Knowledge-based Methods to Help Clinicians Find Answers in MEDLINE

CHARLES A. SNEIDERMAN, MD, PHD, DINA DEMNER-FUSHMAN, MD, PHD, MARCELO FISZMAN, MD, PHD, NICHOLAS C. IDE, MS, THOMAS C. RINDFLESCH, PHD

Abstract

Objectives: 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 article 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 varying extents, were evaluated separately and in combination. Fifteen therapy and prevention questions in three categories (general, intermediate, and specific questions) were searched. The first 10 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, binary preference) and metrics evaluating the number of relevant documents in the first several presented to a physician were used.

Results: Scores (mean average precision = 0.57, binary preference = 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 to answer questions in clinical practice.

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 because of 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 of 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.4,5 Merely using electronic information resources of choice, physicians were not always successful in answering clinical questions.6 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: 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 postprocessing the search results in order to rerank 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)\(^\text{30}\) (alone or in combination with corpus-based methods) to find information for clinical queries of varying complexity.

The first system, SemRep Summarization,\(^\text{11}\) 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,\(^\text{12}\) uses EBM recommendations for finding best answers for questions about treatment and prevention.\(^\text{13}\) Both SemRep and CQA-1.0 rerank a set of MEDLINE citations retrieved using PubMed and clinical query filters.\(^\text{14}\) The third system, Essie,\(^\text{15}\) 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 medical subject headings (MeSH) 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.\(^\text{16}\) A rating scale developed to evaluate the utility of MEDLINE citations in clinical decision making\(^\text{17}\) 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\(^\text{18}\) to matching semantic representation of a patient’s clinical record with that of a MEDLINE citation\(^\text{19}\) and generating personalized summaries in response to physicians’ questions.\(^\text{20}\)

The most thoroughly researched approaches have been automatic query expansion, concept indexing, and retrieval-based feedback. The reported evaluations of these approaches have not shown consistent improvements in satisfying clinicians’ information needs. For example, automatic query expansion using controlled vocabulary terms was shown to improve overall average precision,\(^\text{21}\) but only one third of the queries showed improvement in a study of synonym- or hierarchical thesaurus-based query expansion.\(^\text{22}\) Concept indexing implemented in the SAPHIRE system helped physicians,\(^\text{23}\) whereas it degraded performance for 30 medical questions in a study of effectiveness of conceptual indexing.\(^\text{24}\) 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,\(^\text{25}\) but is also known to degrade performance.\(^\text{26}\)

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.\(^\text{27}\) The effect of unobtrusively providing context-specific links between clinical data and information resources in the form of “infobuttons”\(^\text{28}\) has been examined in several recent studies. Rosenbloom et al.\(^\text{29}\) 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.\(^\text{30}\) observed positive results and increased use of infobuttons over 5.7 years. 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.\(^\text{31}\) observed increased infobutton use with preference for secondary sources (such as Micromedex and UpToDate) that summarize the results of clinical studies and underutilization 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.\(^\text{32}\) 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 (etioloay, 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 reranking system considered in this study is based on SemRep Summarization,\(^\text{33}\) which depends on the semantic natural language processing system SemRep.\(^\text{34,35}\) SemRep identifies semantic predications (relationships) in
biomedical text using underspecified syntactic analysis and structured domain knowledge from the UMLS. SemRep predictions 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 and a part-of-speech tagger. MetaMap 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 prediction 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. Chemicals and Drugs, Disorders, and Population Groups.

SemRep Summarization is an automatic summarization system in the semantic abstraction paradigm. The system takes as input a list of predications extracted by SemRep from biomedical text (MEDLINE citations to be reranked 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. Predications in the summary are linked to the citations from which they were extracted and play an important role in exploiting SemRep Summarization for reranking retrieved citations in this study.

