How Well Do Physicians Use Electronic Information Retrieval Systems?

A Framework for Investigation and Systematic Review

William R. Hersh, MD; David H. Hickam, MD, MPH

Objective.—Despite the proliferation of electronic information retrieval (IR) systems for physicians, their effectiveness has not been well assessed. The purpose of this review is to provide a conceptual framework and to apply the results of previous studies to this framework.

Data Sources.—All sources of medical informatics and information science literature, including MEDLINE, along with bibliographies of textbooks in these areas, were searched from 1966 to January 1998.

Study Selection.—All articles presenting either classifications of evaluation studies or their results, with an emphasis on those studying use by physicians.

Data Extraction.—A framework for evaluation was developed, consisting of frequency of use, purpose of use, user satisfaction, searching utility, search failure, and outcomes. All studies were then assessed based on the framework.

Data Synthesis.—Due to the heterogeneity and simplistic study designs, no meta-analysis of studies could be done. General conclusions were drawn from data where appropriate. A total of 47 articles were found to include an evaluation component and were used to develop the framework. Of these, 21 articles met the inclusion criteria for 1 or more of the categories in the framework. Most use of IR systems by physicians still occurs with bibliographic rather than full-text databases. Overall use of IR systems occurs just 0.3 to 9 times per physician per month, whereas physicians have 2 unanswered questions for every 3 patients.

Conclusions.—Studies comparing IR systems with different searching features have not shown that advanced searching methods are significantly more effective than simple text word methods. Most searches retrieve only one fourth to one half of the relevant articles on a given topic and, once retrieved, little is known about how these articles are interpreted or applied. These studies imply that further research and development are needed to improve system utility and performance.

JAMA. 1998;280:1347-1352

THE USE OF computerized information retrieval (IR) systems by physicians has been increasingly advocated for enhancing the quality of patient care,[1] providing better use of evidence,[2,3] and allowing generalist physicians to keep up with health information.[4] A growing number of products are available at modest cost on computer networks (including the Internet) and CD-ROM. Despite the easy availability of these products, it is not clear how well physicians use them, whether they are cost-effective, or how often they lead to better clinical decisions. The goals of this article are to examine studies of the clinical use of these products by developing a framework for evaluation and to describe prior studies and their limitations.

From the Division of Medical Informatics & Outcomes Research (Drs Hersh and Hickam) and the Department of Medicine (Dr Hersh), Oregon Health Sciences University, and Health Services Research and Development, Department of Medicine, Portland VA Medical Center (Dr Hickam), Portland, Ore.

Reprints: William R. Hersh, MD, Division of Medical Informatics & Outcomes Research, School of Medicine, Oregon Health Sciences University, 3181 SW Sam Jackson Park Rd, BICC, Portland, OR 97201 (e-mail: hersh@ohsu.edu).

METHODS

The data sources for this review were obtained in January 1998 by searching back to 1966. Citations came from the medical informatics and information science literatures via the MEDLINE and Library and Information Science (LISA, RR Bowker, New Providence, NJ) databases. All searching and analysis were performed by the first author (W.R.H.). MEDLINE was searched using the Medical Subject Headings (MeSH) terms *information storage and retrieval*, *information systems*, and *evaluation studies* (exploded to include more specific terms). In addition, bibliographies of textbooks and conference proceedings not indexed in MEDLINE and LISA were searched by hand. Articles were selected if they presented either classifications of IR evaluation studies or results that assessed the use of systems by physicians or medical students. A framework for evaluation was then developed by starting with previously described theoretical models.[5,6] The framework was subsequently modified on the basis of the classification of articles retrieved for this study. After the new framework was developed, the retrieved studies were assessed based on whether they satisfied methodologic criteria defined by the classification.

