Research Paper ■

A Performance and Failure Analysis of SAPHIRE with a MEDLINE Test Collection

WILLIAM R. HERSH, MD, DAVID H. HICKAM, MD, MPH, R. BRIAN HAYNES, MD, PHD, K. ANN MCKIBBON, MLS

Abstract **Objective:** Assess the performance of the SAPHIRE automated information retrieval system.

Design: Comparative study of automated and human searching of a MEDLINE test collection.

Measurements: Recall and precision of SAPHIRE were compared with those attributes of novice physicians, expert physicians, and librarians for a test collection of 75 queries and 2,334 citations. Failure analysis assessed the efficacy of the Metathesaurus as a concept vocabulary; the reasons for retrieval of nonrelevant articles and nonretrieval of relevant articles; and the effect of changing the weighting formula for relevance ranking of retrieved articles.

Results: Recall and precision of SAPHIRE were comparable to those of both physician groups, but less than those of librarians.

Conclusion: The current version of the Metathesaurus, as utilized by SAPHIRE, was unable to represent the conceptual content of one-fourth of physician-generated MEDLINE queries. The most likely cause for retrieval of nonrelevant articles was the presence of some or all of the search terms in the article, with frequencies high enough to lead to retrieval. The most likely cause for nonretrieval of relevant articles was the absence of the actual terms from the query, with synonyms or hierarchically related terms present instead. There were significant variations in performance when SAPHIRE's concept-weighing formulas were modified.

■ **J Am Med Informatics Assoc.** 1994;1:51–60.

MEDLINE is one of the largest and most frequently used online databases in the world. In addition to approximately four million searches carried out annually on the networks of the National Library of Medicine (NLM),[1] MEDLINE can also be searched via several commercial online systems, as well as by many CD-ROM products. Despite the great commercial success of MEDLINE, it has limitations in both indexing and retrieval. Human indexing is expensive; over two million dollars and 44 full-time–equivalent indexers are used by the NLM each year to index MEDLINE.[2] This problem will become exacerbated as more medical text becomes available online. Human indexing is also inconsistent, as shown by Funk and Reid, who looked at MEDLINE references that were, for a variety of reasons, indexed in duplicate.[3] They found that the inter-rater agreement of index term assignments for central concept headings (starred MeSH terms, the most important concepts in the article) was 61%, while for heading–subheading combinations it was only 38%.

For retrieval, searching with most of the available systems is still limited to Boolean search statements phrased in either the Medical Subject Headings (MeSH) vocabulary or individual text words. Slingluff et al. found that MeSH terms are often difficult for users to find and apply in Boolean expressions.[4] Also, MeSH

Affiliations of the authors: Oregon Health Sciences University, Portland, OR (WRH, DHH); McMaster University, Hamilton, Ontario (RBH, KAM).

A version of this work was presented at the Fourteenth Annual Symposium on Computer Applications in Medical CARE (SCAMC), 1991.

Supported in part by National Library of Medicine Grant LM05307.

Correspondence and reprints: William R. Hersh, MD, Biomedical Information Communication Center, Oregon Health Sciences University, 3181 S.W. Sam Jackson Park Road, Portland, OR 97201.

Received for publication: 3/10/93; accepted for publication: 5/19/93.

terms are assigned by indexers as specified by the MEDLARS Indexing Manual,[5] with which few end-users have familiarity. Thus, end-users are often unable to apply the proper MeSH terms for searching.

These limitations of MEDLINE and other information retrieval systems are motivation for automated approaches to information retrieval, utilizing such features as automated indexing, natural-language query input, and ranking retrieved citations by relevance.[2] Early systems that have experimented with these approaches include Salton's SMART system[6] and the NLM's IRX Project.[7] These automated systems use indexing and retrieval based only on individual words. Their limitation is that humans search in terms of concepts, which are combinations of words that take on additional meaning when combined together. For example, the words high, blood, and pressure take on added meaning when they occur together in the phrase *high blood pressure*. An additional problem, however, is that concepts can have synonyms, which sometimes have no words in common, such as *high blood pressure* and *hypertension*.

