Answering Clinical Questions with Knowledge-Based and Statistical Techniques

Dina Demner-Fushman*
University of Maryland, College Park

Jimmy Lin†
University of Maryland, College Park

The combination of recent developments in question-answering research and the availability of unparalleled resources developed specifically for automatic semantic processing of text in the medical domain provides a unique opportunity to explore complex question answering in the domain of clinical medicine. This article presents a system designed to satisfy the information needs of physicians practicing evidence-based medicine. We have developed a series of knowledge extractors, which employ a combination of knowledge-based and statistical techniques, for automatically identifying clinically relevant aspects of MEDLINE abstracts. These extracted elements serve as the input to an algorithm that scores the relevance of citations with respect to structured representations of information needs, in accordance with the principles of evidence-based medicine. Starting with an initial list of citations retrieved by PubMed, our system can bring relevant abstracts into higher ranking positions, and from these abstracts generate responses that directly answer physicians' questions. We describe three separate evaluations: one focused on the accuracy of the knowledge extractors, one conceptualized as a document reranking task, and finally, an evaluation of answers by two physicians. Experiments on a collection of real-world clinical questions show that our approach significantly outperforms the already competitive PubMed baseline.

1. Introduction

Recently, the focus of question-answering research has shifted away from simple fact-based questions that can be answered with relatively little linguistic knowledge to "harder" questions that require fine-grained text analysis, reasoning capabilities, and the ability to synthesize information from multiple sources. General purpose reasoning on anything other than superficial lexical relations is exceedingly difficult because there is a vast amount of world knowledge that must be encoded, either manually or automatically, to overcome the brittleness often associated with long chains of evidence. This situation poses a serious bottleneck to "advanced" question-answering systems. However, the availability of existing knowledge sources and ontologies in certain domains provides exciting opportunities to experiment with knowledge-rich approaches. How might one go about leveraging these resources effectively? How might one integrate

* Department of Computer Science and Institute for Advanced Computer Studies. E-mail: demner@umd.edu.
† College of Information Studies, Department of Computer Science, and Institute for Advanced Computer Studies. E-mail: jimmylin@umd.edu.

Submission received: 4 July 2005; revised submission received: 7 January 2006; accepted for publication: 12 April 2006.

statistical techniques to overcome the brittleness often associated with knowledge-based approaches?

We explore these interesting research questions in the domain of medicine, focusing on the information needs of physicians in clinical settings. This domain is well-suited for exploring the posed research questions for several reasons. First, substantial understanding of the domain has already been codified in the Unified Medical Language System (UMLS) (Lindberg, Humphreys, and McCray 1993). Second, software for utilizing this ontology already exists: MetaMap (Aronson 2001) identifies concepts in free text, and SemRep (Rindflesch and Fiszman 2003) extracts relations between the concepts. Both systems utilize and propagate semantic information from UMLS knowledge sources: the Metathesaurus, the Semantic Network, and the SPECIALIST lexicon. The 2004 version of the UMLS Metathesaurus (used in this work) contains information about over 1 million biomedical concepts and 5 million concept names from more than 100 controlled vocabularies. The Semantic Network provides a consistent categorization of all concepts represented in the UMLS Metathesaurus. Third, the paradigm of evidence-based medicine (Sackett et al. 2000) provides a task-based model of the clinical information-seeking process. The PICO framework (Richardson et al. 1995) for capturing well-formulated clinical queries (described in Section 2) can serve as the basis of a knowledge representation that bridges the needs of clinicians and analytical capabilities of a system. The confluence of these many factors makes clinical question answering a very exciting area of research.

Furthermore, the need to answer questions related to patient care at the point of service has been well studied and documented (Covell, Uman, and Manning 1985; Gorman, Ash, and Wykoff 1994; Ely et al. 1999, 2005). MEDLINE, the authoritative repository of abstracts from the medical and biomedical primary literature maintained by the National Library of Medicine, provides the clinically relevant sources for answering physicians' questions, and is commonly used in that capacity (Cogdill and Moore 1997; De Groote and Dorsch 2003). However, studies have shown that existing systems for searching MEDLINE (such as PubMed, the search service provided by the National Library of Medicine) are often inadequate and unable to supply clinically relevant answers in a timely manner (Gorman, Ash, and Wykoff 1994; Chambliss and Conley 1996). Furthermore, it is clear that traditional document retrieval technology applied to MEDLINE abstracts is insufficient for satisfactory information access; research and experience point to the need for systems that automatically analyze text and return only the relevant information, appropriately summarizing and fusing segments from multiple texts. Not only is clinical question answering interesting from a research perspective, it also represents a potentially high-impact, real-world application of language processing and information retrieval technology—better information systems to provide decision support for physicians have the potential to improve the quality of health care.

Our question-answering system supports the practice of evidence-based medicine (EBM), a widely accepted paradigm for medical practice that stresses the importance of evidence from patient-centered clinical research in the health care process. EBM prescribes an approach to structuring clinical information needs and identifies elements (for example, the problem at hand and the interventions under consideration) that factor into the assessment of clinically relevant studies for medical practice. The foundation of our question-answering strategy is built on knowledge extractors that automatically identify these elements in MEDLINE abstracts. Using these knowledge extractors, we have developed algorithms for scoring the relevance

of MEDLINE citations in accordance with the principles of EBM. Our scorer is employed to rerank citations retrieved by the PubMed search engine, with the goal of bringing as many topically relevant abstracts to higher ranking positions as possible. From this reranked list of citations, our system is then able to generate textual responses that directly address physicians' information needs. We evaluated our system with a collection of real-world clinical questions and demonstrate that our combined knowledge-based and statistical approach delivers significantly better document retrieval and question-answering performance, compared to systems used by physicians today.

This article is organized in the following manner: We start in the next section with an overview of evidence-based medicine and its basic principles. Section 3 provides an overview of MEDLINE, the bibliographic database used by our system, and PubMed, the public gateway for accessing this database. Section 4 describes our system architecture and outlines our conception of clinical question answering as "semantic unification" between query frames and knowledge frames derived from MEDLINE citations. The knowledge extractors that underlie our approach are described in Section 5, along with intrinsic evaluations of each component. In Section 6, we detail an algorithm for scoring the relevance of MEDLINE citations with respect to structured query representations. This scoring algorithm captures the principles of EBM and uses the results of the knowledge extractors as basic features. To evaluate the performance of this citation scoring algorithm, we have gathered a corpus of real-world clinical questions. Section 7 presents results from a document reranking experiment where our EBM scores were used to rerank citations retrieved by PubMed. Section 8 provides additional details on attempts to optimize the performance of our EBM citation scoring algorithm. Answer generation, based on reranked results, is described in Section 9. Answers from our system were manually assessed by two physicians; results are presented in Section 10. Related work is discussed in Section 11, followed by future work in Section 12. Finally, we conclude in Section 13.

2. The Framework of Evidence-Based Medicine

Evidence-based medicine (EBM) is a widely accepted paradigm for medical practice that involves the explicit use of current best evidence, that is, high-quality patient-centered clinical research such as reports from randomized controlled trials, in making decisions about patient care. Naturally, such evidence, as reported in the primary medical literature, must be suitably integrated with the physician's own expertise and patient-specific factors. It is argued by many that practicing medicine in this manner leads to better patient outcomes and higher quality health care. The goal of our work is to develop question-answering techniques that complement this paradigm of medical practice.

EBM offers three orthogonal facets that, when taken together, provide a framework for codifying the knowledge involved in answering clinical questions. These three complementary facets are outlined below.

The first facet describes the four main clinical tasks that physicians engage in (arranged roughly in order of prevalence):

Therapy: Selecting treatments to offer a patient, taking into account effectiveness, risk, cost, and other relevant factors (includes **Prevention**—selecting actions to reduce the chance of a disease by identifying and modifying risk factors).

Diagnosis: This encompasses two primary types:

Differential diagnosis: Identifying and ranking by likelihood potential diseases based on findings observed in a patient.

Diagnostic test: Selecting and interpreting diagnostic tests for a patient, considering their precision, accuracy, acceptability, cost, and safety.

Etiology/Harm: Identifying factors that cause a disease or condition in a patient.

Prognosis: Estimating a patient's likely course over time and anticipating likely complications.

These activities represent what Ingwersen (1999) calls "work tasks." It is important to note that they exist independently of information needs, namely, searching is not necessarily implicated in any of these activities. We are, however, interested in situations where questions arise during one of these clinical tasks—only then does the physician engage in information-seeking behavior. These activities translate into natural "search tasks." For therapy, the search task is usually *therapy selection* (for example, determining which course of action is the best treatment for a disease) or *prevention* (for example, selecting preemptive measures with respect to a particular disease). For diagnosis, there are two different possibilities: in *differential diagnosis*, a physician is considering multiple hypotheses regarding what disease a patient has; in *diagnostic methods selection*, the clinician is attempting to ascertain the relative utility of different tests. For etiology, *cause determination* is the search task, and for prognosis, *patient outcome prediction*.

Terms and the types of studies relevant to each of the four tasks have been extensively studied by the Hedges Project at the McMaster University (Haynes et al. 1994; Wilczynski, McKibbon, and Haynes 2001). The results of this research are implemented in the PubMed Clinical Queries tools, which can be used to retrieve task-specific citations (more about this in the next section).

The second facet is independent of the clinical task and pertains to the structure of a well-built clinical question. The following four components have been identified as the key elements of a question related to patient care (Richardson et al. 1995):

- What is the primary problem or disease? What are the characteristics of the patient (e.g., age, gender, or co-existing conditions)?
- What is the main intervention (e.g., a diagnostic test, medication, or therapeutic procedure)?
- What is the main intervention compared to (e.g., no intervention, another drug, another therapeutic procedure, or a placebo)?
- What is the desired effect of the intervention (e.g., cure a disease, relieve or eliminate symptoms, reduce side effects, or lower cost)?

These four elements are often referenced with the mnemonic PICO, which stands for Patient/Problem, Intervention, Comparison, and Outcome.

Finally, the third facet serves as a tool for appraising the strength of evidence presented in the study, that is, how much confidence should a physician have in the results? Several taxonomies for appraising the strength of evidence based on the type and quality of the study have been developed. We chose the Strength of Recommendations Taxonomy (SORT) as the basis for determining the potential upper bound on the

quality of evidence, due to its emphasis on the use of patient-oriented outcomes and its attempt to unify other existing taxonomies (Ebell et al. 2004). There are three levels of recommendations according to SORT:

- **A-level evidence** is based on consistent, good-quality patient outcome-oriented evidence presented in systematic reviews, randomized controlled clinical trials, cohort studies, and meta-analyses.
- **B-level evidence** is inconsistent, limited-quality, patient-oriented evidence in the same types of studies.
- **C-level evidence** is based on disease-oriented evidence or studies less rigorous than randomized controlled clinical trials, cohort studies, systematic reviews, and meta-analyses.

A question-answering system designed to support the practice of evidence-based medicine must be sensitive to the multifaceted considerations that go into evaluating an abstract's relevance to a clinical information need. It is exactly these three complementary facets that we attempt to encode in a question-answering system for clinical decision support.

3. MEDLINE and PubMed

MEDLINE is a large bibliographic database maintained by the U.S. National Library of Medicine (NLM). This database is viewed by medical professionals, biomedical researchers, and many other users as the authoritative source of clinical evidence, and hence we have adopted it as the target corpus for our clinical question-answering system. MEDLINE contains over 15 million references to articles from approximately 4,800 journals in 30 languages, dating back to the 1960s. In 2004, over 571,000 new citations were added to the database, and it continues to grow at a steady pace. The subject scope of MEDLINE is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering needed by health professionals and others engaged in basic research and clinical care, public health, health policy development, or related educational activities. MEDLINE also covers life sciences vital to biomedical practitioners, researchers, and educators, including aspects of biology, environmental science, marine biology, plant and animal science, as well as biophysics and chemistry.[1]

Each MEDLINE citation includes basic information such as the title of the article, name of the authors, name of the publication, publication type, date of publication, language, and so on. Of the entries added over the last decade or so, approximately 76% have English abstracts written by the authors of the articles—these texts provide the source for answers extracted by our system.

Additional metadata are associated with each MEDLINE citation. The most important of these is the controlled vocabulary terms assigned by human indexers. NLM's controlled vocabulary thesaurus, Medical Subject Headings (MeSH),[2] contains approximately 23,000 descriptors arranged in a hierarchical structure and more than 151,000 Supplementary Concept Records (additional chemical substance names) within a

1 http://www.nlm.nih.gov/pubs/factsheets/medline.html
2 Commonly referred to as MeSH terms or MeSH headings, although technically the latter is redundant.

separate thesaurus. Indexing is performed by approximately 100 indexers with at least bachelor's degrees in life sciences and formal training in indexing provided by NLM. Since mid-2002, the Library has been employing software that automatically suggests MeSH headings based on content (Aronson et al. 2004). Nevertheless, the indexing process remains firmly human-centered.

As a concrete example, an abstract titled "Antipyretic efficacy of ibuprofen vs. acetaminophen" might have the following MeSH headings associated with it:

MH - Acetaminophen/*therapeutic use
MH - Child
MH - Comparative Study
MH - Fever/*drug therapy
MH - Ibuprofen/*therapeutic use

To represent different aspects of the topic described by a particular MeSH heading, up to three subheadings may be assigned, as indicated by the slash notation. In this example, a trained user could interpret from the MeSH terms that the article is about drug therapy for fever and the therapeutic use of ibuprofen and acetaminophen. An asterisk placed next to a MeSH heading indicates that the human indexer interprets the term to be the main focus of the article. Multiple MeSH terms can be notated in this manner.