The framework for IR system evaluation was composed of 6 criteria, which are covered in the 6 parts of the "Results" section:

Table 1.—Comparison of Studies of Measured Use of Information Retrieval Systems in Clinical Settings

Study (Retrieval System Assessed)	No. of Users	Length of Observation, mo	Use per Person-Month
Horowitz et al[7] (MEDLINE via PaperChase)	3654	36	0.3*
Collen and Flagle[8] (MEDLINE and full-text journals via MEDIS)	508	4	6.1*
Markert et al[9] (MEDLINE and full-text journals via MEDIS)	**187**	11	7.2*
Students	75		9.0
Residents	22		.8
Full-time faculty	46		6.7
Volunteer faculty	44		5.0
Haynes et al[10] (MEDLINE via Grateful Med)	**158**	8	2.7
Students	30		3.2
Interns	22		3.0
Residents	45		2.6
Fellows	14		1.5
Attending staff	47		0.9
Abate et al[11] (BRS and Dialog)	43	19	0.7
Hersh and Hickam[12] (MEDLINE via Knowledge Finder, textbooks, QMR, and Yearbook Series)	**31**	10	1.2
Interns	8		0.6
Junior residents	9		1.2
Senior residents	8		1.3
Attending staff	6		2.1

*Use per person-month from these studies was estimated from data reported in the articles.

1. *Frequency of use.* Measurement of IR system access by the users to whom it is made available.
2. *Purpose of use.* The clinical questions posed to IR systems.
3. *User satisfaction.* Measured by instruments such as questionnaires.
4. *Searching utility.* The effectiveness of IR systems for searching.
5. *Search failure.* The retrospective determination of the problems that led to the system not performing as anticipated (known as *failure analysis*).
6. *Outcomes.* Measurement of whether the IR system leads to some type of improved health care delivery outcome, such as better quality or cost-effectiveness.

The initial literature searching focused on finding articles that were on the topic of information or library retrieval systems in medicine. Of the 619 articles assessed for possible inclusion, only 47 were found to have an evaluation component (complete reference list available from authors on request). These articles were then used to develop the above framework. When the framework was completed, each article was assigned to 1 or more of the criteria categories above, and each was assessed to determine whether the methodology was sufficient to warrant generalizable conclusions. Articles were excluded if studies assessed completely outdated technology (eg, studies from the 1960s assessing MEDLINE by queries mailed to a regional library) or used inadequate methodology (eg, extremely small samples of users or queries, exclusive reliance on retrospective questionnaire data, or study designs yielding dubious causal relationships). Of the 47 articles with an evaluation component, 21 met the inclusion criteria of 1 or more of the 6 evaluation factors.[1,7-26]

RESULTS

Frequency of Use

Certainly a valuable measure of a technology such as IR is whether a system is actually used by its intended audience. While several studies have measured use by questionnaire,[27-29] these are potentially subject to recall bias. Hence, this section focuses on studies in which use was directly tracked by the computer system. Six long-term evaluation studies have measured use in clinical settings by direct monitoring, as follows:

1. Horowitz et al[7] made the PaperChase system (Beth Israel Hospital, Boston, Mass) available on terminals available throughout the hospital and clinics at Beth Israel Hospital.
2. Collen and Flagle[8] provided MEDIS (no longer available), a system featuring access to MEDLINE references as well as the full text of several journal titles and textbooks at 7 hospitals nationwide, with a mix of academic and community settings.
3. Markert et al[9] made MEDIS available to students and faculty at 6 hospital sites and via modem from home at Wright State University, Dayton, Ohio.
4. Haynes et al[10] observed the use of Grateful Med (National Library of Medicine [NLM], Bethesda, Md) in 5 sites at McMaster University Medical Centre (Hamilton, Ontario): the emergency department, intensive care unit, ambulatory clinic, and 2 inpatient hospital wards.
5. Abate et al[11] performed a randomized comparative trial making BRS (now Ovid, Ovid Technologies, New York, NY) or Dialog (Dialog, Palo Alto, Calif) available in 3 office-based practices, a clinical pharmacy group, and a university-based family medicine practice in West Virginia.
6. Hersh and Hickam[12] assessed a multiapplication workstation in the General Medicine Clinic at Oregon Health Sciences University, featuring access to Knowledge Finder (KF, Aries Systems, North Andover, Mass), MEDLINE, Stat!-Ref electronic textbooks (Teton Data Systems, Jackson, Wyo), the Yearbook Series (CMC Research, Portland, Ore), and the decision support system Quick Medical Reference (QMR, First Databank, San Francisco, Calif).

