A Comparison of Two Methods for Indexing and Retrieval from a Full-text Medical Database

WILLIAM R. HERSH, MD, DAVID H. HICKAM, MD, MPH

The objective of this study was to compare how well medical professionals are able to retrieve relevant literature references using two computerized literature searching systems that provide automated (non-human) indexing of content. The first program was SAPHIRE, which features concept-based indexing, free-text input of queries, and ranking of retrieved references for relevance. The second program was SWORD, which provides single-word searching using Boolean operators (AND, OR). Sixteen fourth-year medical students participated in the study. The database for searching was six volumes from the 1989 *Yearbook* series. The queries were ten questions generated on teaching rounds. All subjects searched half the queries with each program. After the searching, each subject was given a questionnaire about prior experience and preferences about the two programs. Recall (proportion of relevant articles retrieved from the database) and precision (proportion of relevant articles in the retrieved set) were measured for each search done by each participant. Mean recall was 57.6% with SAPHIRE; it was 58.6% with SWORD. Precision was 48.1% with SAPHIRE vs 57.6% with SWORD. Each program was rated easier to use than the other by half of the searchers, and preferences were associated with better searching performance for that program. Both systems achieved recall and precision comparable to existing systems and may represent effective alternatives to MEDLINE and other retrieval systems based on human indexing for searching medical literature. *Key words*: information retrieval systems; automated indexing; medical literature; MEDLINE. **(Med Decis Making 1993;13:220–226)**

The practice of medicine requires gathering and interpreting information. Knowledge of the medical literature facilitates this process. However, physicians tend to underestimate their information needs[1] and are often not up to date on clinically relevant advances in their specialties.[2] Computer-based systems that provide access to medical knowledge have been proposed as one solution to the problem,[3] yet they have not been used routinely. Some programs are difficult for infrequent users to master,[4] while others do not provide the types of capability and content useful to busy clinicians.[5]

The most successful computer applications to date have been literature-searching programs. Most of these systems use the MEDLINE database, which is maintained by the National Library of Medicine (NLM) and contains seven million bibliographic references to the peer-reviewed medical literature. MEDLINE does not contain the actual text of articles but does contain abstracts for about 60% of the references. More recently, other literature databases have become available, containing the full text of journals, textbooks, and monographs.

Despite the enthusiasm for information-retrieval systems, most users tend to overestimate the quality of searching performance. A study of MEDLINE showed that medical librarians and experienced clinician users retrieve about half of all relevant references in a given search, while casual clinician users retrieve only a fourth.[6] Strategies to improve the yield of literature searching depend upon understanding the interface between the clinician and computer system. However, this interface has received little study. Most existing search systems require the user to employ Boolean operators (e.g., AND, OR) to link together special indexing terms or specific words that may appear in the title or abstract of a citation. Many clinicians find this approach difficult to use effectively.[7] For example, designating the search terms too loosely can result in retrieval of an overwhelmingly large number of citations. Furthermore, using Boolean operators inappropriately can lead to either very few or excessive citations retrieved.[8]

The goal of this study was to compare two different computer systems for searching a full-text database. Full-text systems are becoming increasingly prevalent, offering the advantage over systems such as MEDLINE that they provide more information than is available

Received July 7, 1992, from the Biomedical Information Communication Center (WRH) and the Department of Medicine (WRH, DHH), Oregon Health Sciences University, and the Department of Veterans Affairs Medical Center (DHH), Portland Oregon. Revision accepted for publication October 16, 1992. Supported in part by Grant 9040 from the Medical Research Foundation of Oregon and Grant LM05307 from the National Library of Medicine.

Address correspondence and reprint requests to Dr. Hersh: BICC, Oregon Health Services University, 3181 SW Sam Jackson Park Road, Portland, OR 97201.

