An Evaluation of Interactive Boolean and Natural Language Searching with an Online Medical Textbook

William R. Hersh*
Biomedical Information Communication Center, Oregon Health Sciences University, Portland, OR 97201.
E-mail: hersh@ohsu.edu

David H. Hickam
Health Services Research and Development, Department of Veteran Affairs Medical Center,
Oregon Health Sciences University, Portland, OR 97201

Few studies have compared the interactive use of Boolean and natural language searching systems. We studied the use of three retrieval systems by senior medical students searching on queries generated by actual physicians in a clinical setting. The searchers were randomized to search on two of three different retrieval systems: a Boolean system, a word-based natural language system, and a concept-based natural language system. Our results showed no statistically significant differences in recall or precision among the three systems. Likewise, we found no user preference for any system over the others. In the course of this study we did find, however, a number of problems with traditional measures of retrieval evaluation when applied to the interactive search setting.

Introduction

Despite decades of research in automated text retrieval, it is still unknown how automated retrieval systems, which feature natural language input, automated concept mapping, and/or relevance ranking, fare in real-world searching environments. There is no lack of evaluations of automated systems (Salton & Buckley, 1988, 1990), and some studies have even used user-generated search strategies (Belkin, Cool, Croft, & Callan, 1993; Fuhr & Knorz, 1984; Salton, 1972; Salton, Fox, & Wu, 1983; Turtle & Croft, 1991), but few have involved real-time interaction with a retrieval system. The Blair and Maron (1985) study did feature direct user searching, but assessed only a Boolean system without the features common to automated systems. Three studies have assessed interactive Boolean and natural language searching:

(1) Robertson and Thompson (1990) compared search intermediaries using the CIRT system, showing roughly comparable performance.
(2) Hersh, Buckley, Leone, and Hickam (1994) evaluated a commercial natural language system for accessing MEDLINE and found novice users achieved comparable results with more experienced searchers using traditional MEDLINE searching.
(3) Turtle (1994) compared the two types of searching with expert searchers using the WESTLAW system, finding better results with natural language searching.

In this study, we compared three retrieval systems, one "traditional" (e.g., Boolean searching on text words) and the other two "automated" (e.g., natural language querying with automated term weighting, with one using just words for indexing and the other attempting to map the text into medical concepts). The systems were compared using medical queries of an electronic medical textbook, *Scientific American Medicine* (SAM), which is a multi-authored, three-volume textbook of internal medicine (Rubenstein & Federman, 1990). The motivation for using an online textbook stemmed from our interests in the search needs of professionals, in particular physicians in busy clinical settings. In this environment, bibliographic databases, whether in electronic or print form, are impractical and infrequently used (Curley, Connelly, & Rich, 1990), while textbooks, compendia, and other similar references tend to be used more commonly. Furthermore, the commercial market for electronic textbooks is growing rapidly. Thus, while the results might not generalize to all types of full-text infor-

* To whom all correspondence should be addressed.

Received January 10, 1994; revised April 12, 1994; accepted September 21, 1994.

mation retrieval, assessing electronic versions of commonly used paper references is important.

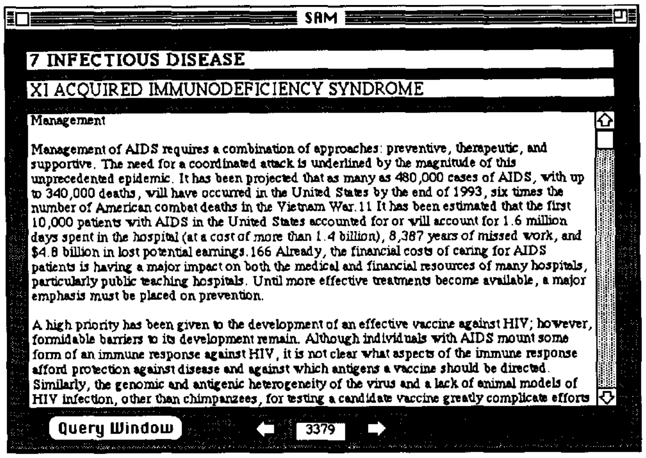
The main objectives of the study were to compare the performance characteristics of and users' preferences for the three different systems. The former was measured by recall and precision, based on topical judgments of relevance. While some question the ability of nonusers to judge the output of queries, i.e., Swanson's "fallacy of delegation" (Swanson, 1977), we have argued that evaluation based on topical relevance is the first of many steps in system evaluation (Hersh, 1994). This type of evaluation is also appropriate in more defined situations, such as the information needs encountered in clinical medical practice. While this is a specific situation whose results may not generalize to all users of information retrieval systems, it is an increasingly important situation in this era of increasing health care costs and explosion of medical knowledge. Users' preferences were assessed using a post-searching questionnaire that had users evaluate each system used as well as compare the different ones used.

There were several secondary objectives for this study. The first was to compare retrieval performance at two levels of relevance. This was done since users often have different needs when using a retrieval system. Whereas some might need access to all information on a given information request (and thus want to look at partially relevant documents), others might need a quick and simple answer to a question (and thus just want to see highly relevant documents). Another secondary objective was to compare how the absolute and relative values of recall and precision differed by changing the type of descriptive statistics used to compare them. A final secondary objective was to assess whether the results of "batch" style retrieval evaluation so prevalent with automated systems correlated with results from real users of the systems.

Three Systems for Information Retrieval

This study used three retrieval systems developed at Oregon Health Sciences University (OHSU) which represent a spectrum of indexing and retrieval paradigms. Each has a relatively common-looking user interface to minimize system look-and-feel as a cause of bias in the results of experiments. For example, each system displays the titles of its retrieved documents in the bottom pane of the search window, allowing any to be viewed by double-clicking on its title. Likewise, the two systems which use relevance ranking of documents initially display the titles of 10 documents. Additional document titles are added in increments of 10 by clicking the **More Documents** button.

All three systems run on the Apple Macintosh computer. Apple's Hypercard is used for the user interface, which communicates with the indexing and retrieval engines that are written in C. Each system also uses Hypercard to display its documents. The interface to the SAM textbook used in this study is shown in Figure 1.



