Barriers to Retrieving Patient Information from Electronic Health Record Data: Failure Analysis from the TREC Medical Records Track

Tracy Edinger, ND, Aaron M. Cohen, MD, MS, Steven Bedrick, PhD, Kyle Ambert, BA, William Hersh, MD
Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Abstract

Objective: *Secondary use of electronic health record (EHR) data relies on the ability to retrieve accurate and complete information about desired patient populations. The Text Retrieval Conference (TREC) 2011 Medical Records Track was a challenge evaluation allowing comparison of systems and algorithms to retrieve patients eligible for clinical studies from a corpus of de-identified medical records, grouped by patient visit. Participants retrieved cohorts of patients relevant to 35 different clinical topics, and visits were judged for relevance to each topic. This study identified the most common barriers to identifying specific clinic populations in the test collection.* ***Methods:*** *Using the runs from track participants and judged visits, we analyzed the five non-relevant visits most often retrieved and the five relevant visits most often overlooked. Categories were developed iteratively to group the reasons for incorrect retrieval for each of the 35 topics.* ***Results:*** *Reasons fell into nine categories for non-relevant visits and five categories for relevant visits. Non-relevant visits were most often retrieved because they contained a non-relevant reference to the topic terms. Relevant visits were most often infrequently retrieved because they used a synonym for a topic term.* ***Conclusions****: This failure analysis provides insight into areas for future improvement in EHR-based retrieval with techniques such as more widespread and complete use of standardized terminology in retrieval and data entry systems.*

Introduction

The use of clinical data for research is a widely anticipated benefit of the electronic health record (EHR) [1]. Clinical data stored in structured fields is relatively straightforward to retrieve and use; however, a large proportion of EHR data is "locked" in textual documents [2]. EHR chart notes are typically stored in text files, which include the medical history, physical exam findings, lab reports, radiology reports, operative reports, and discharge summaries. These records contain valuable information about the patient, treatment, and clinical course. This "free text" data is much more difficult to access for secondary purposes. In order to use this data, we must be able to retrieve records accurately and reliably for a desired patient population, usually through the use of natural language processing (NLP). While NLP has been applied to EHR data for decades, the performance of these systems has been variable across the techniques used, as well as the clinical task [3].

Historically, the field of information retrieval (IR) has studied the retrieval of documents and other content [4]. However, IR has tended to place a greater focus on presenting content to users for human interpretation, rather than on extracting the specific information they contain. This task is typically referred to as information extraction or text mining [5]. IR also has a long tradition of system evaluation, especially involving the use of test collections that contain fixed assemblies of content, query topics, and relevance judgments, a "gold standard" defining which content items are relevant to which topics. Such test collections are important, because they allow direct comparison of results obtained by different IR systems.

The field of IR also has a tradition of advancing knowledge by hosting challenge evaluations, in which the same test collection is used by many groups to compare the efficacy of different approaches. One of the best-known is the Text Retrieval Conference (TREC), an annual challenge evaluation hosted by the US National Institute for Standards & Technology (NIST) [6]. TREC is a long-standing event that allows different tasks and approaches to be assessed in an open, collegial, and comparable manner. Each year, TREC holds a number of "tracks" devoted to different aspects of IR, such as Web searching or cross-language IR. While TREC is focused on general-purpose IR, there have been some tracks dedicated to specific domains, including genomics [7].

In 2011, TREC launched a Medical Records Track (TRECMed) [8] to develop an IR challenge task pertinent to real-world clinical medicine. The track was made possible by access to a large corpus of de-identified medical text from

the University of Pittsburgh Medical Center (http://www.dbmi.pitt.edu/blulab). De-identified clinical documents in the collection are organized according to patient visits. The task in the first year of TRECMed was to retrieve cohorts of patients fitting criteria similar to those specified for participation in clinical studies. Retrieval topics were derived from an Institute of Medicine list prioritizing conditions for comparative effectiveness research [9] and modified to be unambiguous and to generate an appropriate quantity of visits relevant to the tasks. Funding from NIST allowed organization of the topic development and relevance assessment processes of the track.

