Pacific Symposium on Biocomputing 12:328-339(2007)

A COGNITIVE EVALUATION OF FOUR ONLINE SEARCH ENGINES FOR ANSWERING DEFINITIONAL QUESTIONS POSED BY PHYSICIANS

HONG YU

University of Wisconsin-Milwaukee, Department of Health Sciences, 2400 E. Hartford Avenue PO Box 413, Milwaukee, WI 53210, USA

DAVID KAUFMAN

Columbia University, Department of Biomedical Informatics, 622 West, 168th Street VC-5, New York, NY10032, USA

The Internet is having a profound impact on physicians' medical decision making. One recent survey of 277 physicians showed that 72% of physicians regularly used the Internet to research medical information and 51% admitted that information from web sites influenced their clinical decisions. This paper describes the first cognitive evaluation of four state-of-the-art Internet search engines: Google (i.e., Google and Scholar.Google), MedQA, Onelook, and PubMed for answering definitional questions (i.e., questions with the format of "What is X?") posed by physicians. Onelook is a portal for online definitions, and MedQA is a question answering system that automatically generates short texts to answer specific biomedical questions. Our evaluation criteria include *quality of answer*, *ease of use*, *time spent*, and *number of actions* taken. Our results show that MedQA outperforms Onelook and PubMed in most of the criteria, and that MedQA surpasses Google in *time spent* and *number of actions*, two important efficiency criteria. Our results show that Google is the best system for *quality of answer* and *ease of use*. We conclude that Google is an effective search engine for medical definitions, and that MedQA exceeds the other search engines in that it provides users direct answers to their questions; while the users of the other search engines have to visit several sites before finding all of the pertinent information.

1 Introduction

The Internet offers widespread access to health and science information. Although there were a lot of concerns about the quality due to variations in accuracy, completeness, and consistency (1-10), the Internet is having a profound impact on both patients' access to healthcare information (11, 12) and physicians' medical decision making (13). One recent survey of 277 physicians showed that 72% physicians regularly used internet to research medical information and 51% declared that the Internet influenced their healthcare decisions (13).

The Internet may satisfy physicians' information needs by two means. First, it is well-reported that physicians often have questions when caring for their patients (14); the Internet incorporates vast amount of healthcare and scientific information which may provide an excellent resource to answer their questions. Although the quality of the information is still in dispute, studies found that the Internet has increased in quality over years (15). In certain domains, the information presented in the Internet was evaluated to be accurate (16). Secondly, the Internet provides different publicly available search engines and information retrieval systems (e.g., Google and PubMed) that may allow

physicians to efficiently access information. Efficiency is extremely important to physicians as studies found that physicians spent on average two minutes or less seeking an answer to a question, and that if a search took longer, it was likely to be abandoned (14, 17-19). In this study, we report a cognitive evaluation to compare a special purpose biomedical search engine, MedQA with three state-of-the-art search engines with the goal of identifying an optimal system that best suite physicians' information needs.

Specifically, we asked physicians to evaluate Google, MedQA, Onelook, and PubMed for answering definitional questions (i.e., questions with the format of "What is X?"). Google is a popular online search engine (4) and was evaluated to be the best web-search engine for answering medical questions (18). Google offers a wide range of resources and special-purpose search engines such as Google Scholar. Subjects were free to use any of Google tools to conduct their searches. OneLook (http://www.onelook.com/) is a portal for numerous online dictionaries including several medical editions (e.g., Dorland's). A recent study suggested that domain portals were most efficient for accessing healthcare information (20). MedQA automatically analyzed thousands of documents to generate a coherent paragraph-level text to specifically answer an ad-hoc medical question (21). PubMed is frequently searched by physicians at clinical settings (22).