MEDLINE is publicly accessible on the World Wide Web through PubMed, the National Library of Medicine's gateway, or through third-party organizations that license MEDLINE from NLM. PubMed is a sophisticated boolean search engine that allows users to query not only on abstract text, but also on metadata fields such as MeSH terms. In addition, PubMed provides a number of pre-defined "search templates" called Clinical Queries (Haynes et al. 1994; Wilczynski, McKibbon, and Haynes 2001) that allow users to narrow the scope of retrieved articles. These filters are implemented as fixed boolean query fragments (containing restrictions on MeSH terms, for example) that are appended to the original user query. Our experiments involve the use of PubMed to retrieve an initial set of candidate citations for subsequent processing.

4. System Architecture

We view clinical question answering as "semantic unification" between information needs expressed in a PICO-based frame and corresponding structures automatically extracted from MEDLINE citations. In accordance with the principles of EBM, this matching process should be sensitive to the nature of the clinical task and the strength of evidence of retrieved abstracts.

As a concrete example, consider the following clinical question:

In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?

The information need might be formally encoded in the following manner:

Search Task: therapy selection
Problem/Population: acute febrile illness/in children
Intervention: acetaminophen
Comparison: ibuprofen
Outcome: reducing fever

This query representation explicitly encodes the search task and the PICO structure of the clinical question. After processing MEDLINE citations, automatically extracting PICO elements from the abstracts, and semantically matching these elements with the query, a system might produce the following answer:

> Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses.
>
> PMID: 1621668
> Strength of Evidence: grade A

Physicians are usually most interested in outcome statements that assert a patient-oriented clinical finding—for example, the relative efficacy of two drugs. Thus, outcomes can serve as the basis for good answers and an entry point into the full text. The system should automatically evaluate the strength of evidence of the citations supplying the answer, but the decision to adopt the recommendations as suggested ultimately rests with the physician.

What is the best input to a clinical question-answering system? Two possibilities include a natural language question or a structured PICO query frame. We advocate the latter. With a frame-based query interface, the physician shoulders the burden of translating an information need into a frame-based representation, but this provides several advantages. Most importantly, formal representations force physicians to "think through" their questions, ensuring that relevant elements are captured. Poorly formulated queries have been identified by Ely et al. (2005) as one of the obstacles to finding answers to clinical questions. Because well-formed questions should have concretely instantiated PICO slots, a frame representation clearly lets the physician see missing elements. In addition, a structured query representation obviates the need for linguistic analysis of a natural language question, where ambiguities may negatively impact overall performance. We discuss alternative interfaces in Section 12.

Ideally, we would like to match structured representations derived from the question with those derived from MEDLINE citations (taking into consideration other EBM-relevant factors). However, we do not have access to the computational resources necessary to apply knowledge extractors to the 15 million plus citations in the MEDLINE database and directly index their results. As an alternative, we rely on PubMed to retrieve an initial set of hits that we then postprocess in greater detail—this is the standard pipeline architecture commonly employed in other question-answering systems (Voorhees and Tice 1999; Hirschman and Gaizauskas 2001).

The architecture of our system is shown in Figure 1. The query formulator is responsible for converting a clinical question (in the form of a query frame) into a PubMed search query. Presently, these queries are already encoded in our test collection (see Section 6). PubMed returns an initial list of MEDLINE citations, which is then analyzed by our knowledge extractors (see Section 5). The input to the semantic matcher, which implements our EBM citation scoring algorithm, is the query frame and annotated MEDLINE citations. The module outputs a ranked list of citations that have been scored in accordance with the principles of EBM (see Section 6). Finally, the answer generator takes these citations and extracts appropriate answers (see Section 9).

In summary, our conception of clinical question answering as semantic frame matching suggests the need for a number of capabilities, which correspond to the bold outlined boxes in Figure 1: knowledge extraction, semantic matching for scoring

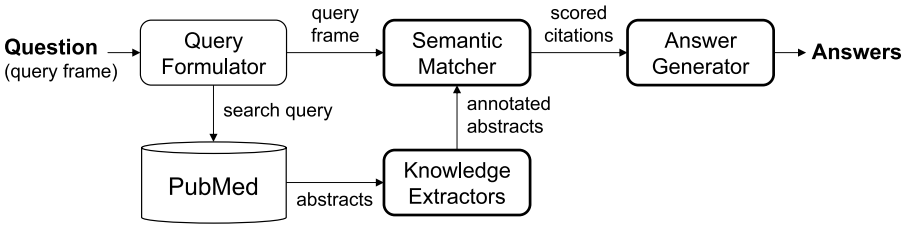

Figure 1
Architecture of our clinical question-answering system.

citations, and answer generation. We have realized all three capabilities in an implemented clinical question-answering system and conducted three separate evaluations to assess the effectiveness of our developed capabilities. We do not tackle the query formulator, although see discussion in Section 12. Overall, results indicate that our implemented system significantly outperforms the PubMed baseline.

5. Knowledge Extraction for Evidence-Based Medicine

The automatic extraction of PICO elements from MEDLINE citations represents a key capability integral to clinical question answering. This section, which elaborates on preliminary results reported in Demner-Fushman and Lin (2005), describes extraction algorithms for population, problems, interventions, outcomes, and the strength of evidence. For an example of a completely annotated abstract, see Figure 2. Each individual PICO extractor takes as input the abstract text of a MEDLINE citation and identifies the relevant elements: Outcomes are complete sentences, while population, problems, and interventions are short noun phrases.

Our knowledge extractors rely extensively on MetaMap (Aronson 2001), a system for identifying segments of text that correspond to concepts in the UMLS Metathesaurus. Many of our algorithms operate at the level of coarser-grained semantic types called Semantic Groups (McCray, Burgun, and Bodenreider 2001), which capture higher-level generalizations about entities (e.g., CHEMICALS & DRUGS). An additional feature we take advantage of (when present) is explicit section markers present in some abstracts. These so-called structured abstracts were recommended by the Ad Hoc Working Group for Critical Appraisal of the Medical Literature (1987) to help humans assess the reliability and content of a publication and to facilitate the indexing and retrieval processes. These abstracts loosely adhere to the introduction, methods, results, and conclusions format common in scientific writing, and delineate a study using explicitly marked sections with variations of the above headings. Although many core clinical journals require structured abstracts, there is a great deal of variation in the actual headings. Even when present, the headings are not organized in a manner focused on patient care. In addition, abstracts of much high-quality work remain unstructured. For these reasons, explicit section markers are not entirely reliable indicators for the various semantic elements we seek to extract, but must be considered along with other sources of evidence.

The extraction of each PICO element relies to a different extent on an annotated corpus of MEDLINE abstracts, created through an effort led by the first author at the National Library of Medicine (Demner-Fushman et al. 2006). As will be described herein, the population, problem, and the intervention extractors are based largely on recognition of semantic types and a few manually constructed rules; the outcome extrac-

tor, in contrast, is implemented as an ensemble of classifiers trained using supervised machine learning techniques (Demner-Fushman et al. 2006). These two very different approaches can be attributed to differences in the nature of the frame elements: Whereas problems and interventions can be directly mapped to UMLS concepts, and populations easily mapped to patterns that include UMLS concepts, outcome statements follow no predictable pattern. The initial goal of the annotation effort was to identify outcome statements in abstract text. A physician, two registered nurses, and an engineering researcher manually identified sentences that describe outcomes in 633 MEDLINE abstracts; a post hoc analysis demonstrates good agreement ($\kappa = 0.77$). The annotated abstracts were retrieved using PubMed and attempted to model different user behaviors ranging from naive to expert (where advanced search features were employed). With the exception of 50 citations retrieved to answer a question about childhood immunization, the rest of the results were retrieved by querying on a disease, for example, diabetes. Of the 633 citations, 100 abstracts were also fully annotated with population, problems, and interventions. These 100 abstracts were set aside as a held-out test set. Of the remaining citations, 275 were used for training and rule derivation, as described in the following sections.

After much exploration, Demner-Fushman et al. (2006) discovered that it was not practical to annotate PICO entities at the phrase level due to significant unresolvable disagreement and interannotator reliability issues. Consider the following segment:

> This double-blind, placebo-controlled, randomized, 3-period, complete block, 6-week crossover study examined the efficacy of simvastatin in adult men and women (N = 151) with stable type 2 DM, low density lipoprotein-cholesterol 100 mg/dL, HDL-C $<$ 40 mg/dL, and fasting triglyceride level $>$ 150 and $<$ 700 mg/dL.

All annotators agreed that the sentence contained the problem, population, and intervention. However, they could not agree on the exact phrasal boundaries of each element, and more importantly, general guidelines for ensuring consistent annotations. For example, should the whole clause starting with *adult men and women* be marked as population, or should type 2 Diabetes Mellitus (*type 2 DM*) be marked up only as the problem? How should we indicate that the cholesterol levels description belongs to 151 subjects of the study, and so forth? This issue becomes important for evaluation because there is a mismatch between annotated ground truth and the output of our knowledge extractors, as we will discuss.

In what follows, we describe each of the individual PICO extractors and a series of component evaluations that assess their accuracy. This section is organized such that the description of each extractor and its evaluation are paired together. Results are reported in terms of the percentage of correctly identified instances, percentage of instances for which the extractor had no answer, and percentage of incorrectly identified instances. The baselines and gold standards for each extractor vary, and will be described individually. The goal of these component evaluations is a general characterization of performance, as we focused the majority of our efforts on the two other evaluations.

5.1 Population Extractor

The PICO framework makes no distinction between the population and the problem, which is rooted in the concept of the population in clinical studies, as exemplified by text such as *POPULATION: Fifty-five postmenopausal women with a urodynamic diagnosis of genuine urinary stress incontinence.* Although this fragment simultaneously describes

the population (of which a particular patient can be viewed as a sample therefrom) and the problem, we chose to separate the extraction of the two elements because they are not always specified together in abstracts (issues with respect to exact boundaries mentioned previously notwithstanding). Furthermore, many clinical questions ask about a particular problem without specifying a population.

Population elements, which are typically noun phrases, are identified using a series of manually crafted rules that codify the following assumptions:

- The concept describing the population belongs to the semantic type GROUP or any of its children. In addition, certain nouns are often used to describe study participants in medical texts; for example, an often observed pattern is "subjects" or "cases" followed by a concept from the semantic group DISORDER.
- The number of subjects that participated in the study often precedes or follows a concept identified as a GROUP. In the latter case, the number is sometimes given in parentheses using a common pattern n = *number*, where "n = " is a shorthand for the number of subjects, and *number* provides the actual number of study participants.
- The confidence that a clause with an identified number and GROUP contains information about the population is inversely proportional to the distance between the two entities.
- The confidence that a clause contains the population is influenced by the position of the clause, with respect to headings in the case of structured abstracts and with respect to the beginning of the abstract in the case of unstructured abstracts.

Given these assumptions, the population extractor searches for the following patterns:

- GROUP ([Nn]=[0–9]+)
  for example, *in 5–6-year-old French children (n = 234), Subjects (n = 54)*
- *number** GROUP
  for example, *forty-nine infants*
- *number** DISORDER* GROUP?
  for example, *44 HIV-infected children*

The confidence score assigned to a particular pattern match is a function of both its position in the abstract and its position in the clause from which it was extracted. If a number is followed by a measure, for example, *year* or *percent*, the number is discarded, and pattern matching continues. After the entire abstract is processed in this manner, the match with the highest confidence value is retained as the population description.

5.2 Evaluation of Population Extractor

Ninety of the 100 fully annotated abstracts in our collection were agreed upon by the annotators as being clinical in nature, and were used as test data for our population extractor. Because these abstracts were not examined in the process of developing the extractor rules, they can be viewed as a blind held-out test set. The output of our popu-

Table 1
Evaluation of the population extractor.

	Correct (%)	Unknown (%)	Wrong (%)
Baseline	53	–	47
Extractor	80	10	10

lation extractor was judged to be correct if it occurred in a sentence that was annotated as containing the population in the gold standard. Note that this evaluation presents an upper bound on the performance of the population extractor, whose outputs are noun phrases. We adopted such a lenient evaluation setup because of the boundary issues previously discussed, and also to forestall potential difficulties with scoring partially overlapping string matches.

For comparison, our baseline simply returned the first three sentences of the abstract. We considered the baseline correct if any one of the sentences were annotated as containing the population in the gold standard (an even more lenient criterion). This baseline was motivated by the observation that the aim and methods sections of structured abstracts are likely to contain the population information—for structured abstracts, explicit headings provide structural cues; for unstructured abstracts, positional information serves as a surrogate.

The performance of the population extractor is shown in Table 1. A manual error analysis revealed three sources of error: First, not all population descriptions contain a number explicitly, for example, *The medical charts of all patients who were treated with etanercept for back or neck pain at a single private medical clinic in 2003.* Second, not all study populations are population groups, as for example in *All primary care trusts in England.* Finally, tagging and chunking errors propagate to the semantic type assignment level and affect the quality of MetaMap output. For example, consider the following sentence:

> We have compared the LD and recombination patterns defined by single-nucleotide polymorphisms in ENCODE region ENm010, chromosome 7p15 2, in Korean, Japanese, and Chinese samples.

Both *Korean* and *Japanese* were mistagged as nouns, which lead to the following erroneous chunking:

> [We] [have] [compared] [the LD] [and] [recombination patterns] [defined] [by single-nucleotide polymorphisms] [in] [ENCODE] [region ENm010,] [chromosome 7p15 2,] [in Korean,] [Japanese,] [and] [Chinese samples.]

This resulted in the tagging of Japanese as a population. Errors of this type affect other extractors as well. For example, *lead* was mistagged as a noun in the phrase *Echocardiographic findings lead to the right diagnosis,* which caused MetaMap to identify the word as a PHARMACOLOGICAL SUBSTANCE (lead is sometimes used as a homeopathic preparation).

