Factors associated with successful answering of clinical questions using an information retrieval system*

By William R. Hersh, M.D.
Division of Medical Informatics and Outcomes Research

Oregon Health Sciences University
3181 S.W. Sam Jackson Park Road
Portland, Oregon 97201

M. Katherine Crabtree, R.N., D.N.Sc.
School of Nursing

Oregon Health Sciences University
Portland, Oregon

David H. Hickam, M.D., M.P.H.
Division of Medical Informatics and Outcomes Research
Oregon Health Sciences University and

Health Services Research and Development
Portland VA Medical Center
Portland, Oregon

Lynetta Sacherek, M.L.S.
Division of Medical Informatics and Outcomes Research

Oregon Health Sciences University
Portland, Oregon

Linda Rose, R.N., M.S.N.
School of Nursing

Oregon Health Sciences University
Portland, Oregon

Charles P. Friedman, Ph.D.
Center for Biomedical Informatics

University of Pittsburgh
Pittsburgh, Pennsylvania

Objectives: Despite the growing use of online databases by clinicians, there has been very little research documenting how effectively they are used. This study assessed the ability of medical and nurse-practitioner students to answer clinical questions using an information retrieval system. It also attempted to identify the demographic, experience, cognitive, personality, search mechanics, and user-satisfaction factors associated with successful use of a retrieval system.

Methods: Twenty-nine students completed questionnaires of clinical and computer experience as well as tests of cognitive abilities and personality type. They were then administered three clinical questions

* This study was supported by Grant LM-06311 from the U.S. National Library of Medicine.

to answer in a medical library setting using the MEDLINE database and electronic and print full-text resources.

Results: Medical students were able to answer more questions correctly than nurse-practitioner students before and after searching, but both had comparable improvements in the number of correct questions before and after searching. Successful ability to answer questions was also associated with having experience in literature searching and higher standardized test-score percentiles.

Conclusions: Medical and nurse-practitioner students obtained comparable benefits in the ability to answer clinical questions from use of the information retrieval system. Future research must examine strategies that improve successful search and retrieval of clinical questions posed by clinicians in practice.

INTRODUCTION

A growing number of health care professionals and students use information retrieval (IR) systems to answer clinical questions. Despite this increasing use, there is very little research documenting how effectively these systems are used. Much previous work, summarized by Hersh and Hickam [1], has focused on measuring quantities of relevant documents using recall and precision. While achieving good recall and precision are important for users, these measures present incomplete information for ascertaining successful use of IR systems. In particular, they do not capture the interactive nature of the actual use of systems [2], tend to focus the assessment on the system and ignore the user [3], and do not necessarily correlate with user success [4].

One area of IR evaluation research has focused on the ability of users to perform tasks with the IR system. The premise has been that the primary objective of the user has been to answer questions or to obtain new knowledge, rather than retrieve relevant documents. The first ''task-oriented'' evaluation of an IR system was performed by Egan et al. and evaluated the ability of students to answer questions about statistics using the SuperBook hypertext system [5]. Others have subsequently used this general approach to evaluate the ability of college students to find information in a textbook about Sherlock Holmes [6] and of medical students to answer questions in an online factual database of microbiology [7, 8]. The interactive track at the Text Retrieval Conference (TREC) has also adopted a task-oriented framework in order to assess how well real users could retrieve information from the TREC test collection [9]. This approach has been used to assess medical students using online textbooks [10] and the MEDLINE database [11].

Assessing only the ability to perform tasks is not enough, however, to understand the factors that influence successful use of IR systems or to gather information about how they can be improved. Research in this area must go further and attempt to identify first the factors associated with successful use of IR systems, and then determine ways that this knowledge can lead to improved systems or improved use. The goal of this study was to expand upon the task-oriented approach and to identify user and system factors associated with successful completion of a task, in this case, the answering of clinical questions by medical and nurse-practitioner (NP) students.

In order to determine the factors associated with successful use of IR systems, one must develop a model that incorporates them. Thus, the first step in this investigation was to develop a model of factors that potentially influence successful use of IR systems. The most comprehensive model to date was developed by Fidel and Soergel [12]. This model included many of the demographic, experience, search-mechanics, and user-satisfaction factors typically measured in human-computer interaction studies. It also included individual characteristics of system users, such as cognitive and personality factors.