The results of physician use in these studies are summarized in Table 1. While the different study situations (ie, different databases, clinical environments, specialties, and period of observation) make direct comparison difficult, overall use of systems is small. This is particularly true since other studies have shown that physicians have an unmet information need for 2 of every 3 patients seen,[30,31] and that about half of these needs can be met by searching MEDLINE.[32] The studies summarized in Table 1 show that medical IR systems are consulted during clinical care only a few times a month or less. While an IR system is not needed to answer every clinical question that arises, clearly these systems are not playing a major role in meeting overall physician information needs.

These studies also address whether databases other than MEDLINE might enhance use of IR systems. The study by Abate et al[11] offered access to all 100 to 200 databases on the BRS and Dialog systems, yet MEDLINE was used 53% to 71% of the time, followed by the bibliographic database EMBASE (14%) and the full-text journals on BRS (11%). Likewise, Hersh and Hickam[12] found that MEDLINE was used substantially more frequently (68%) than electronic textbooks (19%), QMR (6%), or the Yearbook Series (2%). Thus, bibliographic databases, especially MEDLINE, are still the flagship IR applications in health care.

Purpose of Use

Three articles met the inclusion criteria for purpose of use evaluation fac-

tor.[9,10,12] Of the studies that measured use in clinical settings, Haynes et al[10] and Hersh and Hickam[12] attempted to classify users' statements of information need. Both studies found questions of therapy to be most frequent (26%-41%) followed by overview or review (12%-23%). Markert et al[9] and Haynes et al[10] also asked users their reason for searching. The former found that medical students were most likely to search for patient care (87%), whereas full-time faculty searched most frequently for research and/or scholarly writing (72%) and volunteer faculty searched most often for conference or other presentations (64%). Haynes et al[10] reported that 56% of searches were done for patient care, with the rest distributed among rounds (8%), research (7%), teaching (1%), and miscellaneous or unstated reasons (28%).

User Satisfaction

Another method of evaluating the impact of an IR system is to measure user satisfaction. Six articles met the inclusion criteria for user satisfaction.[7-12] In 3 of the 6 studies that measured use frequency, user satisfaction was reported to be high. For example, Collen and Flagle[8] found that 75% of users planned to continue using the system after its initial pilot implementation. Likewise, the clinicians studied by Haynes et al[10] responded in general that the MEDLINE system was convenient and easy to use, that searching was not time-consuming, and that it was preferable to other information sources. In addition, Hersh and Hickam[12] found that the multiapplication system was easy and quick to use. They also found that users felt more comfortable using MEDLINE and improved their skills in finding relevant articles during the study.

Other studies, however, provided data counter to the high degree of user satisfaction reported in the studies described above. It can be seen in Table 1 that system use tended to correlate inversely with length of the study. While this could be an artifact of the discordant study designs, it may also represent a novelty effect, in that system use may fall off when the excitement of a new technology fades. The decline of use over time has been noted elsewhere; Marshall[13] followed up a group of "early adopters" of Grateful Med, revisiting them after 3 years. She found that one-third had given up searching, with the most frequent reasons cited including system too difficult to use (24%), content poor or inappropriate (21%), system too slow (17%), and user too busy to search (14%). Another insight into the value of online searching was the observation by Haynes et al[14] that searching decreased by two thirds when user fees were added. While MEDLINE is now freely available and universally accessible on the Internet, monetary value is certainly one method of assessing value of an IR system, and these results suggest that satisfaction may have been more modest than the self-reported questionnaire data from the studies in Table 1.

Searching Utility

Although the use frequency and satisfaction of users is important, it is also beneficial to understand how effectively users search with IR systems. A total of 10 studies met the inclusion criteria for searching utility.[10,12,15-22] Two approaches to assess effectiveness of searchers and their searching have been used. Probably the most clinically pertinent of these measures is whether systems can actually help clinicians answer questions concerning patient care. Another approach has been to assess the quantity of articles retrieved that are relevant to the search topic.

