Question Answering in Biomedicine

-An evaluation of third party search engines-

Student: Andreea Tutos
Id : 41064739

Supervisor: Diego Molla

Abstract

The internet has become a source of wealth of medical information which is made available to practitioners mainly through search engines. The biological and medical research has experienced rapid development during the last decade, with the number of biomedical literature exponentially increasing to the point where thousands of new articles are published daily world wide. Every day patient treatment decisions are impacted by the ability to locate this information available on different web sites. This report describes the evaluation of the answerability of a set of clinical questions posed by physicians. The clinical questions have been identified as belonging to two categories of the five leaf high level hierarchical Evidence Taxonomy created by Ely and his colleagues: Intervention and Non Intervention. This study employs the services of five search engines (PubMed, Google, MedQA, Answers.com BrainBoost and OneLook) for locating the answers to the question corpus. The evaluation criteria include quality of answers, ranked according to the link's position in the response list and also relevance to the question. The results show that No Intervention questions seem to be easier to answer than Intervention questions and that, for answers extracted from scientific medical articles that follow a set standard structure, there is a probability of 50% that the answer is located in the abstract section.

Acknowledgements

I would like to thank Diego Molla - Aliod, my project supervisor, for his valuable and timely comments provided during the research carried out for this project. His guidance has greatly enhanced the quality of the project work.

Table of Contents

1. Introduction 5
2. Background 7
  2.1. Domain knowledge sources 7
  2.2. Question corpus 9
    2.2.1. PICO format 9
    2.2.2. Question corpus sources 10
  2.3. Search Engines and Question Answering systems 11
    2.3.1. Generic Search Engines and Question Answering systems 11
    2.3.1. Medical Search Engines and Question Answering systems 14
  2.4. Question analysis 17
    2.4.1 Question classification and the Evidence taxonomy 17
    2.4.2 Query analysis 18
  2.5. Answer extraction 19
3. Evaluation Methodology 21
  3.1. Question corpus 21
  3.2. Question processing 23
  3.3. Answer extraction 25
4. Results 26
5. Discussion and conclusions 28
6. Future work 30
References 31
Appendices 33

1. Introduction

Information overload is one of the most widely felt problems in the modern society. Medical professionals face today an overwhelming amount of information they need to be aware of in order to provide high quality medical care. Almost every aspect of patient care relies on searches of published medical articles and evidence from previous work of other physicians. This practice relates to the latest clinical guides that urge physicians to practice Evidence Based Medicine when providing care for their patients (Yu et al, 2005). Evidence Based Medicine implies referring to the best evidence from scientific and medical research that can assist in making decisions about patient care.

A survey of over 100 senior physicians and 625 primary-care physicians revealed that 66% of them reported the volume of scientific information available as unmanageable (Craig et al, 2001). In this context, researchers and physicians have to rely on search engine technologies like Google or PubMed in locating documents relevant to their work. These are Information Retrieval Systems that will typically return a large list of documents as a response to a query.

Physicians have limited time to browse the documents retrieved by Information Retrieval systems. Studies have revealed that any such search task that takes longer than an average of 2 to 3 minutes will be abandoned (Ely et al, 1999). The task of accessing relevant information in the required time frame seems to almost constantly fail, leaving most physicians questions without an answer. An Australasian survey named two main impediments in maximizing the utility of research data: insufficient time, reported by 74% of the study's respondents and limited access to evidence, coming next with 43% (Craig et al, 2001). As a result, automatic methods such as question answering technologies are becoming a real necessity.

In this context, question answering systems are a step beyond conventional Information Retrieval systems as they analyze large collections of documents in order to generate short and concise answers in response to input questions.

Craig et al (2001) defines four steps to follow when attempting to incorporate the best available research evidence in everyday physician's decision making: formulating answerable questions, accessing the best information, evaluating the information's validity and relevance and applying the newly acquired information. Our project addresses the first two steps by aiming to determine the answerability of a set of 50 medical questions sourced from the Parkhurst Exchange[1] website. In order to achieve our goal, these questions have been processed through a list of five selected search engines: PubMed[2], Google, MedQA[3], Answers.com BrainBoost[4] and OneLook[5].

Lowering the barriers to the use of evidence based knowledge has the potential of improving the quality of patient consultation at the point of care.

[1] See http://www.parkhurstexchange.com
[2] See http://www.ncbi.nlm.nih.gov/pubmed/
[3] See http://monkey.ims.uwm.edu:8080/MedQA/
[4] See http://www.answers.com/bb/
[5] See http://www.onelook.com

A US study shows that 49% of family physicians errors are due to a lack of knowledge about medical aspects of the case (Ely et al, 1995). Another study revealed that medical practitioners have two questions for every three patients seen and 40% of their clinical questions were on medical facts (Covell et al, 1985). The research goes even further discovering that family physicians do not pursue answers for 64% of their clinical questions. Out of those questions, 80% have answers that could be located (Ely et al, 1995). All these statistics provided highlight the importance of the research that focuses on developing tools that would help physicians find answers to their clinical questions.

Our project takes one of the first steps towards creating a medical question answering system. Recognizing the importance of Evidence Based Medicine, our project goes further than other referenced studies by not limiting the corpus of questions to definitional questions. Definitional questions have been described as having the format of "What is a/an X?".

A question-answering system available for practitioners at the point of care could significantly improve the quality of patient care. Being able to have access to information about similar symptoms that have already been analyzed and diagnosed by other physicians before would save significant time and also potentially ensure that less patients receive incorrect or low quality medical care.

Biomedicine, which is usually more concerned with theory, knowledge and research, is the foundation of all medical applications, diagnosis and treatments. Research in biomedicine can result in discovering new drugs and a better understanding of complicated diseases. Question answering systems could potentially have a significant impact in this context as well, saving time spent on documents that might not be relevant and helping make the correct connections to previous research findings.

The structure of this document includes Section 2 which describes current studies and concepts related to our project. It details on the Evidence taxonomy hierarchy which is the foundation for our question classification and describes the search engines and question answering systems involved in the study.

Section 3 introduces the evaluation methodology employed in our project. The methodology explains the corpus of questions selection process and the transformation of candidate questions in order to maximize the search outcomes. It also describes the reasons behind the selection of the search engines and presents the answer extraction methodology.

Section 4 presents the findings of our research, the results on the answerability of corpus of questions and statistics on answer location, where available.

Section 5 analyzes the results in an attempt to explain our findings and draws the conclusions on our project's success.

Section 6 sets future paths to follow for the research of our project.

2. Background

During the last few years, research efforts have been largely directed towards Question Answering systems on open domains. Question answering has been mainly driven by TREC[6] (Text Retrieval conference), with aims to provide the infrastructure for large-scale evaluation of text retrieval methodologies. TREC introduced the question answering track in 1999 and has since reported on answering systems performances for factoid, list and scenario questions.

Open domains have had a lot of efforts channeled towards them as researchers believed that they are more difficult to deal with than narrow domains. The biomedical domain received considerably less research efforts, with fewer research groups working on medical specific question answering systems (Zweigenbaum, 2003).

The attempt of answering questions over the biomedical field is considered an exception as its scope is quite broad and not just a simplification of open domain question answering (Zweigenbaum, 2003). An attempt to develop a medical QA system will have to include a rigorous selection of its knowledge sources and a thorough analysis of the questions in order to provide a valid answer.

This section introduces some of the main current methods and techniques used in the biomedical question answering domain and their results so far. It also briefly describes the search engines and question answering systems, grouped according to their area of expertise: generic and medical.

2.1. Domain knowledge sources

The British Medical Journal BMJ[7], an online international medical journal, has under its 'Editorials' section, a motto which reflects a very valid statement: 'Having the evidence in your hand is just a start – but a good one'. This statement emphasizes the importance of source selection and extraction for a medical question answering system.

In his study 'Question Answering in biomedicine', Pierre Zweigenbaum (2003) approaches medical knowledge resources from the perspective of availability and trustworthiness. Due to the applicability in the medical domain, such question answering systems need to address the delicate question of the reliability of their medical resources. They should be able to determine the level of confidence assigned to each of their knowledge sources. Availability is approached in the context that processing of medical information has been around for quite a while and it has a long tradition in processing large sources, with health information gateways offering information for the biomedical domain.

Special attention has been given to the reliability of medical information in Zweigenbaum and Jacquemart's (2003) study. The authors suggest using quality

[6] See http://trec.nist.gov/
[7] See http://www.bmj.com

criteria such as NetScoring[8]. NetScoring defines a way to assess the quality of health internet information by including categories like credibility, accessibility, content, links and ethics. A weight is associated to each criterion and the total of the weighted criteria gives the overall score for a site.

Other authoritative medical resources are mentioned in the EPoCare (Evidence at Point of Care) project at the University of Toronto: **Clinical Evidence[9]** and **Evidence-based On Call[10]**. The article of Niu et al (2003) describes EPoCare as a system that relies on results for clinical problems published in Clinical Evidence and Evidence-based On Call. Their system takes the texts and stores them with XML markups in a XML database. Clinical Evidence website is a reliable medical source for decisions made on treatments and patient care, while Evidence-Based On Call is described as a good source of evidence-based summaries covering 38 on-call medical conditions.

Yu et al (2007) mentions in its introduction another medical knowledge database: UpToDate[11] available to physicians to use. Relating back to our project, it is not practical to refer to this resource as it is not publicly available on line. UpToDate offers some limited access to information as a patient, but not as a physician, it requires periodical paid subscriptions in order to take advantage of its full capabilities.

A knowledge source that seems to be used across all projects mentioned in this review is **MEDLINE[12]**. MEDLINE is a reputable medical repository maintained by the US National Library of Medicine (Demner-Fushman and Lin, 2007). The MEDLINE database includes over 15 million medical articles and has proven to be a well recognized knowledge source across medical question answering studies. Metadata is associated with the citations in the database, including basic elements such as title of article, authors, publication name and type, date of publication and also controlled vocabulary terms associated by human indexers. MEDLINE takes advantage of the services of a large team of professionals trained with at least bachelor's degree in sciences that sustain the indexing process. US National Library of Medicine's controlled vocabulary thesaurus is called MeSH (Medical Subject Headings). MeSH contains a hierarchical structure of approximately 23,000 descriptors and more than 151,000 additional chemical substance names (Demner-Fushman and Lin, 2007). Its role is to provide a consistent way of retrieving information for search strings using different terminology for the same concepts.

After consulting all the knowledge sources described, we have decided to include MEDLINE in our study as the database has a dedicated search engine specifically designed to access it, called PubMed. We will introduce PubMed in the Section 2.3 "Search Engines and Question Answering Systems".

