DEVELOPING SEARCH STRATEGIES FOR DETECTING HIGH QUALITY REVIEWS IN A HYPERTEXT TEST COLLECTION

By

Michael Peter Zacks, M.D.

A THESIS

Presented to the Division of Medical Informatics and Outcomes Research

and the Oregon Health Sciences University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Science

May 1998

School of Medicine
Oregon Health Sciences University

CERTIFICATE OF APPROVAL

[redacted]

William Hersh, M.D.

[redacted]

Joan Ash, PhD

[redacted]

Paul Gorman, M.D.

[redacted]

David Hickam, M.D.

Associate Dean for Graduate Studies

TABLE OF CONTENTS

Acknowledgments

I gratefully acknowledge the support of the Portland VA Medical Center which provided the funding for my Informatics Fellowship without which I could not have completed this work. I also thank William Hersh, M.D. and David Hickam, M.D. for their guidance in conducting this study and in writing this thesis.

Abstract

Objective: To identify search strategies for retrieving high quality review studies about etiology, prognosis, therapy, and diagnosis from the World Wide Web (WWW) medical documents.

Design: Observational study of the performance of search strategies based on terms found in high quality review articles in a collection of hypertext medical documents from the WWW.

Measures: The sensitivity and specificity of search strategies for review articles and for review articles in general and with a specific focus were determined by comparison to a manual review of a collection of hypertext medical documents.

Results: A total of 1058 hypertext medical documents from seven American and Canadian government and academic Web sites were included in the study collection. About 16% of the documents in the collection met the criteria for high quality review documents. Search strategies for review documents were identified that had 87% sensitivity and 95% specificity. Compared to simple strategies combining the term "review" and the article focus, more complex strategies based on terms found in high quality studies were more sensitive in identifying review articles of specific focus. These more complex strategies had a sensitivity of 83% for diagnosis, 85% for therapy, 79% for prognosis, and 88% for etiology, while the simple strategies had a sensitivity of 87%, 73%, 38%, and 46%, respectively. In addition, the more complex strategies were more specific for high quality review articles on diagnosis and therapy.

Conclusion: Search strategies can be identified that enhance retrieval of review documents and review documents of specific focus from a collection of WWW hypertext medical documents.

I. Introduction

Over the last five years , the World Wide Web (WWW) has grown rapidly. Between June 1993 and January 1997, the approximate number of WWW sites grew from 130 to over 650,000.[1] Along with this expansive growth, sites with clinically relevant material have become more common. Hospitals, medical schools, journals, government agencies, and others are posting clinical information on the WWW, most commonly in the form of guidelines, general articles, and reviews rather than original research. Although the exact number of clinical WWW sites is unknown, they are becoming an increasingly used method for dissemination of medical information.[2]

As the WWW is emerging as a new medium for medical knowledge, the importance of indexing and retrieving information in hypertext form has grown. Methods for indexing and retrieval are active areas of commercial research and development as demonstrated by the proliferation of search engines such as Altavista, Yahoo, and Excite. However, the relative performance of different strategies for information retrieval from hypertext is poorly defined.

This introductory section describes the background for an inquiry into strategies to improve medical information retrieval from hypertext. The first section proposes a standard for comparison between retrieval methods by developing a model of relevance in clinical information retrieval. The second section describes the current methods of retrieval on the WWW and three potential strategies for improving retrieval of medical information from hypertext. The third section analyzes how each strategy of improving retrieval described in section two fulfills the model of relevance in clinical searching. The final section outlines the research hypothesis and objectives.

1. A Model of Relevance in Clinical Information Retrieval

In order to compare different methods of retrieval from hypertext, the standards for successful clinical information retrieval first must be defined. The traditional measures of performance in information retrieval are recall and precision. According to this view, documents retrieved are either relevant or irrelevant to a given query. Recall (R) is defined as the number of documents retrieved that are relevant divided by the total number of relevant documents:

$$R = \frac{\text{Number of documents retrieved and relevant}}{\text{Total number of documents that are relevant}}$$

Precision (P) is the number of documents retrieved that are relevant divided by the number of documents retrieved:

$$P = \frac{\text{Number of documents that are retrieved and relevant}}{\text{Number of documents are retrieved}}$$

For example, if a collection of 1000 documents has 300 documents relevant to a query and the query retrieves 100 documents, 50 of which are judged relevant; then the recall is 50/300 (17%) and the precision is 50/100 (50%).

Hersh has described some problems with this simple dichotomous view of relevance in clinical searching.[3] First, the methods for establishing relevance judgments for clinical articles have been inadequate. Research has shown that a variety of factors can affect relevance judgments. These include the order of presentation of the articles, the size of the retrieval set, and the users' knowledge of the material. Thus, different judges may reach opposite conclusions about the relevance of a given article. Second, the fact that an article is about a topic does not guarantee its usefulness in clinical practice. Other factors such as an article's intended audience, level of detail, and quality may also influence its usefulness.

An alternative view of relevance in the clinical setting can be derived from the work of researchers at McMaster University.[4] They have described a method for evaluating the medical literature based on the methodology of studies. They outline specific criteria for clinical relevance of studies about therapy,[5] diagnosis,[6] prognosis,[7] etiology;[8] as well as for overviews ,[9] guidelines ,[10] decision analyses,[11] articles about variations in outcomes,[12] and clinical utilization reviews.[13] These criteria are designed to help the reader to judge whether the methods of the studies are valid, to interpret the results presented, and to determine if these results can be applied to the reader's own patients.

According to the McMaster researchers, a dichotomous classification of documents as being relevant or irrelevant to a query on a topic is insufficient to describe the goals of clinical searching. While clinicians may seek information about a specific topic, they also may have more exact requirements. Clinicians may want a particular document type such as a review or a clinical guideline. They also may want a document that deals with a particular topic focus such as etiology or therapy. In addition, clinical searchers often may have requirements about the quality of the materials they seek: they may not want any article on a topic, but one that meets particular criteria for quality.

