ORIGINAL RESEARCH

Can Emergency Medicine Residents Reliably Use the Internet to Answer Clinical Questions?

Richard Krause, MD*
Ronald Moscati, MD*
Shravanti Halpern, MBBS*
Diane G. Schwartz, MLS†
June Abbas, PhD‡

* State University of New York at Buffalo, Department of Emergency Medicine, Buffalo, New York
† State University of New York at Buffalo, Kaleida Health Libraries, Buffalo, New York
‡ University of Oklahoma, Department of Library and Information Studies, Norman, Oklahoma

Introduction: The study objective was to determine the accuracy of answers to clinical questions by emergency medicine (EM) residents conducting Internet searches by using Google. Emergency physicians commonly turn to outside resources to answer clinical questions that arise in the emergency department (ED). Internet access in the ED has supplanted textbooks for references because it is perceived as being more up to date. Although Google is the most widely used general Internet search engine, it is not medically oriented and merely provides links to other sources. Users must judge the reliability of the information obtained on the links. We frequently observed EM faculty and residents using Google rather than medicine-specific databases to seek answers to clinical questions.

Methods: Two EM faculties developed a clinically oriented test for residents to take without the use of any outside aid. They were instructed to answer each question only if they were confident enough of their answer to implement it in a patient-care situation. Questions marked as unsure or answered incorrectly were used to construct a second test for each subject. On the second test, they were instructed to use Google as a resource to find links that contained answers.

Results: Thirty-three residents participated. The means for the initial test were 32% correct, 28% incorrect, and 40% unsure. On the Google test, the mean for correct answers was 59%; 33% of answers were incorrect and 8% were unsure.

Conclusion: EM residents' ability to answer clinical questions correctly by using Web sites from Google searches was poor. More concerning was that unsure answers decreased, whereas incorrect answers increased. The Internet appears to have given the residents a false sense of security in their answers. Innovations, such as Internet access in the ED, should be studied carefully before being accepted as reliable tools for teaching clinical decision making. [West J Emerg Med. 2011;12(4):442–447.]

INTRODUCTION

The clinical environment in the emergency department (ED) encompasses a wide range of clinical problems. The scope of information needed is therefore broad, and decisions are often made under time constraints. Textbooks are the traditional real-time reference source for emergency clinicians. When time constraints are less critical, information can be obtained in greater depth from medical libraries by using texts, journals, proprietary databases, and professional information specialists. Most EDs provide easy Internet access, and Internet literacy among physicians is commonplace. It is therefore not surprising that emergency clinicians increasingly turn to the Internet as a rapidly accessed and up-to-date source for real-time clinical information.

The perception that the "latest" information is available rapidly via the Internet makes it an attractive information source. Although the Internet's dynamic nature makes it impossible to assess the exact amount of information available,

1 estimate indicated more than 12,000,000,000 pages.[1] Information is typically obtained by using an Internet search engine (ISE) to sort through the vast content available. The ISE Google indexes the greatest number of pages and is the most frequently used ISE by the general public. In February 2008, an estimated 5.9 billion searches originated on Google. This represented 59.2% of all searches.[2] Despite the ease of obtaining information on the Internet, no systematic validation of the information occurs. This is of concern, particularly when searches are unstructured and the information obtained potentially influences medical care and therefore patient safety.

Although many, if not most, computers located in clinical areas maintain links to a variety of medical indices, our observation was that clinicians at our institution also used Google and other general ISEs. In designing a study to focus on the accuracy of finding answers on the Internet, we chose to focus on a single ISE to simplify the objective. The primary objective of the study was to determine the accuracy of emergency medicine (EM) residents' answers to clinical questions when using Google as an ISE.

METHODS

This nonblinded prospective study was designed to determine whether EM residents could identify accurate clinical information by using Google to search the Internet. The study design focused on a single ISE and a single testing sequence common to all of the subjects. The number of residents in the program at a given time is 36; other designs with subgroups taking different tests or using different search resources would have resulted in inadequate statistical power to draw meaningful conclusions. The study was approved by the university institutional review board.