The documents for the task come from the University of Pittsburgh NLP Repository, a repository of 95,702 de-identified clinical reports available for NLP research purposes. The reports were generated from multiple hospitals during 2007 and are grouped into visits consisting of one or more reports from the patient's hospital stay. Reports for each visit are stored as text files and may include medical history, physical exam findings, radiology reports, operative reports, or discharge summaries. Each document is formatted in Extensible Markup Language (XML), with a table that maps one or more reports to a visit. An admission diagnosis, discharge diagnoses, and ICD-9 code(s) are recorded for each visit. These codes are stored at the visit level, rather than at the encounter level, so they must be searched separately from the report documents. The data set contains a total of 17,199 visits.

Based on resources available for relevance judging, a decision was made to develop and judge 35 topics for the test collection, consistent with known observations that 25-50 topics are the minimum required for statistical stability of results in IR test collections [10]. As is common practice in TREC and other challenge evaluations, research groups submitted runs consisting of a ranked list of visits for all 35 topics. Each of the 29 participating research groups was allowed to submit up to eight runs.

A total of 127 runs were submitted, with the results pooled to allow representative sampling for relevance judging. The pooled visits for each topic were judged for relevance by physicians enrolled in Oregon Health & Science University's graduate program in Biomedical Informatics; topics and number of relevant visits are listed in Table 1. Because resources did not allow for exhaustive judging, many of the documents that were retrieved in various runs had not been judged. This led to a decision to use the bpref measure [11] to compare retrieval performance, which only measures recall and precision on documents that have had relevance judgments. Bpref is defined as the inverse of the fraction of judged irrelevant documents that are retrieved before judged relevant documents, and varies from 0 (no retrieval) to 1 (retrieval of all known relevant documents). The best bpref performance in TRECMed was in the range 0.5-0.6, reasonably good performance for a general IR task but not satisfactory for identifying patients in a set of clinical records. Further details of the test collection, including the methods we used for topic development and relevance judging, are in the track overview paper from the TREC 2011 conference proceedings [8].

The goal of the study presented here was to perform a failure analysis on two types of retrieved visits. The first type was those visits that were frequently retrieved but judged to be non-relevant (i.e., false-positive retrievals). The second type was those visits that were infrequently retrieved yet judged to be relevant (false-negative retrievals). For both categories of retrieved visits, we created a list of reasons for the inappropriate frequent or infrequent retrieval, with the goal of determining the reasons for these errors and guiding future system development.

Methods

Using the 127 runs submitted by the participating research groups, we combined each run's top 100 retrieved visits for each topic to produce a list of visits ranked by the number of systems retrieving that visit within the top 100. At the top of this list were visits that were commonly ranked in the top 100 by many systems for a given topic, i.e., commonly highly ranked visits. At the bottom of this list were those rarely ranked within the top 100 by any participating system, the rarely highly ranked visits. At least one system had to rank a visit within its top 100 results for the visit to be considered rarely highly ranked.

Often-retrieved non-relevant visits: Among the commonly highly ranked visits were visits that were incorrectly retrieved and marked as non-relevant by the judges. These visits were problematic for many of the retrieval systems, in that they were often ranked much more highly than they should have been, and can be considered to be common precision errors.

Table 1. TRECMed topic number, description of criteria for visit retrieval, and number of visits judged relevant for each.

Topic Number	Description	Relevant Visits
101	Patients with hearing loss	69
102	Patients with complicated GERD who receive endoscopy	89
103	Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis	7
104	Patients diagnosed with localized prostate cancer and treated with robotic surgery	8
105	Patients with dementia	143
106	Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer	85
107	Patients with ductal carcinoma in situ (DCIS)	9
108	Patients treated for vascular claudication surgically	12
109	Women with osteopenia	119
110	Patients being discharged from the hospital on hemodialysis	95
111	Patients with chronic back pain who receive an intraspinal pain-medicine pump	19
112	Female patients with breast cancer with mastectomies during admission	66
113	Adult patients who received colonoscopies during admission which revealed adenocarcinoma	10
114	Adult patients discharged home with palliative care / home hospice	53
115	Adult patients who are admitted with an asthma exacerbation	33
116	Patients who received methotrexate for cancer treatment while in the hospital	10
117	Patients with Post-traumatic Stress Disorder	22
118	Adults who are received a coronary stent during an admission	50
119	Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes	37
120	Patients admitted for treatment of CHF exacerbation	104
121	Patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix	32
122	Patients who received total parenteral nutrition while in the hospital	24
123	Diabetic patients who received diabetic education in the hospital	33
124	Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma	2
125	Patients co-infected with Hepatitis C and HIV	14
126	Patients admitted with a diagnosis of multiple sclerosis	5
127	Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension	85
128	Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post-op	75
129	Patients admitted with chest pain and assessed with CT angiography	45
130	Children admitted with cerebral palsy who received physical therapy	1
131	Patients who underwent minimally invasive abdominal surgery	96
132	Patients admitted for surgery of the cervical spine for fusion or discectomy	88
133	Patients admitted for care who take herbal products for osteoarthritis	4
134	Patients admitted with chronic seizure disorder to control seizure activity	25
135	Cancer patients with liver metastasis treated in the hospital who underwent a procedure	55