Defining the criteria for quality of clinical material on the WWW has proven to be problematic. In theory, such a set of criteria could help searchers to distinguish more valuable material from the less useful. However, if the criteria cannot be applied reproducibly to select high quality material, they could mislead searchers. As yet, none of the commonly used quality instruments on the WWW has been validated to establish inter-observer reliability and construct validity.[14]

Many of the commonly used rating instruments have sought to define the quality of WWW documents in terms of their physical attributes. For example, Silberg et al. emphasize the importance of providing information about:

1. Authorship: information about the authors, their affiliations, and credentials
2. Attribution: references and sources for all content and copyright information
3. Disclosure: Web site sponsorship and any commercial support or conflict of interest
4. Currency: dates of original posting and updates.[15]

Similarly, the checklist for informational WWW pages created by Widener University emphasizes such criteria as a clearly identifiable sponsor of the page, an address for the sponsor, a list of references of information sources, the lack of advertising content, and the dating of the page's content.[16] Although clearly important, these physical attributes may not be sufficient for clinical searchers to evaluate a document's quality. In addition, clinical searchers may place heavy weight on the actual content of documents.

Another method of defining high quality clinical documents has been promoted by researchers from McMaster University and relies on evaluating the rigor of the study methodology.[17] Specific criteria have been set forth for a variety of article types including review documents [9] and practice guidelines.[10] For example, the primary McMaster criteria for review articles include questions about whether the review addresses a focused clinical topic and whether the inclusion criteria used to select articles are appropriate. There are also secondary criteria that further guide readers in evaluating the methodology of a review (Table 1).

These McMaster criteria have been demonstrated to have high inter-observer reliability.[18] Although initially designed for journal articles, the McMaster criteria also may be applied to documents on the WWW. Indeed, if the WWW is to become a legitimate alternative to print journals, there is reason to argue that the same criteria for quality should apply to both media.

In summary, the McMaster group has proposed four main document characteristics. The first is document topic. The second is document type, such as original study or review. The third is document focus, such as diagnosis or prognosis. The fourth is document quality based on evaluation of the document's methodology.

Table 1: McMaster criteria for high quality review articles

McMasters guides selecting high quality reviews
Primary Guides
1) Did the overview address a focused clinical question?
2) Were the criteria used to select articles for inclusion appropriate?
Secondary Guides
1) Is it unlikely that important relevant studies were missed?
2) Was the validity of the included studies appraised?
3) Were the assessments of the studies reproducible?
4) Were the results similar from study to study?

2. Information Retrieval Methods on the WWW

The WWW is a collection of interconnected files or "pages" residing on host computers or "sites" connected together by the Internet. Each page may have multiple links to other pages located on any computer connected to the Internet. By selecting a link, usually using a mouse, users can navigate between WWW pages. Some pages called index pages have links to many other WWW pages. Other pages may have relatively few or no links. When a group of pages on the same topic is linked together and resides at a single site, it can be thought of as a single document.

Because information of the WWW is dispersed widely among numerous host computers and no master catalogue of all pages exists, locating information on the WWW about a specific topic can vary from simple to challenging. When users know of a site that has information about a topic, they can usually follow links to discover other sites with similar information. However, when users have no idea which site might contain information on a topic, starting a search can prove difficult.

To address this problem, a variety of search engines have been created.[1] These allow users to enter search terms that describe a topic and obtain a list of pages to investigate. Search engines index the WWW by using programs called spiders or web crawlers that navigate the links of the WWW automatically and collect all the pages located. The search engine then creates an index of these pages based on the words they contain. The engines may use some or all of the words in each document

Some search engines employ techniques such as stemming and stop lists to reduce the computational complexity of indexing. Stemming involves the removal of suffixes such as “ing” and “ion” to reduce words to their root form. This groups words of similar meaning together and reduces the index size. Stop lists are collections of words that are very common in English, such as “and”, “a”, and “the”, that a search engine ignores. Indexing these common words is computationally expensive and does not help to discriminate between documents since the words appear in almost every document.

Clinicians using these search engines locate documents by entering a query consisting of one or more terms. Most search engines allow the use of Boolean operators such as “and” and “or” to combine various search terms. The search engine displays a list of pages that contain these search terms. The order of the pages displayed usually depends on the frequency and position of the search terms in the page, although each search engine has its own proprietary algorithm for ordering the search results.

The automated acquisition and indexing of pages by these search engines can make them difficult to use in medical searching.[3] While many WWW pages provide useful, accurate medical information, other pages that contain similar words may be collected from less reputable sources. The indexes created, however, do not distinguish between sources and present all the pages to users who can have trouble establishing the origin of a page.

These difficulties in using the WWW for medical searching have led researchers to propose several broad strategies for improving information retrieval. First, some have attempted to improve retrieval by selecting documents relevant to a particular domain of knowledge such as medicine. Searchers who apply queries to this subset of the WWW may be more likely to obtain relevant, high quality documents since the selection process can filter out other material. Document selection has been used by a variety of researchers interested in providing high quality clinical content on the WWW. For example, Medical Matrix (http://www.medmatrix.org) employs a manual review of WWW sites to select those with clinical usefulness. Links to these sites are provided from the Medical Matrix site.

The strategy of document selection has drawbacks. First, choosing the documents for inclusion in a site requires ongoing human effort. As the WWW continues to grow; the effort required to identify, evaluate, and index clinically relevant sites will also increase. Second, the standards for inclusion in sites that use document selection are not always explicit. Thus, the user may not know exactly how a particular document came to be included and what inclusion means about document quality.