SAPHIRE is an experimental system for automated indexing and retrieval.[8,9] While using many of the features of the word-based programs mentioned above, it indexes and retrieves at the level of full medical concepts. But in contrast to systems that use concept-based human indexing, which can be inconsistent as well as expensive, its concept-recognition is fully automated. In the indexing process, SAPHIRE identifies all concepts in a document. It then ranks the concepts by importance based on weighting algorithm that gives the most weight to concepts that occur frequently in some documents but infrequently in the rest of the collection. The value of this approach has been verified experimentally in word-based systems.[10]

This paper reports a study that had two primary objectives. The first was to assess the performance of SAPHIRE using a test collection of MEDLINE references. Performance was measured by calculating the recall and precision of the results of SAPHIRE's searches using a previously developed set of queries and a set of MEDLINE documents. The second objective was to carry out a failure analysis to look for factors that would allow interpretation of the recall and precision values as well as identify areas where SAPHIRE's performance could be improved. This consisted of: 1) a review of the adequacy of the Metathesaurus vocabulary used for SAPHIRE, 2) a review of the reasons for retrieval of nonrelevant articles and non-retrieval of relevant articles, and 3) experiments modifying the weighting factors used by SAPHIRE.

Background

SAPHIRE's Algorithm

SAPHIRE's weighting of terms is a combination of two factors commonly used in word-based automated systems. The first factor is the inverse document frequency (IDF), a measure of how infrequently a concept occurs in the entire document collection:

$$\mathrm{IDF}_i = \log \frac{(\text{number of documents in collection})}{(\text{number of documents with term i})} + 1 \qquad (1)$$

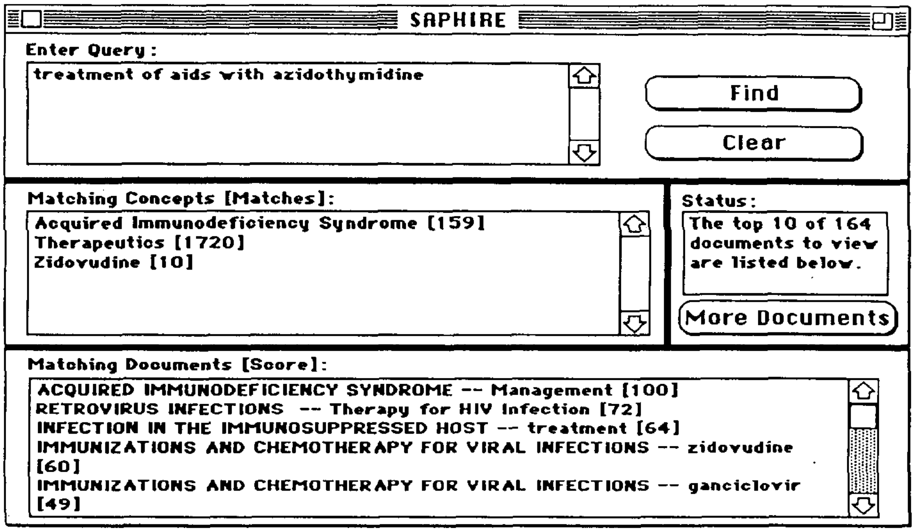
The IDF ensures that concepts that occur widely across a document collection will have less weight, hence importance, than those that occur rarely. The second factor is the term frequency (TF) for a concept in a document:

$$\mathrm{TF}_{ij} = \log(\text{frequency of concept i in document j}) + 1 \qquad (2)$$

These two factors are combined into a single weight, commonly known as the IDF * TF weight:

$$\mathrm{WEIGHT}_{ij} = \mathrm{IDF}_i * \mathrm{TF}_{ij} \qquad (3)$$

SAPHIRE's retrieval process uses the same concept-matching approach to extract concepts from a user's natural-language query. As is shown in Figure 1, the user enters a query, and all matching concepts are



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]

Figure 1 The SAPHIRE user interface. Free text queries are entered in the upper pane of the window. Matching concepts from the query are displayed in the middle pane, along with the numbers of documents in which they occur. Matching documents, ranked for relevance, are shown in the lower pane.

displayed in a list, which can be modified to add new terms or delete existing ones. Each document that has one or more terms from the query list is given a score based on the sum of the weights of all terms appearing in the query and document. This list is then sorted and presented to the user. This sorting of retrieved documents is known as *relevance ranking*; its goal is to point the user to documents that are most relevant, as determined by their increased frequencies of terms from the query.

In order for SAPHIRE to perform concept-based indexing in a broad domain such as biomedicine, a large vocabulary with a great breadth of concepts as well as a great depth of synonyms is required. The vocabulary best suited for this purpose is the Metathesaurus,[8] which is one of the knowledge sources in the Unified Medical Language System project of the NLM.[11] This vocabulary provides over 130,000 concepts from nearly a dozen original medical vocabularies, along with an equal number of synonym forms for those concepts. SAPHIRE's concept-matching process is quite efficient, with a sentence-length query processed in 2–4 seconds and documents processed at the rate of 250–300 kilobytes of text per hour on a Macintosh Quadra 700.