The authors' past research has assessed the association between successful searching by medical students and some of the elements presented in this model [13–16]. These studies have never been able to show any significant association between successful searching and factors such as age, gender, computer experience, time needed to complete a search, number of search terms used, and user satisfaction with the retrieval system.

Other characteristics of IR system users, such as cognitive and personality factors, have not previously been assessed in studies of medical searching. Many studies have assessed the association of cognitive factors with computer skills, with decidedly mixed results, precluding generalization. However, several cognitive factors have been found to be associated with successful use of computer systems in general or retrieval systems specifically:

- Spatial visualization: The ability to visualize spatial relationships among objects has been associated with

retrieval-system performance by nurses [17], ability to locate text in a general retrieval system [18], and ability to use a direct-manipulation (3-D) retrieval system user interface [19].

- Logical reasoning: The ability to reason from premise to conclusion has been shown to improve selectivity in assessing relevant and nonrelevant citations in a retrieval system [20].
- Verbal reasoning: The ability to understand vocabulary has been shown to be associated with the use of a larger number of search expressions and high-frequency search terms in a retrieval system [21].
- Associational fluency: The ability to associate words in meaning or context has been shown to be associated with effectiveness in using retrieval systems [22].

Another measure of cognitive ability is an individual's general knowledge as measured by a standardized test. No studies of medical searching have attempted to assess the association between success at using a retrieval system and general knowledge, though Wildemuth et al. have shown that domain knowledge in microbiology is not associated with improvements in the ability to use an IR system to answer questions in microbiology [23]. As all medical and NP students are required to take standardized tests prior to admission to their programs of study, the results of such tests are available, though one limitation is that they take different tests—the Medical College Admission Test (MCAT) and Graduate Record Examination (GRE), respectively.

Personality factors are most commonly measured using the Myers-Briggs Type Instrument [24]. This test defines four axes of personality type. Although there are no studies assessing IR system usage or performance based on personality attributes, there have been a large number of studies showing their association with a variety of other intellectual tasks, such as learning styles, achievement, and aptitude, as summarized by DiTiberio [25]. The specialization choices health professional students make have also been shown to correlate with certain personality types [26, 27]. For example, individuals pursuing primary care are more likely to have the ''Extraverted Intuition with Introverted Feeling'' (ENFP) personality type while those pursuing surgical specialties are likely to have the ''Introverted Sensing with Extroverted Thinking'' (ISTJ) personality.

The specific research questions address in this study were:

1. How well are health care personnel (in this case, senior medical students and final-year NP students) able to use an IR system to answer clinical questions correctly?
2. What factors are associated with successful use of an IR system to obtain correct answers to clinical questions?

Table 1
Model of factors influencing successful use of an information retrieval system by end users answering clinical questions in a medical library setting using MEDLINE

1. Demographic	**6. Personality**
Student type: medical vs. NP student	Attitudes: Extrovert–Introvert
Age	Processes of perception: Sensing–Intuition
Gender	Processes of judgement: Thinking–Feeling
Ethnicity	Style of dealing with outside work: Judging–Perceiving
2. Computer experience	**7. Pre-search knowledge**
Use productivity software	Pre-search answer
Own home computer	Pre-search certainty
Own modem at home	**8. Certainty of answer**
Use Internet at home	Search certainty
3. Searching experience	**9. Search mechanics**
Literature searching frequency in last year	Time
4. Attitudes toward computers	Number of citations
Practice easier or harder with computers	Used stacks
Enjoy using computers	Search log number of cycles
5. Cognitive	Search log number of articles viewed
Percentile on standardized test (MCAT or GRE)	**10. User satisfaction**
Spatial visualization	QUIS user satisfaction
Logical reasoning	
Verbal reasoning	
Associational fluency	

METHODS

The experiment consisted of building a model of factors related to searching and then carrying out a set of experiments designed to assess which factors are associated with successful searching. This section describes the model, the experimental protocol, the clinical questions, the two searching sessions, the scoring of answers, and the analysis of results.