7 INFECTIOUS DISEASE

XI ACQUIRED IMMUNODEFICIENCY SYNDROME

Management

Management of AIDS requires a combination of approaches: preventive, therapeutic, and supportive. The need for a coordinated attack is underlined by the magnitude of this unprecedented epidemic. It has been projected that as many as 480,000 cases of AIDS, with up to 340,000 deaths, will have occurred in the United States by the end of 1993, six times the number of American combat deaths in the Vietnam War.11 It has been estimated that the first 10,000 patients with AIDS in the United States accounted for or will account for 1.6 million days spent in the hospital (at a cost of more than 1.4 billion), 8,387 years of missed work, and $4.8 billion in lost potential earnings.166 Already, the financial costs of caring for AIDS patients is having a major impact on both the medical and financial resources of many hospitals, particularly public teaching hospitals. Until more effective treatments become available, a major emphasis must be placed on prevention.

A high priority has been given to the development of an effective vaccine against HIV; however, formidable barriers to its development remain. Although individuals with AIDS mount some form of an immune response against HIV, it is not clear what aspects of the immune response afford protection against disease and against which antigens a vaccine should be directed. Similarly, the genomic and antigenic heterogeneity of the virus and a lack of animal models of HIV infection, other than chimpanzees, for testing a candidate vaccine greatly complicate efforts

FIG. 1. *Scientific American Medicine* interface.

BOOLEAN

The first of the three systems is BOOLEAN, which provides word-based Boolean searching comparable to many commercial full-text retrieval packages. The interface adapts the Boolean approach of Grateful Med, an end-user-oriented front-end developed at the National Library of Medicine (NLM) for MEDLINE and other databases. As in Grateful Med, all words on the same line are first connected by logical OR, followed by connection of each nonempty line by logical AND. While this approach limits the complexity of Boolean logic that a searcher might employ, it does provide novices a simple introductory approach. BOOLEAN's user interface is shown in Figure 2. The user enters in terms in up to each of seven rows. The matching document titles are presented in a scrolling list in the lower pane of the search window. They are listed in arbitrary order, as occurs in most commercial full-text retrieval systems. To view a document, the user clicks twice on the document title, showing the document as displayed in Figure 1.

The indexing processes of BOOLEAN (and the word-based SWORD system below) start by identifying each word (as defined by any run of alphanumeric characters and apostrophes) in the document. Words not present on a 250-word stop list (vanRijsbergen, 1979) are stemmed to remove plurals, some common suffixes (*-ed*, and *-ing*), and apostrophes, and are then stored in an inverted file.

SWORD

The second system is SWORD, which features word-based automated indexing and natural language retrieval with relevance ranking, much like the SMART system

(Salton, 1991). Indexing is similar to BOOLEAN, but in addition, each word stem in each document is weighted based on the product of the inverse document frequency and intradocument term frequency. Thus, for each term i in each document j, the inverse document frequency for term i (IDF_i) is:

$$IDF_i = \frac{\log(\text{number of documents in database})}{\log(\text{number of documents with term i})} + 1 \quad (1)$$

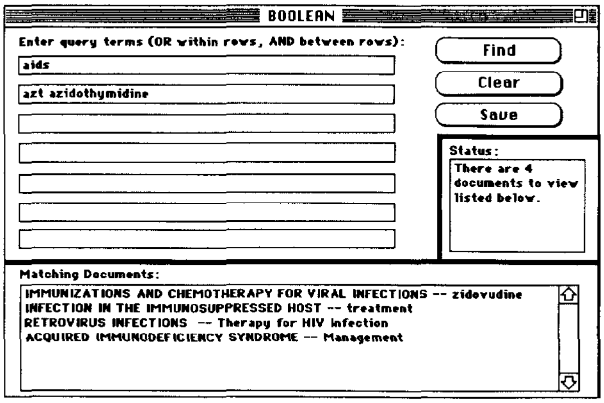
while the intradocument term frequency for term i in document j (TF_{ij}) is:

$$TF_{ij} = \log(\text{frequency of term i in document j}) + 1 \quad (2)$$

giving the weight for term i in document j ($WEIGHT_{ij}$) as:

$$WEIGHT_{ij} = IDF_i * TF_{ij} \quad (3)$$

In SWORD, queries are entered in natural language. Each query word that is not a stop word is stemmed by the rules described above. After the query is processed, the system shows the user which query words were found in the database, not found, or in the stop list. Any document with one or more matching words is given a score based upon the sum of the weights of all words common to the query and document. (In vector space model terminology, this is the inner product of the query and document vectors.) Like SMART, documents are sorted by their score, with the top-ranking document score normalized to 100. Only the top 10 document titles are shown in the scrolling list of the bottom pane initially, with the user able to add more in increments of 10. As with BOOLEAN, the user clicks twice on the document title to view it. SWORD's user interface is shown in Figure 3.



Enter query terms (OR within rows, AND between rows):

aids

azt azidothymidine

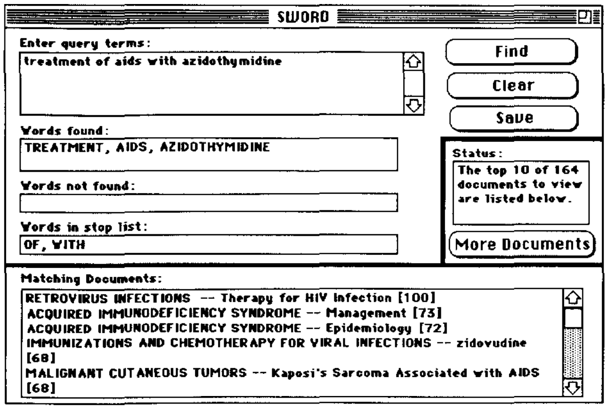
Status:

There are 4 documents to view listed below.

Matching Documents:

IMMUNIZATIONS AND CHEMOTHERAPY FOR VIRAL INFECTIONS -- zidovudine
INFECTION IN THE IMMUNOSUPPRESSED HOST -- treatment
RETROVIRUS INFECTIONS -- Therapy for HIV Infection
ACQUIRED IMMUNODEFICIENCY SYNDROME -- Management

FIG. 2. BOOLEAN interface.



Enter query terms:

treatment of aids with azidothymidine

Words found:

TREATMENT, AIDS, AZIDOTHYMIDINE

Words not found:

Words in stop list:

OF, WITH

Status:

The top 10 of 164 documents to view are listed below.

Matching Documents:

RETROVIRUS INFECTIONS -- Therapy for HIV Infection [100]
ACQUIRED IMMUNODEFICIENCY SYNDROME -- Management [73]
ACQUIRED IMMUNODEFICIENCY SYNDROME -- Epidemiology [72]
IMMUNIZATIONS AND CHEMOTHERAPY FOR VIRAL INFECTIONS -- zidovudine [68]
MALIGNANT CUTANEOUS TUMORS -- Kaposi's Sarcoma Associated with AIDS [68]

FIG. 3. SWORD interface.