Measuring Retrieval Performance

Recall and precision are standard measures of searching performance. Recall is defined as the proportion of relevant articles that are retrieved from the entire collection:

$$\text{Recall} = \frac{\text{articles retrieved and relevant}}{\text{total articles relevant in collection}} \qquad (4)$$

Precision is defined as the proportion of articles relevant from a given search:

$$\text{Precision} = \frac{\text{articles retrieved and relevant}}{\text{total articles retrieved}} \qquad (5)$$

Recall is difficult to measure in large databases such as MEDLINE due to the improbability of knowing the total number of articles relevant for a given search. This problem is overcome by using the measure of relative recall, where the total number of relevant articles is approximated by the total number of relevant articles found by three or more users searching on the same query.

A number of recall and precision studies evaluating MEDLINE have been carried out over the last three decades. The first large-scale study of MEDLINE was done in 1966–68 by Lancaster, who evaluated 302 searches submitted by librarians to the NLM.[12] The mean recall of those searches was 57.7%, while precision was 50.4%. A more recent analysis of MEDLINE was performed by Haynes et al., who divided their subjects into three groups: librarians, expert physicians, and novice physicians.[13] In this study, physicians and medical students on an internal medicine service in an academic medical center who were novices in using MEDLINE originated the search request. Before searching online themselves, they wrote a brief description of the search question; this was later used by physicians expert in the use of MEDLINE and librarians to conduct independent searches. All citations retrieved for each were judged for relevance to that query by a clinician who was expert in the area of the search topic and was unaware of which searcher had retrieved a given citation. Their

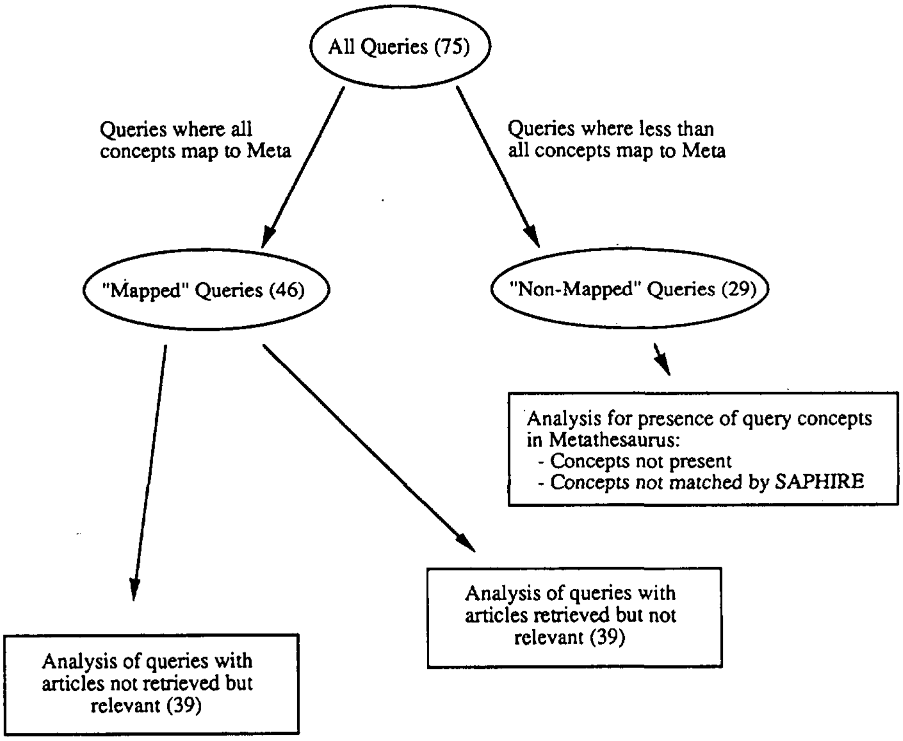

All Queries (75)
Queries where all concepts map to Meta
Queries where less than all concepts map to Meta
"Mapped" Queries (46)
"Non-Mapped" Queries (29)
Analysis for presence of query concepts in Metathesaurus:
- Concepts not present
- Concepts not matched by SAPHIRE
Analysis of queries with articles retrieved but not relevant (39)
Analysis of queries with articles not retrieved but relevant (39)

Figure 2 Failure-analysis schema for mapping and non-mapping queries.

results for 78 queries showed medical librarians to have a mean recall of 48.7% and a mean precision of 57.9%. Expert physician searchers had about the same recall (47.7%) but lower precision (47.1%), while inexperienced physician searchers had much poorer recall (27.0%) and precision (37.1%).