Rarely-retrieved relevant visits: Similarly, within the rarely highly ranked visits, there were records that were correctly retrieved and marked relevant by the TRECMed relevance judges,. These correctly retrieved, rarely highly ranked visits were problematic for the retrieval systems, in that most systems failed to rank them as highly as they should have been, and can be considered to be common recall errors.

For each topic, we examined the top five ranked common precision errors, or false positives, and the bottom five ranked common recall errors, or false negatives. For topics that did not include five relevant documents, we reviewed all relevant visits. If one or more visits were assigned tied system counts, we included all visits having the tied count. Our sample contained 359 visits, with 182 common precision errors and 177 common recall errors.

A review of these two sets of visits per topic enabled us to perform a failure analysis of a sample of the TRECMed results. One author (TE) manually examined the chart notes from each visit to determine why it may have been retrieved incorrectly or missed, iteratively developing a set of codes to categorize the reasons. If our existing code set did not describe the reason for a given error, we developed a new code to describe the reason. Because we did not have access to the actual queries used by the systems in their runs, the codes reflect our best judgment as to why the visit was difficult to find accurately for each topic, based on the topic statements themselves, the concepts contained in the topics, and the content of the visit notes.

The dataset contains documents generated from hospital stays. Each stay, or visit, includes one or more reports. These reports, from each clinical encounter during a visit, are stored as separate text files and may include medical history, physical exam findings, radiology reports, operative reports, or discharge summaries. Each report, or document, is formatted in Extensible Markup Language (XML), with a table that maps one or more reports to a visit. An admission diagnosis, discharge diagnoses, and ICD-9 code(s) are recorded for each visit and are stored at the visit level, rather than at the encounter level. Each document displays the diagnoses in the visit header and a list of all encounters in the visit. For this analysis, the encounter documents were opened one at a time and searched for the topic terms until the reviewer identified a satisfactory explanation for that visit's anomalous retrieval. Once an explanation was found, the reviewer moved on to the next visit.

Results

Our analysis of the common precision or "false positive" errors revealed a wider variety of reasons for incorrect retrieval than we observed for the common recall or "false negatives" errors. The final list contained nine codes for precision errors, and five codes for recall errors (see Figure 1). Several visits appeared to have been judged incorrectly by the TRECMed relevance assessors, accounting for almost 20% of the sample. This is a common occurrence in IR evaluations, and has usually been shown to affect the absolute, but not the relative, comparison of different runs [12].

Figure 1 depicts the distribution of visits and topics for each code category, showing the total number of incorrectly retrieved visits in each category, as well as each category's distinct topics. Irrelevant references were observed to cause the greatest number of incorrect retrievals of non-relevant visits, and use of alternative terms was the most common cause of low retrieval rates of relevant visits. The number of relevant visits with topic terms present but overlooked in the search is most likely due to our error sampling strategy; this is explained in greater detail below.

The specific reasons for relevant visits being infrequently retrieved and non-relevant visits being frequently retrieved are described next, with examples given.

Relevant Visits with Low Retrieval Counts

Topic not present—record is not relevant—disagree with relevance judge
A number of visits judged relevant were found to be irrelevant upon closer examination. There were a variety of reasons for this, including confusion of terms for different body regions and procedures mentioned as future possibilities. These reasons covered the same issues as seen in the often retrieved non-relevant visits discussed in the next section.