SAPHIRE

The third system is SAPHIRE, which features concept-based automated indexing in addition to natural language retrieval and relevance ranking. Both indexing and retrieval are concept based, with concept recognition provided by a concept-matching algorithm (Hersh, 1991). This algorithm takes as its input any string of text, such as a document sentence or a user query, and returns a list of all concepts found, mapped to their canonical or preferred form. This is done by detecting the presence of word-level synonyms between words in concepts (e.g., high and elevated) as well as concept-level synonyms between concepts (e.g., hypertension and high blood pressure).

The concept-matching process is purely semantic, with no syntactic information used. Individual words in the string are stemmed and converted to a canonical representation. Starting with the first word, the algorithm then attempts to find the longest term from the vocabulary that will match all of the successive words. When a term is found, the process repeats from the next word in the string after the term, continuing until the end of the string. The concept-matching algorithm requires a vocabulary of concepts and their synonym forms. The concepts for SAPHIRE's vocabulary originate from a large medical vocabulary created at the NLM called the Metathesaurus, which is a component of the Unified Medical Language System (UMLS) Project (Lindberg, Humphreys, & McCray, 1993). This vocabulary, true to its name, is a thesaurus of terms from various existing medical vocabularies, such as MeSH (used for literature indexing), SNOMED (used for classifying patient findings), and ICD-9 (used for coding diagnoses). Although some reorganization of the vocabulary is required for use by SAPHIRE, no alteration of the content is necessary. This vocabulary allows SAPHIRE to recognize about

130,000 medical concepts and an equal number of synonyms for those concepts.

In SAPHIRE's indexing process, the text to be indexed for each document is passed to the concept-matching algorithm. The indexing terms for each document are the concepts matched, which are weighted with the IDF and TF redefined for concepts. This is analogous to SWORD's indexing, with the difference being that SAPHIRE indexes the concepts instead of words. Thus, for each concept i in each document j, the inverse document frequency for concept i (IDFi) is:

$$\mathrm{IDF}_i = \frac{\log(\text{number of documents in database})}{\log(\text{number of documents with concept i})} + 1 \quad (4)$$

while the intradocument term frequency for concept i in document j (TF_{ij}) is:

$$\mathrm{TF}_{ij} = \log(\text{frequency of concept i in document j}) + 1 \quad (5)$$

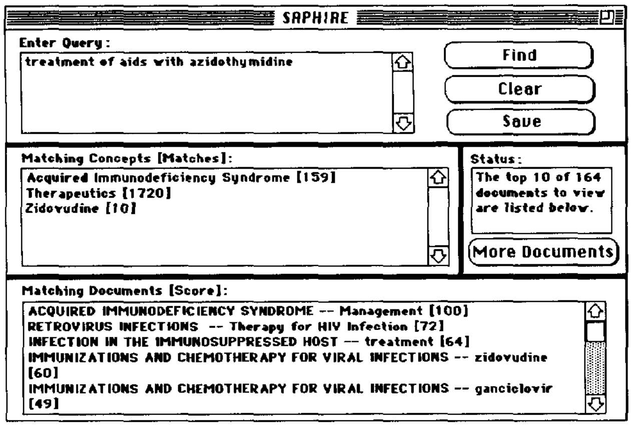
giving the weight for term i in document j (WEIGHT_{ij}) as:

$$\mathrm{WEIGHT}_{ij} = \mathrm{IDF}_i * \mathrm{TF}_{ij} \quad (6)$$

For retrieval, the user enters a natural language query, and the text is passed to the concept-matching algorithm. A wild-card character can be used to have words completed for the user when, for example, the user is unclear on the exact spelling. The algorithm extracts all concepts from the query statement and returns them in a list, which is shown in the middle of the search window and includes the number of documents in which the concept occurs. All words which do not map into concepts are discarded. Each document with concepts in the list then receives a score based on the sum of the weights of terms common to the query and document, as is done in SWORD. The resulting list of matching documents is then sorted, with the weights normalized such that the highest ranking document is given a score of 100. As with the above systems, documents are viewed by double-clicking on their titles. As with SWORD, initially the top 10 document titles are displayed in a scrolling list, with the user able to add more in increments of 10. If the users desire to modify the search, additional terms can be added by entering more text or existing terms can be deleted by double-clicking on a term and verifying its deletion in a dialog box. SAPHIRE's interface is shown in Figure 4.

Previous Evaluations

SAPHIRE and BOOLEAN have been evaluated before in three different studies, though all of the studies only used bibliographic databases. The first two studies did not use interactive searching and assessed only SAPHIRE. In the first study, a collection of 12 queries and 200 abstracts from the AIDSLINE database were used (Hersh & Hickam, 1991). This study showed SAPHIRE to perform inferiorly to regular MEDLINE searching when used as a human indexing substitute in that environment. However, SAPHIRE produced better results with free text queries entered directly than with physician-generated Boolean queries. The second study used a collection of 75 queries and 2,344 documents retrieved from a clinical evaluation of MEDLINE, and showed SAPHIRE to perform comparably to novice and expert clinician users of MEDLINE in batch searching runs (Hersh & Hickam, 1994). The final study compared BOOLEAN (called "SWORD" in that study) and SAPHIRE with a collection of 10 queries and 1,992 extended abstracts from six volumes of the Medical Yearbook Series (Hersh & Hickam, 1992a). Senior medical students searched half of the queries with BOOLEAN and half with SAPHIRE. The results showed near equivalence in recall and precision. A questionnaire about subjective preferences for each system also showed near equivalence, though users preferring one system over the other tended to get better results with the system that they subjectively preferred.



Enter Query:
treatment of aids with azidothymidine

Matching Concepts [Matches]:
Acquired Immunodeficiency Syndrome [159]
Therapeutics [1720]
Zidovudine [10]

Status:
The top 10 of 164 documents to view are listed below.

Matching Documents [Score]:
ACQUIRED IMMUNODEFICIENCY SYNDROME -- Management [100]
RETROVIRUS INFECTIONS -- Therapy for HIV Infection [72]
INFECTION IN THE IMMUNOSUPPRESSED HOST -- treatment [64]
IMMUNIZATIONS AND CHEMOTHERAPY FOR VIRAL INFECTIONS -- zidovudine [60]
IMMUNIZATIONS AND CHEMOTHERAPY FOR VIRAL INFECTIONS -- ganciclovir [49]

FIG. 4. SAPHIRE interface.