A second strategy for improving information retrieval on the WWW is document classification. This strategy involves assigning to WWW documents keywords, usually from a controlled vocabulary such as the Medical Subject Headings (MeSH). This strategy can improve retrieval by grouping documents by medical concepts. Thus, regardless of the terminology used in a document, all documents indexed with the same keyword can be found by a search. Keyword indexing can also be combined with document selection. These two strategies have been effectively combined in CliniWeb (http://www.ohsu.edu/ cliniweb), which allows keyword searching using MeSH.[19]

Despite the advantages of document classification, it also has important drawbacks. First, like document selection, manually assigning keywords to WWW pages will become more difficult as the WWW grows. At present there is no reliable automated method of indexing WWW documents, although experimental systems are being developed.[20] A second problem with document classification is that the fluid nature of the WWW will require continual effort to maintain currency of the links.

A third potential strategy for improving information retrieval on the WWW is query refinement. Here, the user attempts to select query terms that improve the chance of retrieving relevant documents. This strategy has been applied with success to searching on MEDLINE by a group at McMaster University led by Haynes (Table 2).[18] For example, these researchers were able to devise a search strategy that retrieved high quality, original articles about diagnosis with 92% sensitivity and 73% specificity from a collection of MEDLINE documents.

These strategies for query refinement developed for MEDLINE, however, may not translate directly into searching on the WWW. MEDLINE strategies often rely on

Table 2: MEDLINE search strategies for identifying studies of high methodological quality for articles from 1991 and 1986. From Haynes et al [18]

Article Type	Search Strategy
1991	
Diagnosis	(Exp sensitivity a#d specificity) or diagnosis& (px) or diagnostic use (sh) or sensitivity (tw) or specificity (tw)
Treatment	Randomized controlled trial (pt) or drug therapy (sh) or therapeutic use (sh) or random (tw)
Prognosis	Incidence or exp mortality or follow-up studies or mortality (sh) or prognos: (tw) or predict: (tw) or course: (tw)
Etiology	Exp cohort studies or exp risk or (odds (tw) and ratio (tw)) or (relative (tw) and risk (tw) or (case (tw) and control (tw))
1986	
Diagnosis	Diagnosis& (px) or specificity (tw)
Treatment	Random allocation or comparative study or drug therapy (sh) or placebo (tw) or (controlled (tw) and trial (tw))
Prognosis	Prognosis or exp cohort studies or mortality (sh) or (natural (tw) and history (tw)) or predict: (tw) or course (tw)
Etiology	Exp Cohort studies or risk (tw) or causation (tw) or causal: (tw)

Terms are MeSH unless otherwise noted. tw = textword; sh = subject heading; px = subheading, pre-explosion; pt = publication type; # and : denote single and multiple wild-card characters.

the use of controlled vocabulary indexing or employ database fields of MEDLINE, such as publication type, that are not available for WWW documents. It is, therefore, uncertain whether queries can be identified that will retrieve high quality material from the WWW.

3. Retrieval methods and clinical searching

Document selection, document classification, and query refinement differ in their ability to meet the goals of clinical searching (Table 3). Document selection can easily be used to pick out documents about a particular topic, of a given type, or of a certain quality. While document selection also could be used to pick out documents of a particular focus, such as diagnosis, it is unclear how this would be useful. Document selection may not easily scale unless automated methods can be found to choose documents for inclusion. Document classification also could help users to identify articles by topic, type, focus, and even quality. However, at present classification is a manual process that may not easily scale as the size of the WWW increases. Finally, query refinement can pick out articles by topic and has potential to identify articles by type, focus, and quality. But unlike the other two methods, query refinement may easily scale as the WWW grows in size. This is because once a query strategy has been defined it can be used whether there are a thousand articles or a million articles to search. Thus, query refinement has great potential to improve retrieval of medical documents on the WWW.

Table 3: Retrieval methods ability to meet goals of clinical searching

	Topic	Type	Focus	Quality	Able to Scale
Document Selection	Yes	Yes	Maybe	Yes	No
Document Classification	Yes	Yes	Yes	Yes	No
Query Refinement	Yes	Maybe	Maybe	Maybe	Yes

The current study focused on the retrieval of review articles for several reasons. First, casual observation suggests that much of the medical information on the WWW is in the form of tertiary literature such as reviews, guidelines, and general articles. Second, although McMaster criteria exist for reviews and guidelines, no criteria have been created to evaluate the quality of general articles. Third, the McMaster criteria for quality of review articles seem more objective than those for guidelines. The primary quality criteria for reviews are: 1) "Did the article address a focused clinical question?" and 2) "Can you determine criteria used to select articles included in the review?"[9] In contrast, the primary quality criteria for guidelines are: 1) "Were all important options and outcomes clearly specified?" and 2) "Was an explicit and sensible process used to identify, select, and combine evidence?"[10] Answering these later questions may require more knowledge of the article topic than answering the questions for reviews.

4. Research Hypothesis and Objectives

The research hypothesis of the study was:

> Search strategies can be identified that improve retrieval of hypertext review documents of high methodological rigor.

The specific objectives were:

1. To create a collection of hypertext WWW documents.

2. To measure the sensitivity and specificity of three search strategies for selecting high quality review documents from a collection of hypertext medical documents.

3. For reviews that focus on diagnosis, therapy, etiology, and prognosis; to compare the sensitivity and specificity of two search strategies: one simple strategy based on the focus term and the term "review" and one more complex strategy based on the terms found in methodologically sound articles and derived from Haynes et al. [18]

II. Material and Methods

The study compared the retrieval performance of search strategies using a computerized search engine program with a manual review of each document in a collection of hypertext medical documents assembled from the WWW. Search strategies designed to retrieve high quality documents were created using terms that appeared in documents with rigorous study design. These search strategies were treated as diagnostic tests and were compared to a manual review of the documents which served as the "gold standard."

1. Assembling the Hypertext collection

A collection of hypertext medical pages was assembled from those available on the WWW. WWW sites were considered for inclusion if they had large number of medical pages within the site and were judged by the author to be likely to contain pages of high methodological quality. The sites included in the test collection were: the Health Services/Technology Assessment Text (HSTAT) Database (http://text.nlm.nih.gov) containing Agency for Health Care Policy Research Guidelines, National Institute of Health Consensus Statements, Guide to Clinical Preventative Services, AIDS Information Service documents, and Health Technology Reports; the American College of Physicians' ACP Online Annals of Internal Medicine (http://www.acponline.org/ journals/annaltoc.htm); the Cochrane Collaboration Systematic Reviews (no longer available on the WWW); and the Canadian Practice Guideline Infobase (http://www.cma.ca/ cpgs/index.html).