5.3 Problem Extractor

The problem extractor relies on the recognition of concepts belonging to the UMLS semantic group DISORDER. In short, it returns a ranked list of all such concepts within a given span of text. We evaluate the performance of this simple heuristic on segments

Table 2
Evaluation of the problem extractor.

	Correct (%)	Unknown (%)	Wrong (%)
Abstract title	85	10	5
Title + 1st two sentences	90	5	5
Entire abstract	86	2	12

of the abstract varying in length: abstract title only, abstract title and first two sentences, and entire abstract text. Concepts in the title, in the introduction section of structured abstracts, or in the first two sentences in unstructured abstracts, are given higher confidence values due to their discourse prominence. Finally, the highest-scoring problem is designated as the primary problem in order to differentiate it from co-occurring conditions identified in the abstract.

5.4 Evaluation of Problem Extractor

Although our problem extractor returns a list of clinical problems, we only evaluate performance on identification of the primary problem. For some abstracts, MeSH headings can be used as ground truth, because one of the human indexers' tasks in assigning terms is to identify the main topic of the article (sometimes a disorder). For this evaluation, we randomly selected 50 abstracts with disorders indexed as the main topic from abstracts retrieved using PubMed on the five clinical questions described in Sneiderman et al. (2005).

We applied our problem extractor on different segments of the abstract: the title only, the title and first two sentences, and the entire abstract. These results are shown in Table 2. Here, a problem was considered correctly identified only if it shared the same concept ID as the ground truth problem (from the MeSH heading). The performance of our best variant (abstract title and first two sentences) approaches the upper bound on MetaMap performance—which is limited by human agreement on the identification of semantic concepts in medical texts, as established in Pratt and Yetisgen-Yildiz (2003).

Although problem extraction largely depends on disease coverage in UMLS and MetaMap performance, the error rate could be further reduced by more sophisticated recognition of implicitly stated problems. For example, with respect to a question about immunization in children, an abstract about the measles-mumps-rubella vaccination never mentioned the disease without the word *vaccination;* hence, no concept of the type DISEASE OR SYNDROME was identified.

5.5 Intervention Extractor

The intervention extractor identifies both the intervention and comparison elements in a PICO frame; processing of these two frame elements can be collapsed because they belong to the same semantic group. In many abstracts, it is unclear which intervention is the primary one and which are the comparisons, and hence our extractor simply returns a list of all interventions under study.

For interventions, we are primarily interested in entities that may participate in the UMLS Semantic Network relations associated with each clinical task. Restrictions on the semantic types allowed in these relations prescribe the set of possible clinical interventions. For therapy these relations include *treats*, *prevents*, and *carries out*; *diagnoses*

Table 3
Evaluation of the intervention extractor.

	Correct (%)	Unknown (%)	Wrong (%)
Baseline	60	–	40
Extractor	80	–	20

for diagnosis; *causes* and *result of* for etiology; and *prevents* for prognosis. At present, the identification of nine semantic types, for example, DIAGNOSTIC PROCEDURE, CLINICAL DRUG, and HEALTH CARE ACTIVITY, serves as the foundation for our intervention extraction algorithm.

Candidate scores are further adjusted to reflect a few different factors. In structured abstracts, concepts of the relevant semantic type are given additional weight if they appear in the title, aims, and methods sections. In unstructured abstracts, concepts towards the beginning of the abstract text are favored. Finally, the intervention extractor takes into account the presence of certain cue phrases that describe the aim and/or methods of the study, such as *This study examines* or *This paper describes.*

5.6 Evaluation of Intervention Extractor

The intervention extractor was evaluated in the same manner as the population extractor and compared to the same baseline. To iterate, 90 held-out clinical abstracts that contained human-annotated interventions served as ground truth. The output of our intervention extractor was judged to be correct if it occurred in a sentence that was annotated as containing the intervention in the gold standard. As with the evaluation of the population extractor, this represents an upper bound on performance. Results are shown in Table 3.

Some of the errors were caused by ambiguity of terms. For example, in the clause *serum levels of anti-HBsAg and presence of autoantibodies (ANA, ENA) were evaluated, serum* is recognized as a TISSUE, *levels* as INTELLECTUAL PRODUCT, and *autoantibodies* and ANA as IMMUNOLOGIC FACTORS. In this case, however, autoantibodies should be considered a LABORATORY OR TEST RESULT.[3] In other cases, extraction errors were caused by summary sentences that were very similar to intervention statements, for example, *This study compared the effects of 52 weeks' treatment with pioglitazone, a thiazolidinedione that reduces insulin resistance, and glibenclamide, on insulin sensitivity, glycaemic control, and lipids in patients with Type 2 diabetes.* For this particular abstract, the correct interventions are contained in the sentence *Patients with Type 2 diabetes were randomized to receive either pioglitazone (initially 30 mg QD, n = 91) or micronized glibenclamide (initially 1.75 mg QD, n = 109) as monotherapy.*

5.7 Outcome Extractor

We approached outcome extraction as a classification problem at the sentence level, that is, the outcome extractor assigns a probability of being an outcome to each sentence in an abstract. Our preliminary work has led to a strategy based on an ensemble of classifiers, which includes a rule-based classifier, a unigram "bag of words" classifier,

3 MetaMap does provide alternative mappings, but the current extractor only considers the best candidate.

an *n*-gram classifier, a position classifier, an abstract length classifier, and a semantic classifier. With the exception of the rule-based classifier, all classifiers were trained on the 275 citations from the annotated collection of abstracts described previously.

Knowledge for the rule-based classifier was hand-coded, prior to the annotation effort, by a registered nurse with 20 years of clinical experience. This classifier estimates the likelihood that a sentence states an outcome based on cue phrases such as *significantly greater, well tolerated,* and *adverse events.* The likelihood of a sentence being an outcome as indicated by cue phrases is the ratio of the cumulative score for recognized phrases to the maximum possible score. For example, the sentence *The dropout rate due to adverse events was 12.4% in the moxonidine and 9.8% in the nitrendipine group* is segmented into eight phrases by MetaMap, which sets the maximum score to 8. The two phrases *dropout rate* and *adverse events* contribute one point each to the cumulative score, which results in a likelihood estimate of 0.25 for this sentence.

The unigram "bag of words" classifier is a naive Bayes classifier implemented with the API provided by the MALLET toolkit.[4] This classifier outputs the probability of a class assignment.

The *n*-gram based classifier is also a naive Bayes classifier, but it operates on a different set of features. We first identified the most informative unigrams and bigrams using the information gain measure (Yang and Pedersen 1997), and then selected only the positive outcome predictors using odds ratio (Mladenic and Grobelnik 1999). Disease-specific terms, such as rheumatoid arthritis, were then manually removed. Finally, the list of features was revised by the registered nurse who participated in the annotation effort. This classifier also outputs the probability of a class assignment.

The position classifier returns the maximum likelihood estimate that a sentence is an outcome based on its position in the abstract (for structured abstracts, with respect to the results or conclusions sections; for unstructured abstracts, with respect to the end of the abstract).

The abstract length classifier returns a smoothed (add one smoothing) probability that an abstract of a given length (in the number of sentences) contains an outcome statement. For example, the probability that an abstract four sentences long contains an outcome statement is 0.25, and the probability of finding an outcome in a ten sentence–long abstract is 0.92. This feature turns out to be useful because the average length of abstracts with and without outcome statements differs: 11.7 sentences for the former, 7.95 sentences for the latter.

The semantic classifier assigns to a sentence an ad hoc score based on the presence of UMLS concepts belonging to semantic groups highly associated with outcomes such as THERAPEUTIC PROCEDURE or PHARMACOLOGICAL SUBSTANCE. The score is given a boost if the concept has already been identified as the primary problem or an intervention.

The outputs of our basic classifiers are combined using a simple weighted linear interpolation scheme:

$$S_{\text{outcome}} = \lambda_1 S_{\text{cues}} + \lambda_2 S_{\text{unigram}} + \lambda_3 S_{n\text{-gram}} + \lambda_4 S_{\text{position}} + \lambda_5 S_{\text{length}} + \lambda_6 S_{\text{semantic type}} \quad (1)$$

We attempted two approaches for assigning these weights. The first method relied on ad hoc weight selection based on intuition. The second involved a more principled method using confidence values generated by the base classifiers and least squares lin-

4 http://mallet.cs.umass.edu/

ear regression adapted for classification (Ting and Witten 1999), which can be described by the following equation:

$$LR(x) = \sum_{k=1}^{N} \alpha_k P_k(X) \quad (2)$$

P_k is the probability that a sentence specifies an outcome, as determined by classifier k (for classifiers that do not return actual probabilities, we normalized the scores and treated them as such). To predict the class of a sentence, the probabilities generated by n classifiers are combined using the coefficients $(\alpha_0, ..., \alpha_n)$. These values are determined in the training stage as follows: Probabilities predicted by base classifiers for each sentence are represented in an $N \times M$ matrix A, where M is the number of sentences in the training set, and N is the number of classifiers. The gold standard class assignments for each sentence is stored in a vector b, and weights are found by computing the vector α that minimizes $||A\alpha - b||$. The solution can be found using singular value decomposition, as provided in the JAMA basic linear algebra package released by NIST.[5]

5.8 Evaluation of Outcome Extractor

Because outcome statements were annotated in each of the 633 citations in our collection, it was possible to evaluate the outcome extractor on a broader set of abstracts. From those not used in training the outcome classifiers, 153 citations pertaining to therapy were selected. Of these, 143 contained outcome statements and were used as the blind held-out test set. In addition, outcome statements in abstracts pertaining to diagnosis (57), prognosis (111), and etiology (37) were also used.

The output of our outcome extractor is a ranked list of sentences sorted by confidence. Based on the observation that human annotators typically mark two to three sentences in each abstract as outcomes, we evaluated the performance of our extractor at cutoffs of two and three sentences. These results are shown in Table 4: The columns marked AH2 and AH3 show performance of the weighted linear interpolation approach with ad hoc weight assignment at two- and three-sentence cutoffs, respectively; the columns marked LR2 and LR3 show performance of the least squares linear regression model at the same cutoffs. In the evaluation, our outcome extractor was considered correct if the returned sentences intersected with sentences judged as outcomes by our human annotators. Although this is a lenient criterion, it does roughly capture the performance of our knowledge extractor. Because outcome statements are typically found in the conclusion of a structured abstract (or near the end of the abstract in the case of unstructured abstracts), we compared our answer extractor to the baseline of returning either the final two or final three sentences in the abstract (B2 and B3 in Table 4).

As can be seen, variants of our outcome extractor performed better than the baseline at the two-sentence cutoff, for the most part. Bigger improvements, however, can be seen at the three-sentence cutoff level. It is evident that the assignment of weights in our ad hoc model is primarily geared towards therapy questions, perhaps overly so. Better overall performance is obtained with the least squares linear regression model.

5 http://math.nist.gov/javanumerics/jama/

Table 4
Evaluation of the outcome extractor. B = baseline, returns last sentences in abstract; AH = ad hoc weight assignment; LR = least squares linear regression. Statistically significant improvement over the baseline at the 1% level is indicated by ▲.

	2-sentence cutoff (%)			3-sentence cutoff (%)		
	B2	AH2	LR2	B3	AH3	LR3
Therapy	74	75	77	75	95▲	93▲
Diagnosis	72	70	78	75	78	89▲
Prognosis	73	76	79▲	85	87	89
Etiology	64	68	74▲	78	83	88▲

Table 5
Examples of strength of evidence categories based on Publication Type and MeSH headings.

Strength of Evidence	Publication Type/MeSH
Level A(1)	Meta-analysis, randomized controlled trials, cohort study, follow-up study
Level B(2)	Case-control study, case series
Level C(3)	Case report, in vitro, animal and animal testing, alternatives studies

The majority of errors made by the outcome extractor were related to inaccurate sentence boundary identification, chunking errors, and word sense ambiguity in the Metathesaurus.

5.9 Determining the Strength of Evidence

The strength of evidence is a classification scheme that helps physicians assess the quality of a particular citation for clinical purposes. Metadata associated with most MEDLINE citations (MeSH terms) are extensively used to determine the strength of evidence and in our EBM citation scoring algorithm (Section 6).

The potential highest level of the strength of evidence for a given citation can be identified using the Publication Type (a metadata field) and MeSH terms pertaining to the type of the clinical study. Table 5 shows our mapping from publication type and MeSH headings to evidence grades based on principles defined in the Strength of Recommendations Taxonomy (Ebell et al. 2004).