[8] See http://www.chu-rouen.fr/netscoring/netscoringeng.html
[9] See http://clinicalevidence.bmj.com/ceweb/index.jsp
[10] See http://www.eboncall.org
[11] See http://www.uptodate.com
[12] See http://www.ncbi.nlm.nih.gov/pubmed/

2.2. Question corpus

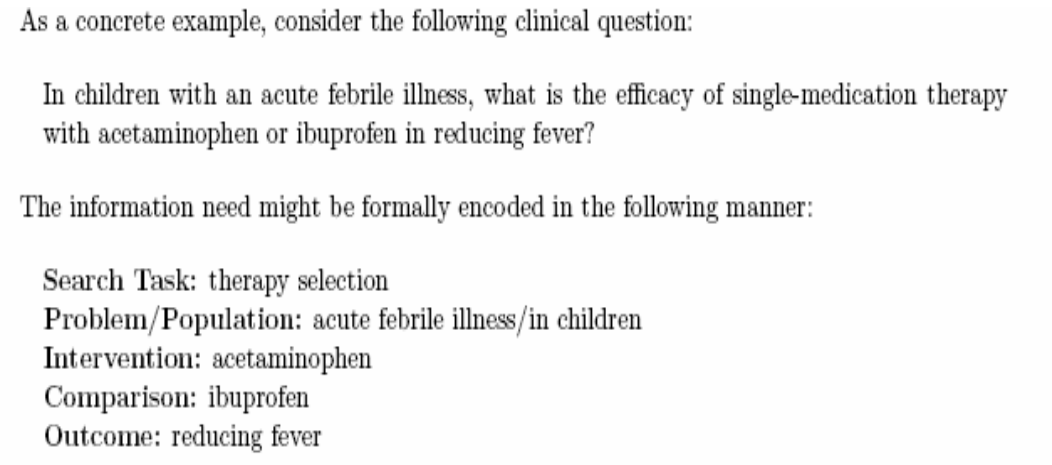
Across different projects and studies, different methodologies are employed to gather or create the corpus of questions that will be the input to the question answering systems.

2.2.1. PICO format

The input question's format is a critical decision that can impact the success of a question answering system. Previous experience has shown that queries poorly formulated are one of the main obstacles in the process of answer location (Demner - Fushman and Lin, 2007). There is some debate around the best input to a clinical question answering system. The options suggested by the reviewed literature are grouped around two categories: natural language questions and PICO questions.

The PICO format has four components that reflect the key aspects of patient care: the primary problem (P), the main intervention (I), the main intervention comparison (C) and the outcome of the intervention (O). In their study, Demner - Fushman and Lin (2007) advocate the PICO format and the advantages of having the information translated into a frame based representation. The two main advantages are that formal representations help physicians "think through" their question ensuring important elements are not missed and that questions already structured eliminate the need for linguistic analysis, avoiding possible ambiguities that may negatively impact answering performance.

In the initial phase of our project we have spent time to determine our best options in relation to questions format and considered the PICO alternative, but we have later on come to the conclusion that we did not have resources with adequate medical knowledge to perform question transformation. An example (Figure 1) of a clinical question in its original natural language format and then transformed into PICO format is presented in Demner - Fushman and Lin's (2007) study.



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

Figure 1: Example of natural language clinical question transformation to PICO format (Demner - Fushman and Lin, 2007).

This is a time consuming process and it requires medical knowledge in order to be able to perform a correct transformation. As none of the two requirements were fully met by the members of our project team, we have decided to build our corpus of questions following the natural language format option.

2.2.2. Question corpus sources

Due to the high degree of complexity of question answering process in the medical domain, some studies aim to enable their question answering systems to address only specific types of questions.

Zweigenbaum and Jacquemart's (2003) mention that they have collected the majority of their questions from clinical students, completed with questions from student textbooks in Stomatology, in an attempt to ensure a balanced coverage of domains. As a result, they had to perform some conversions of questions to a 'canonical form', in order to simplify the process of deriving general question patterns. This process of converting questions was not an option for our project as we did not have the required level of medical knowledge so we had to turn to other options in regards to constructing our corpus of questions.

During our research, we have come accross a few sources of clinical questions. Yu et al (2007) have evaluated the MedQA system based on a set of medical questions sourced from the Clinical Questions Collection website[13]. The website contains thousands of medical questions collected at clinic settings, clearly and explicitly formulated, avoiding complex sentences.

Demner-Fushman and Lin (2007) gathered their medical questions from Parkhurst Exchange website and from The Journal of Family Practice[14].

The corpus of questions of our study has been constructed from the questions and answers list available on the Parkhurst Exchange website. Parkhurst Exchange is a highly regarded medical publishing website based in Canada that includes a collection of over 4800 clinical questions and their answers provided by physicians. Since 1983 when it first started, it continues to develop strong relationships with top physicians across many medical disciplines.

Our reasons behind the final choice of Parkhurst Exchange were that it provided both clinical questions and their answers and also had a very user friendly interface. As shown in Figure 2, from the main website's page, the entire questions and answers collection is only one click away (the "Search Q&A" button).

[13] See http://clinques.nlm.nih.gov
[14] See http://www.jfponline.com/

Figure 2: Parkhurst Exchange web site.

2.3. Search Engines and Question Answering systems

Our work is related to the study of Yu and Kaufman (2007) who conducted a cognitive evaluation of four online engines on answering definitional questions. Our study does not limit the input questions to definitional questions only, exploring further by including questions belonging to the Evidence node in the Evidence Taxonomy, on which we detail in Section 2.4. Yu and Kaufman's evaluation criteria included quality of answers, ease of use, time spent and number of actions taken to locate an answer. Their results showed that PubMed performed poorly, Google was the preferred system for quality of answer and ease of use and MedQA surpassed Google in time spent and number of actions. The findings of this study contradict the work of Berkowitz (Berkowitz, 2002) who concluded in his research that Google performed poorly in regards to quality of answers as it referred to consumer-oriented sites.

2.3.1. Generic Search Engines and Question Answering systems

Google is a widely used web search engine that uses text matching techniques to locate web pages relevant to a user's search. Google's architecture includes a list of features that make it a high-precision search engine. First to be mentioned is the ability to determine quality rankings or PageRanks for each web page based on the link structure of the Web. PageRank offers an objective measure of citation's importance and it proved to be a great resource for prioritizing the results of web keyword searches. Another important characteristic of Google is that

it establishes a relation between the text of links and the pages the links point to, this way making use of anchor's advantages: more accurate descriptions of web pages and access to links to web pages that have not actually been crawled (Brin and Page, 1998).

Google (Figure 3) was included in our study as two different entities: the standard Google and Google pointed towards the PubMed database. Our justification for this approach was the observation that Google returned quite often information from consumer-oriented web sites, as opposed to scientific articles and publications. To make results interpretation more accurate, Google was pointed to search for information against PubMed (MEDLINE) database, ensuring compatibility with the results provided by the PubMed search engine itself.

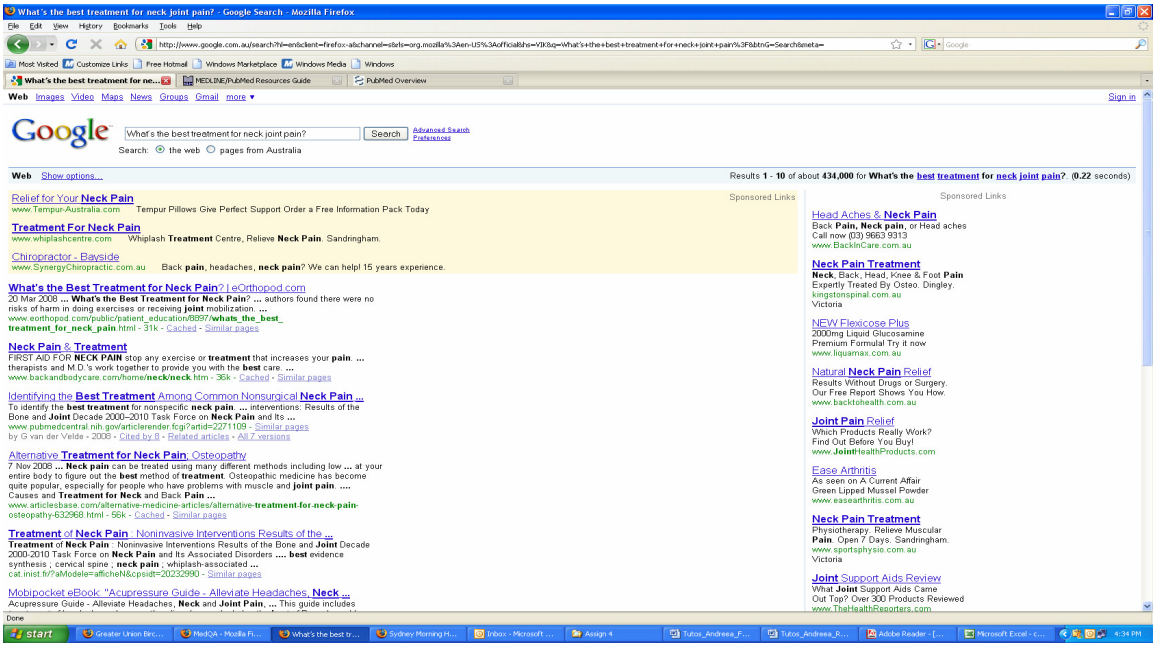


Figure 3: Google main page

Answers.com is a website that can offer answers to categories of questions in twenty domains such as business, health, travel, technology, science, entertainment or arts, all drawn from high-quality sources. Their collection includes over four million answers drawn from over 180 titles from brand-name publishers, together with content created by their own editorial team. The website is operated by the Answers Corporation founded in 1999 and provides patented technology and software tools that present concise information in response to user queries. Answers.com seems to be very effective particularly for "Who/What is" questions.

In December 2005 Answers.com has acquired BrainBoost, as the next step in their development plan to add natural language mechanisms to their existing search capabilities. They have opted for this strategy as they evaluated that existing experience and development efforts at BrainBoost would help them reach the next stage faster than building similar artificial intelligence technology themselves. BrainBoost technology for parsing and answering natural language queries extracts

candidate answers, ranks them heuristically and displays the top ranked results in simple English form. BrainBoost technology compliments Answers.com own technology providing answers to users who enter just search keywords or more complex natural language questions. Figure 4 shows the Answers.com BrainBoost question answering engine main page.