The subjects were residents from an EM residency program. The EM residency has 12 residents per year in a 1- to 3-year format. Residents from all 3 classes during the year 2007 through 2008 were eligible to participate on a voluntary basis. The residents were informed that their decision to participate and performance in answering study questions would not have an impact on their academic standing. The study plan was presented at Grand Rounds 1 week before the study, and all volunteers signed consent forms before participation.

The test, consisting of 71 questions, was developed by 2 of the authors (R.K. and R.M.), who are residency faculty. The questions were clinically oriented and challenging, simulating questions that come up in day-to-day clinical ED practice. They were open ended, and an attempt was made to make them as unambiguous as possible (Table 1). Sources used to verify answers varied, depending on the question. For information on topics not subject to much change over time, we used EM texts such as *Rosen's*, *Tintinalli*, and *Harwood Nuss*. For more-recent topics, we used peer-reviewed journals, including *Annals of Emergency Medicine* and *Academic Emergency Medicine*, as well as other EM and non-EM journals. Other sources for current information included Web sites, such as the Centers for

Disease Control, American College of Emergency Physicians Clinical Guidelines, and eMedicine. Answers were agreed on by both faculty members and validated from 1 or more references. The answers were then considered the gold standard for the study.

The subjects initially completed a demographic questionnaire that included age, gender, year of training, as well as questions about computer and ISE use and familiarity. Each subject was assigned a 3-digit study number, known only by the study team. This provided anonymity for study subjects.

Subjects were given the 71-question test in a written, closed-book setting. This was referred to as the PreTest. Participants were asked to attempt to simulate the real-life ED environment where time is limited. A guideline of 5 minutes per question was suggested but not given as a strict limit. Subjects were asked to answer questions to a reasonable degree of clinical certainty, defined as sufficient confidence to use the information in a clinical setting. The answers were scored as correct, incorrect, or unsure (not sufficiently confident).

The authors scored each PreTest, comparing the subject's answers with the standardized answers. Answers were judged to be incorrect when they were either clearly factually wrong or where, in the authors' opinions, the response given would have caused a medical error if implemented in a clinical setting. The 2 EM faculty discussed questionable answers to decide on correctness. The result was used to create an individual test for each subject, consisting of the questions that had been answered incorrectly or as unsure on the PreTest. The second test was referred to as the Google Test.

Subjects were given their individual Google Test 1 week after the PreTest. This part of the study was conducted in the hospital library computer lab. Subjects were instructed to answer the questions with the help of a computer, using only Google as the ISE. They could then link to whichever sources, including other medical indices, provided by their Google search to obtain the information necessary to answer the question. Subjects were allowed to perform multiple Google searches for individual questions. The participants were once more instructed to use 5 minutes per question as a guideline and to answer the questions as "unsure" if they were not clinically confident of the answer. They were also instructed not to use any search engine other than Google. The individual subject's answers to the Google Test were then scored as before (correct, incorrect, or unsure).

Key logging software captured the search strategy used by the participants. The software tracked how the residents conducted their searches, which ISE features they used, and how much time they spent on each Web page. Whereas it is likely that most Google searches would identify links with accurate clinical information within the vast number of links identified, the ability of the residents to identify accurate information correctly from those search results is key to determining the correct answer. The information obtained from

Table 1. Example test questions.

Name 2 indications for operating room thoracotomy after chest-tube placement (tube thoracostomy)?

What is the preferred glucocorticoid for the treatment of suspected (as opposed to confirmed) acute adrenal insufficiency?

Which commonly used cardiac marker is elevated above the normal range soonest in acute myocardial infarction?

What is the age range in which Legg-Calve-Perthes disease is commonly seen?

List 2 commonly used oral antibiotics to which community-acquired methicillin-resistant *Staphylococcus aureus* isolates are typically sensitive when tested in vitro?

What are the 2 most common specific enteric pathogens that cause diarrhea in AIDS patients?