Using lwp-rget, a Perl module (http://www.linpro.no/lwp/), the WWW pages at each of these sites were downloaded onto a local computer. This program recursively follows the links from a WWW page and downloads all the linked pages. For each site the program was set to exclude pages outside the specified site and the depth of the recursive search was adjusted to download all the relevant pages at each site. The downloaded pages were then modified using a Perl script created by the author that eliminates all links connecting to pages external to hypertext collection (see Appendix 1). This was done to create a closed collection for manual review.

2. Manual review of the files

Each page in the hypertext collection was reviewed manually to determine whether it was part of a larger document. If a page had hypertext links to other pages at the same site that addressed the same topic, the linked pages were classified as a single document. Pages without any links to other documents and index pages were also classified as single documents.

Documents were further classified for format, focus and methodological rigor; and this information was stored in an Access (Microsoft Corporation, Redmond, WA) database created for this purpose. The format categories adapted from Haynes et al. and their corresponding definitions are shown in Table 4.[18] Documents that were identified as reviews were further classified for focus using categories adapted from Haynes et al.[18] Documents could have more than one focus and were assigned all that applied. The focus categories and their corresponding definitions are shown in Table 5.

Table 4: The format categories and their corresponding definitions

Format	Definition
Original Study	Any full-text article in which the investigators made firsthand observations
Review	A full-text article that was designated as a review, that had the word review in its title, or that indicated in its text that its purpose was to review or summarize the literature on a topic
General Article	A general discussion of a topic that did not include original investigations and was not intended to review a topic
Conference Report	An article labeled as such in the headline. This was not used if the article met the criteria for review or original study
Decision Analysis	An article labeled as such in the headline. This was not used if the article met the criteria for review or original study.
Index	A document containing an list of other articles with links to them
Guideline	A document that was designated as a guideline, that had the word guideline in its title, or that indicated its purpose was to provide recommendations for medical practice. This label was not used if the guideline met the criteria for review
Case Report	An original study that included fewer that 10 subjects
Non-English	An article that was written predominantly in a language other than English
Other	An article that did not meet any of the other definitions

Table 5: Focus categories and their corresponding definitions used to classify documents identified as reviews.

Focus	Definition
Etiology	Content pertaining to the causation of a disease
Treatment	Content pertaining to therapy or prevention of disease
Prognosis	Content pertaining to the clinical course of a disease. Articles that discuss clinical course in relation to a treatment were not classified as prognosis
Diagnosis	Content pertaining to the diagnosis of a disease process

Reviews were further classified for methodological rigor. The primary criterion for high quality was the inclusion of a clear description of the methods used to identify and select the studies used as the basis of the review. For example, a document was judged to be of high quality if the methods section contained a complete description of a MEDLINE search strategy and the inclusion criteria used to select studies.

Although the preferred method for establishing the reliability of classifying articles would be to measure inter-observer agreement, there were insufficient resources available to have a second person re-code a sample of the documents. However, some aspects of document classification, such as the determination of article quality and focus, might not be recalled easily by a person coding documents several months after the initial evaluation. Therefore, the test-retest reliability of the classification of article focus and of methodological rigor was assessed for a random sample of 10% of articles several months after the articles were first coded. With the exception of etiology, the kappa coefficient[21] was equal to or above 0.60 in all cases. (Methodological rigor: 0.75, Diagnosis: 1, Therapy: 0.62, Prognosis: 0.60) The kappa coefficient for etiology could not be calculated because the sample had no articles classified as having an etiology focus.

3. Indexing the collection

The hypertext document collection was indexed using the Simple Web Indexing System for Humans, Version 1.1.1 (SWISH) (Available at ftp://ftp.eit.com/pub/web/software/swish). Like other search engines, SWISH creates an index of words in the text from a collection of hypertext documents. It employs a stop list of several hundred common words that are not indexed and accepts user-defined parameters for words to ignore based on their frequency. In this study SWISH was configured to ignore words appearing in more than 400 pages or in more than 90% of the documents. Users of SWISH search for documents by entering query terms that can be combined using

Boolean operators. The results of the search are presented as a relevance-ranked list where documents with the greatest frequency of the search terms appear higher in the list (http://www.eit.com/goodies/software/swish/ or http://askdonna.ask.uni-karlsruhe.de/hppd/hpux/Text/swish-1.1/readme.html).

4. Search Strategies

Using SWISH, search strategies were developed for retrieving review documents in general and review documents of specific focus. The three search strategies for retrieving review documents tested were "review", "medline", and "medline or (search and strategy)." In addition, two search strategies were compared for each of the review focus areas. The first strategy combined the term "review" and the name of the focus. For example, the search strategy for review documents about treatment was "review and treatment." The second strategy combined "review" with terms adapted from the optimal search strategies given by Haynes et al. for MEDLINE documents.[18] The strategies tested were: for diagnosis "(sensitivity or specificity) and review", for therapy "(random or randomized) and review", for prognosis "[(natural and history) or predict or predicts or course] and review", and for etiology "[(odds and ratio) or (relative and risk) or (case and control)] and review".