5.10 Sample Output

A complete example of our knowledge extractors working in unison is shown in Figure 2, which contains an abstract retrieved in response to the following question: "In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?" (Kauffman, Sawyer, and Scheinbaum 1992). *Febrile illness* is the only concept mapped to DISORDER, and hence is identified as the primary problem. *37 otherwise healthy children aged 2 to 12 years* is correctly identified as the population. *Acetaminophen*, *ibuprofen*, and *placebo* are correctly

Antipyretic efficacy of ibuprofen vs acetaminophen
Kauffman RE, Sawyer LA, Scheinbaum ML
Am J Dis Child. 1992 May;146(5):622-5
OBJECTIVE–To compare the antipyretic efficacy of ibuprofen, placebo, and acetaminophen. DESIGN–Double-dummy, double-blind, randomized, placebo-controlled trial. SETTING–Emergency department and inpatient units of a large, metropolitan, university-based, children's hospital in Michigan. PARTICIPANTS–37 otherwise healthy children aged 2 to 12 years$_{\text{Population}}$ with acute, intercurrent, febrile illness$_{\text{Problem}}$. INTERVENTIONS–Each child was randomly assigned to receive a single dose of acetaminophen$_{\text{Intervention}}$ (10 mg/kg), ibuprofen$_{\text{Intervention}}$ (10 mg/kg) (7.5 or 10 mg/kg), or placebo$_{\text{Intervention}}$ (10 mg/kg). MEASUREMENTS/MAIN RESULTS–Oral temperature was measured before dosing, 30 minutes after dosing, and hourly thereafter for 8 hours after the dose. Patients were monitored for adverse effects during the study and 24 hours after administration of the assigned drug. All three active treatments produced significant antipyresis compared with placebo.$_{\text{Outcome}}$ Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses.$_{\text{Outcome}}$ No adverse effects were observed in any treatment group. CONCLUSION–Ibuprofen is a potent antipyretic agent and is a safe alternative for the selected febrile child who may benefit from antipyretic medication but who either cannot take or does not achieve satisfactory antipyresis with acetaminophen.$_{\text{Outcome}}$
Publication Type: Clinical Trial, Randomized Controlled Trial
PMID: 1621668
Strength of Evidence: grade A

Figure 2
Sample output from our PICO extractors.

extracted as the interventions under study. The three outcome sentences are correctly classified; the short sentence concerning adverse effects was ranked lower than the other three sentences and hence below the cutoff. The study design, from metadata associated with the citation, allows our strength of evidence extractor to classify this article as grade A.

6. Operationalizing Evidence-Based Medicine

In our view of clinical question answering, the knowledge extractors just described supply the features on which semantic matching occurs. This section describes an algorithm that, when presented with a structured representation of an information need and a MEDLINE citation, automatically computes a topical relevance score in accordance with the principles of EBM.

In order to develop algorithms that operationalize the three facets of EBM, it is necessary to possess a corpus of clinical questions on which to experiment. Because no such test collection exists, we had to first manually create one. Fortunately, collections of clinical questions (representing real-world information needs of physicians), are

Table 6
Composition of our clinical questions collection.

	Therapy	Diagnosis	Prognosis	Etiology	Total
Development	10	6	3	5	24
Test	12	6	3	5	26

available on-line. From two sources, the *Journal of Family Practice*[6] and the Parkhurst Exchange,[7] we gathered 50 clinical questions, which capture a realistic sampling of the scenarios that a clinical question-answering system would be confronted with. These questions were minimally modified from their original form as downloaded from the World Wide Web. In a few cases, a single question actually consisted of several smaller questions; such clusters were simplified by removing questions more peripheral to the central clinical problem. All questions were manually classified into one of the four clinical tasks; the distribution of the questions roughly follows the prevalence of each task type as observed in natural settings, noted by Ely et al. (1999). The final step in the preparation process was manual translation of the natural language questions into PICO query frames.

Our collection was divided into a development set and a blind held-out test set for verification purposes. The breakdown of these questions into the four clinical tasks and the development/test split is shown in Table 6. An example of each question type from our development set is presented here, along with its query frame:

Does quinine reduce leg cramps for young athletes? (Therapy)
search task: therapy selection
primary problem: leg cramps
co-occurring problems: muscle cramps, cramps
population: young adult
intervention: quinine

How often is coughing the presenting complaint in patients with gastroesophageal reflux disease? (Diagnosis)
search task: differential diagnosis
primary problem: gastroesophageal reflux disease
co-occurring problems: cough

What's the prognosis of lupoid sclerosis? (Prognosis)
search task: patient outcome prediction
primary problem: lupus erythematosus
co-occurring problems: multiple sclerosis

What are the causes of hypomagnesemia? (Etiology)
search task: cause determination
primary problem: hypomagnesemia

6 http://www.jfponline.com/
7 http://www.parkhurstexchange.com/qa/

As discussed earlier, we do not believe that natural language text is the best input for a question-answering system. Instead, a structured PICO-based representation captures physicians' information needs in a more perspicuous manner—primarily because clinicians are trained to analyze clinical situations with this framework.

Mirroring the organization of our knowledge extractors, we broke up the P in PICO into population, primary problem, and co-occurring problems in the query representation. The justification for this will become apparent when we present our algorithm for scoring MEDLINE citations, as each of these three facets must be treated differently. Note that many elements are specified only to the extent that they were explicit in the original natural language question; for example, if the clinician does not specify a population, that element will be empty. Finally, outcomes are not directly encoded in the query representation because they are implicit most of the time; for example, in *Does quinine reduce leg cramps for young athletes?,* the desired outcome, naturally, is to reduce the occurrence and severity of leg cramps. Nevertheless, outcome identification is an important component of the citation scoring algorithm, as we shall see later.

What is the relevance of an abstract with respect to a particular clinical question? Evidence-based medicine outlines the need to consider three different facets (see Section 2), which we operationalize in the following manner:

$$S_{\text{EBM}} = S_{\text{PICO}} + S_{\text{SoE}} + S_{\text{task}} \tag{3}$$

The relevance of a particular citation, with respect to a structured query, includes contributions from matching PICO structures, the strength of evidence of the citation, and factors specifically associated with the search tasks (and indirectly, the clinical tasks). In what follows, we describe each of these contributions in detail.

Viewed as a whole, each score component is a heuristic reflection of the factors that enter into consideration when a physician examines a MEDLINE citation. Although the assignment of numeric scores is based on intuition and may seem ad hoc in many cases, evaluation results in the next section demonstrate the effectiveness of our algorithm. This issue will be taken up further in Section 8.

6.1 Scores Based on PICO Elements

The score of an abstract based on extracted PICO elements, S_{PICO}, is broken into individual components according to the following formula:

$$S_{\text{PICO}} = S_{\text{problem}} + S_{\text{population}} + S_{\text{intervention}} + S_{\text{outcome}} \tag{4}$$

The first component in the equation, S_{problem}, reflects a match between the primary problem in the query frame and the primary problem in the abstract (i.e., the highest-scoring problem identified by the problem extractor). A score of 1 is given if the problems match exactly based on their unique UMLS concept ID as provided by MetaMap. Matching based on concept IDs has the advantage that it abstracts away from terminological variation; in essence, MetaMap performs terminological normalization. Failing an exact match of concept IDs, a partial string match is given a score of 0.5. If the primary problem in the query has no overlap with the primary problem from the abstract, a score of -1 is given. Finally, if our problem extractor could not identify a problem (but the query frame does contain a problem), a score of -0.5 is given.

Co-occurring problems must be taken into consideration in the *differential diagnosis* and *cause determination* search tasks because knowledge of the problems is typically

incomplete in these scenarios. Therefore, physicians would normally be interested in any problems mentioned in the abstracts in addition to the primary problem specified in the query frame. As an example, consider the question *What is the differential diagnosis of chronic diarrhea in immunocompetent patients?* Although chronic diarrhea is the primary problem, citations that discuss additional related disorders should be favored over those that don't. In terms of actual scoring, disorders mentioned in the title receive three points, and disorders mentioned anywhere else receive one point (in addition to the match score based on the primary problem, as discussed).

Scores based on population and intervention, $S_{\text{population}}$ and $S_{\text{intervention}}$ respectively, measure the overlap between query frame elements and corresponding elements extracted from abstracts. A point is given to each matching intervention and matching population. For example, finding the population group *children* from a query frame in the abstract increments the match score; the remaining words in the abstract population are ignored. Thus, if the query frame contains a population element and an intervention element, the score for an abstract that contains the same UMLS concepts in the corresponding slots is incremented by two.

The outcome-based score, S_{outcome}, is simply the value assigned to the highest-scoring outcome sentence (we employed the outcome extractor based on the linear regression model for our experiments). As outcomes are rarely explicitly specified in the original question, we decided to omit them in the query representation. Our citation scoring algorithm simply considers the inherent quality of the outcome statements in an abstract, independent of the query. This is justified because, given a match on the primary problem, all clinical outcomes are likely to be of interest to the physician.

6.2 Scores Based on Strength of Evidence

The relevance score component based on the strength of evidence is calculated in the following manner:

$$S_{\text{SoE}} = S_{\text{journal}} + S_{\text{study}} + S_{\text{date}} \tag{5}$$

Citations published in core and high-impact journals such as *Journal of the American Medical Association* (JAMA) get a score of 0.6 for S_{journal}, and 0 otherwise. In terms of the study type, S_{study}, clinical trials, such as randomized controlled trials, receive a score of 0.5; observational studies, for example, case reports, 0.3; all non-clinical publications, -1.5; and 0 otherwise. The study type is directly encoded in the Publication Type field of a MEDLINE citation.

Finally, recency factors into the strength of evidence score according to the formula:

$$S_{\text{date}} = (year_{\text{publication}} - year_{\text{current}})/100 \tag{6}$$

A mild penalty decreases the score of a citation proportionally to the time difference between the date of the search and the date of publication.

6.3 Scores Based on Specific Tasks

The final component of our EBM score is based on task-specific considerations, as reflected in manually assigned MeSH terms. For search tasks falling into each clinical task, we gathered a list of terms that are positive and negative indicators of relevance.

The task score, S_{task}, is given by:

$$S_{\text{task}} = \sum_{t \in \text{MeSH}} \alpha(t) \tag{7}$$

The function $\alpha(t)$ maps a MeSH term to a positive score if the term is a positive indicator for that particular task type, or a negative score if the term is a negative indicator for the clinical task. Note that although our current system uses MeSH headings assigned by human indexers, manually assigned terms can be replaced with automatic processing if needed (Aronson et al. 2004).

Below, we enumerate the relevant indicator terms by clinical task. However, there is a set of negative indicators common to all tasks; these were extracted from the set of genomics articles provided for the secondary task in the TREC 2004 genomics track evaluation (Hersh, Bhupatiraju, and Corley 2004); examples include *genetics* and *cell physiology.* The positive and negative weights assigned to each term heuristically encode the relative importance of different MeSH headings and are derived from the Clinical Queries filters in PubMed, from the JAMA EBM tutorial series on critical appraisal of medical literature, from MeSH scope notes, and based on a physician's understanding of the domain (the first author).

Indicators for Therapy Tasks. Positive indicators for therapy were derived from the PubMed's Clinical Queries filters; examples include *drug administration routes* and any of its children in the MeSH hierarchy. A score of ± 1 is given if the MeSH descriptor or qualifier is marked as the main theme of the article (indicated via the star notation by human indexers), and a score of ± 0.5 otherwise. If the question pertains to the search task of *prevention*, three additional headings are considered positive indicators: *prevention and control, prevention measures,* and *prophylaxis.*

Indicators for Diagnosis Tasks. Positive indicators for therapy are also used as negative indicators for diagnosis because the relevant studies are usually disjoint; it is highly unlikely that the same clinical trial will study both diagnostic methods and treatment methods. The MeSH term *diagnosis* and any of its children are considered positive indicators. As with therapy questions, terms marked as the major theme get a score of ± 1.0, and ± 0.5 otherwise. This general assignment of indicator terms allows a system to differentiate between questions such as *Does a Short Symptom Checklist accurately diagnose ADHD?* and *What is the most effective treatment for ADHD in children?*, which might retrieve very similar sets of citations.

Indicators for Prognosis Tasks. Positive indicators for prognosis include the following MeSH terms: *survival analysis, disease-free survival, treatment outcome, health status, prevalence, risk factors, disability evaluation, quality of life,* and *recovery of function.* For terms marked as the major theme, a score of $+2$ is given; $+1$ otherwise. There are no negative indicators, other than those common to all tasks previously described.

Indicators for Etiology Tasks. Negative indicators for etiology include therapy-oriented MeSH terms; these terms are given a score of -0.3. Positive indicators for the diagnosis task are weak positive indicators for etiology, and receive a positive score of $+0.1$. The following MeSH terms are considered highly indicative of citations relevant to etiology: *population at risk, risk factors, etiology, causality,* and *physiopathology.* If

one of these terms is marked as the major theme, a score of +2 is given; otherwise, a score of +1 is given.

7. Evaluation of Citation Scoring

The previous section describes a relevance-scoring algorithm for MEDLINE citations that attempts to capture the principles of EBM. In this section, we present an evaluation of this algorithm.

Ideally, questions should be answered by directly comparing queries to knowledge structures derived from MEDLINE citations. However, knowledge extraction on such large scales is impractical given our computational resources, so we opted for an IR-based pipeline approach. Under this strategy, an existing search engine would be employed to generate a candidate list of citations to be rescored, according to our algorithm. PubMed is a logical choice for gathering this initial list of citations because it represents one of the most widely used tools employed by physicians and other health professionals today. The system supports boolean operators and sorts results chronologically, most recent citations first.

This two-stage retrieval process immediately suggests an evaluation methodology for our citation scoring algorithm—as a document reranking task. Given an initial hit list, can our algorithm automatically re-sort the results such that relevant documents are brought to higher ranks? Not only is such a task intuitive to understand, this conceptualization also lends itself to an evaluation based on widely accepted practices in information retrieval.

For each question in our test collection, PubMed queries were manually crafted to fetch an initial set of hits. These queries took advantage of existing advanced search features to simulate the types of results that would be currently available to a knowledgeable physician. Specifically, widely accepted tools for narrowing down PubMed search results such as Clinical Queries were employed whenever appropriate.

As a concrete example, consider the following question: *What is the best treatment for analgesic rebound headaches?* The search started with the initial terms "analgesic rebound headache" with a "narrow therapy filter." In PubMed, this query is:

((“headache disorders”[TIAB] NOT Medline[SB]) OR “headache disorders”[MeSH Terms] OR analgesic rebound headache[Text Word]) AND (randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract])) AND hasabstract[text] AND English[Lang] AND “humans”[MeSH Terms]

Note that PubMed automatically identifies concepts and attempts matching both in abstract text and MeSH headings. We always restrict searches to articles that have abstracts, are published in English, and are assigned the MeSH term *humans* (as opposed to say, experiments on animals)—these are all strategies commonly used by clinicians.