List 4 treatments for high-altitude pulmonary edema.

What is the approximate risk of HIV transmission in percutaneous occupational exposure to HIV-positive blood?

What is the maximum recommended dose of lidocaine without epinephrine for local infiltration (on a weight basis)?

Assuming similar serum levels, does chronic or acute digitalis toxicity have a higher mortality?

Which viral infection causes the most serious problems in the first 6 months after a solid organ transplant?

Is the use of rabies immune globulin considered to be generally safe in pregnancy?

Is cigarette smoking a risk factor for deep venous thrombosis in the Well's Critera?

Name the 5 "major criteria" for the diagnosis of acute rheumatic fever, as described by Jones.

Should antibiotics be used in the emergency department in cases of basilar skull fracture from blunt trauma?

the key logging software is being evaluated by 2 of the team members (D.G.S. and J.A.) with expertise in Library Science and Information Science. These results will be presented in a separate article.

The responses to the demographic questionnaire included gender, age, residency year, and several questions regarding computer use and familiarity with using computer resources. The results were scored by using 5-point Likert scales.[3,4]

For both the PreTest and the Google Test, the faculty members reviewed the residents' answers. Based on the previously validated answers, they were marked as being correct, incorrect, and unsure. These were then totaled for each test and entered into the study database. The overall percentage correct, incorrect, and unsure were then determined for each test.

The primary outcome measure was the percentage correct on the Google Test as an indicator of the accuracy of Google searches to answer clinical questions.

The answers to the demographic questions and the results of the PreTest and Google Test are presented as descriptive variables. The percentage correct answers on the Google Test were then compared with the results of the demographic questionnaire in a univariate fashion as categoric variables by using logistic regression to look for associations between the demographic responses and successful Google searches. All tests were 2-sided and tested at an α level = 0.05 for significance.
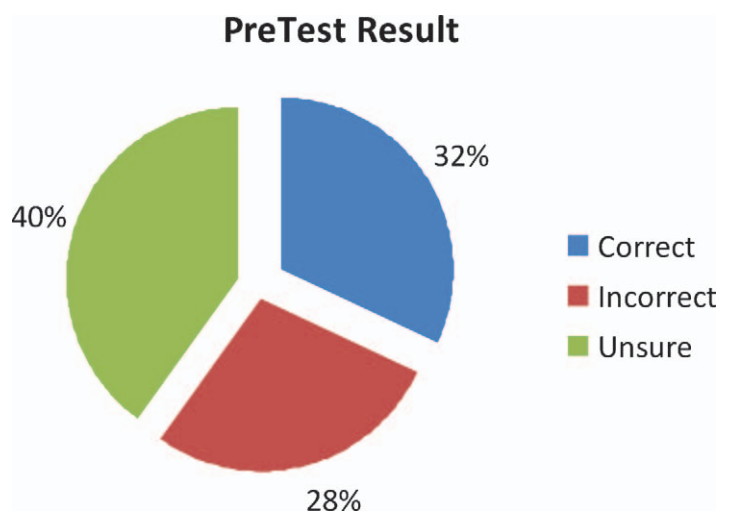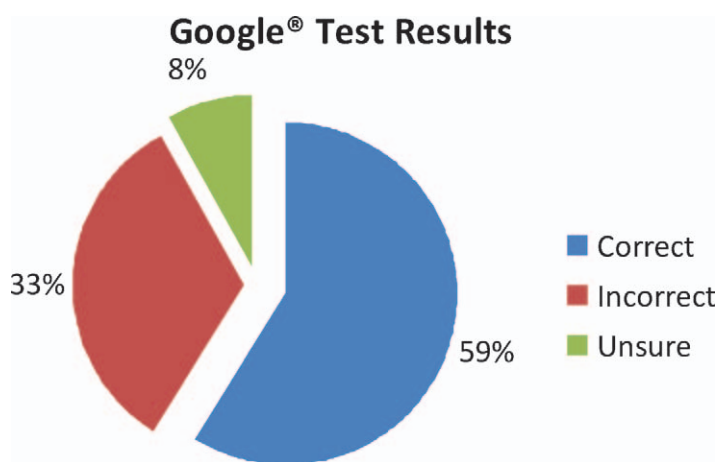
RESULTS

A total of 35 EM residents consented to participate in the study. One resident, who was a study investigator, was ineligible to participate. All 35 residents completed the PreTest.