Answering Clinical Questions.—Assessing how well medical users answer clinical questions with IR systems was first performed by Wildemuth et al,[15] who assessed factors associated with searching success by medical students in factual databases. They found that the number of relevant items retrieved, term overlap (ie, students selecting terms overlapping with those known to lead to retrieval of records containing the answer), and efficiency (as measured by time) had a positive correlation with successful answering of questions, while personal domain (preexisting) knowledge did not.

Hersh et al[16] implemented a question-answer approach to compare 2 different MEDLINE systems that represent the ends of the spectrum in terms of using Boolean searching on human-indexed thesaurus terms (Ovid) vs "natural language" searching on words in the title, abstract, and indexing terms (KF). Medical students were recruited and randomized to 1 of the 2 systems and given 3 clinical questions from a collection of critically appraised topics to answer. The students were able to use each system successfully, with no significant differences in questions correctly answered. On these yes-no questions, students were able to answer about 85% of questions using either system, with most incorrect answers concentrated among a few difficult questions. There was also no difference between systems in time taken to answer the question, number of relevant articles retrieved, or user satisfaction between the systems. Another insight from this study was that the time taken to answer questions using MEDLINE averaged close to 30 minutes, demonstrating the impracticality of its routine use in the busy clinical setting.

Retrieval of Relevant Articles.—The most frequently used measures to assess the effectiveness of searchers and their searching have been recall and precision. These measures estimate the quantity of relevant articles retrieved, although this may not always be the most important aspect of a search done for clinical care. Clinicians may instead be interested in how effectively searches answer clinical questions. Nonetheless, these measures have been used extensively, and many large IR evaluation studies have used them.

For a query, *recall* is the proportion of relevant documents retrieved from the database calculated as the number of relevant documents retrieved in the search divided by the total number of relevant documents in the entire database.

One problem with the measure of recall is that the denominator implies that the total number of relevant documents for a query is known, which is impossible for large databases. In this situation, a measure that approximates recall, called *relative recall*, is used. This measure uses in the denominator the total number of unique relevant documents retrieved in 3 or more different searches on the same topic.[33]

Precision is the proportion of all retrieved documents that are relevant calculated by the number of relevant documents retrieved in the search divided by the number of documents retrieved.

Relevance-Based Measures in Bibliographic Systems.—With the advent of many approaches to accessing MEDLINE and promises about their capability and ease of use, Haynes et al[17] undertook a study comparing the performance and time required of 14 different front ends to the MEDLINE database available in 1986 for 6 clinical topics. They used the same query formulation for each system and found that most systems yielded the same quantity of relevant articles, although there were substantial differences in cost, online time required, and ease of use. This study was repeated with 27 MEDLINE products available in 1994, including online and CD-ROM systems.[18] The repeat study showed more substantial variation than the original study in the numbers of relevant (between 0.8-6.4 per search) and nonrelevant (between 1.1-9.0 per search) articles retrieved. As the queries entered into each system and the underlying MEDLINE database were identical, the marked differences appeared to be due to the divergent features of each system. For example, some systems search against the entire MEDLINE record with words from MeSH terms in the user's query,

Table 2.—Comparison of Novice and Expert Searchers*

Group	Mean No. of Articles Retrieved	Definitely Relevant Only, %		Definitely/Possibly Relevant, %	
		Recall	Precision	Recall	Precision
Clinic physicians using KF	88.8	68.2	14.7	72.5	30.8
Clinic physicians, KF top 15	14.6	31.2	24.8	25.5	43.8
Librarians, full MEDLINE	18.0	37.1	36.1	30.8	59.4
Librarians, text words only	17.0	31.5	31.9	27.0	50.3
Physicians, full MEDLINE	10.9	26.6	34.9	19.8	55.2
Physicians, text words only	14.8	30.6	31.4	24.1	48.4

*Data from Hersh and Hickam.[12] KF indicates Knowledge Finder.

while others limit themselves to the MeSH headings themselves.