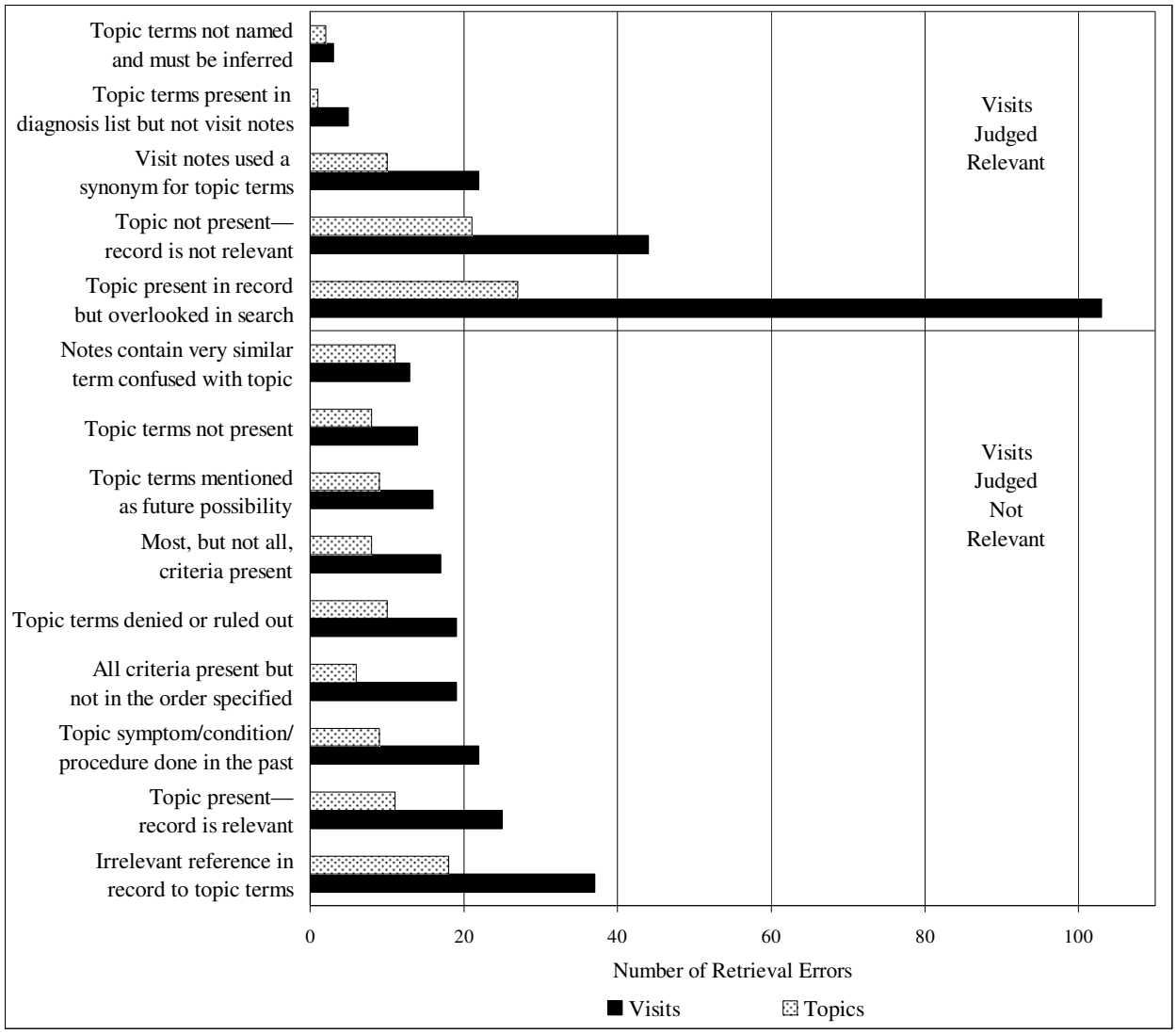

Figure 1. Number of visits and number of topics for each type of incorrect retrieval.