Methods

User Searching

The queries for this study were selected from a collection of over 300 queries generated by residents and faculty using a multi-application computer workstation in the OHSU General Medicine Clinic (Hersh & Hickam, 1992b). Before using any application on this system, users were required to enter a statement about their patient and the information sought. Compliance with this request was incomplete by these busy physician users, so

TABLE 1. Ten sample queries with patient description and information need.

1. Patient: 60-year-old menopausal woman without hormone replacement therapy
   Info need: Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy?
2. Patient: 60-year-old man with disseminated intravascular coagulation
   Info need: pathophysiology and treatment of disseminated intravascular coagulation
3. Patient: 30-year-old with fever, lymphadenopathy, neurologic changes, and rash
   Info need: t-cell lymphoma associated with autoimmune symptoms
4. Patient: 57-year-old man with hypercalcemia secondary to carcinoma
   Info need: effectiveness of gallium therapy for hypercalcemia
5. Patient: 35-year-old man with aids and pancytopenia
   Info need: pancytopenia in aids, workup, and etiology
6. Patient: 68-year-old man with adult-onset diabetes melliitus noted to have thrombocytosis
   Info need: thrombocytosis, treatment and diagnosis
7. Patient: 18-year-old pregnant woman with hyperthyroidism
   Info need: use of beta-blockers for thyrotoxicosis during pregnancy
8. Patient: 35-year-old with advanced metastatic breast cancer
   Info need: chemotherapy advanced for advanced metastatic breast cancer
9. Patient: 63-year-old man with acute renal failure probably 2nd to aminoglycosides/contrast dye
   Info need: acute tubular necrosis due to aminoglycosides, contrast dye, outcome, and treatment
10. Patient: 40-year-old man with cocaine withdrawal
    Info need: cocaine withdrawal management

many queries had less than three words each of patient data and information need. These incomplete queries were not used for the experiment. In addition, a number of these queries were duplicates, since users searched the same topic on more than one application or repeated a query they had done in a previous searching session. After elimination of incomplete and duplicate queries, 106 remained to be used for this study. (Table 1 contains a sample of 10 of these queries.) The database for this study was the 1991 version of SAM, provided by the publisher. Like most textbooks, SAM has a hierarchical organization that lends itself to division into "documents," each of which consists of the lowest subdivision of text that has a heading. The subdivision process converted the 12 megabyte text into 6,623 documents. The title of each document was any heading or subheading excluding the chapter name.

The subjects for this experiment were 21 senior medical students from the OHSU School of Medicine. Each searcher was randomized to use two out of the three systems. With each system, the searcher performed nine queries. This study design permitted each query to be searched once by each system. Twenty queries were actually searched twice with each system in order to provide some overlap information. The searching was done on Apple Macintosh LC computers. The only reference material provided was a drug handbook that allowed participants to look up generic or trade names of drugs. The searching session began with a 30-minute introduction to the two systems being used by the searchers. A basic description of each system's indexing and retrieval mechanisms was given, followed by a demonstration of a sample search. Participants were instructed to search as if they were the physicians in the clinic who generated the questions, seeking to find documents that would provide pertinent information on the patients and their problem. They were instructed to search until they found a few relevant documents on each topic or, if no relevant documents were found, to quit after four to five search cycles. All three systems maintained a log file that timed and recorded every interaction that occurred between user and system.

When done searching, each participant completed a questionnaire asking details on past experience and assessment of the two systems that the searcher was assigned to use. For the latter, there were questions specific to each individual system used as well as questions comparing both against each other. The single-system questions were multiple-choice, while the questions comparing the two systems used a 100 mm analog scale with preference for a given feature of each system towards the end of the scales (Figures 5a, 5b, and 5c).

Relevance Assessment

The 21 searchers retrieved a total of 11,592 documents for the 106 queries. The relevance assessment was done in a two-step process. First, a single reviewer, a physician board certified in internal medicine, quickly scanned all documents retrieved for each query to eliminate all that were definitely nonrelevant. This process identified 1,630 potentially relevant query–document pairs, which were then reviewed in more detail by additional physicians who were also board certified in internal medicine. Documents were classified for relevance by the following criteria:

(1) Not relevant (NR)—document does not provide any pertinent information for the patient and information need described.
(2) Possibly relevant (PR)—document may provide some useful information to the clinician for the patient and information need described.
(3) Definitely relevant (DR)—document provides highly relevant information to the clinician for the patient and information need described.

This process resulted in two possible levels of relevance for a document, DR and definitely or possibly relevant (D+PR).

(a)

1. Mark on the scale below where you felt the two programs compared in ease of use:

I______________________________I______________________________I

SWORD was easier to use	Both were equivalent in ease of use	BOOLEAN was easier to use

2. Mark on the scale below where you felt the two programs compared in obtaining relevant documents:

I______________________________I______________________________I

SWORD obtained more relevant documents	Both were equivalent in obtaining relevant documents	BOOLEAN obtained more relevant docs

3. Mark on the scale below your opinion on SWORD's ranking of documents:

I______________________________I______________________________I

SWORD's ranking of documents was better	The ranking of documents did not matter	It was better not to rank documents (as in BOOLEAN)

(b)

1. Mark on the scale below where you felt the two programs compared in ease of use:

I______________________________I______________________________I

SAPHIRE was easier to use	Both were equivalent in ease of use	BOOLEAN was easier to use

2. Mark on the scale below where you felt the two programs compared in obtaining relevant documents:

I______________________________I______________________________I

SAPHIRE obtained more relevant documents	Both were equivalent in obtaining relevant documents	BOOLEAN obtained more relevant docs

3. Mark on the scale below your opinion of the way to two programs present search terms:

I______________________________I______________________________I

SAPHIRE's presentation of search terms was better	The presentation of search terms did not matter	BOOLEAN's presentation of search terms was better

4. Mark on the scale below your opinion on SAPHIRE's ranking of documents:

I______________________________I______________________________I

SAPHIRE's ranking of documents was better	The ranking of documents did not matter	It was better not to rank documents (as in BOOLEAN)

(c)

1. Mark on the scale below where you felt the two programs compared in ease of use:

I______________________________I______________________________I

SAPHIRE was easier to use	Both were equivalent in ease to use	SWORD was easier to use

2. Mark on the scale below where you felt the two programs compared in obtaining relevant documents:

I______________________________I______________________________I

SAPHIRE obtained more relevant documents	Both were equivalent in obtaining relevant documents	SWORD obtained more relevant docs

3. Mark on the scale below your opinion of the way to two programs present search terms:

I______________________________I______________________________I

SAPHIRE's presentation of search terms was better	The presentation of search terms did not matter	SWORD's presentation of search terms was better

FIG. 5. System comparison portion of questionnaire for comparing SWORD and BOOLEAN (a), SAPHIRE and BOOLEAN (b), and SAPHIRE and SWORD (c).