Our work is related to the work of Berkowitz (2002) (23) in which 14 search engines (e.g., Google and PubMed) were evaluated to answer clinical questions. In that study, *quality of answer* and the overall *time spent* for obtaining an answer were measured. The results showed that Google performed poorly in quality of answer because many of the answers were from consumer-oriented sites and therefore did not incorporate information physicians needed, and that PubMed required a longer time spent for obtaining an answer. The limitations of Berkowitz's study include that it did not measure the cognitive aspects, including interpretation and analysis of number of actions involved for identifying answers. Additionally, all the evaluation was performed subjectively by the author (i.e., Berkowitz) of the article. Our study is based on a randomized controlled cognitive evaluation of four physicians who are not the authors of this article. Additionally, a unique feature of our study is that we provide the evaluation of an advanced, biomedical question answering system, and we compare it to three other state-of-the-art information retrieval systems.

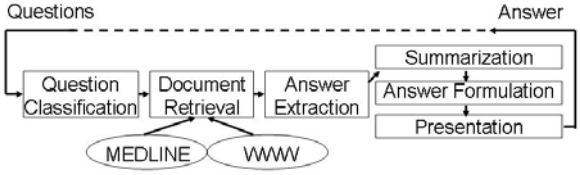

Figure 1: MedQA system architecture

2 MedQA

MedQA is a question answering system that automatically analyzes thousands of documents (both the Web documents and MEDLINE abstracts) to generate a short text

to answer definitional questions (21). In summary, MedQA takes in a question posed by either a physician or a biomedical researcher. It automatically classifies the posed question into a question type for which a specific answer strategy is developed (24, 25). Noun phrases are extracted from the question to be query terms. *Document Retrieval* applies the query terms to retrieve documents from either the World-Wide-Web documents or locally-indexed literature resources. *Answer Extraction* automatically identifies the sentences that provide answers to questions. *Text Summarization* condenses the text by removing the redundant sentences. *Answer Formulation* generates a coherent summary. The summary is then presented to the user who posed the question. Figure 1 shows the architecture of MedQA, and Figure 2 shows MedQA's output of the question "What is vestibulitis?"

Most of the evaluation work on question answering systems (26) focuses on information retrieval metrics. A text corpus and the answer are provided for a question, the evaluation task is to measure the ***correctness*** to extract the text answer from the corpus. None of the studies, to our knowledge, apply ***cognitive*** methods to evaluate human-computer interaction, and to measure efficacy, accuracy and perceived ease of use of a question answering system, and to compare a question answering system to other information systems such as information retrieval systems.

3 Cognitive Evaluation Methods

We designed a randomized controlled cognitive evaluation in order to assess the efficacy, accuracy and perceived ease of use of Google, MedQA, OneLook, and PubMed. The study was approved by the Columbia University Institutional Review Board.

3.1 Question Selection

We manually examined the total of 4,653 questions[1] posed by physicians at various clinical settings (14, 27-29) and found a total of 138 definition questions[2]. We observed that the definitional questions in general fell into several categories including Disease or Syndrome, Drug, Anatomy and Physiology, and Diagnostic Guideline. In order to maximize the evaluation coverage, we attempted to select questions that cover most of the categories.

After preliminary examination, we found that many questions did not yield answers from two or more systems to be evaluated. For example, the question "what is proshield?" did not yield a meaningful answer from three systems (MedQA, OneLook, and PubMed). The objective was to compare different systems, and unanswerable questions present a problem for the analyses because they render such comparisons impossible. On the other hand, if we screened the questions with the four systems, it may introduce bias and a selective exclusion process. We therefore employed an independent information retrieval

[1] The question collection is freely accessible at http://clinques.nlm.nih.gov/

[2] All 138 definitional questions are listed at http://www.dbmi.columbia.edu/~yuh9001/research/definitional_questions.htm.

system, BrainBoost[3], which is a web-based question answering engine that accepts natural language queries. BrainBoost was presented with questions randomly selected from the categories of definitional questions, and the first twelve questions that returned an answer were included in the study. The task was performed by an unbiased assistant who was not privy to the reasons for doing the search. The 12 selected questions are shown in **bold** in http://www.dbmi.columbia.edu/~yuh9001/research/definitional_questions.htm.