Because the model informed the remaining methods of the study, the ''results'' from building it are described here in this methods section. The authors began by eliminating aspects of the Fidel and Soergel model that did not apply to this specific experiment. For example, because the focus was on end-user searching, all factors related to mediated searching could be eliminated. Likewise, because end-user searching on preassigned questions with a single database (MEDLINE) in a library setting was used, the ''variables'' of the search request, database, and setting could be eliminated. The final model of the factors to be assessed related to searching ability is listed in Table 1.

The dependent variable in the model was the ability to answer clinical questions correctly. This variable was operationalized in the study by developing a set of short-answer questions, the answers to which would be obtained by searching MEDLINE. The variables in the model were developed into explicit, measurable data points (Table 2). As described in detail below, some variables were assessed on a per-searcher (taking

Table 2
List of actual factors measured and analyzed in study

Per-searcher data: analyzed versus correct answers per user	
School	M = medical student, N = NP student
Sex	M = male, F = female
Age	In years
Ethnic	White, Hispanic, or Black
Experience	Sum of four variables for computer experience in Table 1
LitSrchYr	Literature searches in last year (1 = 0 searches, 2 = 1–2 searches, 3 = 3–5 searches, 4 = 6–11 searches, 5 = 12+ searches)
Attitude	Sum of two variables for computer attitude in Table 1
StdTest	Percentile on standardized test (MCAT or GRE)
VZ-2	Score on paper-folding test for spatial visualization
RL-1	Score on nonsense syllogisms test for logical reasoning
V-4	Score on advanced vocabulary test for verbal reasoning
FA-1	Score on controlled associations test for associational fluency
EMinusI	Score on Myers-Briggs Extrovert–Introvert axis (E positive)
SMinusN	Score on Myers-Briggs Sensing–Intuition axis (S positive)
TMinusF	Score on Myers-Briggs Thinking-Feeling axis (T positive)
JMinusP	Score on Myers-Briggs Judging–Perceiving axis (J positive)
QUIS	Average score on QUIS
PreScore	Average number of questions correct before searching (0–3)
PostScore	Average number of questions correct after searching (0–3)
Improvement	Average improvement before and after searching (0–3)
Per-question data: analyzed versus correct answer for question	
PreCorrect	True if question answered correctly before searching
PreCertainty	Certainty of answer prior to searching (1 = most, 5 = least)
PostCertainty	Certainty of answer after searching (1 = most, 5 = least)
Order	Order that question was searched (e.g., 1 = first, 3 = third)
Time	Time taken to complete question in minutes
Citations	Number of citations listed in justification of answer
Sets	Number of search sets (search terms or Boolean combinations)
TotDocs	Total number of MEDLINE references viewed
Stacks	True if went to library stacks to answer question
FTDocs	Total number of full-text documents viewed

the total score of the three questions) while others were assessed on a per-question (analyzing each question individually) basis.

The clinical questions for searching were taken from three sources that represented a diverse spectrum of real world versus examination-style information queries. In order to have some questions that were likely to have answers, that authors chose from the Cochrane Database of Systematic Reviews (Update Software, Oxford, U.K.), a collection of clinical reviews for which the topic was exhaustively searched and reviewed. Also, to have questions generated by actual clinicians during the course of their practice, the authors obtained another group from a physician information-needs study that were known to have answers that could be found in MEDLINE [28]. As the authors also wanted to include some traditional examination-style questions, they collected a third group from the *Medical Knowledge Self-Assessment Program* (*MKSAP,* American College of Physicians, Philadelphia, PA), a continuing medical education resource, and converted them from multiple-choice to short-answer form. Because medical knowledge could have changed from the time the questions were developed for their original purpose, the authors verified the answers to each question after the actual searching session, as described below. A sample question from each group is shown in Table 3.

Table 3
A sample question from each of the three groups of questions

Question group	Sample
Cochrane Database of Systematic Reviews	Does antibiotic treatment reduce duration of symptoms in patients with sore throat?
Clinical practice	Is mortality or are complications reduced in advanced atherosclerotic disease by aggressive diet therapy?
Medical Knowledge Self-Assessment Program	A thirty-eight-year-old man with positive results on HIV-antibody testing has had increasing headaches for ten weeks. Results of the neurologic examination and computed tomography with contrast are normal. What is the best next diagnostic test to order?