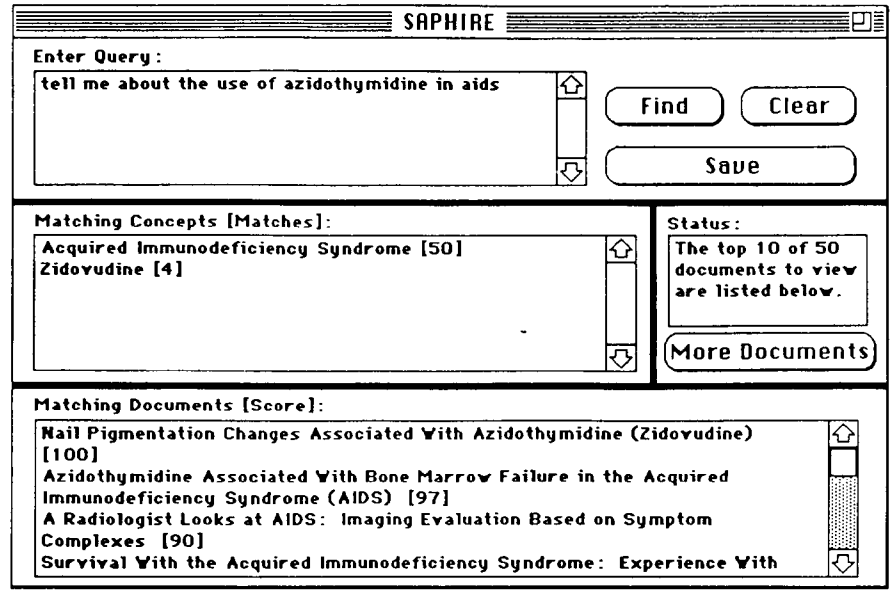

FIGURE 1. SAPHIRE user interface.

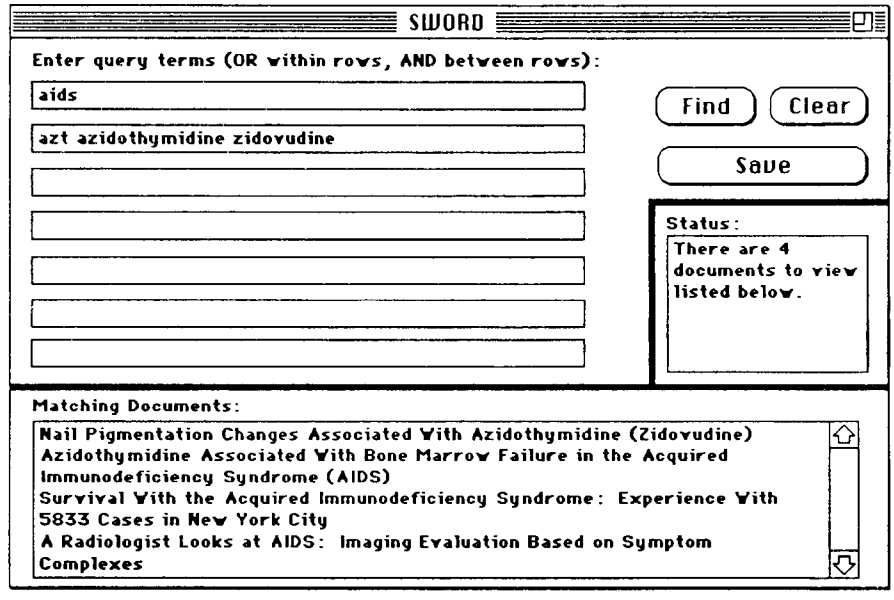

FIGURE 2. SWORD user interface.

in an abstract alone, potentially obviating the need to go to the library to obtain printed articles. A problem with full-text retrieval systems, however, is that they are usually indexed solely by the individual words that occur in each article. In MEDLINE, human indexers assign concepts from the Medical Subject Headings (MeSH) vocabulary to each citation. Although MEDLINE indexers have problems with consistency,[9] searching performance in full-text databases can be poor because words can have different meanings (e.g., *lead* can be a noun or verb), and different words can have the same meaning (e.g., *high, elevated*).