Other evaluation studies have focused on comparing the searching performance of 2 user groups: inexperienced end users who are knowledgeable about the subject domain vs medical librarians or experienced clinician users who are skilled in the advanced features of IR systems. The first comparison of these groups was performed by Haynes et al.[10] In this study, 78 searches were randomly chosen for replication by both a clinician experienced in searching and a medical librarian. During this study, each original ("novice") user had been required to enter a brief statement of information need before entering the search program. This statement was given to the experienced clinician and librarian for searching on MEDLINE. All of the retrievals for each search were given to a subject domain expert who was blinded as to which searcher retrieved which reference. Recall and precision were calculated for each query and averaged. The results showed that experienced clinicians and librarians achieved comparable recall (48%-49%), although the librarians had significantly better precision (58% vs 49%). The novice clinician searchers had lower recall (27%) and precision (38%) than either of the other groups. This study also found, however, that the novice searchers were satisfied with their search outcomes. A follow-up study showed that with minimal training, physicians could improve their performance to the level of experienced searchers by their fourth online search.[19] Providing them with a specially trained clinical preceptor did not improve their searching; physicians without such training improved equally as well after the fourth search.

Hersh and Hickam[12] also carried out a comparison of recall and precision in clinicians and librarians. Novice searchers were also provided with access to MEDLINE via KF, which does not use Boolean (AND and OR) operators, opting instead for natural language searches. The latter allow the user to enter free text and retrieve articles based on the overlap of words in the search statement and article. Hersh and Hickam[12] also compared the performance of the experienced searchers using the full MEDLINE feature set vs just using text words from the title, abstract, and MeSH heading fields. As with Haynes et al,[10] statements of information need were collected online and given to experienced searchers for replication. Likewise, relevance was assessed by clinicians blinded to the searcher.

One problem for this study was that conventional Boolean systems tend to retrieve smaller document sets, but natural language systems such as KF retrieve a large set and rank the output by "relevance" (as measured by frequency of query words in the citation). This makes direct comparison of recall and precision between Boolean and natural language systems difficult. As seen in Table 2, the novice clinicians were able to achieve much higher recall than any of the expert searchers, although they paid a price in precision (and most likely were unwilling to look at all 100 references on the retrieval list anyway). To compare the novice searchers with retrieval of comparable numbers of references retrieved by the experienced searchers, a second set of recall and precision values were calculated with KF's default retrieval lowered to 15, the average size of Boolean retrieval sets. At this level of output, recall and precision were comparable to all groups of expert searchers, with no statistically significant differences. This study, like the one of Haynes et al[10] before it, attempted to assess whether articles rated as relevant were actually able to answer the clinical question that prompted the search, but poor user response precluded a reliable analysis of the data.

Hersh and Hickam[12] also found that there was no benefit of advanced MEDLINE searching features (eg, explosions, subheadings) for experienced clinician or librarian searchers, as each achieved comparable recall and precision using individual word searching with Boolean operators. This latter searching method included searching on the words in the title, abstract, and MeSH terms, indicating that MeSH terms themselves are beneficial but features using them, such as explosions and subheadings, are not universally so.

Relevance-Based Measures in Full-Text Systems.—Full-text searching systems are less well studied. McKinin et al[20] compared searching in 2 collections of full-text journals (MEDIS and the Comprehensive Core Medical Library [CCML], the latter from BRS) and MEDLINE. They took 89 search requests from a university medical library and performed them on each of the 3 systems. Only documents present in all 3 databases were used for recall and precision calculations. Their results showed that full-text searching by word-based Boolean methods led to higher recall (76%-78% vs 41%-42%) at the expense of lower precision (37% vs 55%-62%) when compared with abstract (ie, MEDLINE) searching.

Hersh and Hickam[21] assessed full-text searching of an online medical textbook, comparing the recall and precision of medical students searching *Scientific American Medicine* (Scientific American Inc, New York, NY) with 2 different user interfaces. One interface featured Boolean searching modeled on the NLM's Grateful Med system, while the other used natural language searching similar to what is available in KF. The recall and precision results were virtually identical between the 2 systems.