Comparison of retrieval performances is made more difficult when comparing traditional searching systems with those that use document ranking, such as SAPHIRE, because recall and precision will vary based on how far down the ranked retrieval list the user wishes to look. In general, as more documents are included in the retrieved set, recall will increase while precision will decrease. For this reason, many evaluations of systems using ranking techniques will consist of a recall–precision curve, whose points are the recall and precision values for a given cutoff of the ranked list. Retrieval systems that do not rank documents produce only a single recall–precision point.

A previous evaluation of SAPHIRE showed it to have better retrieval performance than traditional MEDLINE searching for physicians, although the study was limited by several factors.[14] First, all MEDLINE-style Boolean search statements were prepared by users on paper, preventing them from deriving feedback from viewing the retrieved citation that is typically present in an online session. Second, the content material was limited to conference proceedings abstracts exclusively about AIDS, which is not representative of MEDLINE in general. Nonetheless, that study did show that entering the text of the original query into SAPHIRE's natural-language interface led to better recall and precision than were obtained with Boolean search statements generated by physicians.

Methods

A test collection was constructed to serve as a "gold standard" for recall and precision calculations, based on a collection of citations, the actual user queries used to retrieve the citations, and a list of the relevant citations for each query. These citations and judgments were from the MEDLINE evaluation of Haynes and colleagues described above.[13] The initial data consisted of 78 queries, 3,403 citations, and relevance judgments for each citation retrieved by a given query. Since SAPHIRE relies on indexing of the title and abstract, all citations without an abstract were eliminated, leaving a total of 2,344 citations. After elimination of citations without abstracts, three queries were left with no relevant citations. These were also eliminated from the test collection, leaving a total of 75 queries. Recall and precision were then recalculated for each novice, expert, and librarian searcher based on the new 2,344-reference test collection.

SAPHIRE searching was done by entering the user's initial free-text statement of the search subject into SAPHIRE's natural-language interface. The text was largely the same as it occurred in the initial data, with the exception of three spelling corrections and two acronym expansions (when the acronyms were not present in the Metathesaurus). Recall and precision were calculated for SAPHIRE by eliminating all documents below weight cutoffs between 0% and 95% at 5% increments. Levels of recall and precision were compared between each group and SAPHIRE at 60% cutoff, which was the closest point on the SAPHIRE recall–precision graph that was equidistant between novice and experienced physician searchers (see Figure 3). Because the recall and precision distributions were nonparametric, a Wilcoxon signed-rank test was used to test for statistical significance between the groups. The highest possible recall of SAPHIRE (with no weight cutoff) versus librarians was also assessed by a Wilcoxon signed-rank test.

The overall schema of the failure analysis is shown in Figure 2. In order to eliminate the potential bias of a system being evaluated by its own developers, most human judgments in the failure analysis were performed in duplicate, with one of the judges (WRH) involved in the design and implementation of the system and the other (DHH) familiar with the system but not involved in its design. The kappa statistic was used to assess the reliabilities of the various judgments that these two reviewers made. The first step of the failure analysis consisted of assessing the adequacy of the Metathesaurus in providing concepts for indexing and retrieval. Since the SAPHIRE retrieval was done using batch processing of the queries to extract vocabulary terms, the queries needed to be assessed to be certain that the important terms mapped to appropriate vocabulary terms. For the queries that did not map into the Metathesaurus, the medically relevant noun phrases were identified, and the Metathesaurus browser was used in an attempt to find them. The noun phrases were considered present when a potentially SAPHIRE-recognizable variant was found in the Metathesaurus. The noun phrase was not required to represent the actual definition of the term as stated in the Metathesaurus documentation; it only had to map into a semantically meaningful term. This allowed the use of simple terms with potentially many meanings, such as *Association* and *Blood*. A search was also made for concepts that were inappropriately matched by SAPHIRE.

The second step in the failure analysis was to identify the causes of retrieved but not relevant (RNR) articles and not retrieved but relevant (NRR) articles for queries that did map correctly. The reason for excluding non-mapping queries was that the lists of terms resulting from these queries did not represent the complete conceptual contents of the queries. Whereas the interactive use of SAPHIRE permits the user to select all appropriate search terms and eliminate those that are inappropriate, this experiment was non-interactive. Thus, some queries did not map completely into the appropriate set of terms. These queries could lead to erroneous conclusions about the causes of RNR and NRR articles. After eliminating non-mapping queries, a classification scheme to measure the frequencies of the various reasons for inappropriate retrievals was created. In order to normalize the varying numbers of retrieved and/or relevant documents per query, classification was done on a per-query basis. Thus, after all documents were reviewed for a given query, each rater classified the most frequent reasons for RNR and NRR articles.