Information-retrieval systems consist of two components: an indexing system, and a method for using the index to search the database for citations that appear to match the subject of the query. For example, with MEDLINE the indexing system is the MeSH terms assigned to each citation, along with words in the title and abstract (when the abstract is included in the citation). The search method is a Boolean-based program for designating search terms. SAPHIRE is an experimental information-retrieval system that embodies a new approach to indexing and searching.[10] It allows input of questions in free text, along with ranking of articles for relevance. Indexing is done by full medical concepts as opposed to individual words, but it is done by computer, avoiding the inconsistency of human indexers.[9] The indexing process is provided by a *concept-matching algorithm*, whereby different synonyms of a concept can be recognized and are mapped to a common canonical form. For example, *hypertension, high blood pressure*, and *elevated blood*

Table 1 • The Ten Queries for User Searching

1.	How is the diagnosis of pulmonary hypertension made?
2.	What are the complications of steroid therapy?
3.	What are some iatrogenic complications that can occur to liver failure patients when they are hospitalized?
4.	What are the risk factors for cerebrovascular disease?
5.	What is the optimal chemotherapy for non–small-cell carcinoma of the lung and what is its efficacy?
6.	What are the clinical uses for interferon?
7.	What is the treatment for acute myelogenous leukemia?
8.	What are the standard treatment regimens for tuberculosis?
9.	What is the pharmacology of primidone?
10.	Does treatment of hyercholesterolemia lower the risk of myocardial infarction?

pressure are all recognized as the same concept. Concept matching is made possible by the use of the Metathesaurus, which is part of the NLM's Unified Medical Language System (UMLS) Project.[11] The Metathesaurus includes terms from the MeSH vocabulary, which is used by the NLM to index MEDLINE. It also includes synonyms from other vocabularies such as DSM-III, SNOMED, ICD-9, and clinical terms from the COSTAR medical record system.

SAPHIRE's searching interface is shown in figure 1. Users enter a natural-language query and click the "Find" button, from which concepts are extracted by the concept-matching algorithm. Additional concepts can be added to the list by typing more text into the query box, while undesired concepts can be deleted. With the list of concepts, a search of the database is performed. Each matching reference is given a score, with the top score 100. The list is sorted, and the top ten ranking references are displayed. The user can display additional references by clicking the "More Documents" button. Users click the mouse on the title of the reference they wish to see, which is then displayed on the screen.

The second computer program used in this study was SWORD, which provides the type of word-based, Boolean searching found in commercial full-text retrieval systems. In SWORD, the indexing terms for each article are simply a list of all the words that occur in the article. Its searching interface is modeled after Grateful Med,[12] where each individual word on a given line is combined by OR and each line is combined by AND. Like most existing full-text searching programs, relevance ranking is not provided. SWORD (as opposed to other full-text searching programs) was used for this study because its commands and full-screen interface were similar to those of SAPHIRE, thereby minimizing operation of the program as a confounding variable in our results. The searching interface for SWORD is shown in figure 2.

Previous studies of SAPHIRE have suggested improved searching performance over conventional methods for novice searchers,[13,14] although these studies were limited by lack of interactive searching. In this study, we set out to examine SAPHIRE in interactive, real-time use. This not only would provide more realistic comparisons of performance, but also would allow feedback by users.

Methods

A test collection of ten queries and 1,992 documents, with identification of documents relevant to each query, was created. The queries were obtained from a study on clinical questions arising on teaching rounds at the University of Pittsburgh[15] and are shown in table 1. The documents were collected from six volumes of the 1989 *Yearbook* series.[16–21] This series of books is designed to summarize recent medical literature. One volume is published annually for each medical specialty (e.g., internal medicine, family practice, cardiology), containing summaries of all the articles deemed important to the field by the editors for that year. Each summary contains the title, source, and original abstract, augmented by "expert" commentary. A collection of the full text of the *Yearbook* series for three years is available as a CD-ROM product, and the publisher allowed use of some of the text for this study. Each of these summaries was considered to be a document, yielding 1,992 documents from the six volumes used.

A searching population that had clinical medical knowledge and some previous searching experience was recruited by sending all 94 students in the fourth-year class of Oregon Health Sciences University School of Medicine a questionnaire asking about their MEDLINE searching experience and offering remuneration for participation in the study. Of the 55 students who responded, 21 indicated prior experience searching and were invited to participate in the study; of these 21, 16 accepted.