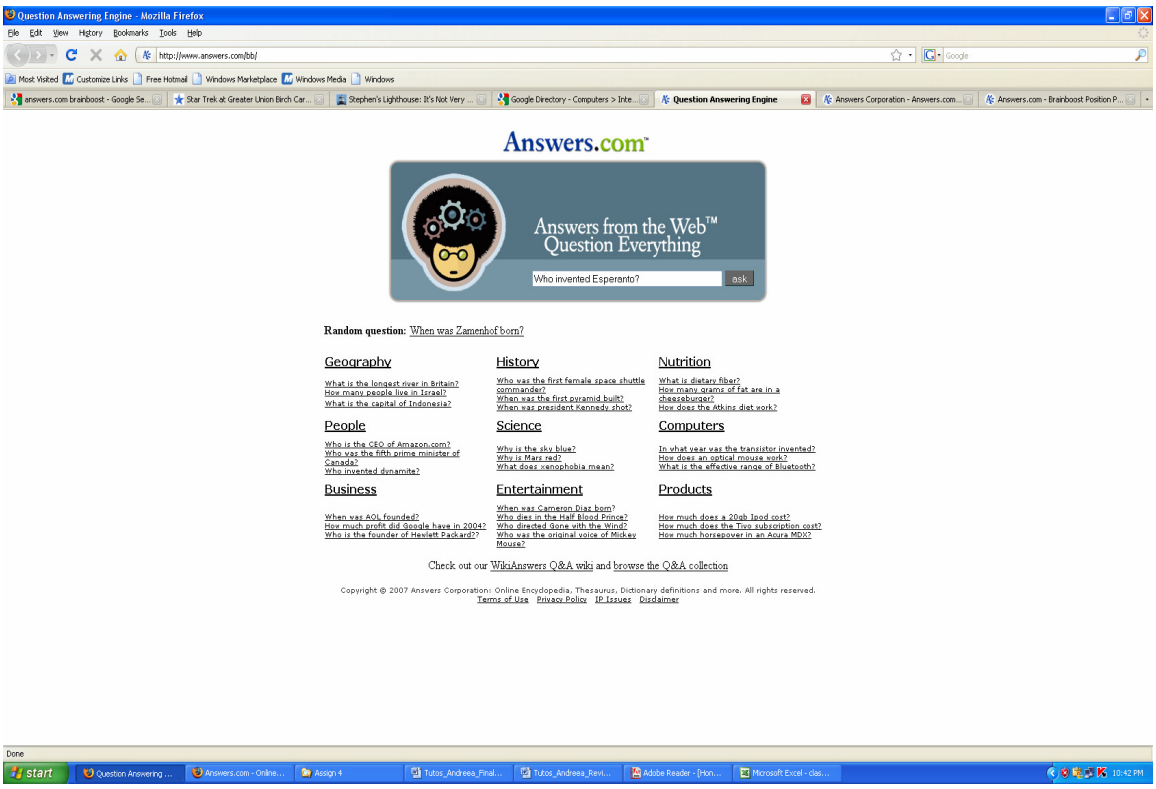


Figure 4: Answers.com BrainBoost search page

OneLook (see Figure 5) free dictionary search service was launched in 1996. It is a dictionary and translation metasearch engine that accesses more than 900 online dictionaries in order to locate the desired definition. It offers the ability to decide on the dictionary to focus on, with choices of domains as medical, art, business, etc. OneLook describes itself as a search engine for words and phrases that also offers a reverse dictionary which enables users to search for words and phrases related to a certain concept.

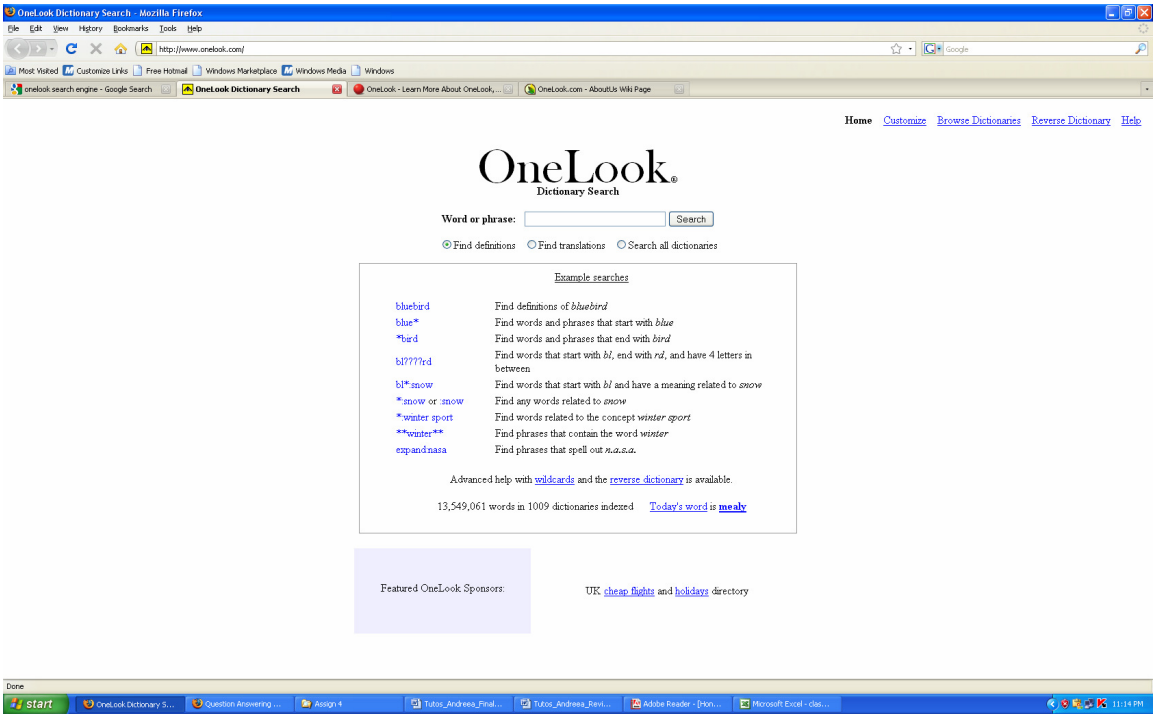


Figure 5: OneLook search page

We have selected the search engines to include in our project based on a few guidelines. They needed to be available online and free of charge and also able to accept natural language questions. Although the initial project plan considered the possibility of transforming the questions into PICO format, this idea was later postponed due to the lack of resources. Without the option of mapping the input questions to the PICO format, selecting search engines that accepted natural language questions became a must.

2.3.2. Medical Search Engines and Question Answering systems

PubMed is a search engine that accesses the MEDLINE database and was developed by the US National Center for Biotechnology Information (NCBI). It accesses citations from biomedical literature and also links to other molecular biology resources. A screen shot of the PubMed home page is shown in Figure 6.

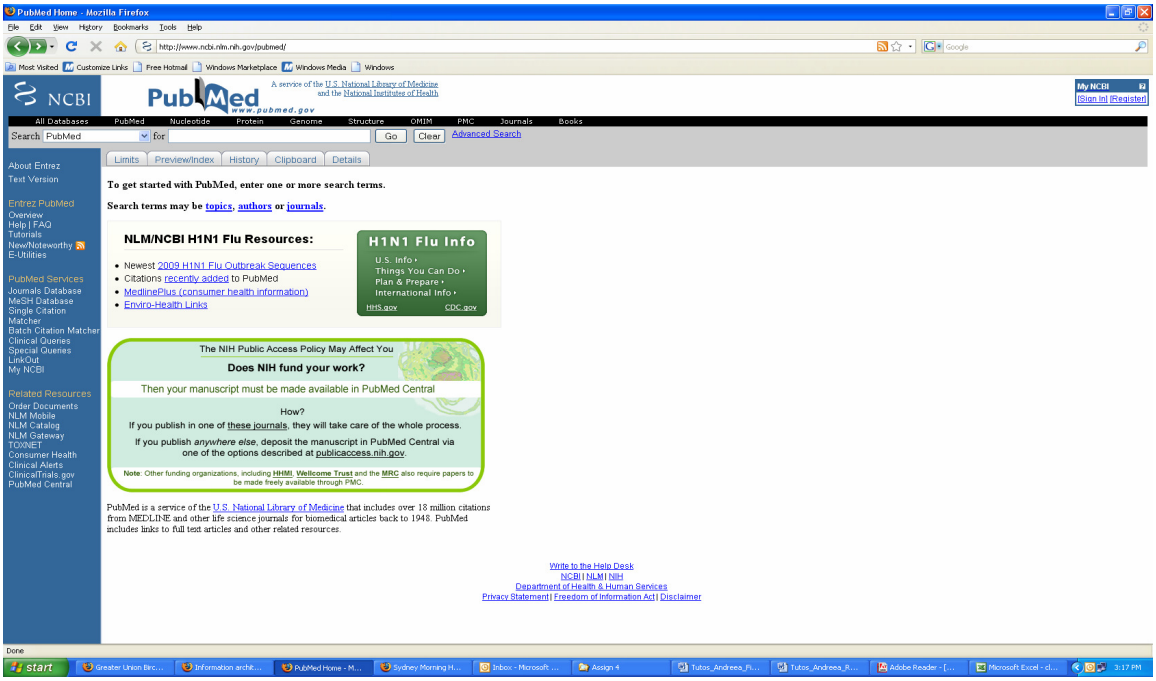


Figure 6: PubMed search engine – home page.

In addition to MEDLINE, which is the main component, PubMed also contains clinical queries, special queries search filters and links to many sites providing full text articles and other related sources.

In their study, Demner-Fushman and Lin (2007) mention the debate around PubMed and its ability to supply relevant answers in a reasonable time frame, arguing that Information Retrieval technology is not performing satisfactorily on MEDLINE. This is in line with the observation of Plikus et al (2006) that concluded that PubMed does not produce well classified search outputs and proposed PubFocus[15] as a web server that helps ranking by adding publication quality attributes. PubFocus calculates and analyzes statistics of the MEDLINE/PubMed search queries taking advantage of additional information from databases that rank scientific journals. It provides a means to determine the quality of each author's research, determining its impact on a particular research field.

After processing our corpus of questions through the selected search engines, including PubMed, we should be able to draw our own conclusions in regards to PubMed's performance and compare our findings with Demner-Fushman and Lin's (2007).

MedQA (Yu et al, 2007) is one of the first developed medical answering systems that respond to definitional questions accessing the MEDLINE records and other World Wide Web collections. It automatically analyzes a large number of electronic documents in order to generate short and coherent answers in response to the input questions. The reason behind deciding on definitional questions is that they are 'more clear-cut' as opposed to other types of clinical questions that can

[15] See http://www.pubfocus.com/

have large variations in their expected answers. Figure 7 shows the MedQA front screen.