In this case, because none of the top 20 results were relevant, the query was expanded with the term *side effects* to emphasize the aspect of the problem requiring an intervention. The final query for the question became:

(((“analgesics”[TIAB] NOT Medline[SB]) OR “analgesics”[MeSH Terms] OR “analgesics”[Pharmacological Action] OR analgesic[Text Word]) AND ((“headache”[TIAB] NOT Medline[SB]) OR “headache”[MeSH Terms] OR headaches[Text Word]) AND (“adverse effects”[Subheading] OR side effects[Text Word])) AND hasabstract[text] AND English[Lang] AND “humans”[MeSH Terms]

The first author, who is a medical doctor, performed the query formulation process manually for every question in our collection; she verified that each hit list contained at least some relevant documents and that the results were as good as could be reasonably achieved. The process of generating queries averaged about 40 minutes per question. The top 50 results for each query were retained for our experiments. In total, 2,309 citations were gathered because some queries returned fewer than 50 hits. The process of generating a "good" PubMed query is not a trivial task, which we have side-stepped in this work by placing a human in the loop. We return to this issue in Section 12.

All abstracts gathered by this process were exhaustively examined for relevance by the first author. It is important to note that relevance assessment in the clinical domain requires significant medical knowledge (in short, a medical degree). After careful consideration, we decided to assess only topical relevance, with the understanding that the applicability of information from a specific citation in real-world settings depends on a variety of other factors (see Section 10 for further discussion). Each citation was assigned one of four labels:

- **Contains answer:** The citation directly contains information that answers the question.
- **Relevant:** The citation does not directly answer the question, but provides topically relevant information.
- **Partially relevant:** The citation provides information that is marginally relevant.
- **Not relevant:** The citation does not provide any topically relevant information.

Because all abstracts were judged, we did not have to worry about impartiality issues when comparing different systems. In total, the relevance assessment process took approximately 100 hours, or about an average of 2 hours per question.

Our reranking experiment compared four different systems:

- The baseline PubMed results.
- A term-based reranker that computes term overlap between the natural language question and the citation (i.e., counted words shared between the two strings). Each term match was weighted by the outcome score of the sentence from which it came (see Section 5.7). This simple algorithm favors term matches that occur in sentences recognized as outcome statements.
- A reranker based on the EBM scorer described in the previous section.
- A reranker that combines normalized scores from the term-based reranker and the EBM-based reranker (weighted linear interpolation).

Questions in the development set were used to debug the EBM-based reranker as we implemented the scoring algorithm. The development questions were also used to tune the weight for combining scores from the term-based scorer and EBM-based scorer; by simply trying all possible values, we settled on a λ of 0.8, that is, 80% weight to the EBM score, and 20% weight to the term-based score. As we shall see later, it is unclear if evidence combination in this simple manner helps at all; for one, it is debatable which metric should be optimized. The test questions were hidden during the system

development phase and served as a blind held-out test set for assessing the generality of our algorithm.

In our experiment, we collected the following metrics, all computed automatically using our relevance judgments:

- **Precision at ten retrieved documents (P10)** measures the fraction of relevant documents in the top ten results.
- **Mean Average Precision (MAP)** is the average of precision values after each relevant document is retrieved (Baeza-Yates and Ribeiro-Neto 1999). It is the most widely accepted single-value metric in information retrieval, and is seen to balance the need for both precision and recall.
- **Mean Reciprocal Rank (MRR)** is a measure of how far down a hit list the user must browse before encountering the first relevant result. The score is equal to the reciprocal of the rank, that is, a relevant document at rank 1 gets a score of 1, 1/2 at rank 2, 1/3 at rank 3, and so on. Note that this measure only captures the appearance of the *first* relevant document. Furthermore, due to its discretization, MRR values are noisy on small collections.
- **Total Document Reciprocal Rank (TDRR)** is the sum of the reciprocal ranks of all relevant documents. For example, if relevant documents were found at ranks 2 and 5, the TDRR would be $1/2 + 1/5 = 0.7$. TDRR provides an advantage over MRR in that it captures the ranks of all relevant documents—emphasizing their appearance at higher ranks. The downside, however, is that TDRR does not have an intuitive interpretation.

For our reranking experiment, we applied the Wilcoxon signed-rank test to determine the statistical significance of the results. This test is commonly used in information retrieval research because it makes minimal assumptions about the underlying distribution of differences. For each evaluation metric, significance at the 1% level is indicated by either $^{\blacktriangle}$ or $^{\blacktriangledown}$, depending on the direction of change; significance at the 5% level is indicated by $^{\triangle}$ or $^{\triangledown}$, depending on the direction of change. Differences that are not statistically significant are marked with the symbol $^{\circ}$.

We report results under two different scoring criteria. Under the lenient condition, documents marked "contains answer" and "relevant" were given credit; these results are shown in Table 7 (for the development set) and Table 8 (for the blind held-out test set). Across all questions, both the EBM-based reranker and combination reranker significantly outperform the PubMed baseline on all metrics. In many cases, the differences are particularly noteworthy—for example, our EBM citation scoring algorithm more than doubles the baseline in terms of MAP and P10 on the test set. There are enough therapy questions to achieve statistical significance in the task-specific results; however, due to the smaller number of questions for the other clinical tasks, those results are not statistically significant. Results also show that the simple term-based reranker outperforms the PubMed baseline, demonstrating the importance of recognizing outcome statements in MEDLINE abstracts.

Are the differences in performance between the term-based, EBM, and combination rerankers statistically significant? Results of Wilcoxon signed-rank tests are shown in Table 11. Both the EBM and combination rerankers significantly outperform the term-based reranker (at the 1% level, on all metrics, on both development and test set), with

Table 7
(Lenient, Development) Lenient results of reranking experiment on development questions for the baseline PubMed condition, term-based reranker, EBM-based reranker, and combination reranker.

	Therapy	Diagnosis	Prognosis	Etiology
Precision at 10 (P10)				
PubMed	0.300	0.367	0.400	0.533
Term	0.520 (+73%)$^{\triangle}$	0.383 (+4.5%)°	0.433 (+8.3%)°	0.553 (+3.8%)°
EBM	0.730 (+143%)$^{\blacktriangle}$	0.800 (+118%)$^{\triangle}$	0.633 (+58%)°	0.553 (+3.7%)°
Combo	0.750 (+150%)$^{\blacktriangle}$	0.783 (+114%)$^{\triangle}$	0.633 (+58%)°	0.573 (+7.5%)°
Mean Average Precision (MAP)				
PubMed	0.354	0.421	0.385	0.608
Term	0.622 (+76%)$^{\blacktriangle}$	0.438 (+4.0%)°	0.464 (+21%)°	0.720 (+18%)°
EBM	0.819 (+131%)$^{\blacktriangle}$	0.794 (+89%)$^{\triangle}$	0.635 (+65%)°	0.649 (+6.7%)°
Combo	0.813 (+130%)$^{\blacktriangle}$	0.759 (+81%)$^{\triangle}$	0.644 (+67%)°	0.686 (+13%)°
Mean Reciprocal Rank (MRR)				
PubMed	0.428	0.792	0.733	0.900
Term	0.853 (+99%)$^{\triangle}$	0.739 (−6.7%)°	0.833 (+14%)°	1.000 (+11%)°
EBM	0.933 (+118%)$^{\triangle}$	0.917 (+16%)°	0.667 (−9.1%)°	1.000 (+11%)°
Combo	0.933 (+118%)$^{\triangle}$	0.917 (+16%)°	1.000 (+36%)°	0.900 (+0.0%)°
Total Document Reciprocal Rank (TDRR)				
PubMed	1.317	1.805	1.778	2.008
Term	2.305 (+75%)$^{\blacktriangle}$	1.887 (+4.6%)°	1.923 (+8.2%)°	2.291 (+14%)°
EBM	2.869 (+118%)$^{\blacktriangle}$	2.944 (+63%)$^{\triangle}$	2.238 (+26%)°	2.104 (+4.8%)°
Combo	2.833 (+115%)$^{\blacktriangle}$	2.870 (+59%)$^{\triangle}$	2.487 (+40%)°	2.108 (+5.0%)°

(a) Breakdown by clinical task

	P10	MAP	MRR	TDRR
PubMed	0.378	0.428	0.656	1.640
Term	0.482 (+28%)°	0.577 (+35%)$^{\blacktriangle}$	0.853 (+30%)°	2.150 (+31%)$^{\triangle}$
EBM	0.699 (+85%)$^{\blacktriangle}$	0.754 (+76%)$^{\blacktriangle}$	0.910 (+39%)$^{\triangle}$	2.650 (+62%)$^{\blacktriangle}$
Combo	0.707 (+87%)$^{\blacktriangle}$	0.752 (+76%)$^{\blacktriangle}$	0.931 (+42%)$^{\blacktriangle}$	2.648 (+61%)$^{\blacktriangle}$

(b) Performance across all clinical tasks

$^{\blacktriangle\blacktriangledown}$ Significance at the 1% level, depending on direction of change.
$^{\triangle\triangledown}$ Significance at the 5% level, depending on direction of change.
°Difference not statistically significant.

the exception of MRR on the development set. However, for all metrics, on both the development set and test set, there is no significant difference between the EBM and combination reranker (which combines both term-based and EBM-based evidence). In the parameter tuning process, we could not find a weight where performance across all measures was higher; in the end, we settled on what we felt was a reasonable weight that improved P10 and MRR on the development set.

Under the strict condition, only documents marked "contains answer" were given credit; these results are shown in Table 9 (for the development set) and Table 10 (for the blind held-out test set). The same trend is observed—in fact, larger relative gains were achieved under the strict scoring criteria for our EBM and combination

Table 8
(Lenient, Test) Lenient results of reranking experiment on blind held-out test questions for the baseline PubMed condition, term-based reranker, EBM-based reranker, and combination reranker.

	Therapy	Diagnosis	Prognosis	Etiology
Precision at 10 (P10)				
PubMed	0.350	0.150	0.200	0.320
Term	0.575 (+64%)▲	0.383 (+156%)°	0.333 (+67%)°	0.460 (+43%)°
EBM	0.783 (+124%)▲	0.583 (+289%)△	0.467 (+133%)°	0.660 (+106%)°
Combo	0.792 (+126%)▲	0.633 (+322%)△	0.433 (+117%)°	0.660 (+106%)°
Mean Average Precision (MAP)				
PubMed	0.421	0.279	0.235	0.364
Term	0.563 (+34%)▲	0.489 (+76%)°	0.415 (+77%)°	0.480 (+32%)°
EBM	0.765 (+82%)▲	0.637 (+129%)△	0.722 (+207%)°	0.701 (+93%)°
Combo	0.770 (+83%)▲	0.653 (+134%)△	0.690 (+194%)°	0.687 (+89%)°
Mean Reciprocal Rank (MRR)				
PubMed	0.579	0.443	0.456	0.540
Term	0.660 (+14%)°	0.765 (+73%)°	0.611 (+34%)°	0.650 (+20%)°
EBM	0.917 (+58%)△	0.889 (+101%)°	1.000 (+119%)°	1.000 (+85%)°
Combo	0.958 (+66%)△	0.917 (+107%)°	1.000 (+119%)°	1.000 (+85%)°
Total Document Reciprocal Rank (TDRR)				
PubMed	1.669	0.926	0.895	1.381
Term	2.204 (+32%)△	1.880 (+103%)°	1.390 (+55%)°	1.736 (+26%)°
EBM	2.979 (+79%)▲	2.341 (+153%)△	2.101 (+138%)°	2.671 (+93%)°
Combo	3.025 (+81%)▲	2.380 (+157%)△	2.048 (+129%)°	2.593 (+88%)°

(a) Breakdown by clinical task

	P10	MAP	MRR	TDRR
PubMed	0.281	0.356	0.526	1.353
Term	0.481 (+71%)▲	0.513 (+44%)▲	0.677 (+29%)°	1.945 (+44%)▲
EBM	0.677 (+141%)▲	0.718 (+102%)▲	0.936 (+78%)▲	2.671 (+98%)▲
Combo	0.688 (+145%)▲	0.718 (+102%)▲	0.962 (+83%)▲	2.680 (+98%)▲

(b) Performance across all clinical tasks

▲▼Significance at the 1% level, depending on direction of change.
△▽Significance at the 5% level, depending on direction of change.
°Difference not statistically significant.

rerankers. Results of Wilcoxon signed-rank tests on the term-based, EBM, and combination rerankers are also shown in Table 11 for the strict scoring condition. In most cases, combining term scoring with EBM scoring does not help. In almost all cases, the EBM and combination reranker perform significantly better than the term-based reranker.

How does better ranking of citations impact end-to-end question answering performance? We shall return to this issue in Sections 9 and 10, which describe and evaluate the answer generation module, respectively. In the next section, we describe more detailed experiments with our EBM citation scoring algorithm.

Table 9
(Strict, Development) Strict results of reranking experiment on development questions for the baseline PubMed condition, term-based reranker, EBM-based reranker, and combination reranker.