Analysis of Search Results

User searching performance was assessed by calculating the recall and precision for each system at two different levels of relevance. (Since only the documents retrieved by at least one of the three systems were assessed for relevance, relative and not absolute recall was measured. For simplicity, relative recall is hereafter referred to as recall.) A document was considered retrieved if it appeared in the retrieval list at the bottom of each system's searching window. In an attempt to find a single score to characterize performance, we combined recall and precision using Meadow's E score (Meadow, 1992):

$$E = 1 - \frac{\sqrt{((1 - \text{Precision}^2) + (1 - \text{Recall}^2))}}{\sqrt{2}} \quad (7)$$

Only queries with at least one definitely relevant document were included in the mean recall and precision calculations, since the recall and precision of queries with zero relevant documents are undefined. (Queries with zero possibly relevant documents were usable if there were one or more definitely relevant documents.) Statistical significance for all tests was measured by repeated measures analysis of variance. Searches performed in duplicate by the same system were not included in the initial analysis so as not to bias the results towards queries searched more frequently. However, the recall and precision values for duplicated searches were analyzed for intragroup correlation to assess their reproducibility.

Batch searching on the original query text was done with only SWORD and SAPHIRE, as BOOLEAN does not have a natural language interface that would permit such studies. As with user searching, SWORD and SAPHIRE performance was assessed using the two different levels of relevance. In addition, original query input into each system was done in two variations: one using only the information query (Q) and the other using the patient description along with information query (P+Q). All of the batch runs were assessed with standard recall-precision tables generated by a program based on the evaluation component of the SMART system. To assess statistical significance, a Wilcoxon signed rank test was used to compare levels of precision at low (0.20), medium (0.50), and high (0.80) levels of recall.

Results

After the 30-minute orientation on the two systems being used by each group of participants, user searching on the set of 18 assigned queries to each took from 2 hours, 10 minutes to 3 hours, 15 minutes. The post-searching questionnaire took 5–10 minutes.

Relevance Assessment

Of the 1,630 potentially relevant query–document pairs, 370 were determined to be PR, while 285 were determined to be DR. There were 24 queries that had no DR documents. Although 16 of these queries had one or more PR documents, the 24 were eliminated from the recall and precision determinations so that direct comparisons could be made between results based on DR or D+PR documents. Figure 6 shows the frequency of DR and D+PR documents per query. About 12% of the doc-

TABLE 2. Mean and standard deviation for per-query documents retrieved, documents retrieved and relevant, percent recall, percent precision, and E for the 84 queries with one or more definitely relevant documents.

	Documents retrieved	Percentage recall	Percentage precision	E
DR documents				
BOOLEAN	42.3 ± 54.3	70.6 ± 34.3	18.9 ± 24.4	0.21 ± 0.15
SWORD	21.7 ± 12.2	75.3 ± 32.2	14.8 ± 13.3	0.21 ± 0.12
SAPHIRE	22.8 ± 16.6	66.9 ± 36.7	16.1 ± 20.2	0.19 ± 0.16
D+PR documents				
BOOLEAN	42.3 ± 54.3	64.2 ± 31.6	28.7 ± 27.4	0.19 ± 0.13
SWORD	21.7 ± 12.2	66.9 ± 26.8	25.6 ± 19.2	0.18 ± 0.12
SAPHIRE	22.8 ± 16.6	61.6 ± 31.2	26.2 ± 22.5	0.18 ± 0.16

uments were reviewed by two relevance judges in order to assess interrater reliability. This was done via a kappa score, which was 0.37, indicating an acceptable degree of reliability (Kramer & Feinstein, 1981).

User Searching

For both levels of relevance, there was a trend towards higher recall and lower precision with SWORD (Table 2). The differences in E were minimal for all three systems. None of the differences in recall, precision, or E reached statistical significance. The standard deviations for all measures were large.

Because the mean number of documents retrieved was so much higher for BOOLEAN, in light of its better precision, we also calculated median values for number of documents retrieved, recall, precision, and E for each system, as shown in Table 3. In contrast to the mean values, the median number of documents retrieved per query was actually lowest for BOOLEAN. This occurred due to a small number of queries for which the BOOLEAN searcher retrieved a very large number of documents. This resulted in a near-zero precision for these small number of queries, which had only minor impact on the mean precision, but had a large impact on average number of documents retrieved. Viewing the data by median values also widened the differences among systems in recall for DR (but not D+PR) documents.

Because of the difference in results for mean and median values, we plotted mean recall against the number of relevant documents per query to see how recall varied based on the number of relevant documents available (Figures 7a–c, 8a–c). For each system and with both levels of relevance, there was a trend for recall to diminish with increasing numbers of relevant documents available. By simple linear regression, this trend was significant for all analyses ($p < .07$).

There were 20 queries searched twice by each system, 14 of which had DR documents. The intraclass correlation coefficient was similar for DR and D+PR documents. Users searching the same query with BOOLEAN had worse-than-chance ($r < 0$) correlation for both recall and precision, while users searching SWORD and SAPHIRE had much better-than-chance correlation for precision ($r > .6$). SWORD users had better than chance for recall ($r > .4$). Thus, reproducibility of search results

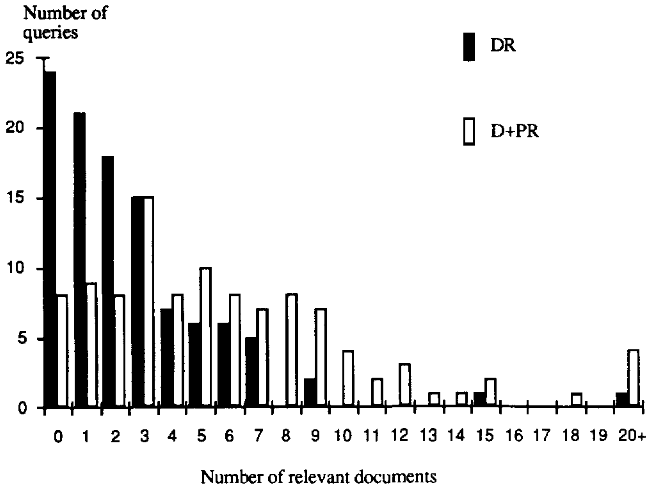


FIG. 6. Frequency of relevant documents per query.