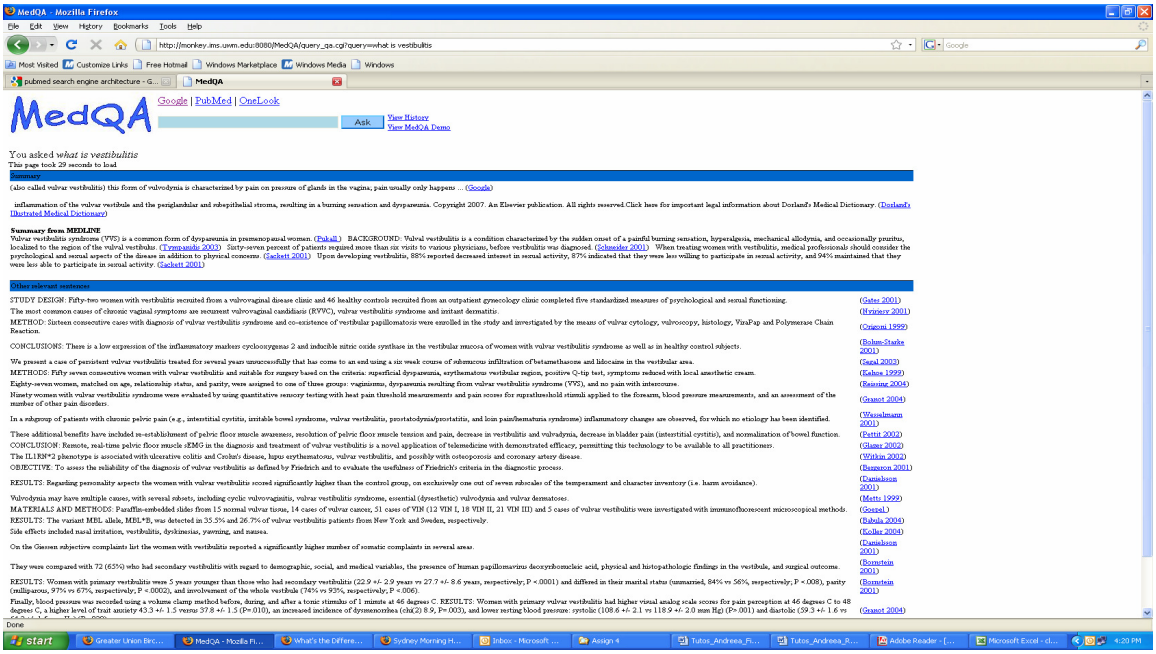


Figure 7: MedQA question answering.

MedQA relies on the IMRAD (Introduction, Methods, Results and Discussion) structure of biomedical articles to determine the relevance of an article to the search query (Yu et al, 2007). The medical question answering system is the first one to integrate four advanced techniques: question analysis, information retrieval, answers extraction and summarization techniques (Lee et al, 2006).

Figure 8 shows MedQA's system architecture.

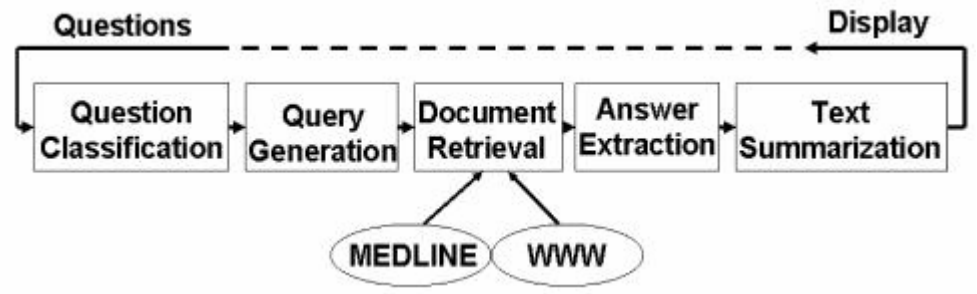


Figure 8: MedQA System Architecture (Lee et al, 2006).

MedQA takes a medical question and relies on the Question Classification module to determine the question type and locate the corresponding answer strategy developed. The classification module applies supervised machine-learning

techniques to classify the input question according to the Evidence Taxonomy described later in this report, under sub-section 2.4.1 "Question classification and the Evidence taxonomy".

During the Query Generation phase, the system uses the shallow parser LT CHUNK in order to identify nouns inside the medical question text. Shallow parsing is a natural language processing technique that partially analyzes a sentence and divides it into series of words that form a grammatical unit. The elements identified by the parser are then applied as the search query (Lee et al, 2006).

The Document Retrieval module uses nouns phrases as query terms to retrieve documents, indexing the MEDLINE database using the LUCENE search engine that returns ranked documents.

Then Answer Extraction module is called to identify sentences that contain answers for the input question. It identifies lexico-syntactic patterns by applying the Unified Medical Language System concepts as candidate terms (Lee et al, 2006).

The Summarization module condenses the text into a shorter version that contains the same information. It uses hierarchical clustering algorithms to group sentences based on similarity and selects the most representative from each cluster (Lee et al, 2006).

2.4. Question analysis

The question analysis phase is highlighted across all studies and articles as a critical stage of the question answering process. It is responsible of classifying the questions in order to determine the information that is the answer to the question and also selecting the elements that help locating candidate documents (Molla and Vicedo, 2009)

2.4.1 Question classification and the Evidence taxonomy

Question classification aims to determine the type of answer a question is expecting and then develop a specific answer strategy for each type. A system needs to understand a question first and then attempt to answer it correctly. The challenges introduced by the medical domain relate to the complexity of medical questions, where determining the answer type is not always enough for assessing what the question is about. Niu et al (2003) mentions the 'question focus' as the extra information contained in the question that is needed to understand the goal.

Our project refers to the Evidence taxonomy created by Ely and his colleagues (Yu and Sable, 2005). This high level, five leaf hierarchy categorizes medical questions that are potentially answerable with evidence. The hierarchy is presented in Figure 9.

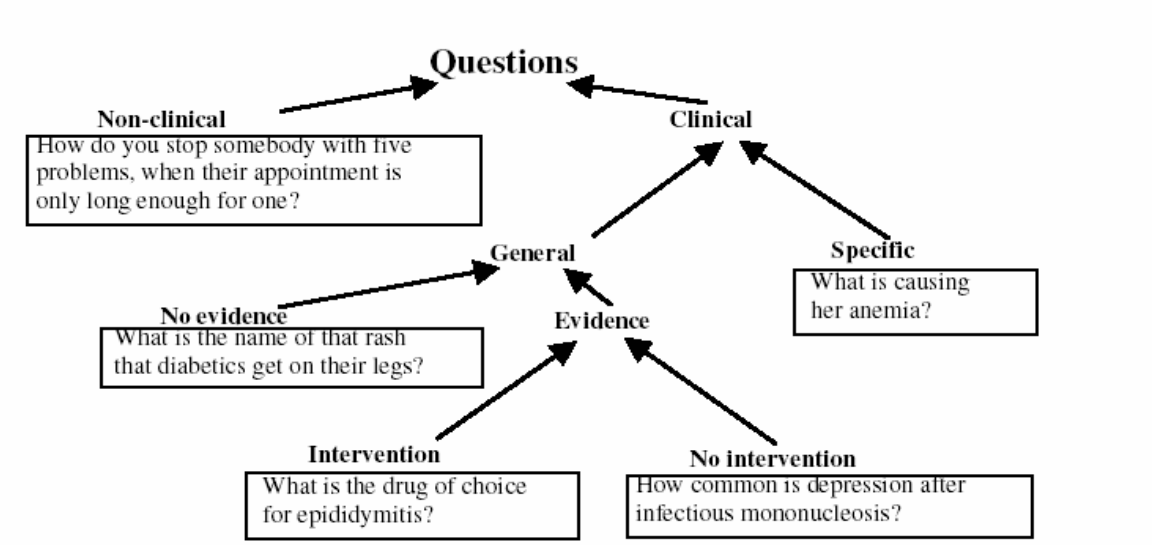


Figure 9: "Evidence Taxonomy" created by Ely and his colleagues, with examples.

Ely and his colleagues have concluded that the Non-clinical, Specific and No Evidence questions are not answerable with evidence, while both categories of Evidence (Intervention and No Intervention) are potentially answerable. Non-clinical questions do not address the specific medical domain and Specific questions require information from the patient personal record.

We have focused on the two evidence categories confirmed by Ely's study as being answerable with evidence (Yu and Sable, 2005): Intervention and Non-Intervention questions. According to the Evidence taxonomy, intervention questions are scenario-based, quite complex and they require complex answers that provide descriptions of possible treatments or recommended drugs. Non-intervention questions usually enquire about medical conditions or drugs, without asking for directions in managing a disease. They generally belong to the family of factoid questions for which short answers are usually expected.

2.4.2 Query analysis

The other facet of question analysis involves extracting existing information in the question that might help with the selection of candidate text extracts.

The processing required here depends greatly on the question format opted for. For PICO formatted questions, the task of analyzing the question is shifted from the system side onto the physician's shoulders. They become responsible with the translation of information and this forces them to thoroughly analyze their questions and notice any missing possible elements or inaccuracies (Demner-Fushman and Lin, 2007). The system implication is that this eliminates the need for a linguistic analysis performed against the natural language question. The workload is this way assigned to the preparation of input questions stage. The system query is then built including the PICO elements of the question except for outcome, which is most of the time implicit.

Molla and Vicedo (2009) describe the two processes commonly involved in building the queries for natural language questions: keyword selection and answer-pattern generation. Keyword selection consists of selecting question terms whose appearance in a text indicates the possibility of locating a candidate answer inside those paragraphs. An example of a technique that performs this task is provided by the MedQA system that calls the LTCHUNK shallow parser in order to identify grammatical units such as noun groups, verbs or verbs groups and apply them as the search query (Lee et al, 2006).

Answer pattern generation is described as the process of generating different combinations of question terms that could lead to several forms of answers. Some of the generated query expressions might not be linguistically accurate, but they would still produce some results when fed to an information retrieval system (Molla and Vicedo, 2009).

