JAMIA PrePrint: Accepted Article. Published August 21, 2007 as doi:10.1197/jamia.M2407

Knowledge-Based Methods to Help Clinicians Find Answers in MEDLINE

Charles A. Sneiderman MD PhD,[a] Dina Demner-Fushman MD PhD,[a] Marcelo Fiszman, MD PhD,[b] Nicholas C. Ide,[a] and Thomas C. Rindflesch PhD[a]

[a]*Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD* [b]*Graduate School of Medicine, University of Tennessee, Knoxville, TN*

Corresponding Author:

Charles Sneiderman

National Library of Medicine

8600 Rockville Pike

Bethesda, MD 20894

Phone 301-435-3253 Fax 301-402-4080 charlie@nlm.nih.gov

Abstract

Objective: Large databases of published medical research can support clinical decision making by providing physicians with the best available evidence. The time required to obtain optimal results from these databases using traditional systems often makes accessing the databases impractical for clinicians. This paper explores whether a hybrid approach of augmenting traditional information retrieval with knowledge-based methods facilitates finding practical clinical advice in the research literature.

Design: Three experimental systems were evaluated for their ability to find MEDLINE® citations providing answers to clinical questions of different complexity. The systems (SemRep, Essie, and CQA-1.0), which rely on domain knowledge and semantic processing to a varying extent, were evaluated separately and in combination. Fifteen therapy and prevention questions in three categories (general questions, intermediate, and specific) were searched. The first ten citations retrieved by each system were randomized, anonymized, and evaluated on a three-point scale. The reasons for ratings were documented.

Measurements: Metrics evaluating the overall performance of a system (Mean Average Precision – MAP, Binary Preference – Bpref) and metrics evaluating the number of relevant documents in the first several presented to a physician were used.

Results: Scores (MAP=0.57, Bpref=0.71) for fusion of the retrieval results of the three systems are significantly ($p < 0.01$) better than those for any individual system. All three systems present three to four relevant citations in the first five for any question type.

Conclusion: The improvements in finding relevant MEDLINE citations due to knowledge-based processing show promise in assisting physicians answer questions in clinical practice.

Key words: MEDLINE, Decision Support Techniques, Information Storage and Retrieval

Introduction

Surveys of physicians consistently show clinical queries unanswered at the time of diagnostic and therapeutic decision-making [1, 2]. Physicians report that they did not pursue answers because of the time required to find information (75%) and due to resource inconvenience (8.3%) [2]. Ely et al. [1] identified five major obstacles in addition to time: 1) difficulty formulating a question; 2) difficulty selecting an optimal search strategy; 3) failure of a resource to cover the topic; 4) uncertainty whether all the relevant evidence had been found and the search could stop; and 5) difficulty synthesizing multiple bits of evidence.

Several possible approaches to addressing these problems have been investigated. Studies of interfaces for structured query formulation report improvements in precision without negative impact on recall, but not necessarily higher user satisfaction or acceptance of the interface [3]. Moreover, training in the searching and appraisal of medical literature are essential for finding satisfactory answers to clinical questions [1, 4]. Merely using electronic information resources of choice, physicians were not always successful in answering clinical questions [5]. The paradigm of evidence-based medicine (EBM) [6] is an important resource in devising solutions to these problems.

In an environment where time and effort are at a premium, the value of MEDLINE for assisting in therapeutic decision making depends not only on well-formed questions but also on algorithms that can improve precision and recall in finding clinically relevant

information [7]. Research indicates that given enough time and skill, clinicians can find answers to their questions in MEDLINE [8]. A recommended strategy is to reduce search results by focusing the question, often by adding more terms to a query. This requires a clinician to invest time in analyzing the information needed to identify search terms describing a clinical situation. An alternative approach is often observed in practice [9]: A clinician under-specifies a search by submitting two or three terms and then selects relevant documents while browsing the results, often using clinical practice guidelines for evaluation. Some of these strategies could be implemented automatically by incorporating domain knowledge in the search and then post-processing the search results in order to re-rank results for relevance to the query.

The research presented here explores the effectiveness of such automatic methods. We evaluate three knowledge-based automatic methods being developed at the National Library of Medicine to assist physicians find clinically relevant information in MEDLINE. The three systems use medical domain knowledge encoded in the Unified Medical Language System® (UMLS®)[10] (alone, or in combination with corpus-based methods) to find information for clinical queries of varying complexity.

The first system, SemRep Summarization [11], uses natural language processing and automatic summarization of MEDLINE citations to find the most relevant information about a clinical query within PubMed retrieval results. The second system, CQA-1.0 [12], uses EBM recommendations for finding best answers for questions about treatment and prevention [13]. Both SemRep and CQA-1.0 re-rank a set of MEDLINE citations

retrieved using PubMed and clinical query filters [14]. The third system, Essie [15], is a probabilistic search engine that uses fine-grained tokenization, concept searching utilizing UMLS-derived synonymy, and phrase searching based on the user's query to find the best MEDLINE citations for answering a clinical question. All three systems use structured domain knowledge from the UMLS to a varying extent and rely directly or indirectly on the MeSH (Medical Subject Headings) controlled vocabulary used to manually index MEDLINE citations.

After providing an overview of the three systems, we concentrate on evaluating them with respect to finding information about treatment and prevention of 15 disorders. A test collection was constructed using the Text REtrieval Conference (TREC) pooling strategy [16]. A rating scale developed to evaluate the utility of MEDLINE citations in clinical decision making [17] was used to compare the performance of the three methods in answering clinical queries. We provide results in several evaluation metrics as a way of predicting the effectiveness of the systems under consideration in different clinical situations. The goal of this work is to explore approaches to increasing the utility of the primary literature with respect to answering clinical questions of varying complexity. Specifically, we investigate whether automatic "understanding" of MEDLINE citations based on medical domain knowledge can provide practical support for clinicians in therapeutic decision-making.