CQA-1.0
Another reranking method is implemented in the prototype clinical question-answering system CQA-1.0. In this system, questions and MEDLINE citations are represented using frames that 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, manually derived patterns and rules, and supervised machine learning techniques to identify the fundamental EBM components listed. The PICO framework (Problem/Patient, Intervention, Comparison, and Outcome) designed to help clinicians formulate clinical questions 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 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 subcategories. 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, 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 reranking 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. 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 one that contains this phrase in the abstract. Essie provides several possibilities for query expansion: exact match, SPECIALIST lexicon-based morphological expansion of terms, and UMLS-based expansion of concepts. Essie was the best-performing search engine in the 2003 TREC Genomics track and one of the best-performing systems in the 2006 TREC Genomics track.

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. 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 docu-
ments 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 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 to predict how the systems would perform in a clinical setting.

**Creating a Test Collection**

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 10 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. 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. 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 were used for reference while judging the abstracts selected for evaluation as described below.

**Assigning Relevance Judgments**

Relevance judgments were generated using the pooling strategy developed for TREC. The top 10 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 nontopical characteristics of key articles identified in Sievert et al. (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 Reranking 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 reranked for each query using SemRep Summarization and the CQA-1.0 system.

---

**Table 1 • Clinical Questions Used to Retrieve MEDLINE Citations**

<table>
<thead>
<tr>
<th>General Questions</th>
<th>Intermediate Questions</th>
<th>Specific Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is the most effective treatment for external genital warts?</td>
<td>1. What is the most effective treatment for ADHD in children?</td>
<td>1. Are antibiotics effective in preventing pneumonia for nursing home patients?</td>
</tr>
<tr>
<td>2. What are the most effective interventions to reduce childhood obesity?</td>
<td>2. Can type 2 diabetes be prevented through diet and exercise?</td>
<td>2. Is methylphenidate useful for treating adolescents with ADHD?</td>
</tr>
<tr>
<td>3. What is the most effective treatment for acute low back pain?</td>
<td>3. What are the best therapies for acute migraine in pregnancy?</td>
<td>3. What is the best treatment for gastroesophageal reflux and vomiting in infants?</td>
</tr>
</tbody>
</table>
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 1,305 documents for the first set, 925 for the second set, 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 reranking 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 reranking 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. In searching with Essie, a strategy similar to PubMed clinical queries, using EBM-related 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 that contain query phrases in the fields observed to be most informative, for example in the title. 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 Fox et al. The contribution of each system to the final ranking was weighted equally. 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). Because clinicians are willing to spend no more than 4 to 5 minutes evaluating search results, it is important that the answer to the question be found in the first few citations retrieved. Metrics that evaluate how soon a user

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, and unprocessed PubMed output (baseline). The trec_eval-8.0 package 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-grade 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 <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.

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. 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 nonrelevant 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. 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:** This measures precision after R documents have been retrieved, where R is the total number of relevant documents for a query.
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 10 retrieved documents (P@10):** Measures the fraction of relevant documents in the top 10 documents retrieved.

3. **Mean Reciprocal Rank (MRR):** The metric used in TREC question answering evaluation.\(^5\) 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.

In Table 3 through 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.

The baseline PubMed performance for intermediate questions (Table 4) was significantly better than 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).

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 three systems (MAP = 0.7839) is particularly successful for this class of questions.

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 10 documents, meaning that three to four of the first five documents retrieved by the evaluated systems (and six to eight of the first 10) 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 reranking; 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

---

<table>
<thead>
<tr>
<th>System</th>
<th>MAP</th>
<th>Bpref</th>
<th>R-prec</th>
<th>MRR</th>
<th>P@5</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>0.1410</td>
<td>0.5148</td>
<td>0.1185</td>
<td>0.6217</td>
<td>0.2400</td>
<td>0.1800</td>
</tr>
<tr>
<td>Essie</td>
<td>0.4403</td>
<td>0.6088</td>
<td>0.3775</td>
<td>1.0000</td>
<td>0.8800</td>
<td>0.8400</td>
</tr>
<tr>
<td>SemRep Summarization</td>
<td>0.2919</td>
<td>0.3703</td>
<td>0.3235</td>
<td>0.7667</td>
<td>0.8400</td>
<td>0.8400</td>
</tr>
<tr>
<td>CQA-1.0</td>
<td>0.4131</td>
<td>0.4686</td>
<td>0.4219</td>
<td>1.0000</td>
<td>0.8400</td>
<td>0.8800</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.4389</td>
<td>0.6897</td>
<td>0.4760</td>
<td>1.0000</td>
<td>0.7200</td>
<td>0.6600</td>
</tr>
</tbody>
</table>
reranking, 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 has identified nonspecific 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 10 documents, semantic reranking 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 reranking 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 U.S. 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 subpopulation to the population of interest.