Thirty-three completed the Google Test. Two were unable to complete the Google Test because of scheduling problems.

The overall results for the PreTest were 32% correct, 28% incorrect, and 40% unsure (Figure 1). The range of correct answers was 16% to 49%. After removing the correctly answered questions, the participants were given their individual Google Tests. The number of questions per resident ranged from 37 to 60, with a median of 49. On the Google Test, 59% (95% confidence interval, 56% to 62%) of the questions were answered correctly, 33% incorrectly, and 8% unsure (Figure 2). The range of correct answers on the Google Test was 36% to 72%.

The results of the demographic questionnaire are presented in Table 2, along with the *P* values for their association with the

Figure 1. PreTest results.

Google® Test Results

Figure 2. Google Test results.

Table 2. Association of demographics with percent correct answers on the Google Test.

	Responses	Number = 33	*P* value
Age range	25–34 years	28	0.03
	35–44 years	5	
Residency year	1st year	12	0.16
	2nd year	11	
	3rd year	10	
Gender	Men	25	0.67
	Women	8	

percentage correct on the Google Test. Table 3 presents the associations of percentage correct answers with the questions on computer confidence and use.

DISCUSSION

The results of our study demonstrated a surprisingly low rate of accuracy for EM residents answering clinical questions by searching Google. The 59% accuracy rate on the Google Test would be unacceptably low if used in a clinical setting. Equally surprising was the 33% incorrect answer rate by using Google searches. We repeatedly emphasized to the participants

that they should answer "unsure" unless they were confident enough of the answer that they would use it in a clinical setting. In fact, on the PreTest, "unsure" was the highest percentage response. However, on the Google Test, the residents marked "unsure" for only 8% of their answers.

The implication is that the residents were overconfident of the information obtained from the Internet. The residents had seen the questions previously on the PreTest and were aware that they were being given the questions again because they were either initially unsure of the answer or had given an incorrect answer on the PreTest. It is surprising then that they would be so confident in their answers on the Google Test. The confidence presumably stems from a perception that the Internet is a reliable source of information.

We also undertook to study the search strategies used by EM residents. The results and analysis of our findings with respect to search strategies will be reported subsequently. The questions used in the study were based on the kinds of information the authors look for daily in their clinical practice and that they have observed EM residents seeking in the clinical setting. We believe the questions are clinically relevant to day-to-day patient care in an ED in the United States. The "correct" answers were determined by 2 experienced American Board of Emergency Medicine certified physicians in active academic clinical practice and were verified by referring to peer-reviewed information sources. Both physicians were sufficiently confident in the accuracy of the answers that they would use the information in clinical practice. Similarly, the EM residents were instructed to answer PreTest and Google Test questions to a degree of certainty that they would feel comfortable using the information for patient care. No strict time limit was placed on EM residents taking the PreTest or the Google Test. Rather, we asked them to spend as much time as they would spend when working in a clinical area. We expected residents to spend approximately 5 minutes per question. They were instructed to use the time either to find an answer that met their internal criteria for clinical certainty, or, as in the "real world" of the ED, to give up using Google and either seek a different source of information or pursue an alternate course of patient care. Thus, we have attempted to include a naturalistic

Table 3. Association of user confidence with percent correct answers on the Google Test.

Variables	Median on 1–5 scale	*P* value
Confidence in using computers	4	0.04
Confidence in using a search engine	4	0.05
Confidence in finding relevant medical information on the Internet	4	0.03
Confidence in reliability of the medical information found on the Internet	3	0.03
Frequency of searching for clinical information	5	0.21
Frequency of using Internet-obtained information for clinical decision making	4	0.48

element into what could be best described as a laboratory exercise.