To obtain subjects for the experiment, senior medical students from Oregon Health Sciences University (OHSU) and NP students from OHSU and University of Portland (UP) were recruited for the study by electronic mail, paper mail, and, in the case of NP students, announcements in classes. Students were offered remuneration of $100 for successful completion of all tasks. The general experimental protocol was to participate in two sessions: a ''large-group'' session, where they would be administered questionnaires and would receive some orientation to MEDLINE and the experiment, followed by a ''hands-on'' session, where they would do the actual searching and answering of questions.

The large-group sessions, consisting of anywhere from three to fifteen subjects at a time, took place in a classroom setting. At each session, subjects were first administered a questionnaire that collected information for the first seven factors listed in Table 2. They next signed a consent form allowing the research team to obtain their standardized test score from their respective deans' offices. Because medical and NP students had taken different standardized tests, the MCAT and GRE respectively, their raw scores were converted to percentiles to allow more direct comparison.

The cognitive attributes were measured by instruments from the Educational Testing Service (ETS) Kit

of Cognitive Factors [29] listed below (ETS mnemonic in parentheses):
1. Paper-folding test to assess spatial visualization (VZ-2)
2. Nonsense syllogisms test to assess logical reasoning (RL-1)
3. Advanced vocabulary test I to assess verbal reasoning (V-4)
4. Controlled associations test to assess associational fluency (FA-1)
The personality attributes were measured via administration of the self-assessment version of the Myers-Briggs Type Instrument.

Following collection of data in the large-group session, the subjects were then provided a brief orientation to the searching task of the experiment. This orientation was followed by a thirty-minute demonstration and hands-on training with six basic MEDLINE searching features: Medical Subject Headings (MeSH), text words, explosions, combinations, limits, and scope notes. These features were chosen because they were the features taught in medical informatics training courses for health care providers offered at OHSU, and the authors believed they comprised the basic skill set for MEDLINE searching by a health care provider. The purpose of providing this instruction was to ensure that subjects had a baseline of skills using MEDLINE. The teaching was done by a medical informaticist experienced in teaching MEDLINE to clinicians.

The final activity of the large-group session was the administration of ten randomly selected questions to each subject to help select for the experiment a group of questions that had varying levels of difficulty. The purpose of this part of the session was to develop a question set that would be large enough to have a variety of question types and levels of difficulty, while still small enough to ensure that each question would be searched many times, resulting in the retrieval of many articles and the likelihood that at least some of the searchers would find the correct answer. The research team had initially collected twenty-four questions from each of the three sources, but decided that ten questions from each would be more appropriate, as the team anticipated recruiting a total of thirty subjects who would search three questions each. This format would allow each question to be searched by three different subjects.

The question culling was performed by giving the subjects ten questions each, selected randomly, and asking them to provide the answer and their certainty of it using a one (most) to five (least) scale. Because the research team did not have the answers ahead of time, the difficulty of each question was determined by the average certainty rating for each student who answered it in this session. In each of the three groups of questions (Cochrane, information-needs study, *MKSAP*), the team sorted the questions from least to most difficult and chose the highest, lowest, and every third ranking question in between. This procedure assured use of the full spectrum of the certainty scale and, presumably, question difficulty.

The individual, hands-on sessions took place anywhere from two to four weeks after the subjects had completed the large-group session. They were encouraged to practice the searching skills taught in the large-group session but were given no other explicit instructions. The sessions took place in the OHSU Library. All searching was done using the Ovid IR system (Ovid Technologies, New York, NY) accessing MEDLINE and a collection of eighty-five full-text journals. The Web version of Ovid was used, including its logging facility with which all search statements could be recorded along with the number of citations presented and viewed by the user in each set. Searching was done using Apple PowerMac computers running Netscape Navigator.

In the hands-on sessions, subjects were assigned one question each, randomly selected, from the three groups of questions described above. They were not assigned a question that they had been asked to answer during the question-culling part of the large-group session. Subjects were limited to one hour per question. Before searching, the subjects were asked the answer and their certainty of it on a one (most certain) to five (least certain) scale for the questions which they would be assigned to search.