5. Calculation of Search Strategy Performance Characteristics

If a search strategy located a page, the page was classified as to format, focus, and rigor according to the manual review of pages. If more than one page from a document was located by the search engine, only the first page was included in the analysis. These classifications of pages were then used to measure the sensitivity and specificity of a search strategy. For example, if search strategy x was tried to find high quality documents about therapy, the results of the search were classified using a two-by-two table shown in Figure 1. The sensitivity of the strategy was calculated from a/(a+c). The specificity of the search strategy was calculated from d/(b+d). The positive likelihood ratio was calculated using the formula:

$$\text{Positive likelihood ratio} = \frac{\text{sensitivity}}{(1 - \text{specificity})}.$$

The negative likelihood ratio was calculated using the formula:

$$\text{Negative likelihood ratio} = \frac{(1- \text{sensitivity})}{\text{specificity}} .$$

In order to understand the reason why high quality documents were not retrieved by search strategies; two high quality, non-retrieved documents were chosen at random for each search strategy for article focus. Potential reasons for non-retrieval evaluated were 1) the required combination of search strategy words was not present in the document and 2) the required combination of search strategy words was present but was not detected by SWISH. To ascertain whether the required combination of search

Figure 1: Sensitivity and specificity of search strategy for therapy using SWISH

	High quality review on therapy according to manual check	Not high quality review on therapy according to manual check
Review found by search strategy using SWISH	a	b
Review not found by search strategy using SWISH	c	d

strategy words was present, the selected documents were searched using the "find in page" command of the Netscape Communicator Browser (Netscape Communications Corp., Mountain View, CA). In addition, for the reviews not found by the search strategies adapted from Haynes et al., other candidate search terms from the appendix of Haynes et al. were tested using the "find in page" command.[18]

6. Statistical Methods

Search strategies were compared using the chi-square test without Yates' correction. Test-retest reliability for binary variables was calculated using the kappa coefficient.[20]

III. Results

A total of 2565 hypertext pages were reviewed. These consisted of 1058 separate documents. Of these, 209 (20%) were review documents, 162 (15%) guidelines, 233 (22%) general articles, 4 (0.4%) decision analyses, 130 (12%) indices, 131 (12%) non-English documents, 75 (7%) conference reports, 76 (7%) original studies, and 38 (3.6%) other (Figure 2).

There were 171 (16%) review documents classified as having high methodological rigor. The breakdown by focus category and methodological rigor is shown in Table 6 . Several strategies for identifying high quality review articles are compared in Table 7 . While "medline" had fairly high specificity, its sensitivity was quite low. The term "review" had fairly high sensitivity for high quality reviews and moderately high specificity.

Figure 2: Formats of articles in the hypertext collection

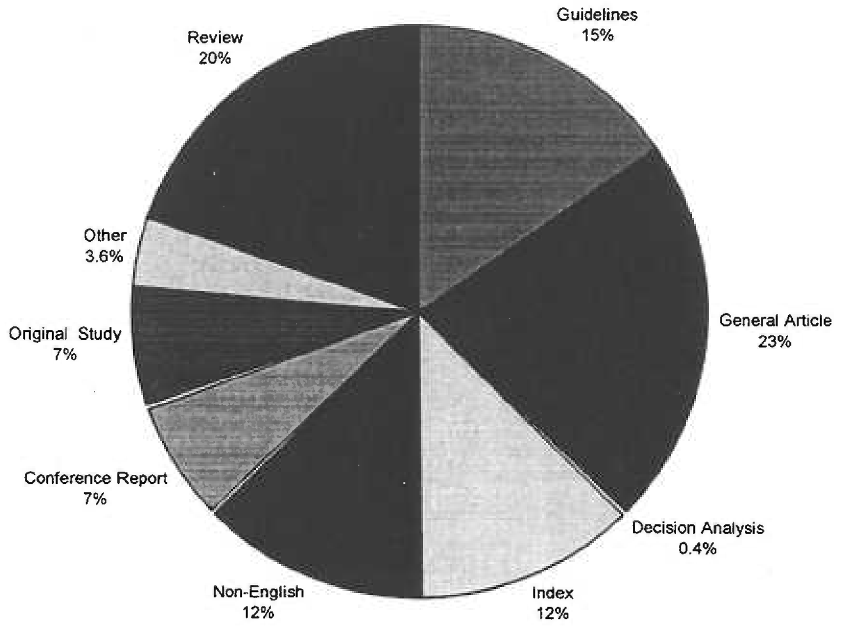


Table 6: Number of review articles in each focus category and those meeting criteria for high methodological rigor

Focus Category	Number of Review Articles for Each Focus Category and those Meeting Criteria for High Methodological Rigor *
Etiology	30
high methodological rigor (%)	24 (80)
Prognosis	66
high methodological rigor (%)	61 (92)
Diagnosis	114
high methodological rigor (%)	97 (85)
Therapy	174
high methodological rigor (%)	142 (82)

*The sum is greater than 209 since review articles could be categorized in more than one focus.

Table 7: Sensitivity, specificity, and likelihood ratios of search strategies for high quality review articles on the WWW

	Sensitivity	Specificity	Positive Likelihood Ratio	Negative Likelihood Ratio
medline	37%	95%	7.4	0.14
medline or (search and strategy)	45%	94%	7.5	0.13
review	87%	70%	2.9	0.34

For each focus category, two strategies were compared for sensitivity and specificity (Table 8). For diagnosis, the strategy adapted from Haynes et al. was superior since it was no worse in sensitivity and had a statistically significant higher specificity. For therapy, the strategy adapted from Haynes et al. was superior in both sensitivity and specificity. Prognosis and etiology demonstrated a tradeoff between more specific simple strategies and more sensitive strategies adapted from Haynes et al. Figure 3 is an example of a search using the strategy "review and therapy." Figures 4 and 5 provide examples of documents that were and were not found in the collection using this search strategy.

