees EHW. 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.xml.