Google is the most popular ISE and is 1 of the most commonly used sources of clinical information by emergency physicians.[5] Google indexes Web pages based largely on relevance and popularity. Search results ranking by Google do not depend on the accuracy of the retrieved information. The results are displayed in an order determined by the Google proprietary PageRank algorithm. PageRank, a copyrighted process owned by Stanford University, is licensed exclusively to Google. Google's description of PageRank reads as follows:

PageRank relies on the uniquely democratic nature of the Web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links that a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."[6]

Google also does not index the entire Internet. Much information resides in proprietary databases, and older information may not exist in an indexed digital format.[7] The result of a typical Google search consists of many Web pages, each with a list of 10 search results ranked in order as determined by PageRank. Most searchers use results from the first few pages, even though a query often returns hundreds of thousands or even millions of results. The user, particularly those with limited time, may never view most of the links. Because the links are not ranked based on the accuracy of the contents, it would not be unusual to view links with incorrect information. Thus, it is logical to question the accuracy of the information obtained as a result of initiating a search on Google.

Very little research has been published concerning the accuracy of Google medical searches. An article by Tang and Ng[8] looked at using Google as a "diagnostic aid." In that article, the authors searched Google for terms they selected from published case records, which they designated "diagnostic cases." From the first 3 to 5 pages of results returned by Google, the authors selected the "three most prominent diagnoses that seemed to fit the symptoms and signs." If 1 of these 3 diagnoses was correct, they regarded the Google search as providing the correct diagnosis. The authors concluded that it is often useful to "Google for a diagnosis," although they acknowledge that many limitations to their findings exist.

Although the types of information emergency physicians and residents seek during clinical shifts has not been widely studied, we have observed in our own clinical practice and teaching that many of the information queries were not for the purpose of arriving at a global diagnosis, but rather for specific pieces of information that would be useful in patient care. Examples are drug doses, drugs of choice for specific

indications, characteristics of diagnostic tests, frequency of certain findings in a disease state of interest, acceptable treatment alternatives, and so on. This also differs considerably from what Tang and Ng[8] studied. In addition, we sought to design a study with a much more definitive and clinically relevant standard for search accuracy. Little is known about the strategies used to search the Web by EM residents. The study of Graber et al[5] is the only reference that specifically addresses this issue, and the object of that study was not concerned with the accuracy of results.

The survey we conducted before administering the tests asked the residents questions regarding their confidence in using computers in general, searching for medical information, and in the reliability of the information retrieved. Each of these parameters had correlations between the level of confidence and the percentage of correct answers on the Google Test. However, other questions, such as how frequently one conducts searches for answers to clinical questions and how frequently answers retrieved from such searches are actually applied in real clinical situations, did not have correlations with the percentage of correct answers. This seems to imply that a group of residents, although less confident than their peers, still conduct and apply the results of these Internet searches in clinical practice. This subgroup also has a higher likelihood of not getting the correct answer from their search and not recognizing that the answer is incorrect.

It is clear that our subjects often retrieved inaccurate information by using Google, yet the residents believed that the information was reliable enough to use in patient care. This may represent a previously unrecognized source of medical error and a threat to patient safety. Many possible explanations exist for this finding. As others have pointed out, the degree of prior knowledge of a subject may influence search strategies and also influence the searcher's ability to arrive at an accurate result.[9] Interestingly, no correlation was found between residency year and Google Test accuracy. However, given the number of residents per year and the wide variability in other types of test scores within a given year, the lack of correlation is not that surprising.