TABLE 3. Median values for per-query documents retrieved, documents retrieved and relevant, percent recall, percent precision, and E for the 84 queries with one or more definitely relevant documents.

	Documents retrieved	Percentage recall	Percentage precision	E
DR documents				
BOOLEAN	15.5	91.65	8.7	0.29
SWORD	20	100	10	0.29
SAPHIRE	18.5	75	10	0.21
D+PR documents				
BOOLEAN	15.5	66.7	19.5	0.17
SWORD	20	66.7	20	0.16
SAPHIRE	18.5	63.1	20	0.15

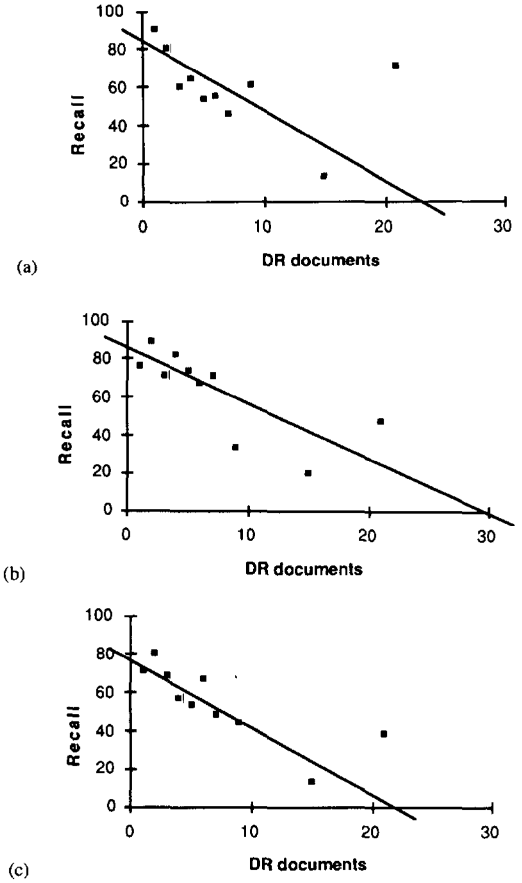


FIG. 7. Recall vs. number of definitely relevant (DR) documents with regression line for (a) BOOLEAN, (b) SWORD, and (c) SAPHIRE.

is much higher for systems which utilize natural language searching and relevance ranking.

Even though there were large individual variations in time spent per query, the average time spent per query was quite similar among the three systems (6.27 minutes for SWORD, 6.38 for SAPHIRE, and 6.39 for BOOLEAN).

There were few differences in user perceptions of the searching systems. Using Fisher's Exact Tests, no statistically significant differences were found among the three systems in users' perception of ease of finding information, ease of designating search terms, or relevance of documents retrieved (Table 4). Likewise, there were no significant differences in perception of the utility of SAPHIRE and SWORD's natural language interface and relevance ranking. The questionnaire results for comparisons between the two systems were measured and compared by single-sample t tests (Table 5). The only statistically significant difference was a preference for the group using SAPHIRE and BOOLEAN to prefer the relevance ranking of SAPHIRE.

Batch Searching

Batch searching results are depicted with recall–precision curves (generated from relevance ranking of batch queries) in Figures 9a and 9b for DR and D+PR documents, respectively, with user searching recall–precision points added for comparison. In general, the differences between the systems (SWORD and SAPHIRE) and the queries (Q vs. P+Q) were small. SWORD did better with P+Q queries than with Q queries, while the opposite occurred for SAPHIRE. This resulted in statistically significant better precision for SWORD with P+Q queries at the 0.20 (DR, p = .06; D+PR, p = .007), 0.50 (DR, p = .01; D+PR, p = .03), and 0.80 (DR, p = .007; D+PR, p = .01) levels of recall. SWORD also obtained higher precision at fixed recall points for Q queries, but the differences were smaller and not statistically significant. The user searching recall–precision points fell within the recall-precision curves for DR documents but not for D+PR documents.

Discussion

A long paper trail of information science opinion has argued the benefits of Boolean versus natural language searching over the last 30 years (Harter, 1992; Robertson & Hancock-Beaulieu, 1992; Swanson, 1988). We per-

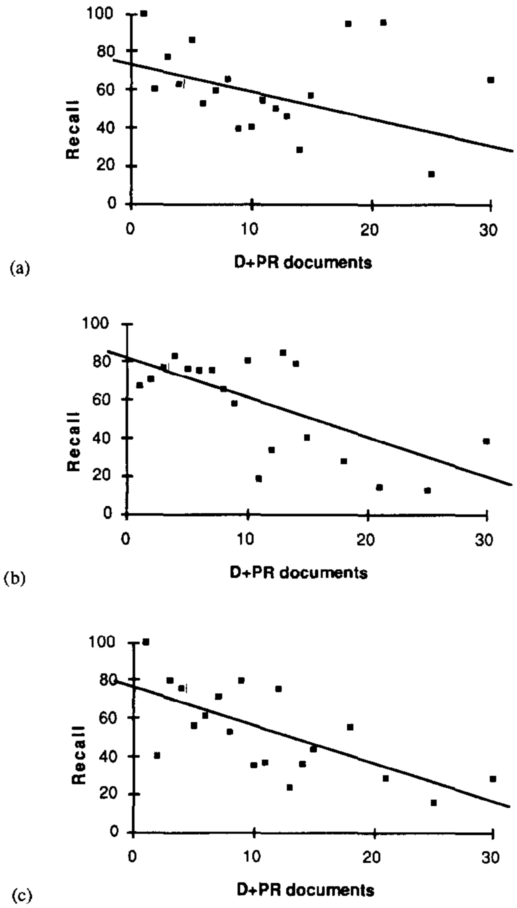


FIG. 8. Recall vs. number of definitely plus possibly relevant (D+PR) documents with regression line for (a) BOOLEAN, (b) SWORD, and (c) SAPHIRE.

TABLE 4. Individual questionnaire results (p values for Fisher's Exact Test across programs).