In order to improve the query performance, Jacquemart and Zweigenbaum (2003) describe the collection of 'named-entities' that in the medical domain correspond to diagnosis, treatments, symptoms or diseases. To increase flexibility they also use morphologically derived words, synonyms hyponyms and hypernyms. This area seems to be very sensitive in the biomedical domain due to the fact that many medical substances or even diseases have various lexical forms. Same concepts are quite often expressed by different terms and expressions. Shi et al (2007) introduce the 'Concept Similarity Module' that tries to eliminate possible causes of query failure due to inadequate sentence similarity. The module identifies synonyms and hypernym – hyponym relations and performs word disambiguation.

The EpoCare project (Niu et al, 2003) has the approach that both a short and a long answer should be prepared for 'wh' questions, yes-no questions and no-answer questions. Their justification goes into the details of patient care needs dictated by various circumstances. The short answer is intended for physicians that need to make a quick decision. The long answer is addressed to situations when clinicians feel there is a need to go into details about other related experiments or evidence. EpoCare treats no-answer questions separately, on the assumption that the system can still provide some related information for physicians to have a starting point.

During our research and evaluation of the selected search engines we have noticed that, even if search engines such as PubMed should already be performing question transformation as part of their internal algorithm, we could still obtain better search results by manually applying some of the processing rules mentioned in this section, such as applying synonyms, hypernyms or hyponyms. We will elaborate on the query transformation later in this report, under section 3.2 "Question processing".

2.5. Answer extraction

Answer extraction identifies relevant sentences that answer the question and are located in the retrieved documents.

The MedQA system described by Yu et al (2007) relies on the IMRAD (Introduction, Methods, Results and Discussion) structure of biomedical articles to determine the relevance of an article. The authors noted that statistics show physicians usually refer to the 'Results' section in order to determine the relevance

of an article. They have also observed that for definitional questions, the relevance of a text can be judged by 'Introduction' and 'Background'.

The authors of MedQA had an interesting approach for extracting answers for their definitional questions. They have taken all the terms in UMLS and 'crawled' the web searching for the corresponding definitions. To perform this they have used the Google: Definition service that provides definitions from different web glossaries. The set of definitions obtained was used to identify patterns of definitional sentences. The result of this approach showed a good quality of definitions from the web, with tests showing that the similarity between definitions retrieved via Google:Definition and MEDLINE abstracts was very high, above a specific predefined threshold (Yu et al, 2007).

Molla and Vicedo (2009) describe a simple method to determine a list of candidate answers. The recommendation is to gather all named entities that exist in the text passage and then remove the ones that do not belong to the answer type, including the named entities in the question. In order to rank answer candidates, a combination of methods is promoted:

a) *Similarity with the question*: A simple way to determine the similarity of two texts is to calculate the number of words that they have in common considering relations such as synonymy, hypernymy or hyponymy. The correct answer is in a sentence which is very similar to the question and has compatibility with the expected answer type.

b) *Answer popularity:* The chances for a string to be the correct answer are directly proportional with the number of times that string was found as a possible answer.

c) *Answer patterns*: There is a finite number of patterns that describe ways of answering a question. The combinations of characters and punctuation marks that form a pattern can be manually determined by observing the expressions that constitute the answers. In order to test a pattern's precision, new searches are performed and statistics computed to determine how often a pattern is able to locate a correct answer.

d) *Answer validation*: Some basic detailed checks can be performed to determine if a candidate answer string is the correct answer. An example is that negative numbers cannot be accepted as answers to questions about age, distances or time elapsed.

The challenges of answer extraction for the medical domain have driven research to new approaches. Niu et al (2003) have found that using roles and role identification they had some success in locating answers to medical questions. The four roles introduced map to the PICO components, as the way information is presented in medical text usually corresponds to this format. Their methodology locates the four roles in the natural language question and candidate answers previously retrieved. To further determine if a candidate answer is the correct answer, the roles in the question are compared with the corresponding roles in the candidate answers. The complexity of this method comes from the fact that some roles correspond to named-entities (patient status, therapy), but some do not (diagnosis description or outcome). Medical named entities can be identified with the support of UMLS, but this process does not apply to non named-entities roles.

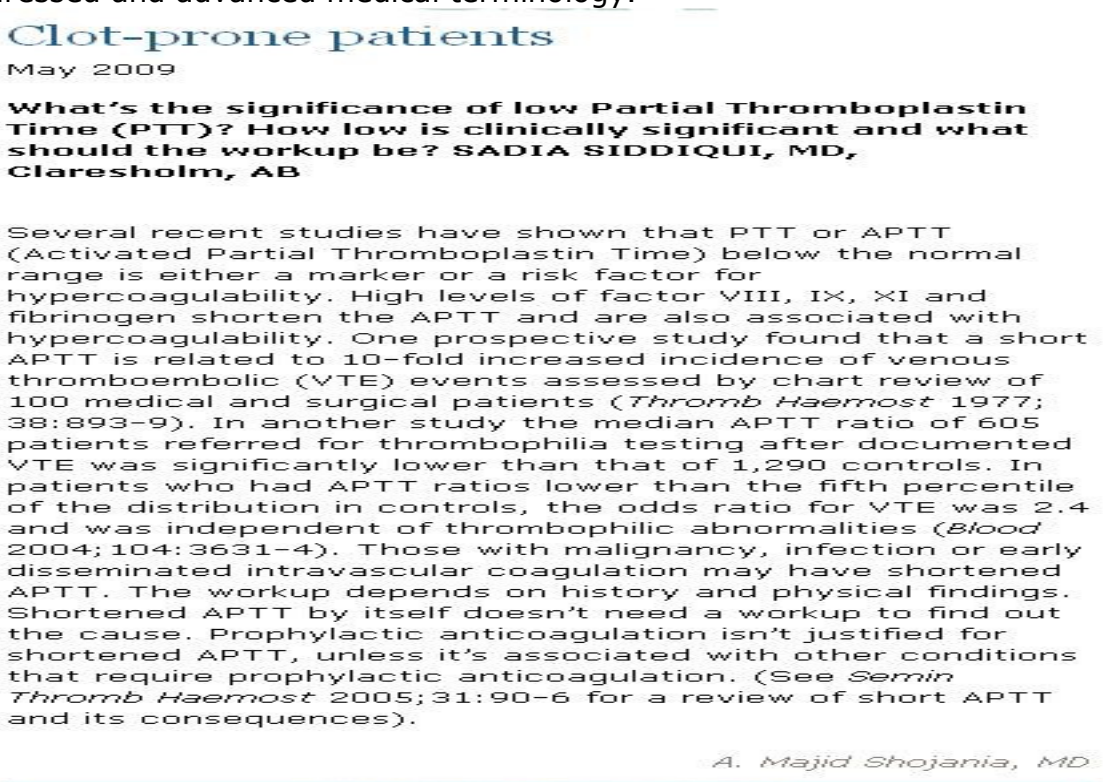
3. Evaluation Methodology

This section describes the steps and corresponding methodologies employed in order to achieve our project's goal. We will detail on the process followed for constructing our corpus of questions, explaining the reasons behind our question choices. We will then present the methodology followed for question processing and answer extraction.

3.1. Question corpus

Constructing the corpus of questions has proved to be a relatively complex process mainly due to the lack of medical background of our project team. To overcome this obstacle, we have decided to select those clinical questions that address relatively simple health issues and have no complicated medical language. The fact that Parkhurst Exchange website, our source of questions, provides both the questions and their answers had a positive impact on the research carried out for our project as it partially compensated the limited medical knowledge in our team.

The question selection process was time consuming, with an average of 6.8 questions and answers browsed per selected question. For a question to be included in our corpus of questions, the criteria to be met was simple medical language describing clinical cases that could be understood by an audience with low to medium medical knowledge. Figure 10 presents a clinical question sourced from the Parkhurst Exchange website and rejected due to the complexity of the medical case addressed and advanced medical terminology.



Clot-prone patients

May 2009

What's the significance of low Partial Thromboplastin Time (PTT)? How low is clinically significant and what should the workup be? SADIA SIDDIQUI, MD, Claresholm, AB

Several recent studies have shown that PTT or APTT (Activated Partial Thromboplastin Time) below the normal range is either a marker or a risk factor for hypercoagulability. High levels of factor VIII, IX, XI and fibrinogen shorten the APTT and are also associated with hypercoagulability. One prospective study found that a short APTT is related to 10-fold increased incidence of venous thromboembolic (VTE) events assessed by chart review of 100 medical and surgical patients (*Thromb Haemost* 1977; 38:893-9). In another study the median APTT ratio of 605 patients referred for thrombophilia testing after documented VTE was significantly lower than that of 1,290 controls. In patients who had APTT ratios lower than the fifth percentile of the distribution in controls, the odds ratio for VTE was 2.4 and was independent of thrombophilic abnormalities (*Blood* 2004;104:3631-4). Those with malignancy, infection or early disseminated intravascular coagulation may have shortened APTT. The workup depends on history and physical findings. Shortened APTT by itself doesn't need a workup to find out the cause. Prophylactic anticoagulation isn't justified for shortened APTT, unless it's associated with other conditions that require prophylactic anticoagulation. (See *Semin Thromb Haemost* 2005;31:90-6 for a review of short APTT and its consequences).

A. Majid Shojania, MD

Figure 10: Example of clinical question not included in our corpus of questions.

The website's medical questions are grouped in over 30 categories such as Psychiatry, Oncology, Pediatrics, Endocrinology, etc. In our selection process we have opted for the 'Browse All' categories option which lists all questions sorted descending based on the date they have entered the collection. A list of examples of selected questions is included in Table 1.

Question	Category
Is watermelon allergenic	No Intervention
When to introduce solids to infants	Intervention
Should family doctors be immunized with Pneumovax and Menactra or Menjugate	Intervention
Can cell phones cause cancer	No Intervention
How much folic acid — 400 µg, 1 mg, 5 mg — is recommended before conception and during pregnancy	Intervention
How to beat recurrent UTIs	Intervention
How to recognize autism in adults	No Intervention
Does skin colour affect vitamin D requirements	No Intervention

Table 1: Example of questions classified according to the Evidence taxonomy.

The source of our question corpus, Parkhurst Exchange, contains mainly clinical questions asked by family doctors. We assumed we should be able to map the selected questions onto the Evidence taxonomy tree. We have analyzed each question selected and found out that they could be classified as belonging to the 'Intervention' and 'Non-Intervention' categories.
Table 1 shows a snapshot of our corpus of questions and categories resulted from the classification process. A complete list of all our selected medical questions is included in Appendix A "Question Corpus".