Our research suggests that searchers who scored higher on the PreTest also scored higher on the Google Test. This finding raises serious questions about whether teaching EM residents to conduct more effective searches will enable the residents to have a higher success rate in answering clinical questions by using an ISE such as Google.[10–12] In many respects, the outcome can have both favorable and unfavorable consequences. Training and education generally result in improved performance, but these findings also indicate that efforts also should focus on improving the residents' knowledge base. The combined protocol could result in searches that produce a higher percentage of correct answers when using an ISE to answer a clinical question. The results also suggest that a more experienced physician, such as an

attending physician, may be more capable of finding the correct answer to a question by using an ISE because of his or her advanced knowledge base and experience. Concomitantly, ED patients and family members, lacking medical knowledge, are vulnerable and more likely to find erroneous medical information when using an ISE to search the Internet.

LIMITATIONS

This was a laboratory study; therefore, caution should be used in translating the results to the clinical setting. In clinical medicine, checks and balances are in place on the use of information; these are not present in the computer laboratory. Pressures to answer questions accurately, as well as time constraints, are different. The actual extent to which residents rely on information from Google searches is not known. Residents may use multiple sources of information and choose between them, based on their prior knowledge and the presumed credibility of the source.

Despite our questioning of residents' use of Internet searches for clinical information and instructions to answer "as if the answer were to be implemented in patient care," no way exists to prove that the answers on the test would translate into actual medical-care errors. No ethical method is known to set up a study in which clinical questions are searched on a computer and then implemented in a real patient-care setting. Our laboratory simulation sought to control the wide variability of the Internet and yet still provide a meaningful estimate as to whether residents can find accurate answers.

Google Scholar (http://scholar.google.com/) is the beta version of a product first released in 2004. It indexes scholarly literature, including peer-reviewed articles, preprints, conference abstracts, theses, and so on. The results of Scholar searches may be more accurate than searches of the general Web using Google, but this has not been verified, and the extent of Google Scholar use by EM residents has not been studied.

CONCLUSION

Technologic innovations, such as Internet access in the ED, should be studied carefully before being accepted as a reliable tool for assisting with clinical decision making. Residents should be instructed to select Internet resources that provide valid, reliable health information. Enlisting the assistance of a health sciences librarian in providing search-strategy training to residents, medical students, and attending physicians can overcome many of the associated pitfalls.

Address for Correspondence: Ronald Moscati, MD, Erie County Medical Center, Department of Emergency Medicine, 462 Grider St, Buffalo, NY 14215. E-mail: moscati@buffalo.edu.

REFERENCES

1. Search Engine Watch. Estimated Internet pages. Search Engine Watch Web site. Available at: www.searchenginewatch.com. Accessed February 2008.
2. Burns E. US core search rankings, February 2008. Search Engine Watch Web site. Available at: http://searchenginewatch.com/3628837. Accessed March 20, 2008.
3. Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *Int J Market Res*. 2008;50:61–78.
4. Likert R. A technique for the measurement of attitudes. *Arch Psychology*. 1932;140:1–55.
5. Graber MA, Randles BD, Ely JW, et al. Answering clinical questions in the ED. *Am J Emerg Med*. 2008;26:144–147.
6. Google Corporation. Page rank technology. Google Web site. Available at: http://www.google.com/corporate/tech.html. Accessed December 15, 2008.
7. Devine J, Egger-Sider F. Beyond Google: the invisible web in the academic library. *J Acad Librarianship*. 2004;30:265–269.
8. Tang H, Ng JH. Googling for a diagnosis–use of Google as a diagnostic aid: internet based study. *BMJ*. 2006;333:1143–1145.
9. Ely JW, Osheroff JA, Chambliss ML, et al. Answering physicians' clinical questions: obstacles and potential solutions. *J Am Med Informat Assoc*. 2005;12:217–224.
10. Schwartz DG, Schwartz SA. MEDLINE training for medical students integrated into the clinical curriculum. *Med Edu*. 1995;29:133–138.
11. Allen MJ, Kaufman DM, Barrett A, et al. Self-reported effects of computer workshops on physicians' computer use. *J Contin Edu Heal Prof*. 2000;20:20–26.
12. Ely JW, Osheroff JA, Maviglia SM, et al. Patient-care questions that physicians are unable to answer. *J Am Med Informat Assoc*. 2007;14:407–414.