	BOOLEAN	SWORD	SAPHIRE	
Finding information was:				
Very easy	2	1	2	
Moderately easy	7	6	6	
Moderately difficult	4	6	5	
Very difficult	1	1	1	($p = .8$)
Designating search terms was:				
Very easy	3	4	4	
Moderately easy	9	7	8	
Moderately difficult	1	3	2	
Very difficult	1	0	0	($p = .8$)
Documents retrieved were:				
Nearly always relevant	1	1	0	
Mostly relevant	8	4	7	
Occasionally relevant	5	9	7	
Never relevant	0	0	0	($p = .5$)
Boolan operators were:				
Always useful	1			
Usually useful	8			
Sometimes useful	4			
Never useful	1			
Natural language interface was:				
Always useful		2	4	
Usually useful		6	6	
Sometimes useful		2	2	
Never useful		4	2	($p = .7$)
Document ranking was:				
Always useful		0	0	
Usually useful		3	5	
Sometimes useful		9	7	
Never useful		2	2	($p = .7$)
Use of medical concepts was:				
Always useful			2	
Usually useful			8	
Sometimes useful			3	
Never useful			1	

formed an actual experiment comparing interactive Boolean and natural language systems, with the results showing comparable performance for three different approaches. In this setting (senior medical students searching an online medical textbook) there were no statistically significant differences among three different indexing and retrieval approaches. The searchers were able to attain high levels of recall with any of the systems. Their precision was low, but this was of less consequence in this test database, which had relatively few relevant documents per query. There was also no difference in mean time to do a query with any of the systems as well as no user preference for any system over the others.

The lack of difference in recall and precision can be

TABLE 5. System comparison questionnaire results based on analog scales.

	Easier to use	Obtained more relevant documents	Preferred ranking over non-ranking
SWORD (0)—BOOLEAN (100)	66	66	59
SAPHIRE (0)—BOOLEAN (100)	41	54	33[a]
SAPHIRE (0)—SWORD (100)	54	51	n/a

Note: The number listed in parentheses next to each system represents the value that would be obtained by marking the end of the scale indicating complete preference for that system's approach. For example, in the only statistically significant result, the average measurement on the scale was 33, indicating that the average response was 33 mm from the SAPHIRE end of the scale and 67 mm from the BOOLEAN end ([a]$p = .02$, for all others $p > .05$).

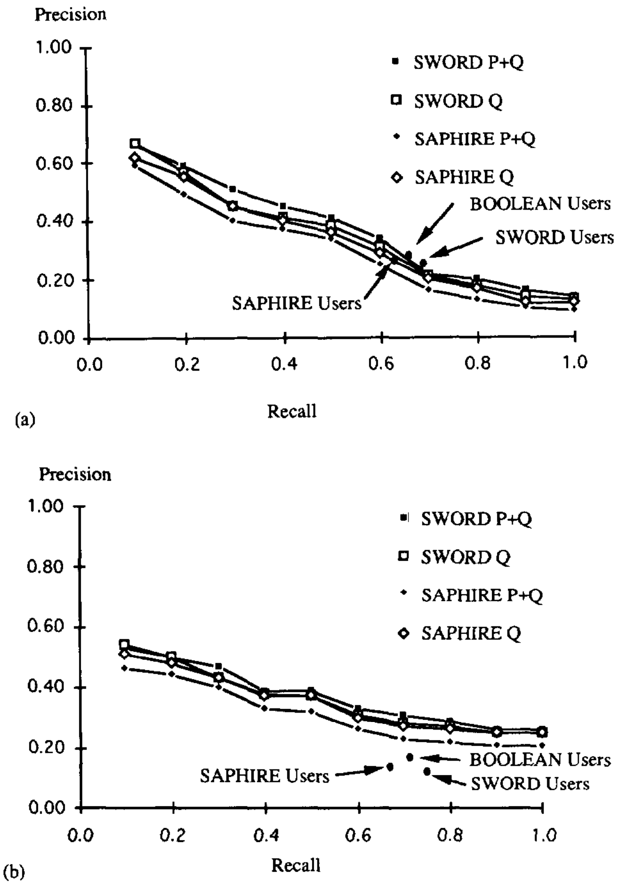


FIG. 9. Recall-precision graph for (a) DR and (b) D+PR documents, with user searching values included. Q indicates batch run with query only, while P+Q indicates with batch run using patient information plus query.

explained in one of two ways: Either there was no difference between the systems, or the difference was not detectable by our experiments. One can try to look for guidance from past similar experiments, in this case CIRT (Robertson & Thompson, 1990), MEDLINE (Hersh et al., 1994), and WESTLAW (Turtle, 1994). Unfortunately, each of these experiments had a different environment, making direct comparison difficult. The most similar experiments were CIRT and WESTLAW, in that they utilized expert searchers and a larger database. WESTLAW showed a benefit for natural language searching, while CIRT did not. The MEDLINE study showed that novice users could use a natural language system for accessing MEDLINE as effectively as more experienced searchers, but it did not look at expert searchers using natural language queries. Nonetheless, with this study, three of the four interactive comparisons of Boolean and natural language searching have shown similar performance, indicating these methods are comparable.

To confound matters, however, the batch searching results found statistically significant differences between SWORD and SAPHIRE in favor of the former. While these differences were of about the same magnitude and in the same direction as the interactive results, they indicate that different results are possible from batch and interactive studies. Given the use of the former to make recommendations about the effectiveness of various approaches to IR (i.e., Salton & Buckley, 1988), further experiments are necessary to determine the validity of results from batch searching experiments.

Before calling for more specific studies to address the divergent results, however, we feel it is more important to address another "finding" of this study, which is the demonstration of the problems in using recall and precision to evaluate interactive retrieval systems. These measures came into use when retrieval systems were in their infancy, in an era when searching was done in batch and by expert intermediaries. In this era of interactive searching directly by end-users, their value must be reassessed. Others have also noted the complexity of interactive retrieval evaluation (Robertson & Hancock-Beaulieu, 1992), and we are already exploring new measures of assessing how well end-users interact with retrieval systems (Hersh, 1994).

Consider some of the basic problems we faced in the use of recall and precision. What, for example, constituted a retrieved document? For this experiment, we defined any document shown in the list of matching documents from any query statement as retrieved. But in this experiment we saw instances of users making mistakes that they were able to correct, such as typing errors or accidental misuse of Boolean logic. The original mistakes, however, led to retrieval of documents by our definition and hence bad recall and/or precision, even when the search was adequate.

We also uncovered some statistical anomalies, which could lead to different conclusions based upon the choice of descriptive statistics. For example, a small number of Boolean queries retrieved a very large number of documents (sometimes due to errors as mentioned in the above paragraph). Much information retrieval research tends to report results as means, which if we did would lead to the conclusion that Boolean searching leads on average to twice as many retrieved documents with no difference in recall. But looking at the median gives the opposite conclusion, which is that fewer documents are retrieved. This discrepancy highlights the need for consistency in results reporting, a sentiment which has been echoed before (Kinnucan, Nelson, & Allen, 1987; Meadow, 1992).