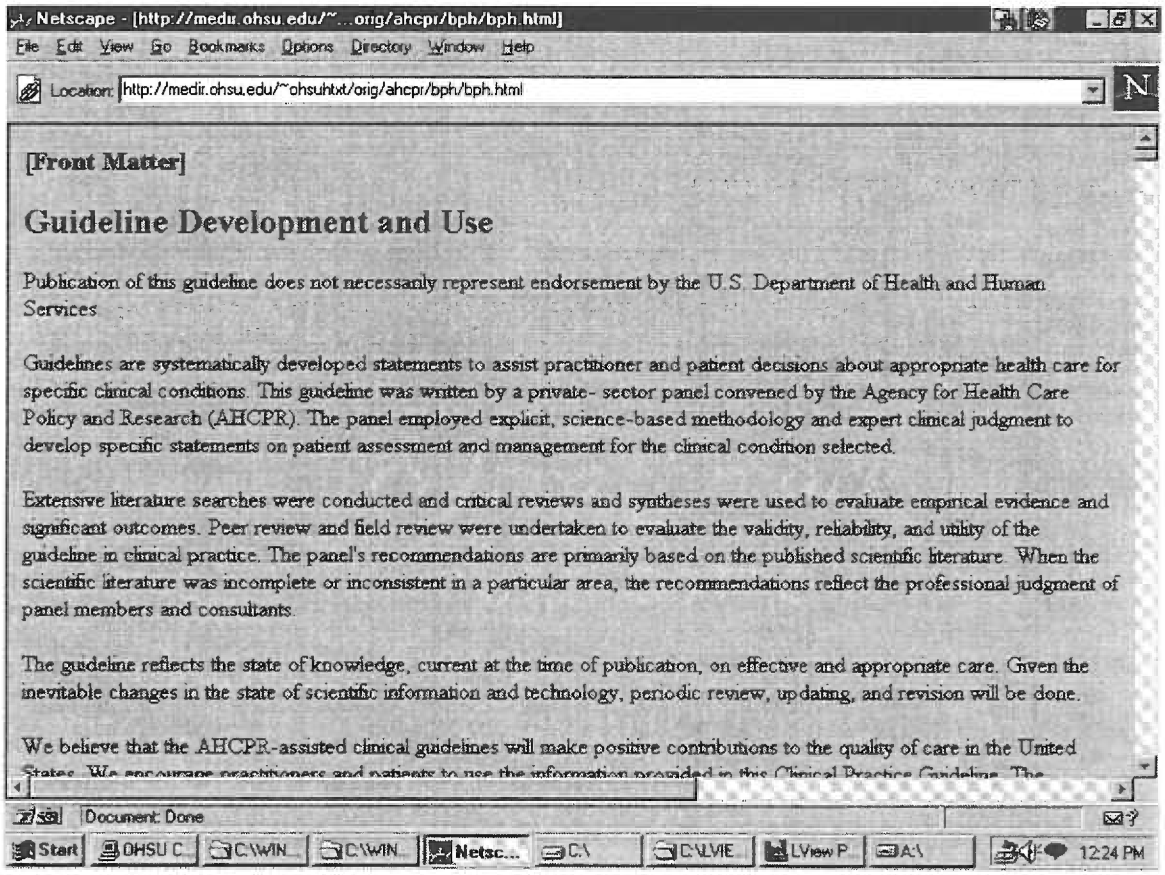
Figure 3: Example of SWISH search for review documents on treatment using the strategy “review and therapy” Truncated at one page.

```
medir% swish -f my_index.swish -w review and therapy

# SWISH format 1.1
search words: review and therapy

1000 /pub9/home/ohsuhtxt/indexed/ahcpr/deprestreat/dep_treat.html "dep_treat.ht7
935 /pub9/home/ohsuhtxt/indexed/ahcpr/deprestreat/dep2ctxt.html "Clinical Guide5
921 /pub9/home/ohsuhtxt/indexed/ahcpr/bph/bph.html "bph.html" 524078
899 /pub9/home/ohsuhtxt/indexed/ahcpr/cancer_pain/cancer_pain.html "cancer_pain3
894 /pub9/home/ohsuhtxt/indexed/ahcpr/otitis/otitis.html "otitis.html" 301487
894 /pub9/home/ohsuhtxt/indexed/ahcpr/stroke_rehab/stroke_rehab.html "stroke_re0
877 /pub9/home/ohsuhtxt/indexed/ahcpr/incont/incont.html "incont.html" 476448
877 /pub9/home/ohsuhtxt/indexed/ahcpr/incont/tempDl34296001.html "tempDl34296001
866 /pub9/home/ohsuhtxt/indexed/ahcpr/low_back/low_back.html "low_back.html" 443
865 /pub9/home/ohsuhtxt/indexed/ahcpr/pain/acute_pain.html "acute_pain.html" 353
862 /pub9/home/ohsuhtxt/indexed/ahcpr/angina/angina.html "angina.html" 383106
846 /pub9/home/ohsuhtxt/indexed/ahcpr/cardia_rehab/cardiac_rehab.html "cardiac_5
844 /pub9/home/ohsuhtxt/indexed/ahcpr/pres_ulcer/pres_ulcer.html "pres_ulcer.ht9
826 /pub9/home/ohsuhtxt/indexed/ahcpr/cataract/tempDl40580.html "tempDl40580.ht1
822 /pub9/home/ohsuhtxt/indexed/ahcpr/cataract/catctxt.html "Clinical Guideline1
806 /pub9/home/ohsuhtxt/indexed/ahcpr/smoking/smoking.html "smoking.html" 327601
785 /pub9/home/ohsuhtxt/indexed/ahcpr/depresdect/depress_detect.html "depress_d1
774 /pub9/home/ohsuhtxt/indexed/ahcpr/heart_failure/failure.html "failure.html"1
774 /pub9/home/ohsuhtxt/indexed/ahcpr/heart_failure/lvdctxt.html "Clinical Prac0
768 /pub9/home/ohsuhtxt/indexed/annals/bestevid.html "Systematic Reviews: Synth7
765 /pub9/home/ohsuhtxt/indexed/ahcpr/ulcer/ulcer.html "ulcer.html" 169288
763 /pub9/home/ohsuhtxt/indexed/annals/systemat.html "Locating and Appraising S3
741 /pub9/home/ohsuhtxt/indexed/hta_htr/91-2.html "91-2.html" 92694
735 /pub9/home/ohsuhtxt/indexed/annals/seekaltr.html "Advising Patients Who See1
723 /pub9/home/ohsuhtxt/indexed/cochrane/text07.htm "Cochrane Reviews" 21606
716 /pub9/home/ohsuhtxt/indexed/hta_htr/94-4.html "94-4.html" 281724
716 /pub9/home/ohsuhtxt/indexed/hta_htr/95-4.html "95-4.html" 281999
714 /pub9/home/ohsuhtxt/indexed/tac/relax.html "relax.html" 140455
695 /pub9/home/ohsuhtxt/indexed/hta_htr/90-1.html "90-1.html" 210708
692 /pub9/home/ohsuhtxt/indexed/cpginfo/0025.html "Guidelines for the emergency1
688 /pub9/home/ohsuhtxt/indexed/annals/corangio.html "Coronary Angiography and 3
685 /pub9/home/ohsuhtxt/indexed/ahcpr/deprestreat/dep2crtxt.html "dep2crtxt.htm3
684 /pub9/home/ohsuhtxt/indexed/annals/dyspepsi.html "Management Strategies for4
```