Subjects were instructed to perform their searching in MEDLINE and then to obtain articles they wanted to read, either in the library stacks or in the full-text collection available online. They were asked to record on paper their answer, the certainty of their answer (on the one-to-five scale), the article or articles that justified their answer, and any article that was looked at in the stacks or in full-text on the screen. Upon completion of the searching, they were administered the Questionnaire for User Interface Satisfaction (QUIS) 5.0 instrument, which measures user satisfaction with a computer system [30]. QUIS provided a score from zero (poor) to nine (excellent) on a variety of user factors, with the overall score determined by averaging responses to each item.

Searching time for each question was measured using a wall clock. All user-system interactions were logged by the Ovid system software. The search logs were processed to count the number of search cycles (each consisting of the entry of a search term or Boolean combination of sets) and the number of full MEDLINE references viewed on the screen.

After all of the hands-on searching sessions were completed, the actual answers to the questions were determined by the research team. This determination was done by assembling all of the articles retrieved for each question and giving them, along with the question, to three members of the study team. The three

first designated an answer individually (blinded to any answers that subjects may have provided) and then worked out their differences by consensus. After the answers were designated, two members of the study team graded the answer forms, resolving any differences by consensus.

After the answers were obtained, each subject was assigned pre-searching and post-searching scores from zero to three indicating the number of questions answered correctly before and after searching. The per-searcher data (items listed in Table 2) were collected from these scores, the instruments from the large-group session, the percentiles from the standardized test scores from the respective deans' offices, and the score on QUIS. The per-question data were obtained from answers on the form completed during and after searching as well as the Ovid searching logs. All data were placed in a Microsoft Excel spreadsheet and transferred to the JMP statistical package for analysis.

Per-searcher data analysis was performed using the post-searching score as the dependent variable. Statistical significance was assessed by using chi-square tests for nominal and ordinal data and one-way analysis of variance (ANOVA) for continuous data. Three statistical comparisons were made for each factor: (1) association between status as medical students versus NP students, (2) association between the factor and the post-searching score, and (3) association between the factor and the improvement in the post-searching score from the pre-searching score (e.g., if a searcher answered one of the three assigned questions correctly before searching and two afterwards, improvement would be +1). A statistically significant result in the first association would indicate a factor had a difference unlikely to be due to chance between medical and NP students. A statistically significant result in the second association would indicate a factor had a non-chance association with a better post-searching score, while one in the third association would indicate a factor had such an association with a better improvement in post-searching score over pre-searching score.

Per-question data analysis used correctness of the answer as the dependent variable. As with the per-searcher data, statistical significance was assessed by using chi-square tests for nominal and ordinal data and one-way ANOVA for continuous data. Because these data were assessed on a per-question level, only the association between the factor and the correct answer was assessed.

RESULTS

Twenty medical students and nine NP students completed the study. They answered three questions each, generating a total of eighty-seven answered questions. One medical and one NP student completed the large-group session but never attended the hands-on searching session; their data were not used in the analysis. Table 4 shows a two-by-two contingency table of total correct and incorrect questions before and after searching, subdivided by student type. Twenty of the eighty-seven questions remained incorrect after searching, including three that the subject had answered correctly initially.

Table 4
Pre- and post-searching questions correct for all subjects, medical students only, and NP students only

	Correct before searching	Incorrect before searching
Correct after searching		
All	36	31
Medical students	29	21
NP students	7	10
Incorrect after searching		
All	3	17
Medical students	2	8
NP students	1	9

The average pre-searching score was 1.4, while the average post-searching score was 2.3, showing an average improvement of 0.9. Medical students had higher scores both before and after searching. Table 5 lists the results of the per-searcher data elements from Table 3. The first three columns of the table show the mean and standard deviation for all subjects, medical students only, and NP students only. The final three columns show statistical significance for (1) the association of differences between medical and NP students, (2) the association between the factor and post-searching score, and (3) the association between the factor and improvement in score after searching as described above.