As noted above, measuring the number of relevant documents yields insight into how successfully systems are used, but does not provide a complete picture. This approach does not measure whether the relevant articles retrieved actually satisfied the user's information need. Clinicians typically do not need anywhere near 100% recall but really only need enough information to answer the question that motivated use of the system. In fact, it has not been determined what difference in recall or precision could be construed as "clinically significant." It has also been found that there is considerable disagreement in blinded relevance judgments.[22]

Search Failure

Another line of research has been to determine why searches do not work well. Four studies met the inclusion criteria for this factor.[20,23-25] Kirby and Miller[23] assessed end-user searching on BRS Colleague at the Medical College of Pennsylvania in Philadelphia. Library users who performed their own searches were offered a free search on the same topic by an intermediary. Users deemed the searches "successful" (39%) or "incomplete" (61%). There were no differences between the 2 categories of searches in terms of time spent or system features used. The successful searches tended to have a simple search statement of 2 to 3

concepts. Incomplete searches were mostly due to problems of "search strategy," such as failure to use MeSH terms (eg, using the text word *hypertension* and retrieving articles that used it in a different context, such as *portal hypertension*) or to specify alternative approaches to formulating the question (eg, not trying new strategies when an initial one was not successful).

Many studies have focused on NLM's Grateful Med, which is designed for end users. A large study of Grateful Med users at the NLM focused on searches retrieving no articles ("no postings").[24] This was found to occur with 27% to 37% of Grateful Med searches. Fifty-one percent of such searches used excessive ANDs, in that no documents contained all of the search terms selected by the searcher. Other reasons for empty sets included inappropriate entering of author names (15%), term misspellings (13%), punctuation or truncation errors (11%), and failed title searches (6%). The investigators did not assess how many no postings occurred due to nothing on the topic being present. Other errors made included the following:

1. Inappropriate use of specialty headings (eg, using the term *pediatrics* to search for children's diseases when it is intended to represent the medical specialty).
2. Incorrect use of subheadings (eg, use of *management*, which refers more to the business sense of the word, instead of *therapy*).
3. Not using related terms, either in the form of text words (eg, adding a term like *cerebr:* or *encephal:* to the MeSH heading *brain*) or MeSH terms (eg, adding terms like *bites and stings* or *dust* to *allergens*).

Walker et al[25] evaluated 172 "unproductive" Grateful Med searches at McMaster University in 1987 and 1988, dividing problems into the categories of search formulation (48%), the Grateful Med software itself (41%), and search failure (11%). While half of search formulation problems were due to no material occurring on the topic, the most common errors were found to be use of low-frequency terms, using general terms instead of subheadings, and excessive use of AND. Problems specific to Grateful Med included inappropriate use of the title line (eg, unwittingly typing a term on the title line, thus limiting retrieval to all articles with that term in the title) and the software's automatic combination of words on the subject line(s) with OR, so that the phrase *inflammatory bowel disease* was searched as *inflammatory OR bowel OR disease.*

Not all failure analyses have looked at bibliographic databases. In their study of full-text retrieval performance described above, McKinin et al[20] also assessed the reasons for full-text retrieval failures. About two thirds of the problems were due to search strategy, in that the concepts from the search were not explicitly present in the document or an excessively restrictive search operator was used. The remaining third were due to natural language problems, such as word variants, more general terms, synonyms, or acronyms used in the documents.

These analyses of failure show that users make frequent errors while searching these systems. Among the approaches to help avoid them include the improved MeSH look-up capability now present in most systems as well as more focused efforts like the COACH system built into the NLM's Internet Grateful Med, which detects the kinds of errors discovered by the NLM analysis described above.[24]

Outcomes

The ultimate measure of an IR system's success, like the success of any other medical intervention, should be its impact on care, such as improved patient outcome or reduced cost, usually via a randomized controlled trial (RCT). Since IR systems are used on an infrequent basis and for heterogeneous purposes, conducting an RCT would be quite difficult. There have been RCTs of other computer systems, namely decision support (or expert) systems, which tend to provide information and/or advice on specific events, such as the diagnosis of abdominal pain, reminders to order specific preventive interventions, or the proper prescription of antibiotics.[34] It is much more difficult, on the other hand, to devise patient or cost outcome measures for an IR system that might be used for all of the above situations and more. If RCTs are to be done with IR systems, they will need to be performed with focused databases and user questions.