All participants searched half of the queries with each program. The participants were randomized as to which of the two sets of five queries was searched first and also as to which program was used on each query set. This led to four different search groups: 1) queries 1–5 with SAPHIRE, followed by queries 6–10 with SWORD; 2) queries 1–5 with SWORD, followed by queries 6–10 with SAPHIRE; 3) queries 6–10 with SWORD, followed by queries 1–5 with SAPHIRE; 4) queries 6–10 with SAPHIRE, followed by queries 1–5 with SWORD.

Before searching, the participants were given a 20–30 minute orientation on using the two systems. They were instructed to search until they had retrieved several relevant references per search or until they had tried four or five times and had not yet retrieved anything relevant. Searching the ten queries took an average of about two hours. After searching, each participant filled out a questionnaire asking: frequency of personal computer use, comfort with computer use,

number of previous MEDLINE searches, number of other databases searched besides MEDLINE, prior knowledge of a Boolean operator, and prior knowledge of MeSH terms. The questionnaire also used five-point Likert scales to assess for each program: overall ease of use, ease in designating search terms, and perception of retrieval success. The searchers were also asked to rate how easy it was to assign Boolean operators. Finally, the searchers were asked to provide comments about what aspects they liked or did not like about the two searching systems.

Relevance judgments for query–document pairs were made for all documents retrieved by any of the 16 searchers. All documents that were about an entirely different topic than the query were eliminated by an initial single judge. Each potentially relevant document was then submitted to one of two other physicians, who made a final decision about its relevance. To assess reliability of these rankings, 10% of the documents were reviewed by both raters.

Searching performance for each program was evaluated using parameters standard to the information science field, recall and precision. Recall is the proportion of relevant documents in a database retrieved by a query:

$$\text{Recall} = \frac{\text{number of documents retrieved and relevant}}{\text{number of relevant documents in database}} \quad (1)$$

If the retrieval of a relevant document is viewed as analogous to a positive test result, then recall is equivalent to sensitivity. However, in a large database, identification of all relevant documents for a given query is difficult. In such cases, relative recall is used. For relative recall, the total number of relevant documents in the database retrieved by three or more searches on the same topic by different searchers is used to approximate the total number of relevant documents. Precision is the proportion of relevant articles in the retrieved set:

$$\text{Precision} = \frac{\text{number of documents retrieved and relevant}}{\text{total number of documents in retrieval set}} \quad (2)$$

Using the test result analogy from above, precision is equivalent to positive predictive value.

Search results in this experiment were evaluated by calculating relative recall and precision for each searcher across all queries used with each program. Analysis of variance was used to measure differences in search results between the two programs. The results for all searches for each program were treated as repeated measures. Order of use of the programs and identity of the searching program were included as between-subjects factors. For both relevant and nonrelevant documents, the numbers of documents retrieved by no searchers, half or less of the searchers, more than half of the searchers, and all searchers were calculated for each program.

Questionnaire results were tabulated and analyzed for associations between user experiences and preferences with the chi-square test for distributions in contingency tables. A sign test was used to compare the two programs for subjective impressions in overall ease of use, ease in designating search terms, and perception of success in retrieving relevant references.

Search results were compared between groups of subjects defined by user experience (with both programs combined) and user preference (for each program). For the former, recall and precision were calculated for groupings generated by questionnaire responses (i.e., those using a personal computer weekly or those having performed more than 50 previous MEDLINE searches.). For the latter, simple recall and precision values could not be used, due to the crossover design of the study, in which half of the searchers used one program with half of the queries and the other program with the other half of the queries. Since the recall and precision values with both programs were markedly lower for the first five queries than the second five queries, we could not directly compare recall and precision values grouped by questionnaire responses. As a result, the raw recall and precision values were converted to rank scores, with 1 being the highest recall (or precision) for a given query and 16 being the lowest. This type of ranking allowed direct comparisons between search results and user preferences.

Results

The 160 queries by the 16 searchers led to retrieval of 2,222 documents, of which 774 were unique. Of these, 132 were relevant and 642 were nonrelevant. Interobserver reliability of the relevance judgments, estimated by the kappa statistic, was 0.51.