Background

Prior work

Previously, several domain knowledge-based automatic approaches to indexing and retrieving scientific publications that contain answers to clinical questions have been explored. The approaches range from developing a set of generic queries for clinical information retrieval [18] to matching semantic representation of a patient's clinical record with that of a MEDLINE citation [19] and generating personalized summaries in response to physicians' questions [20]. The most thoroughly researched approaches have been automatic query expansion, concept indexing, and retrieval-based feedback. The reported evaluations of these approaches have not demonstrated consistent improvements in satisfying clinicians' information needs. For example, automatic query expansion using controlled vocabulary terms was shown to improve overall average precision [21], but only a third of the queries showed improvement in a study of synonym- or hierarchical thesaurus-based query expansion [22]. Concept indexing implemented in the SAPHIRE system helped physicians [23], whereas it degraded performance for 30 medical questions in a study of effectiveness of conceptual indexing [24]. Query expansion based on retrieval feedback (terms from the top few documents retrieved in an initial run are added to the original query) improved average precision [25] but is also known to degrade performance [26].

In addition to using domain knowledge to assist clinicians as they actively seek information, promising results have been obtained through passive query generation using patient records for query formulation [27]. The effect of unobtrusively providing

context-specific links between clinical data and information resources in the form of "infobuttons" [28] has been examined in several recent studies. Rosenbloom et al. [29] found a significant increase in the use of educational materials in the Care Provider Order Entry system when the materials could be accessed through visible hyperlinks (as opposed to menus). Cimino et al. observed positive results and increased use of infobuttons over 5.7 years [30]. The success of context-specific access to knowledge access in this study varies with context and user type. Similar to the study by Cimino et al., Del Fiol et al. [31] observed increased infobutton use with preference for secondary sources (such as Micromedex and UpToDate) that summarize the results of clinical studies and under-utilization of resources providing access to the primary literature (such as MDConsult and PubMed). Although these resources provide valuable general information to the clinician, there remains a need for methods that help find answers to particular questions.

PubMed

PubMed automatically recognizes controlled vocabulary terms matching a user's query with the entries in several translation tables. If a match is found in the MeSH translation table, the term is searched as MeSH (including the MeSH term and any specific terms indented under that term in the MeSH hierarchy) and as a text word. One of the advanced PubMed search options, clinical queries, is a set of filters designed to find clinically relevant and scientifically sound studies [32]. These filters automatically expand queries using predefined sets of terms designed to limit search results to articles addressing one of the four major clinical tasks (etiology, diagnosis, therapy, and prognosis). For each

task, clinical queries provide two search choices: specific (narrow) or sensitive (broad). For example, a “narrow therapy” clinical query augments a user’s query with the following search terms: randomized controlled trial [Publication Type] OR (randomized [Title/Abstract] AND controlled [Title/Abstract] AND trial [Title/Abstract]).

SemRep Summarization

The first re-ranking system considered in this study is based on SemRep Summarization [11, 33], which depends on the semantic natural language processing system SemRep [34, 35]. SemRep identifies semantic predications (relationships) in biomedical text using underspecified syntactic analysis and structured domain knowledge from the UMLS. SemRep predications consist of UMLS Metathesaurus® concepts as arguments and UMLS Semantic Network relations as predicates (relations between the concepts). Analysis begins with an underspecified syntactic parse that relies on the SPECIALIST lexicon [36] and a part-of-speech tagger [37]. MetaMap [38] then matches noun phrases to concepts in the UMLS Metathesaurus® and determines the semantic type for each concept. Concepts are identified as arguments in a predication using syntactic constraints based on dependency grammar rules and semantic constraints imposed by the Semantic Network. Predications representing core aspects of the clinical scenario were central to this study. These predications have predicates such as TREATS, CO-OCCURS_WITH, and OCCURS _IN and arguments belonging to the UMLS semantic groups [39] Chemicals and Drugs, Disorders, and Population Groups.

SemRep Summarization is an automatic summarization system in the semantic abstraction paradigm [40]. The system takes as input a list of predications extracted by

SemRep from biomedical text (MEDLINE citations to be re-ranked in this study). Output is a condensed set of predications that serves as a summary of salient information on a specified topic in the citations processed. The core of the system is a transformation stage that identifies the most important information with respect to the specified topic. The transformation stage relies on four principles: 1) Relevance, which keeps predications on the topic of the summary; 2) Connectivity, which keeps related predications that share an argument with the summary topic; 3) Novelty, which eliminates uninformative predications; and 4) Saliency, which keeps high frequency predications [41]. Predications in the summary are linked to the citations from which they were extracted and play an important role in exploiting SemRep Summarization for re-ranking retrieved citations in this study.

CQA-1.0

Another re-ranking method is implemented in the prototype clinical question answering system CQA-1.0. In this system, questions and MEDLINE citations are represented using frames which capture the fundamental elements of EBM: 1) clinical scenario; 2) clinical task; and 3) strength of evidence. A question frame submitted to the system is used to generate a query and search MEDLINE using PubMed. Retrieved citations are processed with several knowledge extractors and classifiers that rely on a combination of UMLS concept recognition using MetaMap [38], manually derived patterns and rules, and supervised machine learning techniques [12] to identify the fundamental EBM components listed. The PICO framework (Problem/Patient, Intervention, Comparison, and Outcome) designed to help clinicians formulate clinical questions [42] is used to capture the first fundamental component (clinical scenario) in a MEDLINE citation. The

elements of a clinical scenario are identified and extracted by four knowledge extractors. The problem extractor identifies a UMLS concept in the semantic group [39] Disorders, which is the focus of a given study. The population extractor identifies phrases containing numerical expressions and concepts with the semantic type ‘Group’ and its children. The intervention and comparison extractor is based on finding concepts with nine semantic types (for example, ‘Therapeutic or Preventive Procedure’ and ‘Diagnostic Procedure’). Identification of the second fundamental component (clinical task) is based on rules derived from 1) search strategies encoded in PubMed clinical queries, 2) the JAMA EBM tutorial series on critical appraisal of medical literature [43], and 3) MeSH scope notes. The third fundamental component (strength of evidence) is based on the type of clinical study presented in the publication, authority of the journal that published it, and date of publication. Citation scoring and re-ranking with respect to a question are based on 1) matching the question and citation with PICO frames, 2) matching the clinical task that generated the question with the task identified in the clinical study (treatment and prevention, for this study), and 3) the strength of the evidence presented in the study.