Figure 2: MedQA's output of the question "What is vestibulitis?" The output displays an online definition that comes from Dorland's Illustrated Medical Dictionaries, a "summary" that incorporates definitional sentences that are extracted from different PubMed records, and "other relevant sentences" that incorporate other relevant sentences. The parenthetical expression incorporates the last name of the first author and the year of the publication (e.g., (Sackett 2001)); the expression links to the PubMed records from which the preceding sentences are extracted.

3.2 Subjects and Procedure

Four physicians (three females and one male, ages 30's-50's) who were trainees at Department of Biomedical Informatics, Columbia University volunteered to participate in the study. All four physicians have experience using information systems. Each physician was presented with all of the 12 questions selected for inclusion. For each question, the subjects were asked to evaluate two systems in succession and the order of the two systems was counterbalanced. Each subject posed six questions to each of the four systems. The four subjects therefore posed a total of 96 questions (12 x 4 x 2). All evaluation studies were conducted in May, 2006.

[3] http://www.brainboost.com/

After consenting to participate in the study, participants were given written instructions on how to perform the task. They were presented with each question on a cue card and asked to find the text that best answered the question. The order of questions to be presented was randomized. The card also indicated the two systems to be used and their sequence. Once the text was located, they were asked to copy and paste it into a Word document. They were free to continue to search and paste text into the document until they were satisfied that they found the best answer possible. There was a time limit of 5 minutes for each question/system event. We chose 5 minutes as a cutoff because a previous study found that internet users successfully found health information to answer questions in an average of 5 minutes (30). Participants were asked to think-aloud during the entire process. After completing each question evaluation comparing the two systems, they were asked to respond to the following two Likert questions: 1) rate the quality of answer and 2) the ease of use of the system. We employed a five point rating scale from the poorest (1) to the best (5).

We applied Morae usability software system to record the screen activities and audio record a subject's comments for the entire session. Morae provides a video of all screen activity and logs a wide range of events and system interactions including mouse clicks, text entries, web-page changes, and windows dialogue events. It also provides the analyst with the capability to timestamp, code, and categorize a range of video events.

3.3 Analysis

On the basis of a cognitive task analysis (Kaufman et al, 2003; Elhadad, 2005), we identified goals and actions common to all systems. Table 1 shows a list of actions we defined. We also noted system responses (e.g., what was displayed after executing a search), analyzed comments thematically and measured the response times. The protocols were coded by both authors. The total coding time for four subjects was about 30 hours.

Table 1: Actions used to answer questions.

Enter Query: Entering a search term in the search text box provided by the system.
Find Document: An action that involves >10s of time spent examining the retrieved list of documents (e.g., Web documents or PubMed abstracts).
Query Modification: An action that involves modification of the existing query or user-interface (e.g., change from Google to Scholar.Google).
Read Document: An action that involves a subject to spend >10s to read the selected document.
Scroll-Down Document: Scroll down a document to search for the answer.
Search Text Box: A subject applies the "Find" function to locate relevant text.
Select Document: A subject selects and opens a document to examine whether the answer appears in the document
Select Linkout: An action that involves selecting another link from the selected document.
Select Text as Answer: A subject selects the text as the answer to a question.

4 Evaluation Results

In the following section, we present results of the cognitive evaluation. The first part of this section illustrates the processes of question-answering. We also show the coding process used to characterize participants' actions. The second part of this section focuses on a quantitative comparison of the four systems. We include both objective measures

such as actions and response latency, and subjective measures, namely, participants' ratings of the quality of answers as well as their ease of use.

4.1 Illustrations

The following two coding excerpts illustrate the process of question-answering on two pairs of systems, PubMed and MedQA, and OneLook and Google. The excerpts are representative of task performance. The subject was an experienced physician with a master in informatics and was well-versed in performing medical information seeking tasks.

Excerpt 1—PubMed and MedQA **The** subject had completed five questions and was a little more than forty minutes into the session. The question in this excerpt was "What is vestibulitis?" The systems used to find the answer were PubMed and MedQA respectively. The entire segment lasted 6 minutes, of which 4:25 is used to search PubMed and 1:11 to search MedQA.