Topic present in record but overlooked in search

For visits that were relevant to the topic, the topic terms were present in the majority of the low-count visits, and the reason for not retrieving it could not be readily determined. We believe most of these had low counts because they were not ranked highly in the original runs, and they were not included in our lists of the top 100 visits. The system "error" in this case may be that other non-relevant visits were retrieved and ranked higher, rather than an actual failure to retrieve the visit. In other words, non-relevant visits pushed the relevant ones down in the ranking, and therefore the error is in retrieving the other non-relevant visits and ranking them highly. This may be due to topic terms commonly occurring in non-relevant documents.

Visit notes used a synonym or lexical variant for topic terms

The most common feature in under-retrieved relevant visits was the use of different terminology in the notes. A search for patients with hearing loss tended to overlook visits mentioning deafness or hearing aids. Spelling variations and typographical errors also fell into this category.

Topic terms not named in notes and must be inferred
Chart notes for two topics contained information that was equivalent to the desired topic, but did not use the specific term or diagnosis. In searching for patients with anion gap acidosis, it was common to miss visits listing lab values for sodium, potassium, chloride, and bicarbonate that arithmetically indicated anion gap acidosis.

Some of the topic terms present in diagnosis list but not visit notes
Each document displayed a header listing the admission diagnosis, discharge diagnoses, and ICD-9 codes. Although they appear at the top of each document, these headers are stored separately from the encounter notes and must be searched separately from the notes. Visits for patients co-infected with hepatitis C and HIV contained both topic conditions in the diagnosis list but not in the notes. Many encounters were for treatment of one of the conditions, a concomitant condition, or something else entirely, and the notes mentioned only one (or none) of the topic conditions. Systems that did not search both the encounter notes and the diagnosis lists would miss relevant visits.

Non-relevant Visits with High Retrieval Counts

Topic present—record is relevant—disagree with relevance judge
A number of visits judged non-relevant were found to be relevant upon closer examination. The topic terms were present in the notes and appeared to have been overlooked during judging.

Topic terms not present—cannot determine why record was retrieved
A number of visits did not appear to contain any reference to the topic. It is unclear why these would be retrieved by the systems. It is likely that systems were performing query expansion, or some other statistical IR approach that introduced semantic error.

Irrelevant reference in record to some topic terms
The largest number of retrieved non-relevant visits made irrelevant references to topic terms. In the search for patients who underwent minimally invasive abdominal surgery, non-relevant visits were retrieved that mentioned minimally invasive surgery of other parts of the body.

Notes contain very similar term confused with topic
Some visits contained terms very similar to the topic terms. For example, searching for ductal carcinoma in situ (DCIS) often retrieved patients with invasive, rather than in situ, ductal carcinoma.

Topic terms denied or ruled out
Several non-relevant visits appeared to be retrieved because the notes mentioned ruling out a diagnosis or denial of a symptom, a common component of medical records. Visits retrieved for patients with hearing loss included many that noted the "patient denies hearing loss."

Most, but not all, topic criteria present
In some cases the visit contained almost all of the desired criteria. One example was the search for adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes, which also retrieved patients with non-insulin dependent diabetes.

All criteria present but not in the order specified by the topic description
Visits in this category contained all the terms present in the topic but in a different time frame or sequence. In many cases, these were patients who already had a condition or procedure specified by the topic but who were being seen for other conditions. Other errors in this category included visits containing all elements of the topic, but that were still not quite right. For example, one visit retrieved for "patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma" referred to a patient with glaucoma and "blurred vision of unclear etiology in right eye."

Topic symptom/condition/procedure done in the past
Procedures done previously, or pre-existing conditions, also increased the likelihood of retrieving non-relevant visits.

Topic terms mentioned as future possibility
Mentioning a procedure to be done at a later date or a condition to be considered in the future increased the likelihood a visit would be retrieved incorrectly. Several non-relevant visits for patients who received physical therapy (topic 130) or diabetic education (topic 123) mentioned they would be done at a later time: "we are going to consult diabetic education."

Discussion

Several themes emerged from our analysis, with terminology differences, negation, and time aspects having the most impact on retrieval inaccuracy. These areas reflect well-known issues in the general IR field and in the application of IR to medical records [13, 14]. The current analysis demonstrates that these problems also arise during cohort identification in medical text.