Three studies have assessed whether use of libraries or IR systems led to changes in patient care decisions.[1,10,26] Veenstra[26] found that a medical librarian added to the teaching services staff at Hartford Hospital in Connecticut was able to find information that affected patient care 40% to 59% of the time. In their study of Grateful Med introduced in clinical settings, Haynes et al[10] found that clinicians reported use changed the course of patient care 47% of the time. One of the most comprehensive assessments of the impact of MEDLINE used the "critical incident technique," in which users were prompted to recall a recent search that was effective or not.[1] The analysis of this survey focused on the 86% of searches that were deemed effective by a sample of 552 end-user physicians, scientists, and others. The most common impact of the information obtained was to develop an appropriate treatment plan (45%), followed by recognizing or diagnosing a medical problem or condition (22%), implementing a treatment plan (14%), and maintaining an effective patient-physician relationship (10%). These studies imply that IR systems are useful but that the magnitude of their benefit is unknown.

COMMENT

This review has developed a framework and described studies that have assessed the performance of clinical IR systems. Even with the limitations of the studies, it can be seen that IR systems have had a modest but important impact in the health care domain, but that there are many unanswered questions about how well they are used. Clearly some generalizations can be made. For example, current IR systems have had limited use in direct patient care settings, and they are used to meet only a tiny fraction of clinicians' information needs.[30,31,35] This does not mean the systems are not valuable when they are used, but it does challenge developers to implement systems that have more clinically pertinent information and are easier to use. It is certainly possible that changing technologies may increase the use of IR systems, but this has not yet been documented.

While clinicians tend to use nonjournal literature sources (such as textbooks) to answer the majority of their clinical questions,[31,35] most IR system use still occurs with bibliographic databases aiming to identify articles in journals. Whether computer-based information systems are more amenable to bibliographic information or, rather, that adequate nonjournal sources have yet to be developed is not clear. One problem with current approaches to bibliographic searching is the time required to search for articles, find them, and appraise their content. At an average of 30 minutes per question, using the journal literature on a routine basis is impractical for most clinical questions, especially at the point of care.

Most but not all studies have shown that various system factors, such as type of indexing or user interface, do not make much difference. While many systems and their features have their advocates, most studies show that physicians can find information to meet their needs with basic indexing and user interfaces. It is also clear that no matter how skilled or with what type of database, searchers are unlikely to come close to retrieving all of the potentially relevant material on a given topic.

Finally, while health care IR systems are widely distributed and commercially successful, the magnitude of their impact

on health care professionals and patient care has not been well quantified. While it is not necessary to prove benefit to advocate their use (no one argues that textbooks should not be used because their benefit is unproven), there is a need for further research examining the content and delivery methods of IR systems. Despite the limitations in existing medical IR systems, their use is likely to continue to grow, and the technology will continue to evolve. The evaluation framework and results presented in this article show, however, that continued assessment of use and performance are warranted.