44:23 ACTION (Enter Query-PubMed): vestibulitis
44:34 SYSTEM RESPONSE: 251 MEDLINE records returned
44:51 (User) COMMENT: OK, I definitely got some answers that do not apply at all…I have no idea why the first set of returns are coming back with psychological problems, but maybe not true, as a physician just makes assumption of that ENT would be returned, but if I am gynecologist, that probably is what I am looking for. Vulvar vestibulitis, I have no idea what it is. I guess I will go find out because I do not know.
45:22 ACTION SELECT DOCUMENT
45:23 ACTION SELECT FULL-TEXT OUT-LINK
45:24 SYSTEM RESPONSE: Out-link failed
45:25 ACTION SELECT FULL-TEXT OUT-LINK
45:26 SYSTEM RESPONSE: Out-link failed
45:33 ACTION FIND DOCUMENT
COMMENT: No… I can not find any definitions
46:11 ACTION (Query Modification, "vestibulitis") COMMENT: Try vestibulitis only
46:14 SYSTEM RESPONSE: 251 MEDLINE records returned
46:17 ACTION SELECT DOCUMENT COMMENT: Just try this one, surgical treatment of vulvar vestibulitis, this seems to be a good definition
46:29 ACTION SELECT FULL-TEXT 46:39 SYSTEM RESPONSE: Out-link failed
46:40 ACTION SELECT LINKOUT (of the full-text article)
46:41 SYSTEM RESPONSE: Out-link failed
COMMENT: It does not seem to have any outlink, it is only the abstract. The abstract does not give any characteristics of what syndrome is.
47:10 ACTION SELECT TEXT AS ANSWER
47:49 ACTION FIND DOCUMENT
47:57 ACTION SELECT DOCUMENT
48:00 ACTION SELECT FULL-TEXT (PDF FILE)
48:02 ACTION READ DOCUMENT
COMMENT: seems to get pain syndromes
48:48 ACTION SELECT TEXT AS ANSWER
COMMENT: OK, I am going to leave PubMed
49:12 ACTION (ENTER QUERY-MEDQA): What is vestibulitis?
COMMENT: MedQA uses MEDLINE, probably will return the same information, hopefully, it will get other information as well.
49:52 SYSTEM RESPONSE: shown in Figure 2
COMMENT: OK, MedQA pulls back exact the same information, nothing else.
50:23 ACTION SELECT TEXT AS ANSWER

GENERAL COMMENT: I would say that PubMed again all the information was there but was not held in a useful fashion and I need to search all and I have to filter myself…and quality of answer was OK and ease of use is poor because I need to go through everything. MedQA quality of answer is excellent and ease of use is excellent, I do not need to do anything.

Excerpt 2—OneLook and Google The subject had completed nine questions and was a little more than one hour and half into the session. The current question answered was "What is gemfibrozil?" The systems used to find the answer are OneLook and Google, respectively. The entire segment was 5:08 minutes, of which 1:44 is used to search the OneLook system and 2:46 to search Google.