Table 5 shows that the NP students were more likely to be female and older, to have less literature searching experience, to score lower on standardized test percentile, and to score lower on cognitive tests with the exception of associational fluency (FA-1). Their overall results showed both lower pre-searching and post-searching scores. However, their improvement in scores was nearly identical to that of the medical students, implying that they achieved comparable benefit from the IR system as medical students did.

Fewer factors were associated directly with a higher post-searching score. As noted above, medical students had statistically significant higher post-searching scores. Other factors associated with a higher post-searching score included self-reported experience in literature searching, percentile on standardized test, and improvement in number of questions correct after searching. There was also a trend toward positive association with the spatial visualization (VZ-2) score.

No factors achieved statistical significance in association with improvement in post-searching score over

Table 5
Pre-searcher results for factors and statistical comparison for medical versus NP students, association with post-searching score, and association with improvement

Factor	Mean (SD) for all	Mean (SD) for medical students	Mean (SD) for NP students	Difference for medical vs. NP students (*P* value)	Association with post-searching score (*P* value)	Association with improvement (*P* value)
School*				N/A	0.02	NS
Sex*				0.005	NS	NS
Age	33.3 (7.6)	30.4 (5.1)	39.9 (8.2)	0.001	NS	NS
Ethnic*				NS	NS	NS
Experience	5.6 (1.4)	6.0 (1.5)	4.8 (1.0)	NS	NS	NS
LitSrchYr	4.4 (1.0)	4.9 (0.3)	3.3 (1.1)	< 0.001	0.01	NS
Attitude	3.2 (0.6)	3.1 (0.6)	3.4 (0.5)	NS	NS	NS
StdTest	60.8 (21.8)	71.8 (12.6)	36.2 (17.5)	< 0.0001	0.01	NS
VZ-2	12.7 (5.2)	15.2 (3.8)	7.2 (3.3)	< 0.0001	0.06	NS
RL-1	12.4 (8.8)	15.6 (8.1)	5.4 (6.0)	0.001	NS	NS
V-4	21.0 (6.5)	23.6 (5.1)	15.2 (5.4)	0.0002	NS	0.06
FA-1	49.8 (12.8)	51.1 (13.6)	46.8 (11.1)	NS	NS	NS
EMinusI	−0.2 (14.1)	−3.2 (14.5)	6.2 (11.2)	NS	NS	NS
SMinusN	−4.0 (15.1)	−7.2 (16.0)	3.0 (10.4)	NS	NS	NS
TMinusF	1.5 (13.5)	0.8 (14.0)	3.0 (13.0)	NS	NS	0.06
JMinusP	−1.7 (14.6)	0.2 (15.1)	−5.7 (13.3)	NS	NS	NS
QUIS	6.3 (1.3)	6.3 (1.4)	6.4 (1.0)	NS	NS	NS
PreScore	1.4 (0.9)	1.6 (0.8)	0.9 (0.9)	0.03	NS	N/A
PostScore	2.3 (0.7)	2.5 (0.5)	1.9 (0.8)	0.02	N/A	N/A
Improvement	0.9 (0.9)	0.9 (0.9)	1.0 (1.1)	NS	0.02	N/A

* Summaries of these nominal variables are given in the text of the paper. Statistical *P* values < 0.1 are given for all differences.

pre-searching score. However, two factors—higher verbal reasoning (V-4) score and thinking (as opposed to feeling) personality type—showed trends toward significant association with improvement in scores.

Table 6 lists the factors that were analyzed on a per-question basis. Knowing the answer ahead of time was associated with obtaining a correct answer. Also associated with obtaining a correct answer was a higher number of citations that justified the answer, although the magnitude of the difference (2.3 versus 2.1) was small. Higher certainty of correctness showed a trend associated with obtaining a correct answer. The order in which the question was searched (first, second, or third) and time taken for the search had no effect.