SEARCH RESULTS

The two searching programs provided similar search results (table 2). The only statistically significant ($p < 0.05$) difference was a better precision for SWORD on queries 6–10. Table 3 shows the numbers of relevant and nonrelevant documents retrieved by zero, half or less, more than half, and all of the searchers for each program. (Although SWORD had a better average precision, it had a larger absolute number of nonrelevant documents retrieved due to one search where an inappropriate OR led to retrieval of several hundred documents.)

Table 2 • Overall Search Results

Search Method	Mean Recall	Mean Precision
SAPHIRE	57.6	48.1
Queries 1–5	45.5	43.6
Queries 6–10	69.7	52.6*
SWORD	56.6	57.5
Queries 1–5	44.3	43.6
Queries 6–10	68.9	71.1

*Significant difference, $p < 0.05$.

QUESTIONNAIRE RESULTS

The searcher preferences (table 4) revealed that the users were almost evenly divided between designating SWORD and designating SAPHIRE overall easier to use and easier for designating search terms. Half felt SWORD retrieved more relevant references, while most of the remainder felt both were equally good in finding them. All users felt Boolean operators were easy or very easy to use. The searchers' previous experiences are summarized in table 5. The majority of searchers were not personal computer owners and used computers monthly or less. Most felt comfortable or neutral with computers. All had some MEDLINE searching experience, and the sample could be clustered into three groups based on the previous number of MEDLINE searches. The overwhelming majority of searching experience was with OHSU MEDLINE, a free, local MEDLINE subset. (This system offers a simple interface that does not require Boolean operators or MeSH terms, as well as a more advanced interface that uses Booleans and MeSH headings.) Only a minority of searchers had previously heard of Boolean operators or MeSH terms.

There was no significant association between past experience and preference for one program over the other. There was also no association between order of use of the programs and preference for one over the other. The factor most highly associated with preference was which program was used for the group of queries that had better overall recall and precision (queries 6–10). Those who searched these queries with SWORD felt that it was easier to use by a 5–3 margin, whereas those who used SAPHIRE for these queries felt that it was easier to use by a 6–2 margin, a difference that was nearly statistically significant ($p = 0.08$).

A qualitative analysis of the subjects' comments revealed a consistent pattern. Users generally praised SAPHIRE for its ability to be less rigorous in phrasing queries, its capacity to find synonyms, and its relevance ranking. Users did not like, however, the inability to find some terms they thought should be there, the occasional extraneous concepts found, and the tendency of common concepts (e.g., treatment, diagnosis) to lead to retrieval of irrelevant articles. Other aspects commented upon negatively were the default number of articles retrieved (ten; one user thought it should be 20), the inconsistency of relevance ranking, and the lack of Boolean operators. SWORD, on the other hand, was praised for retrieving fewer irrelevant articles, not requiring designation of medical concepts, and seeming to allow greater control over searching through the use of AND and OR. Users did not like its inability to specify synonyms, its lack of relevance ranking, and its retrieval of references where words were used in a different context.

ASSOCIATIONS BETWEEN SEARCH AND QUESTIONNAIRE RESULTS

With the exception of personal computer ownership, which conferred a 9% detriment in recall ($p = 0.03$) and a 9% benefit in precision ($p = 0.13$), there was no experience characteristic that was associated with a significant difference in searching performance (table 5). Table 6 shows rank scores by program preferences. Users who felt SWORD was overall easier to use had better rank scores with SWORD for recall and precision, while those who felt SAPHIRE was overall easier to use had better rank scores with SAPHIRE for recall but with SWORD for precision. The same pattern held for perception of ease in designating search terms. Those who felt SWORD retrieved more relevant references actually had better rank scores for recall with SAPHIRE, although they had better rank scores for precision with SWORD. These results indicate that, in general, users who preferred aspects of a given program obtained better searching results with it (when adjusted for the difference in magnitude of recall and precision for the two query sets).