We admit that the question selection process might have introduced bias in our corpus of questions and there is no guarantee that the proportion of questions reflects real life statistics. We could not locate any research studies to provide statistics on real life clinical question percentages resulted from classification using the Evidence Taxonomy. We could only relate to Ely et al (2000) study that classified 1396 clinical questions and obtained the distribution shown in Table 2, which includes the most common generic question types they have found, covering 63% of their corpus of questions. They have concluded that most frequent questions were "What is the drug of choice for condition X?" with a percentage of 11%, followed by "What is the cause of symptom X?" and "What is indicated in situation X?" question types with 8% each.

In order to correlate their results with our study, we have classified each of the generic questions using the evidence node in the Evidence Taxonomy and we obtained the results in Table 2, column "Category", which translates to the summarized totals of 30% for Intervention questions and 33% for No Intervention questions.

Total questions (n=1396)		
Question	% of questions	Category
What is the drug of choice for condition x?	11%	Intervention
What is the cause of symptom x?	8%	No Intervention
What test is indicated in situation x?	8%	Intervention
What is the dose of drug x?	7%	No Intervention
How should I treat condition x (not limited to drug treatment)?	6%	Intervention
How should I manage condition x (not specifying diagnostic or therapeutic)?	5%	Intervention
What is the cause of physical finding x?	5%	No Intervention
What is the cause of test finding x?	5%	No Intervention
Can drug x cause (adverse) finding y?	4%	No Intervention
Could this patient have condition x?	4%	No Intervention

Table 2: Ely's clinical questions distribution and our classification according to the Evidence taxonomy

After completing the question classification process, the resulted structure of our question corpus was 46% Intervention questions and 54% No Intervention questions. The resulted distribution is relatively close to the percentages we have calculated using the results of Ely et al (2000) and the mapping to the Evidence taxonomy.

3.2. Question processing

Turning knowledge into specific requests for information is not always an easy task. Some information needs are difficult to express and when they can be expressed, the way the question is interpreted influences the delivered answers. Yu et al (2005) mentions that Ely and his colleagues have calculated an average of 2.7 different ways of expressing generic general practitioner clinical questions. The same study mentions the difficult step of explaining the context of the questions to the information source.

Query modification was applied to the corpus questions when running the original question through a search engine did not produce any relevant results. We have defined five levels of processing to be applied to improve search outcomes.

The first level of processing involves introducing synonyms or hyponyms of the medical terms in an attempt to improve the performance of the search. Example: we have replaced "infectious" in "infectious conjunctivitis" with its corresponding hyponym "bacterial" in "bacterial conjunctivitis".

The next level is detecting any abbreviations that might decrease search engines ability to find answers. Example: we have replaced "BP" with "blood pressure".

The third level is adding general medical terms such as 'disease', 'syndrome' or 'condition' to help clarify the target of the search query. Example: "What is shoulder frozen" has been replaced with "What is frozen shoulder syndrome".

The fourth level we have defined implies eliminating additional grammatical terms such as adverbs and prepositions from the original question. Example:

original question "Are there any contraindications to dental office visits in pregnancy" was modified to "Dental office visits in pregnancy".

The fifth level in our processing diagram involves using existing knowledge to transform the question in the attempt to express the medical context. Example: "What is the evidence that antibiotics change the course of the disease in infectious conjunctivitis" became "Are antibiotics recommended for bacterial conjunctivitis".

When question transformation was required, it was performed across all the selected search engines and question answering systems that did not produce any relevant links or answers for the original question. Appendix B "Question processing" includes examples of questions that have required processing and the actual transformation applied.

In order to be able to process the questions through the different levels we have used one of the online medical dictionaries: MedLinePlus[16]. MedLinePlus has extensive information from the National Institute of Health and other trusted sources on over 750 diseases and conditions and is a service offered by the US National Library of Medicine.

A summary of the five levels of question processing is shown in Table 3.

Processing Level	Description	Original Question/Term	Processed Question/Term
1	introduce synonyms/hyper nyms	infectious	bacterial
2	replace abbreviations	BP	blood pressure
3	Introduce general medical terms	What is shoulder frozen	What is shoulder frozen syndrome
4	eliminate additional terms	Are there any contraindications to dental office visits in pregnancy	Dental office visits in pregnancy
5	express medical context	What is the evidence that antibiotics change the course of the disease in infectious conjunctivitis	Are antibiotics recommended for bacterial conjunctivitis

Table 3: Question processing levels

In order to evaluate the efficiency of our question processing and the degree to which each defined level of transformation had a positive impact on the search results, we have analyzed the questions that did not produce any relevant answers when run through the search engines in their original form. We have then determined which level of transformation has been applied in order to get a relevant answer. If after applying a particular level of processing, we have obtained a relevant answer or link, we have flagged that question as being improved by Level x

[16] See http://medlineplus.gov/

of transformation. In order to quantify if there was an improvement, we did not consider the position of the relevant link on the results page and did not try to improve the relevant link position in the list by applying a subsequent level of processing. After computing the results we have reached the conclusion that Level 4 of processing "Eliminate additional terms" was applied with the highest frequency (45.95% of total successful transformations), followed by Level 5 "Express medical context" (27.03% of total successful transformations). The results are presented in Table 4.

Improved search results					
Processing levels	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**
How often level was applied	5.41%	10.81%	10.81%	45.95%	27.03%

Table 4: Question processing results

3.3. Answer extraction

In our attempt to locate answers to our corpus of questions, we have established a limit of 10 first links returned in response to a query. Any other links past this limit, relevant or irrelevant, have been ignored. Any relevant link that refers to a scientific article but does not have an abstract available has been ignored. We have set this rule as usually, if the abstract of the article is not available, the attempt of viewing the full text of the publication fails, requesting a registered username and password.

Most of the search engines included in the study will return a list of links that will then need to be evaluated in order to determine their relevance to the query. This is a time consuming process that MedQA, as a question answering system, manages to overcome by providing a summarized and concise answer. For some instances of our searches, when PubMed returned only one link in response to a query, the abstract was automatically displayed and we were able to locate the answer.

It is important to mention that we had to give special attention to the search of questions through the Google search engine. Due to its remarkable coverage of its engine, Google was able to locate the Parkhurst Exchange answer itself, whenever the original question in the selected corpus of questions was entered as the search string. As this was distorting our search results we have ignored those links and proceeded to the next available relevant entry in the list. If no relevant answers were listed, the transformed question was applied as the search string.

4. Results

In order to evaluate the results of our answers search, we have used a scoring system first referred to in the Text Retrieval Conference (TREC), called Mean Reciprocal Rank (MRR) (Voorhees, 2001). If a link returned by a search was the nth ($n \leq 10$) position in the list of resulted links, and it was evaluated as being relevant to the question using the Parkhurst Exchange answers as a benchmark, it was given a score of 1/n. We have adopted this methodology in order to assess the ranking system of each search engine. The further down the list, the more effort required from the user to locate the answer. Our evaluation includes the "ease of use" component in our scoring system.

In order to evaluate if a summarized answer or a link returned in response to a search query are relevant, we have referred to the answer provided by the Parkhurst Exchange website. We have initially opted for a lenient evaluation, in the sense that a link or summarized answer that was relatively relevant to the question received a score that was giving them a credit lower that the 10^{th} position of a relevant answer in the top 10 list: 1/11. However we have later revised this scoring system as we came to the conclusion that it was possible that this methodology was introducing bias in our evaluation. We have decided to stick with the strict evaluation that only gives credit to links or summarized answers that express the same ideas as the Parkhurst Exchange benchmark answer. This decision was mainly supported by the fact that a limited medical knowledge baggage does not allow a proper evaluation and judgment of diagnosis, drugs and treatments that are related to the search question.

The results of our evaluation are presented in Figures 11 and 12, for the two evidence categories our corpus of questions was mapped to. They have been calculated as an average of scores, per question category and search engine.