Essie

A different approach to finding citations answering clinical questions is implemented in Essie, a probabilistic search engine developed at the National Library of Medicine for the ClinicalTrials.gov database. Essie incorporates a number of strategies aimed at alleviating the need for sophisticated user queries [15]. These strategies include a fine-grained tokenization algorithm that preserves punctuation information, concept searching utilizing UMLS-derived synonymy, and phrase searching based on the user’s query. Citations containing phrases identified in a user’s query are ranked higher than citations

containing individual words comprising the phrase. Position of a matching phrase or term in a citation also influences the rank of a citation with respect to a query. For example, if a phrase is found in the title, the citation is ranked higher than the one that contains this phrase in the abstract. Essie provides several possibilities for query expansion: exact match; SPECIALIST Lexicon-based [36] morphological expansion of terms; and UMLS-based expansion of concepts. Essie was the best performing search engine in the 2003 TREC Genomics track [44] and one of the best performing systems in the 2006 TREC Genomics track [45].

Evaluation strategy

Our evaluation is based on techniques developed over the past 15 years in the framework of TREC – a yearly large-scale evaluation of information retrieval and question answering systems [16]. Traditionally, systems are evaluated using test collections consisting of 1) a corpus of documents (for example, MEDLINE citations), 2) a set of queries or questions (called topics in TREC), and 3) relevance judgments – human assessment of the relevance of each document in the collection to a given topic. Ideally, each document in the corpus would be judged with respect to each topic. Due to the size of modern document collections, such evaluation is not feasible even in the framework of TREC, which leads to an alternative strategy of first selecting a subset of documents to be judged, then assessing the relevance of these documents to the topics, and finally using these relevance judgments to assess the relative performance of the systems. A practical solution to the question of selection of an appropriate small subset of documents is the TREC pooling strategy. Documents to be judged for each topic are contributed to the pool by each information retrieval system participating in the evaluation. In TREC, the

top 75 to 100 documents returned by each system are combined into a set given to the judge. The judged documents are subsequently used to evaluate the relative performance of the contributing systems.

Methods

In exploring the effectiveness of SemRep Summarization, CQA-1.0, and Essie in pinpointing answers to clinical questions in MEDLINE citations, we first created a test collection based on published clinical questions deemed to be of interest to the majority of American family physicians [46] and a set of MEDLINE citations created using the retrieval results of the three experimental systems and the TREC pooling strategy. The citations were judged for relevance to the questions on a three-point scale. Results from the three systems being evaluated were compared against the test collection individually and fused. Several evaluation metrics were computed in order to predict how the systems would perform in a clinical setting.

Creating a test collection

In order to evaluate the performance of the three systems under scrutiny we constructed a test collection consisting of 15 clinical questions along with relevant MEDLINE citations and judgments of their relevance to the questions. The top ten documents returned by each system were added to the pool of documents evaluated by the first author, who did not participate in the development of any of the experimental systems.

Question selection

For the questions, the first author, a practicing family physician, selected 15 queries (Table 1) from the Family Practice Information Network (FPIN) clinical queries

collection, which is published monthly in the Journal of Family Practice and American Family Physician and contains queries typically generated in the daily practice of general medicine [46]. Even if the query did not adhere to the syntactic form of a question (for example, specific queries 4 and 5), the original queries were not modified. The queries selected pertain to therapeutic or preventive interventions for clinical problems and can be regarded as instances of generic clinical questions [47]. We identified two types of clinicians' information needs: general (an overview of a topic) and specific (an exact answer to a focused question). When inspecting the FPIN clinical queries collection, we determined that some questions are intermediate; they do not call for an overview but are not focused enough for an exact answer. The nature of the questions in the FPIN collection warrants exploration of all three question types. Five queries were selected as "general" in that the only element of a clinical scenario in the question was the problem. Five were "intermediate," with clinical scenario elements of population group, intervention, or outcome included with the request for therapy or prevention of a problem. Finally, five were "specific" or "complex," including at least two elements of a clinical scenario selected from population group, intervention, or outcome (in addition to the problem). Our focus on therapy and prevention questions and the intent to evaluate the systems' performance for all levels of difficulty precluded random selection of the questions. Instead, the first author selected five questions of interest to his practice from each level.

In the FPIN collection, each query is accompanied by a published answer derived from a careful process involving a search of the published literature by a medical librarian,

review of that literature and any other sources of evidence by two clinicians trained in evidence-based medicine, and editorial review by FPIN academic family physicians. The MEDLINE citations published in the evidence summaries [48] were used for reference while judging the abstracts selected for evaluation as described below.

Table 1 Clinical Questions used to retrieve MEDLINE citations.