Table 3 • Relevant and Nonrelevant Documents Retrieved by SWORD and SAPHIRE

	Relevant Documents Retrieved		Nonrelevant Documents Retrieved	
	SWORD	SAPHIRE	SWORD	SAPHIRE
Retrieved by no searcher	13	43	187	386
Retrieved by more than zero but less than half of searchers	78	42	433	214
Retrieved by more than half but not all searchers	28	24	25	26
Retrieved by all searchers	13	23	6	16

Table 4 • Program Preference Questionnaire Results

Overall ease of use	
SWORD much easier	2
SWORD somewhat easier	5
Both equally easy	0
SAPHIRE somewhat easier	8
SAPHIRE much easier	1
Ease of designating search terms	
SWORD much easier	4
SWORD somewhat easier	4
Both equally easy	0
SAPHIRE somewhat easier	6
SAPHIRE much easier	2
Retrieved more relevant references	
SWORD definitely retrieved more	3
SWORD probably retrieved more	5
Neither	6
SAPHIRE probably retrieved more	2
SAPHIRE definitely retrieved more	0

Discussion

This study showed essentially equivalent performances for a word-based, Boolean system and a concept-based, free-text query/relevance ranking system, with a trend towards better precision for the former. Compared with recent studies of MEDLINE searching showing 50% recall for expert searchers and 27% for novices,[6] the results for this relatively novice group of searchers were actually quite good. There are three possible explanations for these findings: 1) this database was particularly well suited for searching, 2) these queries were more likely to lead to good searching results, or 3) the qualities of the two searching programs led to good performance. There are reasons why any of the above might be true. Since the *Yearbook* series is a "value-added" information resource, with expert commentary following the original abstract for each article, it is possible that this extra text conferred an advantage to the indexing process, leading to better retrieval results. This database might also be amenable to good searching due to its relatively small size. It is also possible that, since these queries arose from medical student teaching rounds, they were more amenable to being answered, and thus there were more relevant references in the medical literature. Finally, both systems were implemented on the Apple Macintosh computer, and their simple, easy-to-use interfaces probably contributed to better performance.

Our results are consistent with the well-known observation that different searchers tend to retrieve different sets of documents.[22] For both methods, only one-third of all relevant articles were retrieved by more than half of the searchers. In addition, the majority of nonrelevant articles retrieved were obtained by less than half of the searchers.

There was no user-experience factor that affected overall results with both methods, other than diminished recall and improved precision for those who owned personal computers. Those who felt SAPHIRE was overall easier to use, designated search terms more easily, and retrieved more relevant references had overall better recall and precision with both methods. The strongest predictor of satisfaction with one program over the other was which program was used for queries 6–10, which had over a 20% better recall and 10–30% better precision with either method. This probably reflects the simpler nature of these queries, especially query 9, for which every searcher had 100%

Table 5 • Overall Search Results by Experience Characteristics of the 16 Subjects

	Mean Recall	Mean Precision
Own a personal computer		
Yes (n = 4)	50.9	**59.7**
No (n = 12)	59.2	**50.5**
Frequency of computer use		
Weekly (n = 3)	57.7	**56.8**
Monthly or less (n = 13)	57.0	**51.9**
Comfort with computers		
Comfortable (n = 9)	57.7	**54.3**
Neutral–uncomfortable (n = 7)	56.3	**50.9**
Previous MEDLINE searches		
<10 (n = 3)	62.6	**54.5**
20–30 (n = 7)	53.1	**56.3**
50–100 (n = 6)	59.0	**47.9**
Heard of "Boolean operator"		
Yes (n = 5)	57.3	**52.5**
No (n = 11)	57.0	**52.9**
Heard of "MeSH term"		
Yes (n = 5)	56.2	**51.8**
No (n = 11)	57.5	53.3

Table 6 • Search Results Ranked by Perceptions of Programs

	Recall		Precision	
	SWORD	SAPHIRE	SWORD	SAPHIRE
Overall easier to use				
SWORD	8.6	9.3	6.1	8.7
SAPHIRE	8.3	7.9	8.2	10.5
Easier to designate search terms				
SWORD	8.8	9.1	6.9	9.4
SAPHIRE	8.1	7.6	8.0	10.1
Retrieved more relevant references				
SWORD	9.6	8.7	6.1	8.8
Neither	8.3	8.5	9.4	11.1
SAPHIRE	6.0	8.3	6.4	9.4

recall. In addition, three of these five queries are about treatment of a disease, which is generally a simple and successful type of search.