	Therapy	Diagnosis	Prognosis	Etiology
Precision at 10 (P10)				
PubMed	0.130	0.133	0.100	0.253
Term	0.230 (+77%)°	0.217 (+63%)°	0.233 (+133%)°	0.293 (+16%)°
EBM	0.350 (+170%)△	0.350 (+163%)°	0.267 (+167%)°	0.293 (+16%)°
Combo	0.350 (+170%)△	0.367 (+175%)°	0.300 (+200%)°	0.313 (+24%)°
Mean Average Precision (MAP)				
PubMed	0.088	0.108	0.058	0.164
Term	0.205 (+134%)°	0.142 (+32%)°	0.090 (+54%)°	0.246 (+50%)°
EBM	0.314 (+260%)°	0.259 (+140%)°	0.105 (+79%)°	0.265 (+62%)°
Combo	0.301 (+244%)°	0.248 (+130%)°	0.129 (+122%)°	0.273 (+67%)°
Mean Reciprocal Rank (MRR)				
PubMed	0.350	0.453	0.394	0.367
Term	0.409 (+17%)°	0.581 (+28%)°	0.528 (+34%)°	0.700 (+91%)°
EBM	0.675 (+93%)△	0.756 (+67%)°	0.444 (+13%)°	0.800 (+118%)°
Combo	0.569 (+63%)△	0.676 (+49%)°	0.833 (+111%)°	0.700 (+91%)°
Total Document Reciprocal Rank (TDRR)				
PubMed	0.610	0.711	0.568	0.721
Term	0.872 (+43%)°	1.022 (+44%)°	0.804 (+42%)°	1.224 (+70%)°
EBM	1.434 (+135%)▲	1.601 (+125%)°	0.824 (+45%)°	1.298 (+80%)°
Combo	1.282 (+110%)▲	1.502 (+111%)°	1.173 (+106%)°	1.241 (+72%)°

(a) Breakdown by clinical task

	P10	MAP	MRR	TDRR
PubMed	0.153	0.105	0.385	0.653
Term	0.240 (+57%)△	0.183 (+75%)°	0.527 (+37%)△	0.974 (+49%)▲
EBM	0.328 (+115%)▲	0.264 (+152%)▲	0.693 (+80%)▲	1.371 (+110%)▲
Combo	0.340 (+123%)▲	0.260 (+148%)▲	0.656 (+71%)▲	1.315 (+101%)▲

(b) Performance across all clinical tasks

▲▼ Significance at the 1% level, depending on direction of change.
△▽ Significance at the 5% level, depending on direction of change.
°Difference not statistically significant.

8. Optimizing Citation Scoring

A potential, and certainly valid, criticism of our EBM citation scoring algorithm is its ad hoc nature. Weights for various features were assigned based on intuition, reflecting our understanding of the domain and our knowledge about the principles of evidence-based medicine. Parameters were fine-tuned during the system implementation process by actively working with the development set; however, this was not done in any systematic fashion. Nevertheless, results on the blind held-out test set confirm the generality of our citation scoring algorithm.

Table 10
(Strict, Test) Strict results of reranking experiment on blind held-out test questions for the baseline PubMed condition, term-based reranker, EBM-based reranker, and combination reranker.

	Therapy	Diagnosis	Prognosis	Etiology
Precision at 10 (P10)				
PubMed	0.108	0.017	0.000	0.080
Term	0.192 (+77%)°	0.133 (+700%)°	0.033 °	0.140 (+75%)°
EBM	0.233 (+115%)°	0.167 (+900%)°	0.100 °	0.200 (+150%)°
Combo	0.258 (+139%)$^{\triangle}$	0.200 (+1100%)°	0.100 °	0.220 (+175%)°
Mean Average Precision (MAP)				
PubMed	0.061	0.024	0.015	0.050
Term	0.082 (+36%)°	0.118 (+386%)°	0.086 (+464%)°	0.086 (+74%)°
EBM	0.109 (+80%)°	0.091 (+276%)°	0.234 (+1442%)°	0.159 (+220%)°
Combo	0.120 (+99%)°	0.107 (+339%)°	0.224 (+1372%)°	0.165 (+232%)°
Mean Reciprocal Rank (MRR)				
PubMed	0.282	0.073	0.031	0.207
Term	0.368 (+31%)°	0.429 (+488%)°	0.146 (+377%)°	0.314 (+52%)°
EBM	0.397 (+41%)°	0.431 (+490%)°	0.465 (+1422%)°	0.500 (+142%)°
Combo	0.556 (+97%)$^{\triangle}$	0.422 (+479%)°	0.438 (+1331%)°	0.467 (+126%)°
Total Document Reciprocal Rank (TDRR)				
PubMed	0.495	0.137	0.038	0.331
Term	0.700 (+41%)°	0.759 (+454%)°	0.171 (+355%)°	0.596 (+80%)°
EBM	0.807 (+63%)°	0.654 (+377%)°	0.513 (+1262%)°	0.946 (+186%)°
Combo	0.969 (+96%)$^{\triangle}$	0.698 (+409%)°	0.479 (+1172%)°	0.975 (+195%)°

(a) Breakdown by clinical task

	P10	MAP	MRR	TDRR
PubMed	0.069	0.045	0.190	0.328
Term	0.150 (+117%)$^{\blacktriangle}$	0.092 (+105%)$^{\blacktriangle}$	0.346 (+82%)$^{\blacktriangle}$	0.632 (+93%)$^{\blacktriangle}$
EBM	0.196 (+183%)$^{\blacktriangle}$	0.129 (+187%)$^{\blacktriangle}$	0.433 (+127%)$^{\blacktriangle}$	0.765 (+133%)$^{\blacktriangle}$
Combo	0.219 (+217%)$^{\blacktriangle}$	0.138 (+207%)$^{\blacktriangle}$	0.494 (+160%)$^{\blacktriangle}$	0.851 (+159%)$^{\blacktriangle}$

(b) Performance across all clinical tasks

$^{\blacktriangle\blacktriangledown}$ Significance at the 1% level, depending on direction of change.
$^{\triangle\triangledown}$ Significance at the 5% level, depending on direction of change.
°Difference not statistically significant.

In the development of various language technology applications, it is common for the first materialization of a new capability to be rather ad hoc in its implementation. This is a reflection of an initial attempt to understand both the problem and solution spaces. Subsequent systems, with a better understanding of the possible technical approaches and their limitations, are then able to implement a more principled solution. Because our clinical question-answering system is the first of its type that we are aware of, in terms of both depth and scope, it is inevitable that our algorithms suffer from some of these limitations. Similarly, our collection of clinical questions is the first test collection of its type that we are aware of. Typically, construction of formal models is only made possible by the existence of test collections. We hope that our work sheds new insight on question answering in the clinical domain and paves the way for future work.

Table 11
Performance differences between various rerankers.

	P10	MAP	MRR	TDRR
Development Set				
EBM vs. Term	+45.0% ▲	+30.8% ▲	+6.7% °	+23.3% ▲
Combo vs. Term	+46.7% ▲	+30.4% ▲	+9.1% °	+23.2% ▲
Combo vs. EBM	+1.2% °	−0.3% °	+2.3% °	−0.1% °
Test Set				
EBM vs. Term	+40.8 ▲	+40.1% ▲	+38.3% ▲	+37.3% ▲
Combo vs. Term	+43.2 ▲	+40.0% ▲	+42.1% ▲	+37.8% ▲
Combo vs. EBM	+1.7 °	−0.1% °	+2.7% °	+0.3% °

(a) Lenient Scoring

	P10	MAP	MRR	TDRR
Development Set				
EBM vs. Term	+36.4% ▲	+43.8% ▲	+31.3% °	+40.7% ▲
Combo vs. Term	+41.6% ▲	+41.9% ▲	+24.5% °	+35.0% △
Combo vs. EBM	+3.8% °	−1.3% °	−5.2% °	−4.1% °
Test Set				
EBM vs. Term	+30.8 °	+40.4% △	+24.9% °	+20.9% °
Combo vs. Term	+46.2 ▲	+50.0% ▲	+42.8% △	+34.6% △
Combo vs. EBM	+11.8 △	+6.8% °	+14.3% °	+11.3% °

(b) Strict Scoring

▲▼ Significance at the 1% level, depending on direction of change.
△▽ Significance at the 5% level, depending on direction of change.
°Difference not statistically significant.

In addition, there are some theoretical obstacles for developing a more formal (say, generative) model. Most methods for training such models require independently and identically distributed samples from the underlying distribution—which is certainly not the case with our test collection. Moreover, the event space of queries and documents is extremely large or even infinite, depending on how it is defined. Our training data, assumed to be samples from this underlying distribution, is extremely small compared to the event space, and hence it is unlikely that popular methods (e.g., maximum likelihood estimates) would yield an accurate characterization of the true distribution.

Furthermore, many techniques for automatically setting parameters make use of maximum likelihood techniques—which do not maximize the correct objective function. Maximizing the likelihood of generating the training data does not mean that the evaluation metric under consideration (e.g., mean average precision) is also maximized—this phenomenon is known as metric divergence.

Nevertheless, it is important to better understand the effects of parameter settings in our system. This section describes a few experiments aimed at this goal.

The EBM score of a MEDLINE citation is the sum of three separate components, each representing a facet of evidence-based medicine. This structure naturally suggests a modification to Equation (3) that weights each score component differently:

$$S_{\text{EBM}} = \lambda_1 S_{\text{PICO}} + \lambda_2 S_{\text{SoE}} + (1 - \lambda_1 - \lambda_2) S_{\text{task}} \quad (8)$$

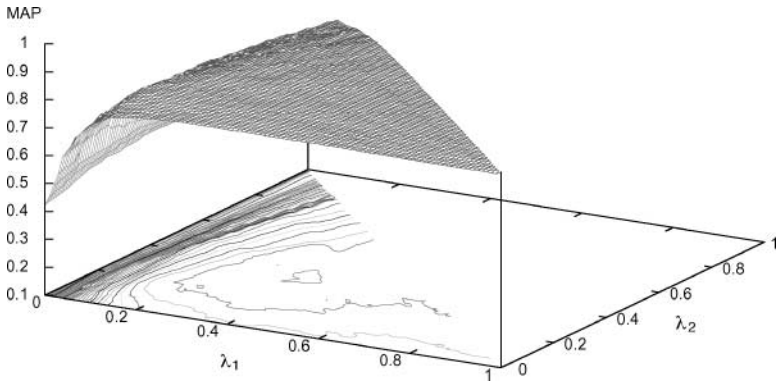

Figure 3
The MAP performance surface for λ_1 and λ_2.

Table 12
Results of optimizing λ_1 and λ_2 on therapy questions.

	P10	MAP	MRR	TDRR
Development Test				
Baseline	0.730	0.819	0.933	2.869
Optimized	0.760 (+4.1%)°	0.822 (+0.4%)°	0.933 (+0.0%)°	2.878 (+0.3%)°
Test Test				
Baseline	0.783	0.765	0.917	2.979
Optimized	0.783 (+0.0%)°	0.762 (−0.4%)°	0.917 (+0.0%)°	2.972 (−0.2%)°

°Difference not statistically significant.

The parameters λ_1 and λ_2 can be derived from our development set. For therapy questions, we exhaustively searched through the entire parameter space, in increments of hundredths, and determined the optimal settings to be $\lambda_1 = 0.38$, $\lambda_2 = 0.34$ (which was found to slightly improve all metrics). The performance surface for mean average precision is shown in Figure 3, which plots results for all possible parameter values on the development set. Numeric results are shown in Table 12. It can be seen that optimizing the parameters in this fashion does not lead to a statistically significant increase in any of the metrics. Furthermore, these gains do not carry over to the blind held-out test set. We also tried optimizing the λ's on all questions in the development set. These results are shown in Table 13. Once again, differences are not statistically significant.

Why does parameter optimization not help? We believe that there are two factors at play here: On the one hand, parameter settings should be specific to the clinical task. This explains why optimizing across all question types at the same time did not improve performance. On the other hand, there are too few questions of any particular type to represent an accurate sampling of all possible questions. This is why parameter tuning on therapy questions did not significantly alter performance. These experiments point to the need for larger test collections, which is an area for future work.

Table 13
Results of optimizing λ_1 and λ_2 on all questions.

	P10	MAP	MRR	TDRR
Development Test				
Baseline	0.699	0.754	0.910	2.650
Optimized	0.707 (+1.2%)°	0.755 (+0.1%)°	0.918 (+0.9%)°	2.660 (+0.4%)°
Test Test				
Baseline	0.677	0.718	0.936	2.671
Optimized	0.669 (−1.1%)°	0.716 (−0.3%)°	0.936 (+0.0%)°	2.662 (−0.3%)°

°Difference not statistically significant.

Table 14
Results of assigning uniform weights to the EBM score component based on the clinical task.

	P10	MAP	MRR	TDRR
Development Test				
Baseline	0.699	0.754	0.910	2.650
$\alpha(t) = \pm 1$	0.690 (−1.2%)°	0.738 (−2.1%)°	0.927 (+1.9%)°	2.646 (−0.2%)°
Test Test				
Baseline	0.677	0.718	0.936	2.671
$\alpha(t) = \pm 1$	0.627 (−7.4%)$^{\triangledown}$	0.681 (−5.2%)°	0.913 (−2.4%)°	2.519 (−5.7%)°

$^{\triangle\triangledown}$ Significance at the 5% level, depending on direction of change.
°Difference not statistically significant.

Another component of our EBM citation scoring algorithm that contains many ad hoc weights is S_{task}, defined in Equation (7) and repeated here:

$$S_{\text{task}} = \sum_{t \in \text{MeSH}} \alpha(t) \tag{9}$$

The function $\alpha(t)$ maps a particular MeSH term to a weight that quantifies the degree to which it is a positive or negative indicator for the particular clinical task. Because these weights were heuristically assigned, it would be worthwhile to examine the impact they have on performance. As a variant, we modified $\alpha(t)$ so that all MeSH terms were mapped to ± 1; in other words, we did not encode granular levels of "goodness." These results are shown in Table 8. Although performance dropped across all metrics, none of the differences were statistically significant except for P10 on the test set.