References

1. Lindberg D, Siegel E, Rapp B, Wallingford K, Wilson S. Use of MEDLINE by physicians for clinical problem solving. *JAMA.* 1993;269:3124-3129.
2. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA.* 1992;268:2420-2425.
3. Sackett D, Richardson W, Rosenberg W, Haynes R. *Evidence-Based Medicine: How to Practice and Teach EBM.* New York, NY: Churchill Livingstone; 1997.
4. Fletcher R, Fletcher S. What is the future of internal medicine? *Ann Intern Med.* 1993;119:1144-1145.
5. Lancaster F, Warner A. *Information Retrieval Today.* Arlington, Va: Information Resources Press; 1993.
6. Fidel R, Soergel D. Factors affecting online bibliographic retrieval: a conceptual framework for research. *J Am Soc Inf Sci.* 1983;34:163-180.
7. Horowitz GL, Jackson JD, Bleich HL. PaperChase: self-service bibliographic retrieval. *JAMA.* 1983;250:2494-2499.
8. Collen M, Flagle C. Full-text medical literature retrieval by computer: a pilot test. *JAMA.* 1985;254:2768-2774.
9. Markert R, Parisi A, Barnes H, et al. Medical student, resident, and faculty use of a computerized literature searching system. *Bull Med Libr Assoc.* 1989;77:133-137.
10. Haynes R, McKibbon K, Walker C, Ryan N, Fitzgerald D, Ramsden M. Online access to MEDLINE in clinical settings. *Ann Intern Med.* 1990;112:78-84.
11. Abate M, Shumway J, Jacknowitz A. Use of two online services as drug information sources for health professionals. *Methods Inf Med.* 1992;31:153-158.
12. Hersh W, Hickam D. The use of a multi-application computer workstation in a clinical setting. *Bull Med Libr Assoc.* 1994;82:382-389.
13. Marshall J. The continuation of end-user on line searching by health professionals: preliminary survey results. In: Proceedings of the Medical Library Association Annual Meeting; May 18-24, 1990; Detroit, Mich.
14. Haynes R, Ramsden M, McKibbon K, Walker C. Online access to MEDLINE in clinical settings: impact of user fees. *Bull Med Libr Assoc.* 1991;79:377-381.
15. Wildemuth B, deBliek R, Friedman C, File D. Medical students' personal knowledge, searching proficiency, and database use in problem solving. *J Am Soc Inf Sci.* 1995;46:590-607.
16. Hersh W, Pentecost J, Hickam D. A task-oriented approach to information retrieval evaluation. *J Am Soc Inf Sci.* 1996;47:50-56.
17. Haynes R, McKibbon K, Walker C, et al. Computer searching of the medical literature: an evaluation of MEDLINE searching systems. *Ann Intern Med.* 1985;103:812-816.
18. Haynes R, Walker C, McKibbon K, Johnston M, William A. Performance of 27 MEDLINE systems tested by searches with clinical questions. *J Am Med Inform Assoc.* 1994;1:285-295.
19. Haynes RB, Johnston ME, McKibbon KA, Walker CJ, William AR. A program to enhance clinical use of MEDLINE: a randomized controlled trial. *Online J Curr Clin Trials* [serial online]. May 11, 1993;doc 56.
20. McKinin E, Sievert M, Johnson E, Mitchell J. The MEDLINE full-text research project. *J Am Soc Inf Sci.* 1991;42:297-307.
21. Hersh W, Hickam D. An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *J Am Soc Inf Sci.* 1995;46:478-489.
22. Hersh W. Relevance and retrieval evaluation: perspectives from medicine. *J Am Soc Inf Sci.* 1994;45:201-206.
23. Kirby M, Miller N. MEDLINE searching on Colleague: reasons for failure or success of untrained users. *Med Ref Serv Q.* 1986;5:17-34.
24. Kingsland L, Harbourt A, Syed E, Schuyler P. COACH: applying UMLS knowledge sources in an expert searcher environment. *Bull Med Libr Assoc.* 1993;81:178-183.
25. Walker C, McKibbon K, Haynes R, Ramsden M. Problems encountered by clinical end users of MEDLINE and Grateful Med. *Bull Med Libr Assoc.* 1991;79:67-69.
26. Veenstra R. Clinical medical librarian impact on patient care: a one-year analysis. *Bull Med Libr Assoc.* 1992;80:19-22.
27. Poisson E. End-user searching in medicine. *Bull Med Libr Assoc.* 1986;74:293-299.
28. Ludwig L, Mixter J, Emanuelle M. User attitudes toward end-user literature searching. *Bull Med Libr Assoc.* 1988;76:7-13.
29. Wallingford K, Humphreys B, Selinger N, Siegel E. Bibliographic retrieval: a survey of individual users of MEDLINE. *MD Comput.* 1990;7:166-171.
30. Covell D, Uman G, Manning P. Information needs in office practice: are they being met? *Ann Intern Med.* 1985;103:596-599.
31. Gorman P, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. *Med Decis Making.* 1995;15:113-119.
32. Gorman P, Ash J, Wykoff L. Can primary care physicians' questions be answered using the medical literature? *Bull Med Libr Assoc.* 1994;82:140-146.
33. Hersh W. *Information Retrieval: A Health Care Perspective.* New York, NY: Springer-Verlag; 1996.
34. Johnson M, Langton K, Haynes R, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. *Ann Intern Med.* 1994;120:135-142.
35. Curley S, Connelly D, Rich E. Physicians' use of medical knowledge resources: preliminary theoretical framework and findings. *Med Decis Making.* 1990;10:231-241.