This hands-on study of two end-user literature-searching systems showed that they can be viable methods for indexing and retrieval of textual databases in the biomedical domain. Further study with different (and larger) query and document sets will be needed to assess where this approach fits into the overall medical-information–retrieval picture. Ongoing research will include experimentation with more powerful methods of concept recognition,[23] development of larger test collections, and deployment of systems in actual clinical settings.

The authors acknowledge the assistance of Jerome Oscheroff, MD, of the University of Pittsburgh, who provided clinical questions used for searching; CMC Research Inc., of Tigard, Oregon, which provided the text for searching from its *Yearbooks on Disk* CD-ROM product; and Gregg Coodley, MD, and Terry Morrow, MD, of the Division of General Medicine at Oregon Health Sciences University, who judged relevance of retrieved articles.

References

1. Covell D, Uman G, Manning P. Information needs in office practice: are they being met? Ann Intern Med. 1985;103:596–9.
2. Williamson J, German P, Weiss R, Skinner E, Bowes F. Health science information management and continuing education of physicians. Ann Intern Med. 1989;110:151–60.
3. Shortliffe E. Computer programs to support clinical decision-making. JAMA. 1987;258:61–6.
4. Kirby M, Miller N. MEDLINE searching on colleague: reasons for failure or success of untrained users. Medical Reference Services Quaterly. 1986;5:17–34.
5. Haynes R, Ramsden M, McKibbon K, Walker C, Ryan N. A review of medical education and medical informatics. Acad Med. 1989;64:207–12.
6. 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.
7. Slingluff D, Lev Y, Eisan A. An end-user search service in an academic health sciences library. Medical Reference Services Quarterly. 1985;4:11–21.
8. Borgman C. Why are online catalogs hard to use? Lessons learned from information retrieval studies. J Am Soc Information Sci. 1986;37:387–400.
9. Funk M, Reid C. Indexing consistency in MEDLINE. Bull Med Lib Assoc. 1983;71:176–83.
10. Hersh W, Greenes R. SAPHIRE: an information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. Comput Biomed Res. 1990;23:405–20.
11. Hersh W. Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. Med Decis Making. 1991;11(suppl):S120–4.
12. Haynes R, McKibbon K. Grateful Med. MD Computing. 1987;4(5):47–57.
13. Hersh W, Hickam D. A comparative analysis of retrieval effectiveness for three methods of indexing AIDS-related abstracts. Proc 54th Annual Meeting of the American Society for Information Science. 1991;211–25.
14. Hersh W, Hickam D, Haynes R, McKibbon K. Evaluation of SAPHIRE: a concept-based approach to information retrieval. Proc 15th Annual Symposium on Computer Applications in Medical Care, Washington, DC. 1991;804–8.
15. Oscheroff J, Forsythe D, Buchanan B, Bankowitz R, Blumenfeld B, Miller R. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med. 1991;114:576–81.
16. Bragg D, ed. Yearbook of Diagnostic Radiology. Chicago: Yearbook Publishers, 1989.
17. DeJong R, Currier R, ed. Yearbook of Neurology and Neurosurgery. Chicago: Yearbook Publishers, 1989.
18. Hollister L, ed. Yearbook of Drug Therapy, Chicago: Yearbook Publishers, 1989.
19. Rakel R, ed. Yearbook of Family Practice. Chicago: Yearbook Publishers, 1989.
20. Rogers D, ed. Yearbook of Medicine. Chicago: Yearbook Publishers, 1989.
21. Schlant R, ed. Yearbook of Cardiology. Chicago: Yearbook Publishers, 1989.
22. Saracevic T. Individual differences in organizing, searching, and retrieving information. Proc 54th Annual Meeting of the American Society for Information Science. 1991;82–6.
23. Evans D, Hersh W, Monarch I, Lefferts R, Handerson S. Automatic indexing of abstracts via natural language processing using a simple thesaurus. Med Decis Making. 1991;11(suppl):S108–15.