1:31:08 ACTION (ENTER QUERY-ONELOOK): gemfibrozil
COMMENT: I know I am looking into medication, Gemfibrozil, I know that I have the advantage of what I am looking for.
31:32 SYSTEM RESPONSE: 4 matching dictionaries in General and 4 matching dictionaries in Medicine
COMMENT: So I get of course a General definition and Medicine related match. I will go my favorite Wikipedia first
31:51 SYSTEM RESPONSE Web Page Changes--Wiki…
COMMENT: it returns out-links…
COMMENT: Unfortunately, the Wikipedia isn't so good because it gives me more or less an outline of a whole set of other links that I would have to go find in order to get specific information. I am going back from Wikipedia and go to Medical online dictionaries, I am going to try Online Medical Dictionary first.
32:20 SYSTEM RESPONSE Web Page Changes –Online Medical Dictionary
COMMENT: I got absolutely useless information. I am going to Stedman's and Stedman's is not working, I found it out before. I go to Dorland's, Dorland's Medical Dictionary…
32:30 SYSTEM RESPONSE Web Page Changes – Dorland's Medical Dictionary
32:52: ACTION SELECT TEXT AS ANSWER
COMMENT: I get gemfibrozil … which is medication used to lower serum lipid level by decreasing triglyceride, it is just one line definition. I would say that it is probably acceptable, but if I have spent the time with the Wikipedia following the out-links, I probably would be able to find more information.
33:30 ACTION (ENTER QUERY-SCHOLAR.GOOGLE): gemfibrozil
COMMENT Now I am going to Scholar.Google
33:43 SYSTEM RESPONSE Web Page Changes – Google returned three article links
33:58 ACTION SELECT DOCUMENT (a full-text article)
34:10 ACTION READ DOCUMENT
COMMENT: On my first look on the medication…
34:32 ACTION SELECT TEXT AS ANSWER
COMMENT: I get quite a good description of the effects of new medication along with …
34:40 ACTION PULLUP PDF FILE
34:48 ACTION SCROLL-DOWN DOCUMENT
34:55 ACTION SELECT TEXT AS ANSWER
COMMENT: looks great…along with appropriate bibliography…With Google, with Google again, I got lucky, find an article very quickly, given me the best information about the medication.
35:50 ACTION (ENTER QUERY-GOOGLE): gembibrozil
COMMENT: let's see what happened if I go Google itself as appose to Google Scholar.
36:05 SYSTEM RESPONSE Web Page Changes – (Google returns 1,330,000 hits)
COMMENT: I got Medicine.com dictionary
36:16 ACTION SELECT TEXT AS ANSWER

COMMENT: I got some very good information. …which is more an overview, put gemfibrozil in the context with other medications for lowering serum lipid levels, so I would get a more understanding from this perspective and therefore Google general as oppose to Google.Scholar is actually a better choice as the Google search engine.
GENERAL COMMENT: For this study, Onelook I would say, was able to give me the definition which was OK in terms of quality, ease of use was poor because either that a lot of out-links are not working, or that the out links link to useless information. Google in this instance the quality of answer is definitely good, excellent, and ease of use in this instance, again is excellent, right answer comes from the top.

The two excerpts show that the pattern of actions employed by participants reflects the nature of interactions supported by each system. For example, subjects would iteratively search PubMed until they found a satisfactory answer. As a consequence, they would examine multiple documents (necessitating find link and Linkout actions), only a few of which were relevant. The subjects typically searched for full-text articles as the Linkout actions. The iterative nature of the search was also evidenced by the number of actions pertaining to query modification, searching the text box and document selection.

Table 2 lists a summary of the comments made by subjects throughout the evaluation. Our results show that Google received more favorable comments than complaints. Both MedQA and OneLook received some good comments and some complaints. PubMed was generally criticized and was not given any favorable comments.

Table 2: A summary of comments of different systems (**D** for disadvantages and **A** for advantages)

Google (D) retrieves back a lot of links (to the question "What is cubital tunnel syndrome?"). Most of links seem to relate to individual cases of the diseases, not necessarily definitions.
Google (D): One needs to search and evaluate the definitions in Google.
Google (A) retrieves both patient (Google) and physician-centric (scholar. Google) information.
Google (A): Scholar.Google is much faster because it is the second link, while in PubMed the evaluator has to search through a lot of other articles.
MedQA (D) needs to type in 'What is' versus a direct query.
MedQA (D) takes a considerable longer time to respond than other systems.
MedQA (A) returns all the context otherwise the evaluator has to search manually. It is only one step and gets exactly needed.
MedQA (A) gives answer (to the question "What is Popper?") that Onelook did not, which is that the drug is injectable, which is important to know for a physician.
Onelook (D) pulls all links. It lets the user to guess which link contains a comprehensive answer. Sometimes, the links are broken. It is a matter of luck to get to the right links.
Onelook (D) answer quality is poor. It has a terrible user-interface. It shows two ugly photos.
Onelook (A) definition has more content than PubMed.
PubMed (D) is not a good resource for definitions.
PubMed (D) is not useful. It takes forever to find information.