Table 6
Per-question factors and statistical association with correct answers

Numerical data	Incorrect	Correct	*P* value
PreCorrect*			0.002
PreCertainty	2.7 (1.3)	2.6 (1.3)	NS
PostCertainty	1.6 (0.9)	1.6 (0.8)	0.08
Order	2.0 (0.8)	2.1 (0.8)	NS
Time	31.2 (16.1)	30.4 (16.0)	NS
Citations	2.1 (1.3)	2.3 (1.3)	0.01
Sets	9.1 (6.8)	8.7 (6.9)	NS
TotDocs	8.4 (7.0)	8.0 (6.8)	NS
Stacks*			NS
FTDocs	0.9 (1.4)	0.8 (1.3)	NS

* A summary of these nominal variables is given in the text of the paper.

DISCUSSION

The research questions addressed by this study included an assessment of how well medical and NP students could use an IR system to improve their ability to answer clinical questions correctly, and what factors were associated with that ability. This study showed that both medical and NP students who used a state-of-the-art MEDLINE access system were able to improve their ability to answer questions. While the medical students had better overall ability to answer questions, they also had higher baseline knowledge. Both groups showed that the IR system improved their rate of correct answers.

The most significant factors associated with successful question answering were being a medical student, knowing the answer ahead of time, having a higher standardized test score, and having more literature searching experience. There were no cognitive or personality factors or measures of computer experience or attitude associated with improved success at using retrieval systems. Whether the medical students' greater ability to answer questions related to inherent intellectual ability or their advanced training relative to NP students was not discernable from these data.

A more important question to study, however, was how well IR systems improve the ability of users to answer questions over their baseline knowledge. By looking at improvement in scores before and after searching, the results showed that medical and NP

students benefited equally from the IR system. There were no significant differences in any other factor related to improvement in searching, including demographic variables, cognitive abilities, personality traits, or user satisfaction.

Another goal of this study was to learn about factors associated with searching success, and how they can be used to build better systems or improve user training. While it is reassuring that all users could benefit in their clinical practice from searching, the data did not uncover factors that could be used to guide improvements in systems that would lead to further benefits for users. Obviously, improving the literature-searching experience of users would enhance their ability to use IR systems more effectively. However, the results of this study did not indicate any interventions that would improve users' abilities to answer clinical questions by searching.

There were some limitations to the study. The use of students, albeit in late stages of their training, limited the generalizability of the results beyond others at their level of clinical training. In future studies, community practitioners will also be included. This study was also limited by taking place in a laboratory setting, in that behaviors in the pursuit of actual clinical knowledge in the real clinical setting might be different than those exhibited in this controlled environment. However, the ability to use a defined set of tasks and questions provided a benefit that could not be obtained in the real clinical setting. A final limitation of the study was not incorporating the notion of users judging the quality of evidence for the information they obtained. Many advocate that being able to judge the evidence is a critical skill [31], and this task will be incorporated into future studies.

While this study supplies some insight into the factors associated with successful searching, it does not by itself provide all the answers. It does, however, provide a methodology for researchers to begin probing what aspects of the end-user searching process could have interventions for improvement, such as systems with better features or abilities of experts to train beginners better. Future studies, already underway, by the present authors, will employ more subjects, more questions, assessment of evidence-based approaches, and clinicians more advanced in their training.

These initial results nonetheless demonstrate that IR systems are beneficial for both types of clinical practitioners by improving the ability to answer clinical questions. While those with higher baseline knowledge achieve higher absolute benefit, clinicians from the entire spectrum improve their knowledge equally. The continuing challenge is to build more effective systems and to teach users how to use them for maximum benefit.