The final part of the failure analysis was a comparison of weighting schemes different than the original formula used, IDF * TF. A variety of other weighting measures were used individually and combined. The same experiments using the original natural-language query were run with a version of SAPHIRE modified to allow these different weights. These included the collection frequency (CF):

$$\mathrm{CF}_i = \log(\text{frequency of concept i in collection}) + 1 \qquad (6)$$

the non-logarithmic term frequency (NLTF): $$\mathrm{NLTF}_{ij} = \text{frequency of concept i in document j} \qquad (7)$$

the binary term frequency (BF): $$\mathrm{BIN}_{ij} = 1 \text{ if concept i in document j, 0 otherwise} \qquad (8)$$

and the document normalization factor, which controls for length of the document:

$$\mathrm{NORM}_j = \log(\text{number of indexing items in document j}) \qquad (9)$$

Table 1 ■

Recall and Precision Values of the Four Search Groups, for All Queries and Subgrouped by Whether the Queries Had All Medically Relevant Noun Phrases Mapped into Metathesaurus Terms

	Recall			Precision		
	All Queries	Map	Non-Map	All Queries	Map	Non-Map
Librarians	52.6			59.8		
		53.8	51.0		58.4	62.4
Expert clinicians	51.3			46.6		
		54.7	45.9		46.7	46.5
Novice clinicians	42.3			39.7		
		41.5	43.5		42.5	35.3
SAPHIRE*	45.5			41.8		
		47.5	43.2		44.3	35.4

*At 60% weight cutoff, the closest point on the SAPHIRE recall–precision graph equidistant between novice and experienced physician searchers.

Table 2 ■

Recall and Precision Values for SAPHIRE Searching at Different Weight Cutoffs

Weight	*Recall*	*Precision*	*Weight*	*Recall*	*Precision*
0	77.4	11.4	50	55.1	33.7
5	77.4	11.4	55	48.0	38.1
10	77.1	11.4	60	45.5	41.8
15	76.4	11.6	65	37.1	44.6
20	75.5	12.6	70	32.6	47.3
25	74.9	14.1	75	28.5	51.4
30	73.1	17.2	80	24.3	52.2
35	70.5	21.5	85	17.9	51.9
40	65.7	26.5	90	14.8	52.8
45	60.1	30.7	95	12.1	55.0

Each of the different weighting measures provided a ranked list of documents from which recall and precision were calculated. Since these comparisons were between different modifications of SAPHIRE, a continuous measurement of recall vs. precision was used. By converting precision values to false-positive rates, receiver operating characteristic (ROC) curves were created, which allowed comparison of performances across the entire spectrum of weighted documents. Areas under the ROC curves summarize the performances of individual indexing methods for capturing documents relevant to a query. A separate ROC curve was constructed for each query–method pair, and the ROC areas were calculated using the trapezoidal technique.[15] Analysis of variance was used to compare ROC areas among weighting methods.

Results

SAPHIRE's searching performance was similar to that of the physicians using regular MEDLINE. The recalculated values of recall and precision for the novice physicians, the expert physicians, the librarians, and SAPHIRE at 60% weight cutoff are summarized in Table 1. The mean recall and precision values for SAPHIRE at different weight cutoffs are summarized in Table 2 and plotted in a recall–precision graph in Figure 3. At the 60% weight cutoff using the individual search method, SAPHIRE's improvements over the original physicians in recall (p = 0.5) and precision (p = 0.7) were *not statistically significant*. Similarly, the advantages in recall (p = 0.3) and precision (p = 0.3) for expert physician searchers over SAPHIRE were not statistically significant. SAPHIRE achieved the highest recall of all groups, although for SAPHIRE to achieve recall equal to that of librarians (55.1% vs. 52.6%, respectively, at weight cutoff 50%), there was a significant cost in terms of