Overlap in terminology between conditions or procedures

Irrelevant chart references to a term in the topic were associated with retrieval of the greatest number of non-relevant visits. Knowledge and use of the correct medical terminology in the search may increase retrieval accuracy. This could be accomplished by including all the different ways of referring to a condition or procedure and eliminating conditions or procedures that are almost the same as the one desired. For example, a search for patients with hearing loss should include “deaf,” “deafness,” “hearing aids,” and “cochlear implant” in the search criteria.

Some terms are commonly used to refer to specific organs or body parts without actually naming them. For example, a search for ductal carcinoma in situ should distinguish between mammary ductal carcinoma and biliary ductal carcinoma. This search should also distinguish between cancer that has not spread to surrounding tissues (in situ) and cancer that has. Distinguishing between conditions that share terms in their description will require the intelligent use of medical terminology to distinguish relevant from irrelevant phrases. Simply assigning weights to overlapping words regardless of their context was seen to lead to incorrect results.

Negation detection

Successful searches must also incorporate a way of distinguishing between a symptom or condition and the denial or ruling out of that symptom or condition. Our analysis found 19 visits, or 5% of our sample, that were incorrectly retrieved because of this. Negation statements are very common components of medical records: the physical exam will contain references to symptoms denied by the patient, the medical history may include conditions not present in the patient or patient's family, and procedure and chart notes may refer to conditions that have been considered and ruled out. This is a known issue in retrieving information from medical texts, and tools are available to overcome this barrier [15].

Data integrity and lexical variants of terms

Ensuring that clinical data is entered accurately and consistently is necessary. The ability to retrieve data accurately is dependent in large part on the quality of the data. Variations in terminology, spelling errors, and the completeness of the clinical record can all be improved by emphasizing the importance of quality data entry in the clinic and by incorporating spell checking in the EHR. Encouraging consistency in term use and descriptions would further enhance usability.

Time factors or sequencing of terms

Identifying the desired time frame presents another critical factor in correctly retrieving records. Symptoms or procedures done in the past, or being considered for the future treatment, were often confused with those in the present. Our analysis found several instances in which the patient's condition changed dramatically during the visit, resulting in visits that did not correspond to the admission or pre-existing diagnoses. In a few cases, the patient died, rendering orders for post-discharge treatment irrelevant. Incorporating temporal information into query and document analysis may prevent certain classes of retrieval error. Allowing the search to be constructed to exclude past or future conditions or procedures would limit the search results to current issues.

Age proved to be a difficult component for a number of topics. One topic specified "children admitted with cerebral palsy who received physical therapy." All of the non-relevant visits reviewed for this topic referred to adults matching the rest of the topic specifications. The de-identification process used for this data set changed age to a range of ages. This is seen in one record that reports that "the patient is a **AGE[in 20s]-year-old..." This may not be an issue when searching clinical records that have not been de-identified. For this particular case, age could be handled by including terms such as "infant", "month-old," “child,” or “teenager,” or by searching for the phrases produced through the de-identification process.

Incorporating knowledge of logical data constructs

Systems must be designed to search both free-text chart notes as well as more structured data such as diagnosis and medication lists. Searching both areas will enhance retrieval for those conditions listed in only one place. This may be especially relevant for visits where patients have chronic conditions, as in the Hepatitis C and HIV topic, but are seeking treatment for acute illness. This is particularly challenging in the case of the TRECMed data set, since that level of structure is not explicitly encoded in the encounter documents, which are provided as flat text files.

Limitations

There were several limitations in the design and execution of this study. The method of reviewing visits to determine the reason for incorrect retrieval produced one source of potential error. The classification schema for error types was generated using an iterative process that included the feedback and consensus of all the authors. However, only one author reviewed the system retrieval data in detail, and therefore some visits may be misclassified either due to operator error or ambiguity in the source documents of schema itself. Furthermore, once the most likely reason for incorrect retrieval was identified, the reviewer recorded the reason and moved on to the next visit, not evaluating the possibility that there may be more than one reason for incorrect retrieval for an individual document.

Another limitation of this study resulted from the fact that we did not have access to the search strategy or implementation details of each IR system from which runs were submitted. It would not be possible to determine with certainty the reason an individual system correctly or incorrectly retrieved a document without access to these internal system details. It is likely that in some cases our assigned error types do not correspond precisely to the reason that an individual system incorrectly retrieved or missed a document. As mentioned in the results section, a number of non-relevant visits were retrieved although they did not contain any obvious relevance to the topic. We presume that many of these were retrieved incorrectly as a result of the query expansion used. In these visits, the reason for incorrect retrieval cannot be confirmed without access to the code.