Notes taken during evaluation identified additional nonspecific criteria used to assess usefulness of the citation to a clinician. The rater (who considers himself a typical primary care physician) evaluated the utility of a citation and used nontopical cues present in the citation, as well as “world knowledge.” For example, 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.

Citations that were judged to be B grade (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.

Citations that were judged to be C grade (not helpful in answering the clinical query) also involved 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 judged as A grade 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.

**Table 4 • Results for Intermediate Questions**

<table>
<thead>
<tr>
<th>System</th>
<th>MAP</th>
<th>Bpref</th>
<th>R-prec</th>
<th>MRR</th>
<th>P@5</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>0.3640</td>
<td>0.3996</td>
<td>0.3387</td>
<td>0.8400</td>
<td>0.6400</td>
<td>0.5600</td>
</tr>
<tr>
<td>Essie</td>
<td>0.2986</td>
<td>0.3707</td>
<td>0.3255</td>
<td>0.8000</td>
<td>0.6800</td>
<td>0.6400</td>
</tr>
<tr>
<td>SemRep</td>
<td>0.3624</td>
<td>0.3972</td>
<td>0.4162</td>
<td>1.0000</td>
<td>0.8000</td>
<td>0.7200</td>
</tr>
<tr>
<td>Summarization</td>
<td>CQA-1.0</td>
<td>0.4395</td>
<td>0.5468</td>
<td>0.4367</td>
<td>0.8667</td>
<td>0.7200</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.4897</td>
<td>0.6333</td>
<td>0.5423</td>
<td>0.7677</td>
<td>0.6000</td>
<td>0.4600</td>
</tr>
</tbody>
</table>

1See American Academy of Family Physicians Policy and Advocacy for one definition of a typical family doctor.

**Table 5 • Results for Specific (Complex) Questions**

<table>
<thead>
<tr>
<th>System</th>
<th>MAP</th>
<th>Bpref</th>
<th>R-prec</th>
<th>MRR</th>
<th>P@5</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>0.4125</td>
<td>0.2944</td>
<td>0.4527</td>
<td>0.8667</td>
<td>0.6400</td>
<td>0.5800</td>
</tr>
<tr>
<td>Essie</td>
<td>0.5356</td>
<td>0.6358</td>
<td>0.5394</td>
<td>1.0000</td>
<td>0.8800</td>
<td>0.8600</td>
</tr>
<tr>
<td>SemRep</td>
<td>0.2158</td>
<td>0.2338</td>
<td>0.2520</td>
<td>0.8000</td>
<td>0.7200</td>
<td>0.5000</td>
</tr>
<tr>
<td>Summarization</td>
<td>CQA-1.0</td>
<td>0.6286</td>
<td>0.7195</td>
<td>0.6008</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.6677</td>
<td>0.6572</td>
<td>0.5775</td>
<td>0.9000</td>
<td>0.8800</td>
<td>0.8200</td>
</tr>
<tr>
<td>Fusion 3</td>
<td>0.7839</td>
<td>0.8001</td>
<td>0.6866</td>
<td>1.0000</td>
<td>0.9200</td>
<td>0.8400</td>
</tr>
</tbody>
</table>
though the number and range of clinical queries and citations retrieved are too small for any definitive conclusion, it seems that the semantic processing alone may be less helpful in finding relevant citations than the hybrid approach of combining topical semantically identified factors and nonsubject features associated with MEDLINE citations. The Essie search engine performed significantly better than the baseline overall for general searches of therapy for disease. The 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 postprocess MEDLINE citations to enable increasingly sophisticated methods for ranking retrieval for the clinician is promising.

References


32. Fisman M, Rindflesch TC, Kilicoglu H. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. Proc AMIA Symp 2003;239–43.