Figure 4: Example of a document that was found by the search strategy “review and therapy”


Netscape - [http://medir.ohsu.edu/~...orig/ahcpr/bph/bph.html]
File Edit View Go Bookmarks Options Directory Window Help
Location: http://medir.ohsu.edu/~ohsuhtxt/orig/ahcpr/bph/bph.html
N
Document: Done
Start
OHSU C...
C:\WIN...
C:\WIN...
Netsc...
C:\
C:\LVIE...
LView P...
A:\
12:24 PM

[Front Matter]

Guideline Development and Use

Publication of this guideline does not necessarily represent endorsement by the U.S. Department of Health and Human Services.

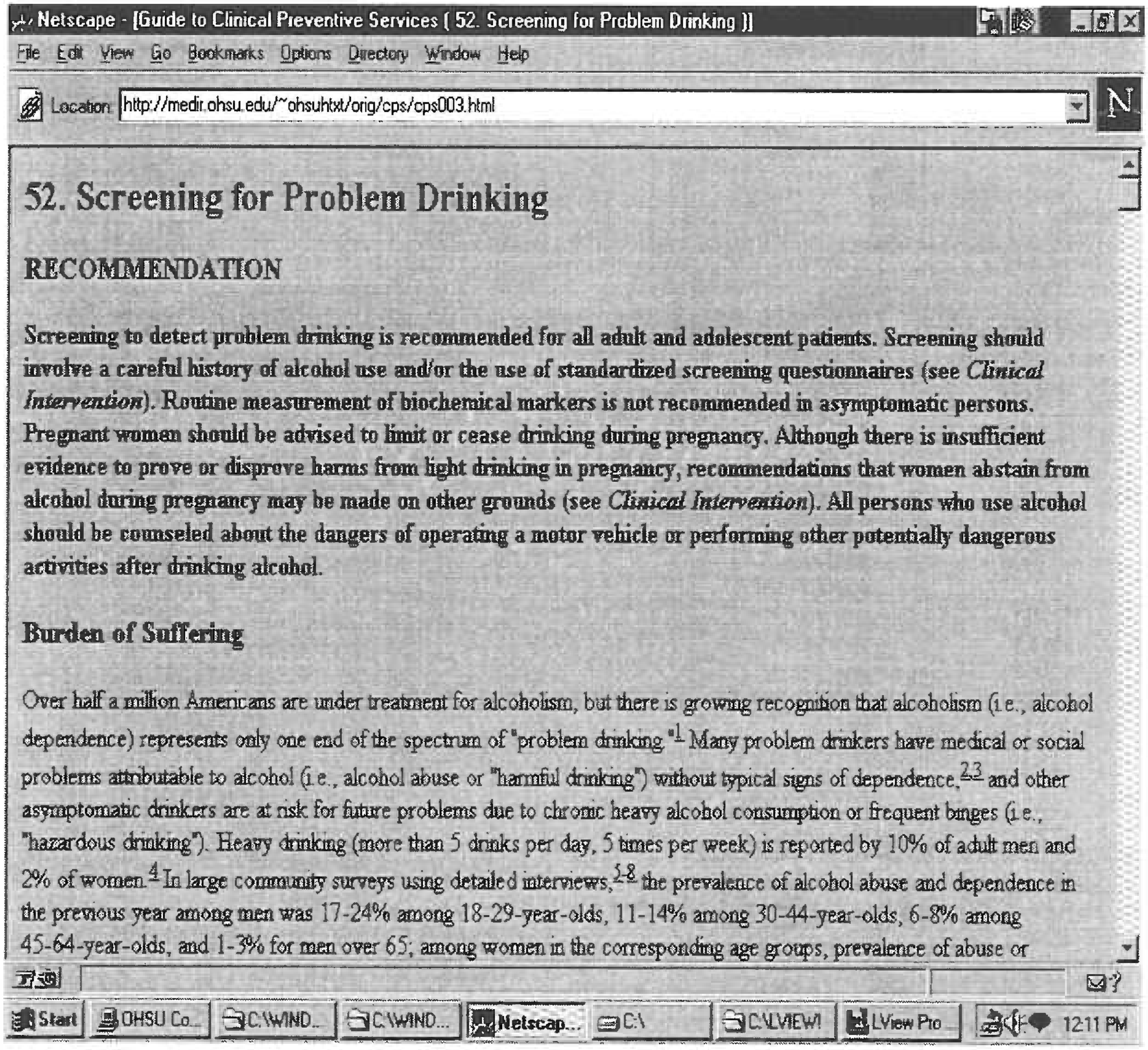
Guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical conditions. This guideline was written by a private- sector panel convened by the Agency for Health Care Policy and Research (AHCPR). The panel employed explicit, science-based methodology and expert clinical judgment to develop specific statements on patient assessment and management for the clinical condition selected.