General Questions	
1.	What is the most effective treatment for external genital warts?
2.	What are the most effective interventions to reduce childhood obesity?
3.	What is the most effective treatment for acute low back pain?
4.	What is the best approach to treatment of osteoporosis?
5.	What are effective treatments for panic disorder?
Intermediate Questions	
1.	What is the most effective treatment for ADHD in children?
2.	Can type 2 diabetes be prevented through diet and exercise?
3.	What are the best therapies for acute migraine in pregnancy?
4.	Do steroid injections help with osteoarthritis of the knee?
5.	What is the best antiviral agent for influenza infection?
Specific Questions	
1.	Are antibiotics effective in preventing pneumonia for nursing home patients?
2.	Is methylphenidate useful for treating adolescents with ADHD?
3.	What is the best treatment for gastroesophageal reflux and vomiting in infants?
4.	Antiviral Agents for Pregnant Women with Genital Herpes.
5.	Intravenous Fluids for Children with Gastroenteritis.

Assigning relevance judgments

Relevance judgments were generated using the pooling strategy developed for TREC [16]. The top ten documents from each system were collected and duplicates were removed. The titles and abstracts of MEDLINE citations were printed in random order and given to the first author. The non-topical characteristics of key articles identified in [49] (authors and their institutional affiliations, or document types) were removed so that the judgments could be based only on the content of the abstract text. The abstracts were rated on a three-point scale: "A" – leads to an answer (definitely useful in clinical decision-making for the question); "B" – might lead to an answer (relevant, but not sufficient to make a decision); "C" – not relevant (not useful for clinical decision-making). In addition to randomized blinded evaluation of the citations, the first author documented the reasons for rating. Analysis of these reasons for rating provides information about features that make a citation particularly useful in decision support.

Retrieving and re-ranking MEDLINE citations

Each FPIN question was used to search MEDLINE with PubMed and Essie, limited to no later than the date of the FPIN answers for each question. Essie returns relevance-ranked output directly. The chronologically ordered citations from PubMed were subsequently re-ranked for each query using SemRep Summarization and the CQA-1.0 system.

For the strategies based on SemRep Summarization and CQA-1.0, an initial PubMed search strategy was to use the narrow therapy clinical queries filter and the clinical terms identified in a given question. For example, the clinical term in the FPIN question "What is the best approach to treatment of osteoporosis?" is *osteoporosis*. The addition of the

PubMed clinical queries filter to this term yields the following query: *(osteoporosis[MeSH Terms] OR osteoporosis[Text Word]) AND (randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract]))*. If the initial search yielded no results, the search was repeated with the clinical queries filter replaced with the following limits: citations with abstracts, restricted to human studies written in English. Two of the intermediate questions and all specific questions required this substitution. A total of 1305 documents for the first set, 925 for the second, and 959 for the third set were retrieved from MEDLINE using PubMed. Unranked PubMed results were used as a baseline against which experimental results were compared.

In exploiting SemRep Summarization for re-ranking retrieved citations, predications were extracted from the MEDLINE citations retrieved for each FPIN query. After summarizing the predications, the citations from which the predications were extracted were promoted as being more highly relevant to the query based on how closely and how frequently arguments in those predications matched Metathesaurus concepts extracted from the query.

The CQA-1.0 re-ranking algorithm promotes citations in which the automatically identified problems and interventions match those in the question; patient oriented outcomes are identified with strong confidence, the task matches that of the questions, the study population is large, and the strength of evidence is high [50].

In searching with Essie, a strategy similar to PubMed clinical queries, using EBM- and therapy-related terms (such as *therapeutic use, clinical trial,* etc.) was applied. Unlike the clinical queries filters, this strategy promotes EBM-oriented citations without reducing the number of retrieved citations. Essie core document ranking promotes citations which contain query phrases in the fields observed to be most informative, for example in the title [51]. To take advantage of UMLS synonymy, UMLS-based expansion of concepts was used in the search. Essie returned 2,500 citations in the first set, 896 in the second, and 673 in the third.

Fusion of results

In addition to evaluation of individual systems, the ranked results generated by each system were merged using fusion. Fusion was based on the rank order assigned to a document by each system, rather than on scores. This is because the systems either do not score documents or generate scores for ranking purposes only (that is, scores represent neither the similarity of a citation and the query nor the system's confidence in the relevance of a citation to the query). This approach relies on document overlap, which for SemRep, CQA-1.0 and the baseline PubMed retrieval constitutes the whole result set. The results were merged using the fusion approach proposed by [52]. The contribution of each system to the final ranking was weighted equally.

Evaluation

Five sets of output were evaluated as part of this study: the ranked output from each of the systems under consideration, the fused output from all three, and unprocessed PubMed output (baseline). The trec_eval-8.0 package [53] was used to evaluate the

results. The systems were evaluated under two conditions: “strict,” considering only citations graded “A” in the three-point scale evaluation to be relevant, and “soft,” considering both three-point scale “A” and “B” grade citations relevant to the question. Because the relative ranking of the systems with respect to the baseline is identical under both conditions we present and discuss the results of the soft evaluation. The differences in retrieval results between systems were compared using a Wilcoxon signed ranks test for all metrics. P-values less than 0.05 were considered significant. The Wilcoxon signed ranks test is used when the values in the two results being compared are naturally paired (for example, the same set of documents is ranked by two systems), and the relative magnitude as well as the direction of the differences is considered [54].

Two classes of evaluation metrics were used to account for two different information needs experienced by clinicians, one general and the other focused. The first type of information need is reflected in our general questions and corresponds to a situation in which a clinician might need an overview of a topic. In this scenario, a clinician would be interested in both precision (the percentage of the retrieved citations that are relevant) and recall (the percentage of the relevant documents that are retrieved.) Evaluation metrics that reflect this need are:

1) *Mean Average Precision (MAP):* For multiple topics, it is the mean of the average precision scores for each of the topics. The average precision score for a single topic is computed by averaging the precision after each relevant document is retrieved [16]. This metric has recall and precision components and is widely-

accepted in information retrieval as reflecting the level of performance a user should expect for a new topic retrieved using a system that achieves a given MAP value.