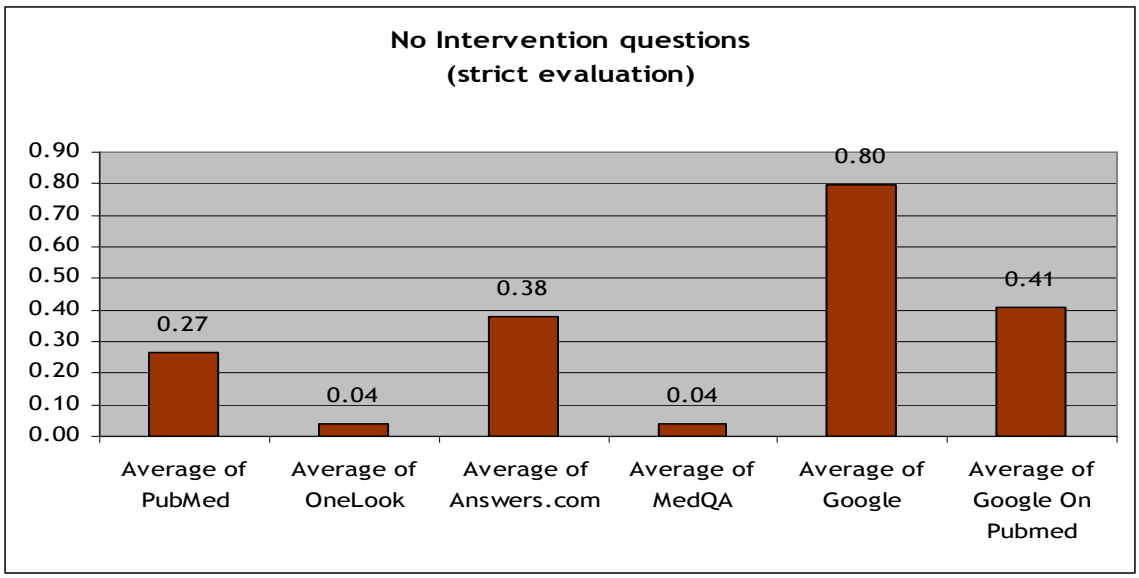


Figure 11: No Intervention questions scores

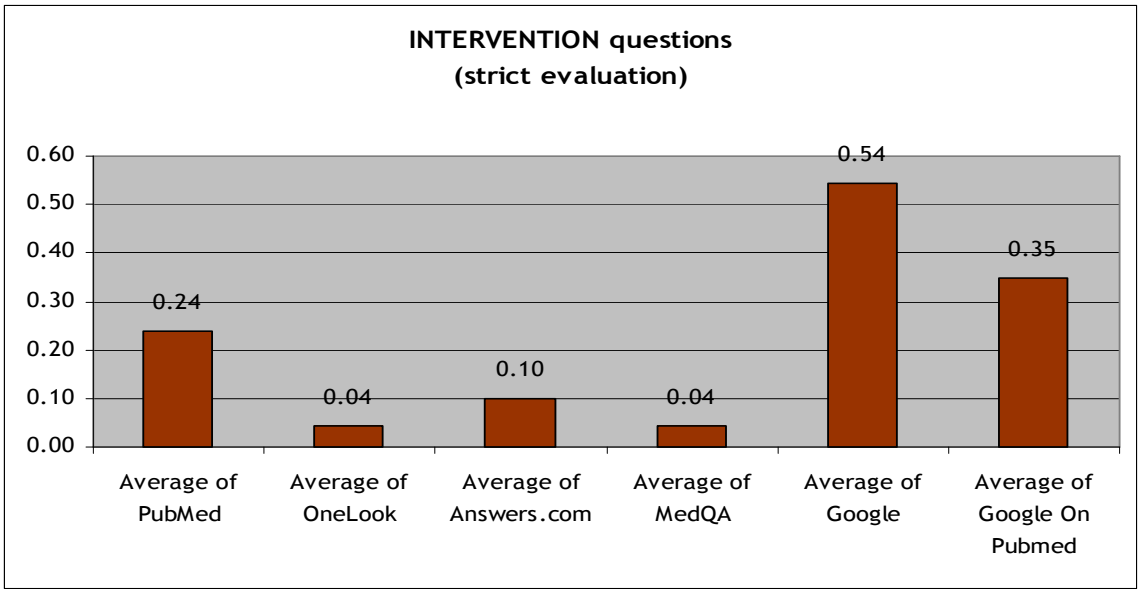


Figure 12: Intervention questions scores

After processing the 50 medical questions through all the selected search engines, we have obtained a total of 119 answers. The results of the actual location of the answer in a scientific article are shown in Figure 13. Our results show that the answer can be located in one of the sections: abstract, results, conclusions, recommendations, purpose or methods.

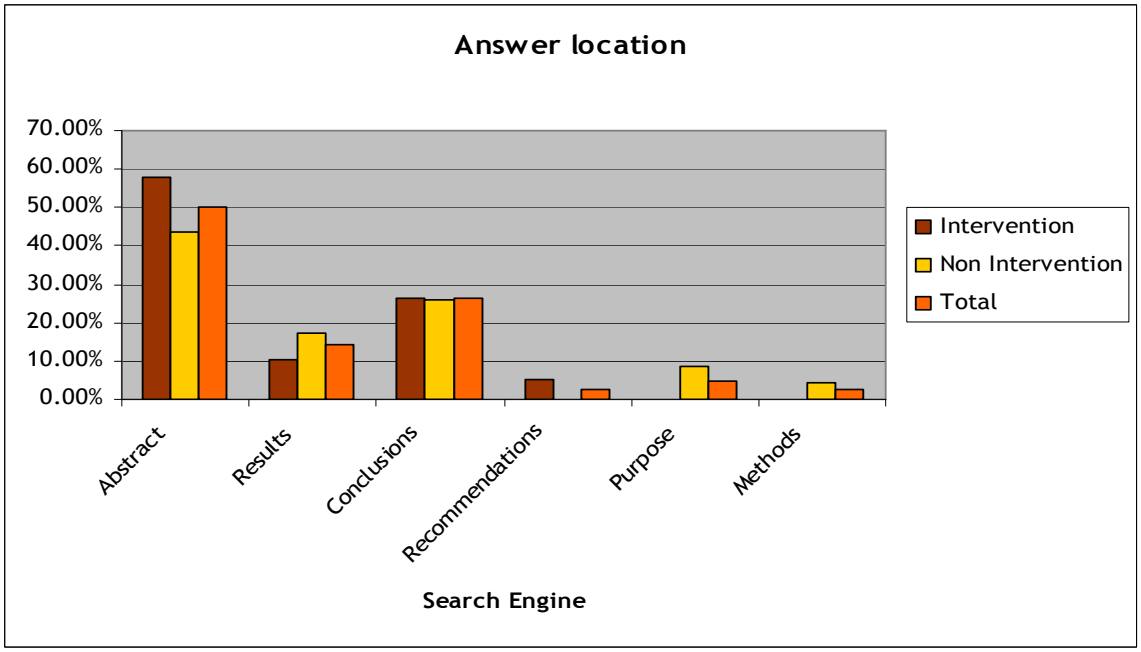


Figure 13: Answer location in scientific articles

The results above do not refer to answers located in consumer oriented websites which do not follow a set document structure. They have been obtained after analyzing the answers extracted from medical scientific articles which represent 34% of our total number of answers.

5. Discussion and conclusions

As represented in Table 5, our results show that Google has the best performance for both Intervention and Non-Intervention questions. Google on PubMed also has the second place for both Intervention and No Intervention questions, proving that Google still seems to be one of the best search engines.

Intervention	Source	Position
	Google	1
	Google On Pubmed	2
	PubMed	3
	OneLook	4
	MedQA	4
	Answers.com	5
No Intervention	Google	1
	Google On Pubmed	2
	Answers.com	3
	PubMed	4
	OneLook	5
	MedQA	5

Table 5: Search engines and answering systems ranking

PubMed was outperformed by Google on PubMed which is sort of a surprise. Analyzing the detailed results, this is mainly due to PubMed returning the relevant links further down in the list and consequently obtaining a lower score than Google on PubMed. The conclusion is that the ranking algorithm adopted by PubMed is not performing as well as Google's. This is in line with the observation of Plikus et al (2006) that concluded that PubMed does not produce well classified search outputs and proposed PubFocus to help ranking by adding publication quality attributes. Table 6 provides some examples of rankings for the same link in PubMed as opposed to Google on PubMed. It is quite obvious that in those instances Google assigned a better score than PubMed for the same relevant link.

Question No	Question	Category	PubMed ranking	Google on PubMed ranking
19	When should moles be removed?	Intervention	7	2
43	What can be done for a patient with persistent (non-typhoid) *Salmonella* in stool, despite 2 antibiotics	Intervention	7	4
46	Is it a good idea to take an ASA before an extended period of air travel?	Intervention	5	4

Table 6: Google on PubMed versus PubMed ranking for the same relevant link

Analyzing PubMed search engine behavior we noticed that even if MEDLINE benefits from MeSH, the controlled vocabulary thesaurus, with the role of providing a consistent way of retrieving information for search strings using different terminology for the same concepts, it still underperforms in retrieving relevant answers for medical questions using acronyms. An example is the original question "Is it a good idea to take ASA before an extended period of air travels" in which ASA is the medical acronym for "acetylsalicylic acid". Although we sort of expected that Google, as a generic search engine, would not be able to handle the acronym, PubMed was also not able to translate it and could not provide any answers until we have manually processed the question and replaced the abbreviation in the question with the explicit chemical substance name.

It also looks like PubMed lacks the abilities of processing natural language questions as our research showed that it would benefit a lot from manual question processing of applying our Level 4 "Eliminate additional terms" transformation. An example is the original question "What's the best antihistaminic for mild acute urticaria in infants and children?" for which PubMed could not locate an answer until we have transformed it into "antihistaminic for mild acute urticaria in children". Google also needed to receive the processed question as an input in order to be able to retrieve any relevant links.

MedQA obtained one of the worst scores, but this was mainly due to the fact that the online link was not always up and running. Our attempt to rerun all questions through the answering system has proved even more unsuccessful, as we have not been able to obtain any answers due to the error message we seemed to constantly receive from the MedQA website ("HTTP status 404 - / qaseam / answer.seam " – The requested resource is not available).

Analyzing the performance of the generic search engines and question answering systems, we conclude that Google performed better on No Intervention questions with an average score of 0.8 as opposed to 0.54 for Intervention questions. Answers.com also proved better on No Intervention questions than Intervention questions. The results prove that a non-medical oriented search engine has more difficulties on producing answers for scenario-based, complex medical questions. Overall Answers.com performed much better than OneLook and this could be explained by the fact that the Answers.com engine we have referred to incorporates BrainBoost technology for natural language processing.

OneLook statistics show that the search engine barely managed to answer two questions out of the 50 included in our corpus questions: one intervention question and one no intervention questions. We could state then that these results show that OneLook is currently not suitable as a potential technology for medical answering systems.

Analyzing the overall results, 4 out of 6 search engines across both groups managed to handle No Intervention questions better than Intervention questions,

confirming the claim that the complexity of the query has an impact on the results. We have also found out that all the questions in our corpus of questions are answerable with current technology, which we consider to be an important finding for future medical question answering systems as it implies that the current technology is ready and available to serve question answering purposes.

Going further to the actual location of the answer in medical articles, we have determined that the probability of the answer to be located in the Abstract section of an article is 50%, Conclusions section 26.19% and Results section 14.29%. This gives a good indication on the areas a question answering medical system should look most of the time for answers to ad-hoc queries.

6. Future work

Our study results have been compiled on a small set of 50 questions and we admit this might introduce some bias in our process. Our results will have to be confirmed and compared to the performance obtained on a larger corpus of questions (over 200). For a more confident evaluation, we recommend having medically trained teams performing the answers search and evaluation.

We would like to include the evaluation of PubMed enhanced with the ranking system provided by PubFocus in our future work. We would then compare these results with the previous PubMed performance and evaluate the extent of the improvement.

Another challenging research topic we would like to follow-up is determining the actual quality of the answers returned by a particular search engine or question answering system. The answers quality topic is of special interest in the medical question answering domain as critical patient care decisions might be taken based on the information provided. We would like to use NetScoring, described earlier on in this report, to quantify quality using categories such as credibility, accessibility, content and ethics. The results of this exercise will help determining if the conclusion of the study of Berkowitz (2002) that stated that Google performs poorly in regards to quality of answers because it refers to consumer-oriented websites when locating relevant links, had some support.

Another item on our list of future work is obtaining permission from Dr. Ely and his team to access the 200 questions used in their study (Ely et al, 2000). We aim to perform an evaluation of those questions through our selected search engines and question answering systems. We feel this exercise would be very beneficial for our research. Firstly we would be able to have an indication on the accuracy of our question classification process of dividing questions into Intervention and No Intervention categories. We could not locate any detailed explanation or examples of questions grouped into the two categories to confirm that our understanding of the two nodes in the Evidence taxonomy is correct.

The second task we would like to perform on Ely's corpus of questions is determining if those questions that were marked as unanswerable mainly due to the extended duration of the search process (over a set maximum time limit) were actually answerable.

References

Berkowitz L, 2002. *Review and evaluation of Internet-based clinical reference tools for physicians*. White Paper commissioned by UpToDate

Brin Sergey, Page Lawrence, 1998. *An Anatomy of a Large-scale Hypertextual Web Search Engine.* Proceedings of the 7th International World Wide Web Conference, Page 107 – 117.