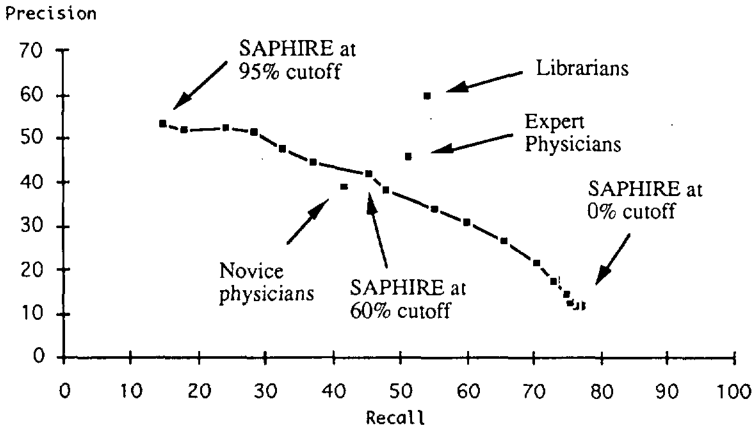


Figure 3 Recall-precision graph for SAPHIRE searching. The individual points for original, expert, and librarian searchers are also shown.

Table 3 ■

Mapping and Non-mapping Queries

Mapping queries

1. Stroke treatment with calcium channel blockers
   Calcium Channel Blockers
   Therapeutics
   Stroke

2. The association between AIDS and neuropathy
   Peripheral Nerve Diseases
   Acquired Immunodeficiency Syndrome
   Association

Non-mapping queries

1. Usefulness of latex agglutination antibody test in viral meningitis*
   Meningitis, viral
   Testing
   Antibodies
   Agglutination
   Latex

2. Treatment of Henoch–Schoenlein purpura†
   Purpura
   Therapeutics

* *Latex agglutination antibody test* should be a single concept.

† *Henoch–Schoenlein purpura* is not in the Metathesaurus, although the less common variants *Henoch purpura* and *Schoenlein–Henoch purpura* are.

precision (33.7% vs. 59.8%, $p < 0.02$). Nonetheless, SAPHIRE's highest possible recall was 82.8%, a figure unmatched by librarians using regular MEDLINE (vs. 52.6%, $p < 0.001$).

In the evaluation of medically important query noun phrases mapping into Metathesaurus concepts, a correctly mapping query was defined as one for which the two analysts agreed. Examples of mapping and non-mapping queries are shown in Table 3. Correct mapping occurred for 46 (61%) of the queries. The interobserver reliability was high (kappa = 0.61). There was no statistically significant difference in recall or precision between the sets of mapping and non-mapping queries for any of the search groups (see Table 1). Both analysts generated 34 noun phrases from the 29 non-mapping queries. Of these queries, 25 had one non-mapping noun phrase, three had two non-mappings, and one had three non-mappings. Of the 34 non-mapping concepts, 24 (71%) had no recognizable variant that could be found in the Metathesaurus. There were a total of 22 (29%) queries with concepts that could not be found in the Metathesaurus. There was only one instance of a concept mapping into a term inappropriately (*inhaled* mapping into *Aspiration*).

For the 46 queries that correctly mapped into concepts, 39 (84.8%) had RNR articles. Table 4 shows the categories of RNR articles, tallied on a per-query basis. For most of these queries, a majority of the nonrelevant articles contained some of the concepts in the query, with a weight high enough (>60%) to lead to retrieval. For some queries, a majority of articles had all of the search concepts present. The major variation between the two raters was the tendency of one to judge the retrieved articles for several queries to be actually relevant. The kappa statistic between the raters was 0.43. There were also 39 que-

Table 4 ■

Categorization of Articles Retrieved by the Two Reviewers but Not Relevant (RNR)

	Reviewer 1	Reviewer 2
Majority of articles are actually relevant	4	13
Majority of articles have all concepts but these are not central to query	11	3
Majority of articles have some concepts only but focus not central to query	24	23
Majority of articles have inappropriately matched concepts from query	0	0

Table 5 ■

Categorization of Articles Not Retrieved by the Two Reviewers but Relevant (NRR)

	Reviewer 1	Reviewer 2
Majority of articles are actually not relevant	6	6
Majority of articles have all concepts present but weighting too low	1	1
Majority of articles have one or more major concepts not present	7	5
Majority of articles have concept in form not matchable by SAPHIRE:		
a. Syntactic problem (noun as verb, etc.)	0	0
b. Semantic problem (synonym form not in Metathesaurus)	12	14
c. Hierarchical problem (parent or child term present)	7	7
d. Both semantic and hierarchical problems	5	7
e. Spelling or punctuation problem	2	0

Table 6 ■

Mean Precision Values at Different Recall Levels for Different Weighting Formulas*