The series of experiments described herein help us better understand the effects of parameter settings on abstract reranking performance. As can be seen from the results, our algorithm is relatively invariant with respect to the choice of parameters, confirming that our primary contribution is the EBM-based approach to clinical question

answering, and that our performance gains cannot be simply attributed to a fortunate choice of parameters.

9. From Scoring Citations to Answering Questions

The aim of question-answering technology is to move from the "hit list" paradigm of information retrieval, where users receive a list of potentially relevant documents that they must then browse through, to a mode of interaction where users directly receive responses that satisfy their information needs. In our current architecture, fetching a higher-quality ranked list is a step towards generating responsive answers.

The most important characteristic of answers, as recommended by Ely et al. (2005) in their study of real-world physicians, is that they focus on bottom-line clinical advice—information that physicians can directly act on. Ideally, answers should integrate information from multiple clinical studies, pointing out both similarities and differences. The system should collate concurrences, that is, if multiple abstracts arrive at the same conclusion—it need not be repeated unless the physician wishes to "drill down"; the system should reconcile contradictions, for example, if two abstracts disagree on a particular treatment because they studied different patient populations. We have noted that many of these desiderata make complex question answering quite similar to multi-document summarization (Lin and Demner-Fushman 2005b), but these features are also beyond the capabilities of current summarization systems.

It is clear that the type of answers desired by physicians require a level of semantic analysis that is beyond the current state of the art, even with the aid of existing medical ontologies. For example, even the seemingly straightforward task of identifying similarities and differences in outcome statements is rendered exceedingly complex by the tremendous amount of background medical knowledge that must be brought to bear in interpreting clinical results and subtle differences in study design, objectives, and results; the closest analogous task in computational linguistics—redundancy detection for multi-document summarization—seems easy by comparison. Furthermore, it is unclear if textual strings make "good answers." Perhaps a graphical rendering of the semantic predicates present in relevant abstracts might more effectively convey the desired information; see, for example, Fiszman, Rindflesch, and Kilicoglu (2004). Perhaps some variation of multi-level bulleted lists, appropriately integrated with interface elements for expanding and hiding items, might provide physicians a better overview of the information landscape; see, for example, Demner-Fushman and Lin (2006).

Recognizing this complex set of issues, we decided to take a simple extractive approach to answer generation. For each abstract in our reranked list of citations, our system produces an answer by combining the title of the abstract and the top three outcome sentences (in the order they appeared in the abstract). We employed the outcome scores generated by the regression model. No attempt was made to synthesize information from multiple citations. A formal evaluation of this simple approach to answer generation is presented in the next section.

10. Evaluation of Clinical Answers

Evaluation of answers within a clinical setting involves a complex decision that must not only take into account topical relevance (i.e., "Does the answer address the information need?"), but also situational relevance (e.g., Saracevic 1975, Barry and Schamber

1998). The latter factor includes many issues such as the strength of evidence, recency of results, and reputation of the journal. Clinicians need to carefully consider all these elements before acting on any information for the purposes of patient care. Within the framework of evidence-based medicine, the physician is the final arbiter of *how* clinical answers are integrated into the broader activities of medical care, but this complicates any attempt to evaluate answers generated by our system.

In assessing answers produced by our system, we decided to focus only on the evaluation of topical relevance—assessors were only presented with answer strings, generated in the manner described in the previous section. Metadata that would contribute to judgments about situational relevance, such as the strength of evidence, names of the authors and the journal, and so on, were purposefully suppressed. Our evaluation compared the top five answers generated from the original PubMed hit list and the top five answers generated from our reranked list of citations. Answers were prepared for all 24 questions in our development set.

We recruited two medical doctors (one family practitioner, one surgeon) from the National Library of Medicine to evaluate the textual answers. Our instructions clearly stated that only topical relevance was to be assessed. We asked the physicians to provide three-valued judgments:

- A plus (+) indicates that the response directly answers the question. Naturally, the physicians would need to follow up and examine the source citation in more detail.
- A check ($\surd$) indicates that the response provides clinically relevant information that may factor into decisions about patient treatment, and that the source citation was worth examining in more detail.
- A minus (−) indicates that the response does not provide useful information in answering the clinical question, and that the source citation was not worth examining.

We purposely avoided short linguistic labels for the judgments so as to sidestep the question of "What exactly is an answer to a clinical question?" Informally, answers marked with a plus can be considered "actionable" clinical advice. Answers marked with a check provide relevant information that may influence the physician's actions.

We adopted a double-blind study design for the actual assessment process: Answers from both systems were presented in a randomized order without any indication of which system the response came from (duplicates were suppressed). A paper printout, containing each question followed by the blinded answers, was presented to each assessor. We then coded the relevance judgments in a plain text file manually. During this entire time, the key that maps answers to systems was kept in a separate file and hidden from everyone, including the authors. All scores were computed automatically without human intervention.

Answer precision was calculated for two separate conditions: Under the strict condition (Table 15), only "plus" judgments were considered good; under the lenient condition (Table 16), both "plus" and "check" judgments were considered good. As can be seen, our EBM algorithm significantly outperforms the baseline under both the strict and lenient conditions, according to both assessors. On average, the length of answers generated from the original PubMed list of citations was 90 words; answers generated from the reranked list of citations averaged 87 words. Answers from both sources

Table 15
Strict answer precision (considering only "plus" judgments).

	Therapy	Diagnosis	Prognosis	Etiology	All
Assessor 1					
Baseline	.160	.233	.333	.480	.267
EBM	.260 (+63%)	.367 (+58%)	.333 (+0%)	.600 (+25%)	.367 (+37%)
Assessor 2					
Baseline	.040	.233	.200	.400	.183
EBM	.200 (+400%)	.300 (+29%)	.266 (+33%)	.560 (+40%)	.308 (+68%)

Table 16
Lenient answer precision (considering both "plus" and "check" judgments).

	Therapy	Diagnosis	Prognosis	Etiology	All
Assessor 1					
Baseline	.400	.300	.533	.520	.417
EBM	.640 (+60%)	.567 (+89%)	.400 (−25%)	.640 (+23%)	.592 (+42%)
Assessor 2					
Baseline	.240	.267	.333	.440	.300
EBM	.520 (+117%)	.600 (+125%)	.400 (+20%)	.560 (+27%)	.533 (+78%)

were significantly shorter than the abstracts from which they were extracted (250 word average for original PubMed results and 270 word average for reranked results).

To give a feel for the types of responses that are generated by our system, consider the following question:

> What is the best treatment for analgesic rebound headaches?

The following is an example of a response that received a "plus" judgment:

> Medication overuse headache from antimigraine therapy: clinical features, pathogenesis and management: Because of easy availability and low expense, the greatest problem appears to be associated with barbiturate-containing combination analgesics and over-the-counter caffeine-containing combination analgesics. The best management advice is to raise awareness and strive for prevention. Reduction in headache risk factors should include behavioural modification approaches to headache control earlier in the natural history of migraine.

This answer was accepted by both physicians because it clearly states that specific analgesics are most likely to cause the problem, and gives a direct guideline for preventive treatment.

In contrast, the following response to the same question received a "check":

> Does chronic daily headache arise de novo in association with regular use of analgesics? Regular use of analgesics preceded the onset of daily headache in 5 patients by a mean of 5.4 years (range, 2 to 10 years). In 1 patient, the onset of daily headache preceded regular use of analgesics by almost 30 years. These findings suggest that individuals with primary headache, specifically migraine, are predisposed to developing chronic daily headache in association with regular use of analgesics.

Although this answer provides information about the risks and causes of the headaches, neither prevention nor treatment is explicitly mentioned. For these reasons this response was marked as potentially leading to an answer, but not as containing one.

To summarize, we have presented a simple answer generation algorithm that is capable of supplying clinically relevant responses to physicians. Compared to PubMed, which does not take into account the principles of evidence-based medicine, our question-answering system represents a leap forward in information access capabilities.[8]

11. Related Work and Discussion

Clinical question answering is an emerging area of research that has only recently begun to receive serious attention. As a result, there exist relatively few points of comparison to our own work, as the research space is sparsely populated. In this section, however, we will attempt to draw connections to other clinical information systems (although not necessarily for question answering) and related domain-specific question-answering systems. For an overview of systems designed to answer open-domain factoid questions, the TREC QA track overview papers are a good place to start (Voorhees and Tice 1999). In addition, there has been much work on the application of linguistic and semantic knowledge to information retrieval; see Lin and Demner-Fushman (2006a) for a brief overview.

The idea that clinical information systems should be sensitive to the practice of evidence-based medicine is not new. Based on analyses of 4,000 MEDLINE citations, Mendonça and Cimino (2001) have studied MeSH terms associated with the four basic clinical tasks of therapy, diagnosis, prognosis, and etiology. The goal was to automatically classify citations for task-specific retrieval, similar in spirit to the Hedges Project (Haynes et al. 1994; Wilczynski, McKibbon, and Haynes 2001). Cimino and Mendonça reported good performance for etiology, diagnosis, and in particular therapy, but not prognosis. Although originally developed as a tool to assist in query formulation, Booth (2000) pointed out that PICO frames can be employed to structure IR results for improving precision. PICO-based querying in information retrieval is merely an instance of faceted querying, which has been widely used by librarians since the introduction of automated retrieval systems (e.g., Meadow et al. 1989). The work of Hearst (1996) demonstrates that faceted queries can be converted into simple filtering constraints to boost precision.

The feasibility of automatically identifying outcome statements in secondary sources has been demonstrated by Niu and Hirst (2004). Their study also illustrates the importance of semantic classes and relations. However, extraction of outcome statements from secondary sources (meta-analyses, in this case) differs from extraction of outcomes from MEDLINE citations because secondary sources represent knowledge that has already been distilled by humans (which may limit its scope). Because secondary sources are often more consistently organized, it is possible to depend on certain surface cues for reliable extraction (which is not possible for MEDLINE abstracts in general). Our study tackles outcome identification in primary medical sources and demonstrates that respectable performance is possible with a feature-combination approach.

8 Although note that answer generation from the PubMed results also requires the use of the outcome extractor.

The literature also contains work on sentence-level classification of MEDLINE abstracts for non-clinical purposes. For example, McKnight and Srinivasan (2003) describe a machine learning approach to automatically label sentences as belonging to introduction, methods, results, or conclusion using structured abstracts as training data (see also Lin et al. 2006). Tbahriti et al. (2006) have demonstrated that differential weighting of automatically labeled sections can lead to improved retrieval performance. Note, however, that such labels are orthogonal to PICO frame elements, and hence are not directly relevant to knowledge extraction for clinical question answering. In a similar vein, Light, Qiu, and Srinivasan (2004) report on the identification of speculative statements in MEDLINE abstracts, but once again, this work is not directly applicable to clinical question answering.

In addition to question answering, multi-document summarization provides a complementary approach to addressing clinical information needs. The PERSIVAL project, the most comprehensive study of such techniques applied on medical texts to date, leverages patient records to generate personalized summaries in response to physicians' queries (McKeown, Elhadad, and Hatzivassiloglou 2003; Elhadad et al. 2005). Although the system incorporates both a user and a task model, it does not explicitly capture the principles of evidence-based medicine. Patient information is no doubt important to answering clinical questions, and our work could certainly benefit from experiences gained in the PERSIVAL project.

The application of domain models and deep semantic knowledge to question answering has been explored by a variety of researchers (e.g., Jacquemart and Zweigenbaum 2003, Rinaldi et al. 2004), and was also the focus of recent workshops on question answering in restricted domains at ACL 2004 and AAAI 2005. Our work contributes to this ongoing discourse by demonstrating a specific application in the domain of clinical medicine.

Finally, the evaluation of answers to complex questions remains an open research problem. Although it is clear that measures designed for open-domain factoid questions are not appropriate, the community has not agreed on a methodology that will allow meaningful comparisons of results from different systems. In Sections 9 and 10, we have discussed many of these issues. Recently, there is a growing consensus that an evaluation methodology based on the notion of "information nuggets" may provide an appropriate framework for assessing the quality of answers to complex questions. Nugget F-score has been employed as a metric in the TREC question-answering track since 2003, to evaluate so-called definition and "other" questions (Voorhees 2003). A number of studies (e.g., Hildebrandt, Katz, and Lin 2004) have pointed out shortcomings of the original nugget scoring model, although a number of these issues have been recently addressed (Lin and Demner-Fushman 2005a, 2006b). However, adaptation of the nugget evaluation methodology to a domain as specific as clinical medicine is an endeavor that has yet to be undertaken.

12. Future Work

The design and implementation of our current system leaves many open avenues for future exploration, one of which concerns our assumptions about the query interface. Previously, a user study (Lin et al. 2003) has shown that people are reluctant to type full natural language questions, even after being told that they were using a question-answering system and that typing complete questions would result in better performance. We have argued that a query interface based on structured PICO frames will yield better-formulated queries, although it is unclear whether physicians would invest

the upfront effort necessary to accomplish this. Issuing extremely short queries appears to be an ingrained habit of information seekers today, and the dominance of World Wide Web searches reinforce this behavior. Given these trends, physicians may actually prefer the rapid back-and-forth interaction style that comes with short queries. We believe that if systems can produce noticeably better results with richer queries, users will make more of an effort to formulate them. This, however, presents a chicken-and-egg problem: One possible solution is to develop models that can automatically fill query frames given a couple of keywords—this would serve to kick-start the query generation process.

The astute reader will have noticed that the initial retrieval of abstracts in our study was performed with high-quality manually crafted queries (that were part of the test collection). Although this was intended to demonstrate the performance of our EBM citation scoring algorithm with respect to a strong baseline, it also means that we have omitted a component in the automatic question-answering process. Translating a clinical question into a good PubMed query is not a trivial task—in our experiments, it required an experienced searcher approximately 40 minutes on average per question. However, it is important to note that query formulation in the clinical domain is not a problem limited to question-answering systems, but one that users of all retrieval systems must contend with.