These limitations would be unlikely to have a significant effect on the conclusions of the study because the goal of this analysis was to identify general challenges and trends in the retrieval task across the set of submitted systems rather than perform a precise error analysis for any individual system. The goal of this study was to determine challenges in the task based on the characteristics of the data. The data examples representing challenging situations were identified by looking at common failures across the submitted systems. For these reasons, in spite of the limitations, we remain confident that our analysis identifies valid areas to address that will improve the ability to more effectively retrieve and utilize clinical records for secondary purposes such as research and surveillance.

Conclusions

The EHR is a valuable source of data for enhanced patient care, research, and quality improvement. The usefulness of this data depends on the ability to locate information as accurately as possible. This error analysis provides insight into areas for future improvement in EHR-based retrieval systems.

In our work, the most common sources of retrieval error included irrelevant chart references to a term, variation in terminology and spelling, lack of distinction between different conditions with similar names, lack of distinction between past, present, and future conditions or procedures; and failure on the part of IR systems to distinguish between the presence or denial of a symptom or condition. All of these sources of error will have to be addressed by future improvements in EHR systems and retrieval capabilities for them.

This analysis also highlights areas where changing EHR data entry could have a beneficial impact on retrieval. The EHR is an important resource for biomedical research and healthcare improvement. Improvement in patient record information retrieval is only one area where clinical text processing has the potential to advance medical research and improve healthcare. Additional work is necessary to realize the full potential of the valuable information stored in the EHR.

References

1. Safran C, Bloomrosen M, Hammond WE, Labkoff SE, Markel-Fox S, Tang P, et al., *Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper.* Journal of the American Medical Informatics Association, 2007. 14: 1-9.

2. Hripcsak G, Friedman C, Anderson PO, DuMouchel W, Johnson SB, and Clayton PD, *Unlocking clinical data from narrative reports: a study of natural language processing.* Annals of Internal Medicine, 1995. 122: 681-688.
3. Stanfill MH, Williams M, Fenton SH, Jenders RA, and Hersh WR, *A systematic literature review of automated clinical coding and classification systems.* Journal of the American Medical Informatics Association, 2010. 17: 646-651.
4. Hersh WR, *Information Retrieval: A Health and Biomedical Perspective (3rd Edition)*. 2009, New York, NY: Springer.
5. Cohen AM and Hersh WR, *A survey of current work in biomedical text mining.* Briefings in Bioinformatics, 2005. 6: 57-71.
6. Voorhees EM and Harman DK, eds. *TREC: Experiment and Evaluation in Information Retrieval*. 2005, MIT Press: Cambridge, MA.
7. Hersh W and Voorhees E, *TREC genomics special issue overview.* Information Retrieval, 2009. 12: 1-15.
8. Voorhees EM and Tong RM. *Overview of the TREC 2011 Medical Records Track. The Twentieth Text REtrieval Conference Proceedings (TREC 2011).* 2011. Gaithersburg, MD: National Institute for Standards and Technology.
9. Anonymous, *Initial National Priorities for Comparative Effectiveness Research*. 2009, Institute of Medicine: Washington, DC.
10. Buckley C and Voorhees E. *Evaluating evaluation measure stability. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000. Athens, Greece: ACM Press. 33-40.
11. Buckley C and Voorhees EM. *Retrieval evaluation with incomplete information. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2004. Sheffield, England: ACM Press. 25-32.
12. Voorhees EM. *Variations in relevance judgments and the measurement of retrieval effectiveness. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998. Melbourne, Australia: ACM Press. 315-323.
13. Nadkarni PM, Ohno-Machado L, and Chapman WW, *Natural language processing: an introduction.* Journal of the American Medical Informatics Association, 2011. 18: 544-551.
14. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, and Uzuner O, *Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions.* Journal of the American Medical Informatics Association, 2011. 18: 540-543.
15. Chapman WW, Bridewell W, Hanbury P, Cooper GF, and Buchanan BG, *A simple algorithm for identifying negated findings and diseases in discharge summaries.* Journal of Biomedical Informatics, 2001. 34: 301-310.