2) *Binary Preference (Bpref):* A preference-based measure that depends on the number of documents that were judged as non-relevant that were retrieved with higher rank than relevant documents. This distinguishes Bpref from MAP, which is determined by the ranks of the relevant documents in the result set and makes no distinction between documents explicitly judged as not relevant and documents that are not judged [55]. This measure is reported to be more stable than MAP with incomplete judgments, which is probably the case for the pilot studies presented below.

3) *R precision:* measures precision after R documents have been retrieved, where R is the total number of relevant documents for a query.

The second type of information need experienced by clinicians corresponds to a situation in which an exact answer to a well-focused question is required (reflected in our specific questions). As clinicians are willing to spend no more than 4-5 minutes evaluating search results [56], it is important that the answer to the question be found in the first few citations retrieved. Metrics that evaluate how soon a user will see the answer and how many relevant citations are at the top of the retrieval results list are:

1) *Precision at five retrieved documents (P@5):* measures the fraction of relevant documents in the top five documents retrieved.

2) *Precision at ten retrieved documents (P@10):* measures the fraction of relevant documents in the top ten documents retrieved.

3) *Mean Reciprocal Rank (MRR):* is the metric used in TREC question answering evaluation [57]. It quantifies the "expected search length" and is computed as the mean of the individual questions reciprocal ranks. The reciprocal rank of the top relevant document is the reciprocal of the rank at which the first relevant document was found.

Results

Table 2 summarizes the results of the exploration of the differences between the three experimental approaches to document ranking for clinical question answering. PubMed results are used as a reference point to provide a comparison of the experimental retrieval approaches with the state-of-the-art baseline (which includes the clinical queries filters). The table also presents the fusion results for the experimental systems. The best results for individual systems and the best fused results are shown in bold.

Overall (Table 2), CQA-1.0 performs best with respect to the baseline. Fusion of the three systems also performs well overall and outperforms CQA-1.0 for general questions.

Table 2 Results for all questions.

System	MAP	Bpref	R-prec	MRR	P@5	P@10
PubMed	0.3058	0.4029	0.3033	0.7761	0.5067	0.4400
Essie	0.4248	0.5384	0.4141	0.9333	0.8133	0.7800
SemRep Summ.	0.2900	0.3338	0.3306	0.8556	0.7867	0.6867
CQA-1.0	**0.4938**	**0.5783**	**0.4865**	**0.9556**	**0.8533**	**0.8733**
Fusion	**0.5708**	**0.7077**	**0.5683**	0.9222	0.7467	0.6533

In Tables 3 - 5, results are presented categorized by the complexity of the question and from the point of view of how well evaluated systems perform in response to general versus focused information needs. For general questions (Table 3), there is no single trend discernible. As noted, MAP, Bpref, and R-prec are likely to be most valuable for evaluating general questions as expressing a general information need. Essie and CQA-1.0 significantly outperformed PubMed according to MAP, but not Bpref. Fusion does well for Bpref.

Table 3 Results for general questions.

System	MAP	Bpref	R-prec	MRR	P@5	P@10
PubMed	0.1410	0.5148	0.1185	0.6217	0.2400	0.1800
Essie	**0.4403**	**0.6088**	0.3775	**1.0000**	**0.8800**	0.8400
SemRep Summ.	0.2919	0.3703	0.3235	0.7667	0.8400	0.8400
CQA-1.0	0.4131	0.4686	**0.4219**	**1.0000**	0.8400	**0.8800**
Fusion	0.4389	**0.6897**	**0.4760**	**1.0000**	0.7200	0.6600

The baseline PubMed performance for intermediate questions (Table 4) was significantly better then for the general questions. The experimental approaches did not significantly improve on the baseline for these questions, according to the measures reflecting a general information need (MAP, Bpref, and R-prec). SemRep Summarization and CQA-1.0 did better on the measures reflecting a more focused information need (SemRep Summarization for MRR and P@5; CQA-1.0 for P@10).

Table 4 Results for intermediate questions.

System	MAP	Bpref	R-prec	MRR	P@5	P@10
PubMed	0.3640	0.3996	0.3387	0.8400	0.6400	0.5600
Essie	0.2986	0.3707	0.3255	0.8000	0.6800	0.6400
SemRep Summ.	0.3624	0.3972	0.4162	**1.0000**	**0.8000**	0.7200
CQA-1.0	**0.4395**	**0.5468**	**0.4367**	0.8667	0.7200	**0.7600**
Fusion	**0.4897**	**0.6333**	**0.5423**	0.7677	0.6000	0.4600

The baseline is higher for specific questions; however, the experimental approaches apparently benefited from additional details provided in the complex questions (Table 5). The CQA-1.0 system, specifically designed to handle questions in the EBM-recommended form, benefited most among individual systems, scoring particularly well on MRR, P@5, and P@10. Fusion also does well on these measures in response to a focused information need. CQA-1.0 also did well according to MAP for the complex questions, 0.6286. However, the difference between CQA-1.0 and Essie is not statistically significant. Fusion of the results for the 3 systems (MAP=0.7839) is particularly successful for this class of questions.

Table 5 Results for specific ("complex") questions.