Nevertheless, there are three potential solutions to this problem: First, although there is an infinite variety of clinical questions, the number of query types is bounded and far smaller in number; see Huang, Lin, and Demner-Fushman (2006) for an analysis. In a query interface based on PICO frames, it is possible to identify a number of prototypical query frames. From these prototypes, one can generate query templates that abstract over the actual slot fillers—this is the idea behind Clinical Queries. Although this method will probably not retrieve citations as high in quality as custom-crafted queries, there is reason to believe that as long as a reasonable set of citations is retrieved, our system will be able to extract relevant answers (given the high accuracy of our knowledge extractors and citation scoring algorithm). The second approach to tackling this problem is to bypass PubMed altogether and index MEDLINE with another search engine. Due to the rapidly changing nature of the entire MEDLINE database, experiments for practical purposes would most likely be conducted on a static subset of the collection, for example, the ten-year portion created for the TREC 2004 genomics track (Hersh, Bhupatiraju, and Corley 2004). Recent results from TREC have demonstrated that high performance ad hoc retrieval is possible in the genomics domain (Hersh et al. 2005), and it is not a stretch to imagine adopting these technologies for clinical tasks. Using a separate search engine would provide other benefits as well: Greater control over the document retrieval process would allow one to examine the effects of different indexing schemes, different query operators, and techniques such as query expansion; see, for example, Aronson, Rindflesch, and Browne (1994). Finally, yet another way to solve the document retrieval problem is to eliminate that stage completely. Recall that our two-stage architecture was a practical expediency, because we did not have access to the computing resources necessary to pre-extract PICO elements from the entire MEDLINE database and directly index the results. Given access to more resources, a system could index identified PICO elements and directly match queries against a knowledge store.

Finally, answer generation remains an area that awaits further exploration, although we would have to first define what a good answer should be. We have empirically verified that an extractive approach based on outcome sentences is actually quite satisfactory, but our algorithm does not currently integrate evidence from multiple

abstracts; although see Demner-Fushman and Lin (2006). Furthermore, the current answer generator does not handle complex issues such as contradictory and inconsistent statements. To address these very difficult challenges, finer-grained semantic analysis of medical texts is required.

13. Conclusion

Our experiments in clinical question answering provide some answers to the broader research question regarding the role of knowledge-based and statistical techniques in advanced question answering. This work demonstrates that the two approaches are complementary and can be seamlessly integrated into algorithms that draw from the best of both worlds. Explicitly coded semantic knowledge, in the form of UMLS, and software for leveraging this resource—for example, MetaMap—combine to simplify many knowledge extraction tasks that would be far more difficult otherwise. The respectable performance of our population, problem, and intervention extractors, all of which use relatively simple rules, provides evidence that complex clinical problems can be tackled by appropriate use of ontological knowledge. Explicitly coded semantic knowledge is less helpful for outcome identification due to the large variety of possible "outcomes;" nevertheless, knowledge-rich features can be combined with simple, statistically derived features to build a good outcome classifier. Overall, this work demonstrates that the application of a semantic domain model yields clinical question answering capabilities that significantly outperform presently available technology, especially when coupled with traditional statistical methods (classification, evidence combination, etc.).

We have taken an important step in building a complete question-answering system that assists physicians in the patient care process. Our work demonstrates that the principles of evidence-based medicine can be computationally captured and implemented in a system, and although we are still far from operational deployment, these positive results are certainly encouraging. Information systems in support of the clinical decision-making process have the potential to improve the quality of health care, which is a worthy goal indeed.

Acknowledgments

We would like to thank Dr. Charles Sneiderman and Dr. Kin Wah Fung for the evaluation of the answers. For this work, D. D-F. was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine. For this work, J. L. was supported in part by a grant from the National Library of Medicine, where he was a visiting researcher during the summer of 2005. We would like to thank the anonymous reviewers for their valuable comments. J. L. would like to thank Kiri and Esther for their kind support.

References

Ad Hoc Working Group for Critical Appraisal of the Medical Literature. 1987. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine*, 106:595–604.

Aronson, Alan R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceeding of the 2001 Annual Symposium of the American Medical Informatics Association (AMIA 2001)*, pages 17–21, Portland, OR.

Aronson, Alan R., James G. Mork, Clifford W. Gay, Susanne M. Humphrey, and Willie J. Rogers. 2004. The NLM Indexing Initiative's Medical Text Indexer. In *Proceedings of the 11th World Congress on Medical Informatics*

(MEDINFO 2004), pages 268–272, San Francisco, CA.

Aronson, Alan R., Thomas C. Rindflesch, and Allen C. Browne. 1994. Exploiting a large thesaurus for information retrieval. In *Proceedings of RIAO 1994: Intelligent Multimedia Information Retrieval Systems and Management*, pages 197–216, New York.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press, New York.

Barry, Carol and Linda Schamber. 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34(2/3):219–236.

Booth, Andrew. 2000. Formulating the question. In Andrew Booth and Graham Walton, editors, *Managing Knowledge in Health Services*. Library Association Publishing, London, England.

Chambliss, M. Lee and Jennifer Conley. 1996. Answering clinical questions. *The Journal of Family Practice*, 43:140–144.

Cogdill, Keith W. and Margaret E. Moore. 1997. First-year medical students' information needs and resource selection: Responses to a clinical scenario. *Bulletin of the Medical Library Association*, 85(1):51–54.

Covell, David G., Gwen C. Uman, and Phil R. Manning. 1985. Information needs in office practice: Are they being met? *Annals of Internal Medicine*, 103(4):596–599.

De Groote, Sandra L. and Josephine L. Dorsch. 2003. Measuring use patterns of online journals and databases. *Journal of the Medical Library Association*, 91(2):231–240.

Demner-Fushman, Dina, Barbara Few, Susan E. Hauser, and George Thoma. 2006. Automatically identifying health outcome information in MEDLINE records. *Journal of the American Medical Informatics Association*, 13(1):52–60.

Demner-Fushman, Dina and Jimmy Lin. 2005. Knowledge extraction for clinical question answering: Preliminary results. In *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*, pages 1–10, Pittsburgh, PA.

Demner-Fushman, Dina and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 841–848, Sydney, Australia.

Ebell, Mark H., Jay Siwek, Barry D. Weiss, Steven H. Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. 2004. Strength of Recommendation Taxonomy (SORT): A patient-centered approach to grading evidence in the medical literature. *The Journal of the American Board of Family Practice*, 17(1):59–67.

Elhadad, Noemie, Min-Yen Kan, Judith Klavans, and Kathleen McKeown. 2005. Customization in a unified framework for summarizing medical literature. *Journal of Artificial Intelligence in Medicine*, 33(2):179–198.

Ely, John W., Jerome A. Osheroff, Mark H. Ebell, George R. Bergus, Barcey T. Levy, M. Lee Chambliss, and Eric R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319:358–361.

Ely, John W., Jerome A. Osheroff, M. Lee Chambliss, Mark H. Ebell, and Marcy E. Rosenbaum. 2005. Answering physicians' clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association*, 12(2):217–224.

Fiszman, Marcelo, Thomas C. Rindflesch, and Halil Kilicoglu. 2004. Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT/NAACL 2004 Workshop on Computational Lexical Semantics*, pages 76–83, Boston, MA.

Gorman, Paul N., Joan S. Ash, and Leslie W. Wykoff. 1994. Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association*, 82(2):140–146.

Haynes, R. Brian, Nancy Wilczynski, K. Ann McKibbon, Cynthia J. Walker, and John C. Sinclair. 1994. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the American Medical Informatics Association*, 1(6):447–458.

Hearst, Marti A. 1996. Improving full-text precision on short queries using simple constraints. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR 1996)*, pages 217–232, Las Vegas, NV.

Hersh, William, Ravi Teja Bhupatiraju, and Sarah Corley. 2004. Enhancing access to the bibliome: The TREC genomics track. In *Proceedings of the 11th World Congress on Medical Informatics (MEDINFO 2004)*, pages 773–777, San Francisco, CA.

Hersh, William, Aaron Cohen, Jianji Yang, Ravi Teja Bhupatiraju1, Phoebe Roberts, and Marti Hearst. 2005. TREC 2005 genomics track overview. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD.

Hildebrandt, Wesley, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*, pages 49–56, Boston, MA.

Hirschman, Lynette and Robert Gaizauskas. 2001. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300.

Huang, Xiaoli, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of PICO as a knowledge representation for clinical questions. In *Proceeding of the 2006 Annual Symposium of the American Medical Informatics Association (AMIA 2006)*, pages 359–363, Washington, D.C.

Ingwersen, Peter. 1999. Cognitive information retrieval. *Annual Review of Information Science and Technology*, 34:3–52.

Jacquemart, Pierre and Pierre Zweigenbaum. 2003. Towards a medical question-answering system: A feasibility study. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *The New Navigators: From Professionals to Patients*, volume 95 of *Actes Medical Informatics Europe, Studies in Health Technology and Informatics*. IOS Press, Amsterdam, pages 463–468.

Kauffman, Ralph E., L. A. Sawyer, and M. L. Scheinbaum. 1992. Antipyretic efficacy of ibuprofen vs acetaminophen. *American Journal of Diseases of Children*, 146(5):622–625.

Light, Marc, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of the BioLink 2004 Workshop at HLT/NAACL 2004*, pages 17–24, Boston, MA.

Lin, Jimmy and Dina Demner-Fushman. 2005a. Automatically evaluating answers to definition questions. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 931–938, Vancouver, Canada.

Lin, Jimmy and Dina Demner-Fushman. 2005b. Evaluating summaries and answers: Two sides of the same coin? In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 41–48, Ann Arbor, MI.

Lin, Jimmy and Dina Demner-Fushman. 2006a. The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 99–106, Seattle, WA.

Lin, Jimmy and Dina Demner-Fushman. 2006b. Will pyramids built of nuggets topple over? In *Proceedings of the 2006 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2006)*, pages 383–390, New York.

Lin, Jimmy, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BioNLP'06)*, pages 65–72, New York.

Lin, Jimmy, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. 2003. What makes a good answer? The role of context in question answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*, pages 25–32, Zürich, Switzerland.

Lindberg, Donald A., Betsy L. Humphreys, and Alexa T. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291.

McCray, Alexa T., Anita Burgun, and Olivier Bodenreider. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, pages 216–220, London, England.

McKeown, Kathleen, Noemie Elhadad, and Vasileios Hatzivassiloglou. 2003. Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proceedings of the 3rd ACM/IEEE Joint Conference on Digital Libraries (JCDL 2003)*, pages 159–170, Houston, TX.

McKnight, Larry and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *Proceeding of the 2003 Annual Symposium of the American*

Medical Informatics Association (AMIA 2003), pages 440–444, Washington, D.C.

Meadow, Charles T., Barbara A. Cerny, Christine L. Borgman, and Donald O. Case. 1989. Online access to knowledge: System design. *Journal of the American Society for Information Science*, 40(2):86–98.

Mendonça, Eneida A. and James J. Cimino. 2001. Building a knowledge base to support a digital library. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, pages 222–225, London, England.

Mladenic, Dunja and Marko Grobelnik. 1999. Feature selection for unbalanced class distribution and Naïve Bayes. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, pages 258–267, Bled, Slovenia.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*, pages 145–152, Boston, MA.

Niu, Yun and Graeme Hirst. 2004. Analysis of semantic classes in medical text for question answering. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*, pages 54–61, Barcelona, Spain.

Pratt, Wanda and Meliha Yetisgen-Yildiz. 2003. A study of biomedical concept identification: MetaMap vs. people. In *Proceeding of the 2003 Annual Symposium of the American Medical Informatics Association (AMIA 2003)*, pages 529–533, Washington, D.C.

Richardson, W. Scott, Mark C. Wilson, James Nishikawa, and Robert S. Hayward. 1995. The well-built clinical question: A key to evidence-based decisions. *American College of Physicians Journal Club*, 123(3):A12–A13.

Rinaldi, Fabio, James Dowdall, Gerold Schneider, and Andreas Persidis. 2004. Answering questions in the genomics domain. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*, pages 46–53, Barcelona, Spain.

Rindflesch, Thomas C. and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477.

Sackett, David L., Sharon E. Straus, W. Scott Richardson, William Rosenberg, and R. Brian Haynes. 2000. *Evidence-Based Medicine: How to Practice and Teach EBM*, second edition. Churchill Livingstone, Edinburgh, Scotland.

Saracevic, Tefko. 1975. Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343.

Sneiderman, Charles, Dina Demner-Fushman, Marcelo Fiszman, and Thomas C. Rindflesch. 2005. Semantic characteristics of MEDLINE citations useful for therapeutic decision-making. In *Proceeding of the 2005 Annual Symposium of the American Medical Informatics Association (AMIA 2005)*, page 1117, Washington, D.C.

Tbahriti, Imad, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. 2006. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the MEDLINE digital library. *International Journal of Medical Informatics*, 75(6):488–495.

Ting, Kai Ming and Ian H. Witten. 1999. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289.

Voorhees, Ellen M. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, Gaithersburg, MD.

Voorhees, Ellen M. and Dawn M. Tice. 1999. The TREC-8 question answering track evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 83–106, Gaithersburg, MD.

Wilczynski, Nancy, K. Ann McKibbon, and R. Brian Haynes. 2001. Enhancing retrieval of best evidence for health care from bibliographic databases: Calibration of the hand search of the literature. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, pages 390–393, London, England.

Yang, Yiming and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997)*, pages 412–420, Nashville, TN.