REFERENCES

1. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? a framework for investigation and review of the literature. J Am Med Assoc 1998 Oct 21;280(15):1347–52.
2. Swanson DR. Historical note: information retrieval and the future of an illusion. J Am Soc Info Sci 1988 Mar;39(2): 92–8.
3. Harter SP. Psychological relevance and information science. J Am Soc Info Sci 1992 Oct;43(9):602–15.
4. Hersh WR. Relevance and retrieval evaluation: perspectives from medicine. J Am Soc Info Sci 1994 Apr;45(3):201–6.
5. Egan DE, Remde JR, Gomez JM, Landauer TK, Eberhardt J, Lochbaum CC. Formative design-evaluation of Superbook. ACM Transactions on Information Systems 1989 Jan;7(1):30–57.
6. Mynatt BT, et al. Hypertext or book: which is better for answering questions? In Proceedings of Computer-Human Interface 92, 1992:19–25.
7. Wildemuth BM, de Bliek R, Friedman CP, File DD. Medical students' personal knowledge, searching proficiency, and database use in problem solving. J Am Soc Info Sci 1995 Sep;46(8):590–607.
8. Friedman CP, Wildemuth BM, Muriuki M, Gant SP, Downs SM, Twarog RG, de Bliek R. A comparison of hypertext and Boolean access to biomedical information. In: Proceedings of the 1996 Annual AMIA Fall Symposium. Washington, DC: Hanely-Belfus, 1996:2–6.
9. Lagergren E, Over P. Comparing interactive information retrieval systems across sites: the TREC-6 interactive track matrix experiment. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia: ACM Press, 1998:164–72.
10. Hersh WR, Elliot DL, Hickam DH, Wolf SL, Molnar A, Leichtenstien C. Towards new measures of information retrieval evaluation. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, WA: ACM Press, 1995:164.
11. Hersh W, Pentecost J, Hickam D. A task-oriented approach to information retrieval evaluation. J Am Soc Info Sci 1996 Jan;47(1):50–6.
12. Fidel R, Soergel D. Factors affecting online bibliographic retrieval: a conceptual framework for research. J Am Soc Info Sci 1983;34(3):163–80.
13. Hersh. Towards new measures of information retrieval evaluation, op. cit.
14. Hersh. A task-oriented approach to information retrieval evaluation, op. cit.
15. Hersh WR, Buckley C, Leone TJ, Hickam DH. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval. Dublin, Ireland: Springer-Verlag, 1994:192–201.
16. Hersh WR, Hickam DH. An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. J Am Soc Info Sci 1995 Aug;46(7):478–89.
17. Staggers N, Mills ME. Nurse-computer interaction: staff performance outcomes. Nursing Research 1994 May–Jun;43(3):144–50.

18. GOMEZ LM, EGAN D, BOWERS C. Learning to use a text editor: some learner characteristics that predict success. Human-Computer Interaction 1986;2(1):1–23.
19. SWAN RC, ALLAN J. Aspect windows, 3-D visualization, and indirect comparisons of information retrieval systems. In: Proceedings of the 21st Annual International ACM Special Interest Group in Information Retrieval. Melbourne, Australia: ACM Press, 1998:173–81.
20. ALLEN BL. Cognitive differences in end-user searching of a CD-ROM index. In: Proceedings of the 15th Annual International ACM Special Interest Group in Information Retrieval. Copenhagen, Denmark: ACM Press, 1992:298.
21. IBID.
22. DUMAIS ST, SCHMITT DG. Iterative searching in an online database. In: Proceedings of the Human Factors Society 35th Annual Meeting, 1991:398–403.
23. WILDEMUTH, op. cit.
24. MYERS IB, MCCAULLEY MH. Manual: a guide to the development and use of the Myers-Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press, 1985.
25. DITIBERIO JK. Education, learning styles, and cognitive styles, in MBTI applications: a decade of research on the Meyers-Briggs Type Indicator. Hammer AL, ed. Palo Alto, CA: Consulting Psychologists Press, 1996:123–66.
26. JAIN VK, LALL R. Nurses' personality types based on the Myers-Briggs type indicator. Psychological Reports 1996 Jun; 78(3 pt. 1):938.
27. LESTER WM, WOLOSCHUK W, JUSCHAKA B, MANDIN H. Assessing the psychological types of specialists to assist students in career choice. Acad Med 1995 Oct;70(10):932–3.
28. GORMAN PN, HELFAND M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Med Dec Making 1995 Apr–Jun;15(2):113–9.
29. EKSTROM RB, FRENCH JW, HARMON HH. Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service, 1976.
30. CHIN JP, DIEHL VA, NORMAN KL. Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of CHI '88—Human Factors in Computing Systems. New York, NY: ACM Press, 1988: 213–8.
31. SACKETT DL, RICHARDSON WS, ROSENBERG W, HAYNES RB. Evidence-based medicine: how to practice and teach EBM. New York, NY: Churchhill Livingstone, 1997.

Received January 2000; accepted May 2000