Covell D G, Uman G C, Manning P R. 1985. *Information needs in office practice: are they being met?* Annual Intern Medicine, vol 103, p596-599

Craig Jonathan C, Irwig Les M, Stockler Martin R, 2001. *Evidence-based medicine: useful tools for decision making.* The Medical Journal of Australia, 174: 248-253.

Demner-Fushman Dina, Lin J Jimmy. 2007. *Answering clinical questions with Knowledge-based and Statistical Techniques.* Computational Linguistics, 33(1):63-103(2007)

Ely J W, Levinson W, Elder N C, Mainous A G,Vinson D C. 1995 , *Perceived causes of family physician's errors*. Journal of family practice, 40(4):337—44

Ely J W, Osheroff J A; Ebell M H; Bergus G R; Levy B T; Chambliss M Ll. 1999. *Analysis of questions asked by family doctors regarding patient care*. British Medical Journal, 319(7206):358–61

Ely John W, Osheroff, Jerome A, Gorman Paul N, Ebell Mark H, Chambliss M Lee, Pifer Eric Af, Stavri P Zoe. 2000. A taxonomy of generic clinical questions: classification study. BMJ 321:429-432.

Huang Xiaoli, Lin Jimmy, and Dina Demner-Fushman. 2006. *Evaluation of PICO as a Knowledge Representation for Clinical Questions*. AMIA Annu Symp Proc.: 359–363.

Lee Minsuk, Cimino James, Zhu Hai Ran, Sable Carl, Shanker Vijay, Ely John and Yu Hong. 2006. *Beyond Information Retrieval-Medical Question Answering.* AMIA Annual Symposium Proc.: 469–473.

Molla Diego, Vicedo Jose L. 2009. *Question Answering*. Draft chapter of Nitin Jotsingani and Fred Damerau (Ed.), "Handbook of Natural Language Processing", Second Edition, CRC Press. To be published.

Niu Yun, Hirst Graeme, McArthur Gregory, Rodriguez-Gianolli Patricia, 2003. *Answering Clinical Questions with Role identification;* Proc. ACL, Workshop on Natural Language Processing in Biomedicine

Plikus V Maxim, Zhang Zina, Chong Cheng-Ming. 2006. *PubFocus:semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithms.* BMC Bioinformatics.

Shi Zhongmin, Melli Gabor, Wang Yang, Liu Yudong, Gu Baohua, Kashani Mehdi M, Sarkar Anoop and Popowich Fred. 2007. *Question Answering Summarization of Multiple Biomedical Documents*. Canadian Conference on AI, volume 4509 of Lecture Notes in Computer Science, page284-295. Springer.

Voorhees Ellen, 2001. *"The TREC question answering track",* Natural Language Engineering, 7(4):361-378, Cambridge University Press.

Yu Hong, Kaufman David. 2007. *A cognitive evaluation of four online search engines for answering definitional questions posed by physicians*. Pacific Symposium on Biocomputing, page 328-339

Yu Hong, Lee Minsuk, Kaufman David, Ely W John, Osheroff Jerome A, Hripcsak George and Cimino James J. 2007. *Development, implementation and a cognitive evaluation of a definitional question answering system for physicians*; Journal of Biomedical Informatics 40(3):236-251.

Yu Hong, Sable Carl, Zhu Hai Ran. 2005. *Classifying Medical Questions based on an Evidence Taxonomy.* AAAI Workshop on Knowledge and Reasoning for Answering Questions

Zweigenbaum Pierre. 2003. *Question answering in biomedicine*; Proc. EACL 2003, Workshop on NLP for Question Answering, Budapest

Zweigenbaum Pierre, Jacquemart Pierre. 2003. *Towards a Medical Question-Answering system: A feasibility study*, Studies in health technology and informatics, Vol. 95 , pp. 463-468.

Appendices

Appendix A. Question Corpus

Q no	Question	Category
1	Is it necessary to avoid egg before 1 year of age?	Intervention
2	When should babies first eat fish and eggs?	Intervention
3	Is watermelon allergenic?	No Intervention
4	When to introduce solids to infants?	Intervention
5	Should family doctors be immunized with Pneumovax and Menactra or Menjugate?	Intervention
6	Are there any contraindications to dental office visits in pregnancy?	No Intervention
7	Can cell phones cause cancer?	No Intervention
8	What's the latest advice regarding children who suck their fingers?	No Intervention
9	How much folic acid — 400 µg, 1 mg, 5 mg — is recommended before conception and during pregnancy?	Intervention
10	How to beat recurrent UTIs?	Intervention
11	How to recognize autism in adults ?	No Intervention
12	Does skin colour affect vitamin D requirements?	No Intervention
13	Is whispering OK in laryngitis management?	No Intervention
14	Is there an age minimum for laser eye correction?	No Intervention
15	Is it advisable to ask patients with allergies to eggs not to take the flu shot?	No Intervention
16	Can you get Coxsackie, or foot and mouth disease, more than once?	No Intervention
17	Can antiperspirants cause breast Cancer?	No Intervention
18	At what age should a young child be referred to a speech therapist?	No Intervention
19	When should moles be removed?	Intervention
20	How long can women continue OCPs?	No Intervention
21	When is conjunctivitis viral or bacterial?	No Intervention
22	When to stop giving iron for mild anemia?	Intervention
23	Can migraine prophylaxis reduce the stroke risk?	Intervention
24	What's the evidence that antibiotic drops change the course of the disease in conjunctivitis that's suspected to be infectious?	No Intervention
25	Are calcium supplements bad for the heart?	No Intervention
26	Could kids with constant nosebleeds have a bleeding disorder?	No Intervention
27	If a person has a BP of 100/60 mm Hg and feels well, is it a problem?	No Intervention
28	Is low magnesium a cause for concern?	No Intervention
29	What is is left shoulder "frozen"?	No Intervention
30	What does C-reactive protein predict?	No Intervention
31	What are the causes for unexplained weight loss?	No Intervention
32	What is the management of male osteoporosis?	Intervention
33	How can a fractured nose be diagnosed?	No Intervention
34	Is there any scientific evidence for cod liver oil providing relief from osteoarthritis?	No Intervention
35	Why is there a difference between vaccination in the gluteal vs deltoid regions?	No Intervention
36	Can children who are taking antibiotics for recurrent otitis media or other types of chronic infection be given childhood immunizations (MMR, DPTP, etc.).	Intervention
37	When sould patients forego flying?	No Intervention
38	Is melatonin use safe in sleep deprivation?	Intervention
39	When to do a follow-up X-RAY after pneumonia?	Intervention
40	Is there such an entity as 'liver pain in non-enlarged liver' with chronic disease?	No Intervention
41	What's the current best treatment for DVT in a person with pancytopenia?	Intervention
42	With patients on long-term analgesics such as oxycodone, when should driving be stopped?	Intervention
43	What can be done for a patient with persistent (non-typhoid) *Salmonella* in stool, despite 2 antibiotics	Intervention
44	What's the best treatment for neck joint pain?	Intervention
45	How do you re-immunize the patient who in the past failed to complete the primary series of hep A and B, as well as HPV vaccinations?	Intervention
46	Is it a good idea to take an ASA before an extended period of air travel?	Intervention
47	What's the best antihistaminic for mild acute urticaria in infants and children?	Intervention
48	After excluding serious causes of constipation in infants and newborns, how should one proceed in family practice to treat a baby?	Intervention
49	Can I give steroids to a child that had a chickenpox vaccine 3-4 days ago?	Intervention
50	What's the best strategy to treat cellulitis of the lower leg aggravated by venous stasis?	Intervention

Appendix B. Question Processing

Original question	Transformed question
When should babies first eat fish and eggs?	fish and egg allergy in infants
When to introduce solids to infants	introduction of solids to infants
What's the latest advice regarding children who suck their fingers	children who suck their fingers
Are there any contraindications to dental office visits in pregnancy?	Dental office visits in pregnancy
How much folic acid — 400 µg, 1 mg, 5 mg — is recommended before conception and during pregnancy	recommended dose of folic acid before pregnancy
How to beat recurrent UTIs	how to treat recurrent UTIs?
How to recognize autism in adults	How to diagnose autism in adults
When is conjunctivitis viral or bacterial	how to distinguish between bacterial and viral conjunctivitis?
When to stop giving iron for mild anemia	When to stop taking iron for mild anemia
What's the evidence that antibiotic drops change the course of the disease in conjunctivitis that's suspected to be infectious?	antibiotic results for bacterial conjunctivitis
Are calcium supplements bad for the heart	calcium supplements increase cardiovascular risk
If a person has a BP of 100/60 mm Hg and feels well, is it a problem?	is low blood pressure a problem?
Is low magnesium a cause for concern	low magnesium
What is is left shoulder "frozen"?	What is frozen shoulder syndrome
What are the causes for unexplained weight loss?	Unexplained weight loss
What is the management of male osteoporosis?	male osteoporosis treatment
How can a fractured nose be diagnosed?	diagnose fractured nose
Is there any scientific evidence for cod liver oil providing relief from osteoarthritis?	cod liver oil in osteoarthritis
Why is there a difference between vaccination in the gluteal vs deltoid regions?	gluteal muscle or deltoid muscle injection
Can children who are taking antibiotics for recurrent otitis media or other types of chronic infection be given childhood immunizations (MMR, DPTP, etc.).	childhood immunisation for children taking antibiotics
What's the current best treatment for DVT in a person with pancytopenia?	deep vein thrombosis treatment for patient with pancytopenia
With patients on long-term analgesics such as oxycodone, when should driving be stopped?	patients on long-term analgesics oxycodone and driving
What can be done for a patient with persistent (non-typhoid) Salmonella in stool, despite 2 antibiotics	persistent Salmonella in stool after antibiotics
What's the best treatment for neck joint pain?	best treatment for neck joint pain
How do you re-immunize the patient who in the past failed to complete the primary series of hep A and B, as well as HPV vaccinations?	How to re-immunize for missed hep A and B, as well as HPV vaccinations
Is it a good idea to take an ASA before an extended period of air travel?	acetylsalicylic acid and air travel
What's the best antihistaminic for mild acute urticaria in infants and children?	antihistaminic for mild acute urticaria in children
After excluding serious causes of constipation in infants and newborns, how should one proceed in family practice to treat a baby?	constipation treatment in infants and newborns
Can I give steroids to a child that had a chickenpox vaccine 3-4 days ago?	steroids contraindication after chickenpox vaccine