4.2 Quantitative Evaluation

The results show that the subjects did not find answers to a single question in Google ("Dawn's phenomenon"), 3 questions in Onelook ("epididymis appendix", "heel pain syndrome", and "Ottawa knee rules"), 3 questions in MedQA ("epididymis appendix", "Ottawa knee rules," and "paregoric"); and 2 questions in PubMed ("epididymis appendix" and "paregoric"). Both MedQA and Onelook acknowledged "no results found" and returned no answers if such an event occurs, while both PubMed and Google

returned a list of documents even if a subject could not identify the definitions from the documents within the 5 minutes of time limit.

We observed that none of the subjects used Google:Definition as the service to identify definitions; instead, they applied the query terms in either Google or Scholar.Google. We also observed that subjects gave the poorest score (i.e., 1) for quality of answer when both MedQA or OneLook returned no answers, and a better score (i.e., 2~3) when a search engine (e.g., Google or PubMed) returns a list of documents, even if the subject could not find any answers from the documents within 5 minutes of time limit. Subjects commented that even documents that do not contain answers frequently provided some knowledge about the answers. For example, subjects found "popper" is a drug although there were no details of definitions found. On the contrary, the subjects typically gave a good score for ease of use when MedQA and OneLook returned no answers.

Table 3 presents descriptive statistics of the subjective and objective measures. In general, Google was the preferred system as reflected both in the quality of the answer and ease of use ratings. MedQA achieved the second highest ratings in both measures. OneLook received the lowest ratings for quality of answer and PubMed was rated the worst in terms of ease of use. If we excluded the poor scores when MedQA did not return any answer, the quality of answer for MedQA went up to 4.5.

Table 3: Average score and (standard deviation) of *quality of answer* and *ease of use* and average *time spent* (in second) and *action taken*.

	Google	**MedQA**	**Onelook**	**PubMed**
Time Spent	69.6 (6.9)	59.1 (57.7)	83.1 (63.6)	182.2 (85.8)
Number of Actions	4.4 (3.0)	2.1 (2.0)	6.5 (7.7)	10.3 (5.7)
Quality of Answer	4.90 (0.15)	2.92 (0.24)	2.77 (0.08)	2.92 (0.88)
Ease of Use	4.75(0.29)	4.0 (0.24)	3.9 (0.32)	2.36 (0.88)

While the processing time to obtain an answer was almost instantaneous for Google, Onelook, and PubMed, the average time spent for MedQA to obtain an answer to the 10 answerable questions was 15.8±7.1 seconds. MedQA was nevertheless the fastest system on average for a subject to obtain the definition. For measuring the average time spent, we excluded the cases in which MedQA and Onelook returned no answer.

The subjects, on average, spent more time searching PubMed than any of the other systems. In fact, the average PubMed search required more than three times the amount of time required to search MedQA. This is at least partly due to the complexity of the interaction. This is borne out by the fact that participants needed more than 10 actions in using PubMed to answer the question, whereas they only required 2 actions on average when they used MedQA. PubMed provides a range of affordances (e.g., limits, MeSH) that supports iterative searching. Although this is a powerful tool, it also increases complexity of the task and user cognitive load. MedQA offers the simplest mode of interaction because it eliminates several of the steps (e.g., upload documents, search text and selectively access relevant information in document) involved in searching for information. The results of the commercial search engines, Google and Onelook, fell in

between MedQA and PubMed. However, as evidenced by the high standard deviations, there was significant variability between questions.