	Mean Precision at 20% Recall	Mean Precision at 50% Recall	Mean Precision at 80% Recall
IDF	0.53	0.41	0.29
TF	0.49	0.37	0.20
CF	0.44	0.32	0.15
NLTF	**0.31**	**0.23**	**0.15**
BIN	0.48	0.33	0.17
IDF * TF	0.54	0.41	0.27
IDF * TF/CF	0.54	0.42	0.30
IDF * CF	0.52	0.41	0.28
TF * CF	0.52	0.40	0.26
IDF/NORM	0.52	0.41	0.26
TF/NORM	0.51	0.38	0.21
CF/NORM	0.40	0.29	0.14
NLTF/NORM	0.33	0.24	0.15
BIN/NORM	0.45	0.34	0.18
IDF * TF/NORM	0.54	0.42	0.28
(IDF * TF/CF)/NORM	**0.54**	**0.43**	**0.29**
IDF * CF/NORM	0.48	0.37	0.20
TF* CF/NORM	0.53	0.42	0.26

*See text for explanation.

ries with NRR articles, though they were not the same as the RNR set. Table 5 shows the categories of NRR articles, with the most common reasons for failure being lack of recognition of synonyms of indexing terms. In this analysis, the major disagreements between the assessors came with classifications of concepts not being found because synonyms were not recognized, or because hierarchically related terms were present instead, or both, which reflects the cooccurrence of these problems. The kappa statistic for the analysis with all the categories was 0.46. When the synonym, hierarchical, and synonym/hierarchical categories were merged, it improved to 0.57.

Weighting parameters are compared in Table 6, in which the best ((IDF * TF/CF)/NORM) and worst (NLTF) formulas are indicated in boldface. Of the parameters used individually, the inverse document frequency performed best, while the non-logarithmic term frequency performed worst. Several combination formulas showed improvement over SAPHIRE's original IDF * TF, the best being the combination of four factors:

$$\text{BEST-WEIGHT}_{ij} = \frac{\text{IDF}_i \ast \text{TF}_{ij}}{\text{CF}_i \ast \text{NORM}_j} \qquad (10)$$

Discussion

The results of this study lead to several conclusions. First, SAPHIRE's performance in retrieval of MEDLINE records is equivalent to that of physicians using regular MEDLINE techniques. In recall and precision, SAPHIRE was slightly better than novice physicians using MEDLINE but somewhat worse than expert physicians, though none of the differences was statistically significant. SAPHIRE did not perform as well as librarians, although it did achieve higher total recall when all citations were retrieved, but at a severe expense to precision. The conclusions from the failure analysis listed below yield insight into how SAPHIRE's performance can be improved, with future research aiming to improve precision without sacrificing recall.

There were some limitations of this study that could have led to SAPHIRE's appearing to perform worse than in actual use. One limitation was that the articles in the test collection created for this study were only those that could be retrieved from MEDLINE using MeSH and text-word queries. Given the previous work by McKibbon et al,[16] which showed that MEDLINE queries done by three different searchers often led to retrieval of disparate sets of references, it is likely that there are additional relevant MEDLINE references for each query that were not part of this test set. Some of those references could be retrieved by SAPHIRE only. Another limitation was that an article was considered retrieved by librarians or clinicians only if it was viewed on the screen during the searching session. Depending on whether the articles not viewed were relevant or not, this could have improved recall or hampered precision. A final limitation was that SAPHIRE was not used interactively for these experiments. In interactive searching, a user would most likely refine his or her search, especially if the initial search yielded few relevant articles. On the other hand, librarians might also perform better under more usual circumstances: they had only the clinicians' questions to work from, without the opportunity for clarification or interaction with the clinicians.

The second conclusion is that the Metathesaurus as used by SAPHIRE was incomplete for a substantial number of queries. Over one-fourth of the queries had medically significant noun phrases that had no representation in the Metathesaurus. While there are probably methods to get around the absence of these topics, such as the use of broader terms or combinations of terms, these are less likely to be used by more inexperienced users. This represents a challenge to the designers of the Metathesaurus, since a goal of the project is to create a comprehensive access to medical terminology. Of course, the lack of mapping noun phrases did not have any effect on SAPHIRE's performance in this study, indicating that the incomplete lists of concepts still led to the same level of retrieval performance. This may have been an artifact of the test collection, which was not evenly distributed among all medical topics, but clustered around topics related to terms in the query, with articles likely to be inappropriately retrieved by SAPHIRE eliminated from the set.