System	MAP	Bpref	R-prec	MRR	P@5	P@10
PubMed	0.4125	0.2944	0.4527	0.8667	0.6400	0.5800
Essie	0.5356	0.6358	0.5394	**1.0000**	0.8800	0.8600
SemRep Summ.	0.2158	0.2338	0.2520	0.8000	0.7200	0.5000
CQA-1.0	**0.6286**	**0.7195**	**0.6008**	**1.0000**	**1.0000**	**0.9800**
Fusion all	0.6677	0.6572	0.5775	0.9000	0.8800	0.8200
Fusion 3	**0.7839**	**0.8001**	**0.6866**	**1.0000**	0.9200	0.8400

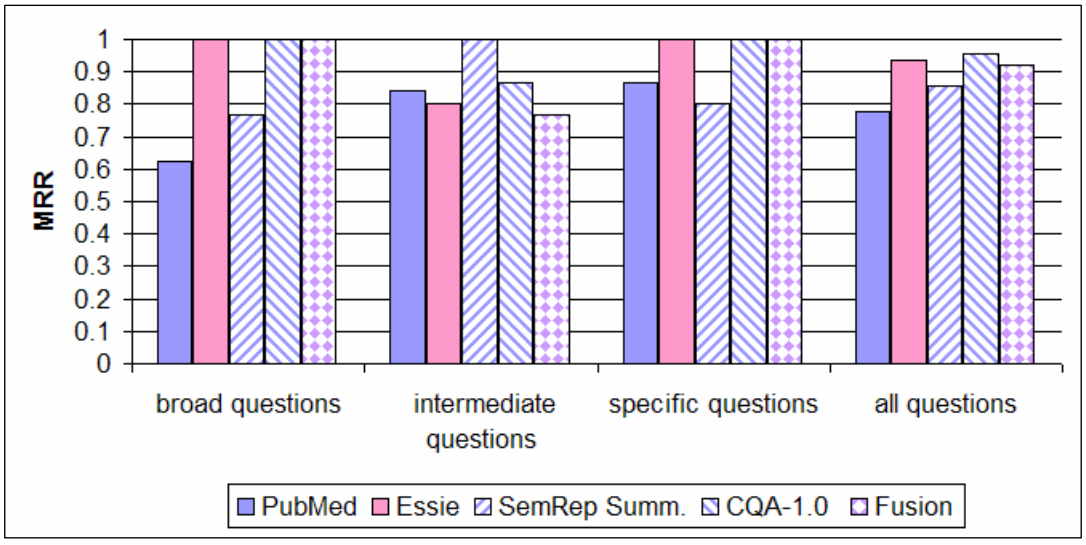

Figure 1. Mean reciprocal rank of the first relevant document retrieved by each method.

In terms of finding answers to specific questions, all experimental methods were successful in promoting relevant documents to the higher ranks, achieving MRR from 0.86 to 0.96 (Figure 1); 79% to 85% precision at five retrieved documents; and 69% to 87% precision at ten documents, meaning that three to four of the first five documents retrieved by the evaluated systems (and six to eight of the first ten) provide information that potentially or definitely leads to answers to a clinical question.

Discussion

Because of the size of the pool and the number of questions, the results of this exploration are promising but not definitive. For 15 questions, the CQA-1.0 improvement over PubMed is statistically significant ($p < 0.01$), and so is the improvement of the fused results of SemRep, Essie and CQA-1.0 over the individual systems and the baseline. Mean average precision is not always improved by semantic re-ranking; that is, well-formed PubMed queries provide respectable recall and precision, and thus a good overview of the information landscape for a given topic. Semantic re-ranking, however, improves the rate of finding answers to specific questions.

External knowledge

The three systems evaluated rely on UMLS domain knowledge to manipulate semantic content in MEDLINE citations. Such content includes: 1) the number of subjects; 2) comparison of multiple therapies; 3) placebo control; and 4) comparative cost of interventions. Previous research [49] has identified non-semantic characteristics of articles as being important in identifying key articles. These include methodological rigor, authors and their institutional affiliations, document types, and population studied.

Our research suggests that such cues, which are used in Essie and CQA-1.0, but not in SemRep, contribute to performance. Judging by the reciprocal rank of the top retrieved document, and precision at five and ten documents, semantic re-ranking is necessary when a clinician is interested in (or has time for) only the first few citations. However, using Essie might preclude the need for re-ranking for general and intermediate questions.

Yet another type of key element identified in this study requires external knowledge in addition to semantic processing. These characteristics include: 1) availability of a therapy for the local practitioner community (e.g. approval by the US Food and Drug Administration or availability in a community environment) and 2) applicability of the study results more generally, for example, extending the results of a clinical trial conducted in a sub-population to the population of interest.

Notes taken during evaluation identified additional non-semantic criteria used to assess usefulness of the citation to a clinician. The rater (who considers himself a typical primary care physician[1]) evaluated the utility of a citation and used non-topical cues present in the citation, as well as "world knowledge." For example, for the query "what is the most effective treatment for ADHD in children?" a citation entitled "Attention-deficit hyperactivity disorder in children and youth: a quantitative systematic review of the efficacy of different management strategies" was judged as "A" (leads to an answer; definitely useful in clinical decision-making for the question) with the assumption that a

[1] See American Academy of Family Physicians Policy & Advocacy [58] for one definition of a typical family doctor.

systematic review was exhaustive of the published literature for efficacy. In contrast, for the query "what is the best antiviral agent for influenza infection?" a citation "Efficacy and safety of oseltamivir in treatment of acute influenza: a randomized control trial" was judged "A" even though the comparisons were to placebo only. The rater believes that a citation with comparisons to various treatment methods is unlikely to appear in the primary literature.

Judgments which were rated as "B" (not sufficient to answer the query but helpful in medical decision making) also need to be qualified by an understanding that the rater assumed the knowledge level of the typical primary care physician. Thus for the query "what is the best treatment for gastroesophageal reflux and vomiting in infants?" a citation not related to therapy entitled "The infant with chronic vomiting: the value of the upper GI series" was retrieved by the probabilistic search method. It was rated "B" because the rater thought that most primary care physicians might not know that "in a study of 344 otherwise healthy infants referred to pediatric gastroenterologists for chronic vomiting "findings other than gastroesophageal reflux were seen in only 2 patients …(0.6%)" and that knowledge might influence a "best therapy" decision.