5 Discussion

The evaluation results show that Google was the best system for quality of answer (4.90) and ease of use (4.75). Recall the highest score for both criteria was 5. The results indicate that the Internet resources incorporate reliable medical definitions, and Google allows subjects to readily access those reliable definitions. This is in contrast to numerous other studies that found Internet information to be of poor quality in the medical context (1-10). However, there are significant differences between our study and the others. First, our study evaluated a more general type of question; namely, definitional questions, while the other works examined more specific medical questions (e.g., "What further testing should be ordered for an asymptomatic Thyroid Nodule solitary thyroid nodule, with normal TFTs?" in (23)). Secondly, physicians would evaluate Google high if they found answers from some sites even if other sites did not provide answers to the questions. In other studies, precision (i.e., the number of hits that provide answers divided by the total retrieved top N hits) plays an important role for measuring the quality. For example, one study (31) concluded that Google hits were of a poor quality because only one link out of five contained relevant information. Lastly, in our study, the quality of answer was estimated by aggregating information from multiple Web pages. Other studies evaluated the quality of each Web page to answer a specific question; such evaluation will certainly lead to a much poorer rating of the Internet because one evaluation study (32) concluded that information were typically scattered across multiple sites: most of the Web pages incorporate information either in depth or in breath, and few Web sites combine both depth and breath.

Our results show that OneLook came in the 3rd in most of the evaluation criteria. We observed that the evaluators frequently expressed frustrations of failed out-links and non-specific, general definitions that are of little value to physicians. We show that PubMed performed worst in almost all criteria. Unlike Google which assigns weights to the returned documents, PubMed returns a list of documents in a chronological order in which the most recent publications appear first. The most relevant documents in PubMed may never appear at the top; and therefore it usually takes a user significant time to identify answers. Previous research showed that it took an average of more than 30 minutes for a healthcare provider to search for answer from PubMed, which meant that "information seeking was practical only 'after hours' and not in the clinical setting" (22).

Finally, we found that MedQA in general outperformed all search engines except for Google. In addition, MedQA out-performed Google in *time spent* and *number of action*s, two important efficiency criteria for obtaining an answer. Although it took less than a second for Google to retrieve a list of relevant documents based on a query keyword and it took an average of 16 seconds for MedQA to generate a summary, the average time spent for a subject to identify a definition was 59.1±57.7 seconds for MedQA, which was faster than 69.6±6.9 seconds for Google. This is due to the fact that information is scattered across the web (32). A subject typically needs to visit multiple web pages for an

answer. One can never be certain when a link will lead to useful information. This is a relative disadvantage for Google as compared to MedQA.

6 Conclusion

We evaluated four search engines; namely, Google, MedQA, OneLook, and PubMed, for their quality and efficiency to answer definitional questions posed by physicians. We found that Google was in general the preferred system and PubMed performed poorly. We also found that MedQA was the best in terms of time spent and number of actions needed to answer a question. It would be ideal if a powerful search engine such as Google could be integrated with an advanced question-answering system to yield timely and precise response to address users' specific information needs.

Although we are encouraged by the findings, this research is best viewed as formative. The conclusions are limited by a number of factors. These include the fact that only four physicians participated in the evaluation. Future research would need to include a larger and more diverse sample of clinicians with different levels of domain expertise and degrees of familiarity with information retrieval systems. In this study, we introduced a novel cognitive method for the in-depth study of the question answering process. The method would have to be validated in different contexts. Finally, the scope of the system (answering definitional questions) is rather narrow at this point and we would want to conduct similar comparisons with different questions types. In general, the results of this work suggest that MedQA presents a promising approach for clinical information access.

Acknowledgement: We thank three anonymous reviewers for valuable comments.

References

1.Purcell G. The quality of health information on the internet. BMJ. 2002;324(7337):557-8.
2.Jadad AR, Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? JAMA. 1998 Feb 25;279(8):611-4.
3.Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor--Let the reader and viewer beware. JAMA. 1997 Apr 16;277(15):1244-5.
4.Glennie E, Kirby A. The career of radiography: information on the web. Journal of Diagnostic Radiography and Imaging. 2006;6:25-33.
5.Childs S. Judging the quality of internet-based health information. Performance Measurement and Metrics. 2005;6(2):80-96.
6.Griffiths K, Christensen H. Quality of web based information on treatment of depression: Cross sectional survey. BMJ. 2000;321:1511-15.
7.Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res. 2001 Dec;16(6):671-92.
8.Benigeri M, Pluye P. Shortcomings of health information on the Internet. Health Promot Int. 2003 Dec;18(4):381-6.
9.Wyatt J. Commentary: measuring quality and impact of the WWW. BMJ. 1997;314:1879.
10.McClung HJ, Murray RD, Heitlinger LA. The Internet as a source for current patient information. Pediatrics. 1998 Jun;101(6):E2.