Another statistical problem is that for queries with low numbers of relevant documents, recall can be an unstable measure. For example, a query with only one relevant document can only have a recall of 0% or 100%, while a query with only two relevant documents can only have a recall of 0%, 50%, or 100%, and so forth. This led us to compare the effect of the number of relevant documents on recall. We found that the average (mean) recall was higher when the number of relevant documents was lower.

Also a problem in the analysis of searching results was the variation of individual searchers. This was assessed by computing intraclass correlation coefficients for queries that were replicated by different searchers. We found that correlation was poor for BOOLEAN searchers but quite good for SWORD and SAPHIRE searchers. These latter two systems utilize natural language input and relevance ranking with a controlled retrieval output size, either of which might explain the better correlation of results. Users of natural language queries might be more likely to type the information need statement directly into the query box, while the controlled output set results in retrieval of a more consistent number of documents. This indicates that differences in individual searching need to be considered when comparing different retrieval paradigms, such as Boolean versus natural language.

Another issue related to recall and precision that could not be addressed by this study design (but is quite important to real users of systems) is, how many relevant documents are enough? The traditional user of a retrieval system in a library often seeks "everything" on a topic, yet the professional, such as a physician, is more likely to just want the answer to a question. Thus, if one document answers the question, then the total number of relevant documents (and hence recall) is meaningless. This implies the need for additional measurements of performance for interactive retrieval evaluation, such as measurements of knowledge gained.

The issue of number of relevant documents needed by the user was addressed partially by one of the secondary objectives of this study, which was the comparison of results based on different levels of relevance. We found for the most part that results did not differ in magnitude or direction whether considering definitely or D+PR documents. Whether this is generally applicable requires further study.

In conclusion, we showed that medical students can search an electronic textbook effectively using Boolean or natural language retrieval systems. They achieved comparable recall and precision with each of the three systems. We also found a number of problems with traditional evaluation measures and the statistics used to describe them in the interactive setting. Work must continue on identifying better measures of retrieval system performance as well as more standardized reporting of results. We have already been devising a general approach to this problem (Hersh, 1994), and in particular are currently investigating how physicians increase their knowledge and decrease their uncertainty with an IR system in a simulated practice environment. These and other studies should lead to better knowledge about the role and behavior of retrieval systems in the hands of end-users.

Acknowledgments

This work was supported by grant LM 05307 from the National Library of Medicine. The authors also thank Scientific American (New York, NY) for providing the text of *Scientific American Medicine* for this study.

References

Belkin, N., Cool, C., Croft, W., & Callan, J. (1993). Effect of multiple query representations on information retrieval system performance. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th Annual International ACM Special Interest Group in Information Retrieval* (pp. 339–346). Pittsburgh, PA: ACM Press.

Blair, D., & Maron, M. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the Association for Computing Machinery, 28,* 289–299.

Curley, S., Connelly, D., & Rich, E. (1990). Physicians use of medical knowledge resources: Preliminary theoretical framework and findings. *Medical Decision Making, 10,* 231–241.

Fuhr, N., & Knorz, G. (1984). Retrieval test evaluation of a rule-based automatic indexing (AIR/PHYS). In C. vanRijsbergen (Ed.), *Research and development in information retrieval* (pp. 391–408). Cambridge: Cambridge University Press.

Harter, S. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43,* 602–615.

Hersh, W. (1991). Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. *Medical Decision Making, 11(supp),* S120–S124.

Hersh, W. (1994). Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science, 45,* 201–206.

Hersh, W., Buckley, C., Leone, T. J., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In W. B. Croft & C. J. van Rijsberger (Eds.), *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval* (pp. 192–201). London: Springer-Verlag.

Hersh, W., & Hickam, D. (1991). A comparative analysis of retrieval effectiveness for three methods of indexing AIDS-related abstracts. In J. M. Griffiths (Ed.), *Proceedings of the 54th Annual Meeting of the American Society for Information Science* (pp. 211–225). Washington, DC: Learned Information.

Hersh, W., & Hickam, D. (1992a). A comparison of two methods for indexing and retrieval from a full-text medical database. In D. Shaw (Ed.), *Proceedings of the 55th Annual Meeting of the American Society for Information Science* (pp. 221–230). Washington, DC: Learned Information.

Hersh, W., & Hickam, D. (1992b). Impact of a computerized information system in a university general medicine clinic. *Clinical Research, 40,* 567A.

Hersh, W., & Hickam, D. (1994). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association, 1,* 51–60.

Kinnucan, M., Nelson, M., & Allen, B. (1987). Statistical methods in information science research. In M. Williams (Ed.), *Annual review of information science and technology* (pp. 147–178). Amsterdam: Elsevier.

Kramer, M., & Feinstein, A. (1981). Clinical biostatistics: LIV. The biostatistics of concordance. *Clinical Pharmacology and Therapeutics, 29,* 111–123.

Lindberg, D., Humphreys, B., & McCray, A. (1993). The unified medical language system project. *Methods of Information in Medicine, 32,* 281–291.

Meadow, C. (1992). *Text information retrieval systems.* San Diego: Academic Press.

Robertson, S., & Hancock-Beaulieu, M. (1992). On the evaluation of IR systems. *Information Processing and Management, 28,* 457–466.

Robertson, S., & Thompson, C. (1990). Weighted searching: The CIRT experiment. In K. P. Jones (Ed.), *Informatics 10: Prospects for intelligent retrieval* (pp. 153–166). York: ASLIB.

Rubenstein, R., & Federman, D. (1990). *Scientific american medicine.* New York: Scientific American.

Salton, G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science, 23*(2), 75–84.

Salton, G. (1991). Developments in automatic text retrieval. *Science, 253,* 974–980.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41,* 288–297.

Salton, G., Fox, E., & Wu, H. (1983). Extended boolean information retrieval. *Communications of the Association for Computing Machinery, 26,* 1022–1036.

Swanson, D. (1977). Information retrieval as a trial-and-error process. *Library Quarterly, 47,* 128–148.

Swanson, D. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39,* 92–98.

Turtle, H. (1994). Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In W. B. Croft & C. J. van Rijsberger (Eds.), *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval* (pp. 212–220). London: Springer-Verlag.

Turtle, H., & Croft, W. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems, 9,* 187–222.

vanRijsbergen, C. (1979). *Information retrieval.* London: Butterworth.