Judgments which were rated "C" (not helpful in answering the clinical query) also made some assumptions regarding utility to the decision-maker. For the query "in children with acute vomiting and diarrhea (gastroenteritis), does treatment with intravenous fluids improve recovery compared with oral rehydration therapy (ORT)?" a citation entitled "Ondansetron decreases vomiting associated with acute gastroenteritis: a randomized

control trial" was rated "C" because the study population reported included only children who had been assigned to intravenous fluid therapy. The information may have been new and helpful in treatment of the disorder, but was not helpful in the decision called for in the query.

Limitations

As previously mentioned, the results of our exploration should be interpreted taking into account several limitations, including the modest size of the pool of judged documents and the number of questions. In addition, our findings pertain to answering therapy questions only. Answering questions about other clinical tasks could provide additional insights. Our evaluation is based on one expert opinion. Although our evaluator is a residency-trained and board-certified practicing family physician, his opinions most probably differ somewhat from the opinions of other family doctors. Because of the exploratory nature of this investigation we did not evaluate the spectrum of opinions of family practitioners regarding the relevancy of the citations in our test collection.

Implications and future work

This study presents some evidence that the burden of overcoming several of the major obstacles [1] in practicing evidence-based medicine could be alleviated by integrating into information retrieval systems the domain knowledge in the UMLS and the EBM principles. Unless connected to an electronic patient record, automatic methods cannot be used for the initial step of formulating an information need nor (under any circumstances) for the final steps of appraising the evidence and making a clinical decision. However,

automatic methods could address the challenging task of determining an optimal search strategy. A system might first provide the clinician with a pick-list for selecting question-type, for example, an overview of best available treatments for a given condition. The system could then use a predetermined optimal search strategy for the question type chosen.

Our study suggests several areas for further exploration. We are currently developing question templates for submitting therapy questions to our systems. We plan to expand these to accommodate other types of clinical questions, including those involving diagnosis, prognosis, and cost effectiveness. Uncertainty about finding all relevant evidence could be mitigated by using optimal recall-oriented strategies. Subsequently, the difficulty of synthesizing and appraising all evidence found could be addressed by presenting aggregated search results to the clinician (using SemRep summaries and patient-oriented outcomes extracted by CQA-1.0, for example).

Conclusion

We investigated three knowledge-based systems for assisting clinicians find answers to questions in MEDLINE. Although the number and range of clinical queries and citations retrieved are too small for any definitive conclusion, it appears that the semantic processing alone may be less helpful in finding relevant citations than the hybrid approach of combining topical semantically identified factors and non-subject features associated with MEDLINE citations. The Essie search engine performed significantly better than the baseline overall for “general” searches of “therapy for disease”. CQA-1.0 clinical question answering system performed significantly better than the baseline for “complex” queries involving population groups, outcomes, and comparison of

intervention. A fusion of the three approaches (SemRep, Essie, and CQA-1.0) outperformed the baseline and each approach taken separately for all types of questions. The significance of any of these methods to point-of-care decision support remains unknown, but the increasing ability to post-process MEDLINE citations to enable increasingly sophisticated methods for ranking retrieval for the clinician is promising.

Acknowledgment

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

1. Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, Pifer EA. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ. 2002 Mar 23;324(7339):710-6.
2. D'Alessandro DM, Kreiter CD, Peterson MW. An evaluation of information-seeking behaviors of general pediatricians. Pediatrics. 2004 Jan;113(1 Pt 1):64-9.
3. Booth A, O'Rourke AJ, Ford NJ. Structuring the pre-search reference interview: a useful technique for handling clinical questions. Bull Med Libr Assoc. 2000 Jul;88(3):239-46.
4. Fozi K, Teng CL, Krishnan R, Shajahan Y. A study of clinical questions in primary care. Med J Malaysia. 2000 Dec;55(4):486-92.
5. McKibbon KA, Fridsma DB. Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. J Am Med Inform Assoc. 2006 Nov-Dec;13(6):653-9.
6. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. 2nd Ed. Edinburgh: Churchill Livingstone, 2000.
7. Geissbuhler A, Miller RA. Clinical application of the UMLS in a computerized order entry and decision-support system. Proc AMIA Symp. 1998;:320-324.

8. Demner-Fushman D, Hauser SE, Humphrey SM, Ford GM, Jacobs JL, Thoma GR. MEDLINE as a Source of Just-in-Time Answers to Clinical Questions. Proc AMIA Symp. 2006;:190-194.
9. Hauser SE, Demner-Fushman D, Ford GM, Jacobs JL, Thoma GR. Preliminary Comparison of Three Search Engines for Point of Care Access to MEDLINE Citations. Proc AMIA Symp. 2006;945.
10. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.
11. Fiszman M, Rindflesch TC, Kilicoglu H. 2004. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics. 2004;76-83.
12. Demner-Fushman D, Lin J. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Computational Linguistics. 2007;33(1):63-104.
13. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1993 Dec 1;270(21):2598-601.
14. Haynes RB, Wilczynski N. Finding the gold in MEDLINE: clinical queries. ACP J Club. 2005 Jan-Feb;142(1):A8-9.
15. Ide NC, Loane RF, Demner-Fushman D. Essie: A Concept Based Search Engine for Structured Biomedical Text. J Am Med Inform Assoc. 2007 May-June;14(3):253-263
16. Harman, D. Evaluation techniques and measures. In Proceedings of the 4th Text REtrieval Conference (TREC-4). 1996;A6–A14.