The third conclusion is that there were recurring patterns as to why nonrelevant articles were retrieved and relevant articles were not retrieved. For the former (RNR), there were enough (and sometimes all) of the query concepts present to cause retrieval, even though the article was not relevant. The problem of the presence of some or all of the search terms in a RNR article is a major challenge to automated indexing systems, which do not have the advantage of human indexers who, however inconsistent, can choose indexing terms focused on the main topics. There are many potential avenues for attacking this problem, some of which we are already investigating. One of these is to apply automatic indexing to more of the text, up to and including the full text of the document. By including more text, the relevant terms are likely to be repeated, leading to their obtaining higher weights in the indexing process, resulting in their contributing more score for the document to a query in which they are used. Another is to apply more knowledge about the semantics of an article, such as the importance of certain terms in various parts of the article. For example, the introduction of a paper might be best for gleaning the focus of the paper, whereas the methods section of a paper might be best for finding terms related to the study design or intervention.

The majority of queries related to non-retrieved articles that were relevant (NRR) had either lack of recognition of synonyms, presence of a hierarchically related term, or both. It is encouraging that most NRR articles had synonyms or hierarchical terms that could lead to retrieval. The synonym aspect of this problem will be helped by future versions of the Metathesaurus that offer more complete lists of synonyms. The hierarchical problem is one that has been addressed by others,[17] and its solution could be implemented in SAPHIRE.

The final conclusion is that there were large differences in system performance based on choice of weighting formula. The best single weighting measure was the inverse document frequency, verifying a long-known principle from information science that a term's rarity in other documents is a better factor in discriminating between documents than its frequency count in a particular document. Of the different measures of term frequency, those that used logarithmic modification of the frequency performed better than the one that did not. In fact, the worst performance was provided by the non-logarithmic term frequency, indicating that, at least in SAPHIRE, simple term counts were detrimental. The best performance for a weighting formula occurred with the combination of factors shown in equation 10.

In summary, while some might argue that the best approach to enhancing search results is to elevate the searching skills of physicians up to that of librarians, SAPHIRE provides a feasible alternative. Furthermore, it is not a static program, and the current level of performance can be improved by new approaches. For example, as described above, the use of full-text documents may lead to greater precision, while the leverage of hierarchical relationships and/or better synonym lists may lead to better recall. There are also other techniques that look promising, such as the selective use of natural-language processing techniques that allow improved concept recognition[18] and tracing of article references of recent papers to find relevant articles more precisely. As each of these enhancements is implemented, the effect on performance will be evaluated, with the ultimate goal of determining the best strategies for information retrieval in the biomedical domain.

References ■

1. Siegel E, Cummings M, Woodsmall R. Bibliographic retrieval systems. In: Shortliffe E, Perreault L, ed. Medical Informatics: Computer Applications in Health Care. Reading, MA: Addison-Wesley, 1990:434–65.
2. Hersh W, Greenes R. Information retrieval in medicine: state of the art. MD Comput. 1990;7:302–11.
3. Funk M, Reid C. Indexing consistency in MEDLINE. Bull Med Library Assoc. 1983;71:176–83.
4. Slingluff D, Lev Y, Eisan A. An end-user search service in an academic health sciences library. Med Reference Serv Q. 1985;4(1):11–21.
5. Charen T. MEDLARS Indexing Manual, Parts I and II. Springfield, VA: National Technical Information Service, 1983.
6. Salton G. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.
7. Harman D, Benson D, Fitzpatrick L, Huntzinger R, Goldstein C. IRX: an information retrieval system for experimentation and user applications. SIGIR Forum. 1988;22:2–10.
8. 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.
9. Hersh W, Greenes R. SAPHIRE: an information retrieval environment featuring concept-matching, automatic indexing,

and probabilistic retrieval. Comput Biomed Res. 1990;23:405–20.

10. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Proc Manage. 1988;24:513–23.
11. Lindberg D, Humphreys B. The UMLS knowledge sources: tools for building better user interfaces. In: Miller R, ed. Proceedings of the Fourteenth Annual Symposium on Computers in Medical Care. Washington, DC: IEEE Computer Society Press, 1990:121–125.
12. Lancaster F. Evaluation of the MEDLARS Demand Search Service. Bethesda, MD: National Library of Medicine, 1968.
13. 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.
14. Hersh W, Hickam D. A comparison of retrieval effectiveness for three methods of indexing medical literature. Am J Med Sci. 1992;303:292–300.
15. Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148:839–43.
16. McKibbon K, Haynes R, Dilks CW, et al. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searchers. Comput Biomed Res. 1990;23:583–93.
17. Rada R, Bicknell E. Ranking documents with a thesaurus. J. Am Soc Info Sci. 1989;40:304–10.
18. 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.