11.Sacchetti P, Zvara P, Plante MK. The Internet and patient education--resources and their reliability: focus on a select urologic topic. Urology. 1999 Jun;53(6):1117-20.
12.Gemmell J, Bell G, Lueder R, Drucker S, Wong C. MyLifeBits: fulfilling the Memex vision. Proceedings of the 10th ACM international conference on Multimedia; France. pp. 235-8.
13.Podichetty V, Booher J, Whitfield M, Biscup R. Assessment of internet use and effects among healthcare professionals: a cross sectional survey. Postgrad Med J. 2006;82:274-9.
14.Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. BMJ. 1999 Aug 7;319(7206):358-61.
15.Pandolfini C, Bonati M. Follow up of quality of public oriented health information on the world wide web: systematic re-evaluation. BMJ. 2002 Mar 9;324(7337):582-3.
16.Sandvik H. Health information and interaction on the internet: a survey of female urinary incontinence. BMJ. 1999 Jul 3;319(7201):29-32.
17.Alper B, Stevermer J, White D, Ewigman B. Answering family physicians' clinical questions using electronic medical databases. J Fam Pract 2001;50(11):960-5.
18.Jacquemart P, Zweigenbaum P. Towards a medical question-answering system: a feasibility study. Stud Health Technol Inform. 2003;95:463-8.
19.Takeshita H, Davis D, Straus S. Clinical evidence at the point of care in acute medicine: a handheld usability case study. Proceedings of the human factors and ergonomics society 46th annual meeting; 2002. p. 1409-13.
20.Bhavnani S, Bichakjian C, Johnson T, Little R, Peck F, Schwartz J, et al. Strategy Hubs: Domain Portals to help Find Comprehensive Information. JASIST. 2006;57(1):4-24.
21.Lee M, Cimino J, Zhu H, Sable C, Shanker V, Ely J, et al. Beyond information retrieval - Medical question answering. AMIA. Washington DC, USA; 2006.
22.Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P, et al. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. J Am Med Inform Assoc. 2002. 9(3):283-93.
23.Berkowitz L. Review and Evaluation of Internet-based Clinical Reference Tools for Physicians: UpToDate; 2002.
24.Yu H, Sable C. Being Erlang Shen: Identifying answerable questions. Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasoning for Answering Questions 2005.
25.Yu H, Sable C, Zhu H. Classifying Medical Questions based on an Evidence Taxonomy. . Proceedings of the AAAI 2005 Workshop on Qusetion Answering in Restricted Domains; 2005.
26.Voorhees E, Tice D. The TREC-8 question answering track evaluation. TREC; 2000.
27.Ely JW, Osheroff JA, Ferguson KJ, Chambliss ML, Vinson DC, Moore JL. Lifelong self-directed learning using a computer database of clinical questions. J Fam Pract. 1997 Nov;45(5):382-8.
28.Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc. 2005 Mar-Apr;12(2):217-24.
29.D'Alessandro DM, Kreiter CD, Peterson MW. An evaluation of information-seeking behaviors of general pediatricians. Pediatrics. 2004 Jan;113(1 Pt 1):64-9.
30.Eysenbach G, Kohler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ. 2002 Mar 9;324(7337):573-7.
31.Berland GK, Elliott MN, Morales LS, Algazy JI, Kravitz RL, Broder MS, et al. Health information on the Internet: accessibility, quality, and readability in English and Spanish. JAMA. 2001 May 23-30;285(20):2612-21.
32.Bhavnani S. Why is it difficult to find comprehensive information? Implications of information scatter for search and design: Research Articles. Journal of the American Society for Information Science and Technology. 2005;56(9):989-1003.