17. Sneiderman C, Demner-Fushman D, Fiszman M, Rindflesch TC. Semantic Characteristics of MEDLINE Citations Useful for Therapeutic Decision-making. Proc AMIA Symp. 2005;1117.
18. Cucina RJ, Shah MK, Berrios DC, Fagan LM. Empirical formulation of a generic query set for clinical information retrieval systems. Medinfo. 2001;10(Pt 1):181-5.
19. Mendonça EA, Johnson SB, Seol Y, Cimino JJ. Analyzing the semantics of patient data to rank records of literature retrieval. Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. 2002;3:69-76.
20. Elhadad N, Kan MY, Klavans J, McKeown K. Customization in a unified framework for summarizing medical literature. Journal of Artificial Intelligence in Medicine. 2005;33(2):179–198.
21. Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. Proc AMIA Annu Fall Symp. 1997;485-9.
22. Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. Proc AMIA Symp. 2000;344-8.
23. Hersh W, Hickam DH, Haynes RB, McKibbon KA. Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. Proc Annu Symp Comput Appl Med Care. 1991;808-12.
24. Voorhees EM. Natural language processing and information retrieval. In: Pazienza MT, editor. Information Extraction: Towards scalable, adaptable systems. New York: Springer; 1999. p. 32-48.

25. Srinivasan P. Retrieval feedback in MEDLINE. J Am Med Inform Assoc. 1996 Mar-Apr;3(2):157-67.
26. Ruthven I, Lalmas M. A survey on the use of relevance feedback for information access systems Knowledge Engineering Review. 2003;18(2):p. 95-145.
27. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. J Am Med Inform Assoc. 1998 Jan-Feb;5(1):62-75.
28. Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. Proc AMIA Annu Fall Symp. 1997;528-32.
29. Rosenbloom ST, Geissbuhler AJ, Dupont WD, Giuse DA, Talbert DA, Tierney WM, Plummer WD, Stead WW, Miller RA. Effect of CPOE user interface design on user-initiated access to educational and patient information during clinical care. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):458-73.
30. Cimino JJ. Use, Usability, Usefulness, and Impact of an Infobutton Manager. Proc AMIA Symp. 2006;151-5.
31. Del Fiol G, Rocha RA, Clayton PD. Infobuttons at Intermountain Healthcare: Utilization and Infrastructure. Proc AMIA Symp. 2006;180-4.
32. Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. Medinfo. 2001;10(Pt 1):390-3.
33. Fiszman M, Rindflesch TC, Kilicoglu H. Integrating a Hypernymic Proposition Interpreter into a Semantic Processor for Biomedical Texts. Proc. AMIA Symposium. 2003 Nov;239-243.

34. Rindflesch TC , Fiszman M. The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. Journal of Biomedical Informatics. 2003;36(6):462-77.
35. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature, in Medical informatics: Advances in knowledge management and data mining in biomedicine, H. Chen, et al., Editors. 2005, Springer-Verlag.;399-422.
36. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994;235-9.
37. Smith L, Rindflesch T, and Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. Bioinformatics. 2004;20(14):2320-1.
38. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17-21.
39. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. Medinfo. 2001;10(Pt 1):216-20.
40. Hahn U, Mani I. The challenges of automatic summarization. Computer. 2000;33(11):29-36.
41. Hahn U, Reimer U. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In: Mani I and Maybury MT, eds. Advances in Automatic Text Summarization. Cambridge: MIT Press, 1999;215-232.

42. Richardson W, Wilson MC, Nishikawa J, Hayward RSA. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club. 1995;123:A-12.
43. Guyatt GH, Rennie D. Users' guides to the medical literature. JAMA. 1993 Nov 3;270(17):2096-7.
44. Hersh W, Bhupatiraju RT, Corley S. Enhancing access to the Bibliome: the TREC Genomics Track. Medinfo. 2004;11(Pt 2):773-7.
45. Hersh W, Cohen AM, Roberts P, Rekapalli HP, TREC 2006 Genomics Track overview, The Fifteenth Text Retrieval Conference - TREC 2006:68-87.
46. Family Physicians Inquiries Network [Online]. 2007 [cited 2007 Jun 1]; Available from: http://www.fpin.org
47. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, Stavri PZ. A taxonomy of generic clinical questions: classification study. BMJ. 2000 Aug 12;321(7258):429-32.
48. Ward D, Meadows SE, Nashelsky JE. The role of expert searching in the Family Physicians' Inquiries Network (FPIN). J Med Libr Assoc. 2005 January; 93(1): 88–96.
49. Sievert ME, McKinin EJ, Johnson ED, Reid JC, Mitchell JA. Beyond relevance--characteristics of key papers for clinicians: an exploratory study in an academic setting. Bull Med Libr Assoc. 1996 Jul;84(3):351-8.
50. Lin J, Demner-Fushman D. The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine. In: SIGIR '06: Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in

information Retrieval; 2006 August 06 – 11; Seattle, Washington, USA. ACM Press, New York, NY. 2006;99-106.

51. Demner-Fushman D, Hauser S, Thoma G. The Role of Title, Metadata and Abstract in Identifying Clinically Relevant Journal Articles. AMIA Annu Symp Proc. 2005;191-5.
52. Fox EA, Shaw JA. Combination of multiple searches. In Proceedings of the 2nd Text REtrieval Conference (TREC-2). 1994;243–252.
53. Text REtrieval Conference (TREC) [Online]. 2007 [cited 2007 Feb 15]; Available from: http://trec.nist.gov/trec_eval/
54. Siegel S, Castellan N. Nonparametric statistics for the behavioral sciences. 2nd Ed. New York: McGraw-Hill, 1988.
55. Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in information Retrieval; 2004 July 25-29; Sheffield, England. ACM Press, New York, NY. 2004;25-32.
56. Alper BS, White DS, Ge B. Physicians answer more clinical questions and change clinical decisions more often with synthesized evidence: a randomized trial in primary care. Ann Fam Med. 2005 Nov-Dec;3(6):507-13.
57. Voorhees E. The TREC Question Answering Track. Natural Language Engineering. 2001;7(4):361-378.
58. Family Medicine [Online]. 2007 [cited 2007 Jun 1]; Available from: http://www.aafp.org/online/en/home/policy/policies/f/familymedicine.html