Extensive literature searches were conducted and critical reviews and syntheses were used to evaluate empirical evidence and significant outcomes. Peer review and field review were undertaken to evaluate the validity, reliability, and utility of the guideline in clinical practice. The panel's recommendations are primarily based on the published scientific literature. When the scientific literature was incomplete or inconsistent in a particular area, the recommendations reflect the professional judgment of panel members and consultants.

The guideline reflects the state of knowledge, current at the time of publication, on effective and appropriate care. Given the inevitable changes in the state of scientific information and technology, periodic review, updating, and revision will be done.

We believe that the AHCPR-assisted clinical guidelines will make positive contributions to the quality of care in the United States. We encourage practitioners and patients to use the information provided in this Clinical Practice Guideline. The

Figure 5: Example of a document that was not found by the search strategy “review and therapy”.


Netscape - [Guide to Clinical Preventive Services (52. Screening for Problem Drinking)]
File Edit View Go Bookmarks Options Directory Window Help
Location: http://medir.ohsu.edu/~ohsuhtxt/orig/cps/cps003.html
N
Start
OHSU Co...
C:\WIND...
C:\WIND...
Netscap...
C:\
C:\LVIEW\
LView Pro
12:11 PM

52. Screening for Problem Drinking

RECOMMENDATION

Screening to detect problem drinking is recommended for all adult and adolescent patients. Screening should involve a careful history of alcohol use and/or the use of standardized screening questionnaires (see *Clinical Intervention*). Routine measurement of biochemical markers is not recommended in asymptomatic persons. Pregnant women should be advised to limit or cease drinking during pregnancy. Although there is insufficient evidence to prove or disprove harms from light drinking in pregnancy, recommendations that women abstain from alcohol during pregnancy may be made on other grounds (see *Clinical Intervention*). All persons who use alcohol should be counseled about the dangers of operating a motor vehicle or performing other potentially dangerous activities after drinking alcohol.

Burden of Suffering

Over half a million Americans are under treatment for alcoholism, but there is growing recognition that alcoholism (i.e., alcohol dependence) represents only one end of the spectrum of "problem drinking."[1] Many problem drinkers have medical or social problems attributable to alcohol (i.e., alcohol abuse or "harmful drinking") without typical signs of dependence,[2,3] and other asymptomatic drinkers are at risk for future problems due to chronic heavy alcohol consumption or frequent binges (i.e., "hazardous drinking"). Heavy drinking (more than 5 drinks per day, 5 times per week) is reported by 10% of adult men and 2% of women.[4] In large community surveys using detailed interviews,[5-8] the prevalence of alcohol abuse and dependence in the previous year among men was 17-24% among 18-29-year-olds, 11-14% among 30-44-year-olds, 6-8% among 45-64-year-olds, and 1-3% for men over 65; among women in the corresponding age groups, prevalence of abuse or improved by adding additional terms. The effect of these additions on specificity, however, remains uncertain.

Discussion

This study involved the creation of a collection of hypertext documents from American and Canadian WWW sites judged likely to contain high quality reviews. Despite the selection criteria, only 16% of the documents in the collection met the standards for high quality reviews. While original articles were not evaluated for quality in this study, they made up only 7% of the collection. Therefore, even assuming that all the original studies were of high quality, less than a quarter of what is currently available on the WWW at selected sites meets even minimal standards for scientific papers of high methodological quality. This suggests that the WWW needs to evolve considerably before clinicians can rely upon it to supply high quality information.

The majority of the hypertext documents in the collection (65%) were reviews, guidelines, general articles, and conference reports. This implies that high quality medical WWW sites differ in composition from the printed medical literature where 20% or more of the articles are original research.[22] Thus, research on improving information retrieval on the WWW is more likely to be useful if it focuses on summary articles rather than original publications.

This study also showed that more complex search strategies based on methodological terms could be identified that had higher sensitivity than more simple search strategies. The main reason for the success of the more complex strategies is that they contain words that are found in high quality studies but not in other studies. These words may appear in the description of the study design. For example, for high quality reviews about therapy, the terms "random" and "randomized" may appear where the paper describes the inclusion criteria for studies. These words may also appear in the references to the papers used for the study or in other places. In these cases, the success of the strategy may be due to the high correlation between the description of the review methods and the appearance of specific terms in the paper's references or elsewhere.

The findings of the present study are in some ways similar to those of Haynes et al. for a collection of MEDLINE documents.[18] That study identified a set of search strategies for original articles based on the article focus. Unlike the current study, however, Haynes et al. did not apply these strategies to WWW documents. This would prove difficult since many of the MEDLINE strategies rely on database fields and MeSH terminology that are not available for WWW documents. In addition, Haynes et al. did not attempt to identify strategies for review articles of different foci. The later goal is particularly important on the WWW since many of the medical documents are not original literature, but tertiary literature such as review documents or guidelines.

The search strategies presented in this study could prove useful to clinicians who are looking for medical review documents on the WWW that meet at least one major criteria for methodological quality, while filtering out documents with weaker designs. The more complex strategies based on the terms found in high quality articles were superior for retrieving reviews that focus on diagnosis and treatment. By contrast, the complex strategies for prognosis and etiology were more sensitive, but less specific, than the corresponding simple strategies. Moreover, the overall positive likelihood ratios were lower for the more complex strategies for reviews on etiology and prognosis. This suggests a need for further research to identify better strategies for these foci.

If the suggested search strategies are used by clinicians, they would need to evaluate the retrieved studies to identify those most suited to the clinical situation. In addition, like any diagnostic test, these strategies will produce both false positive and false negative results. Thus, searchers cannot be guaranteed that all documents retrieved by a search strategy will meet the methodological criteria for high quality reviews or that additional studies that failed to be retrieved by the search strategy do not exist.

The success of these search strategies is also highly dependent on the pre-test probability or prevalence of high quality material. The 16% prevalence of high quality reviews in this test collection is probably higher than the percentage of high quality documents of the WWW as a whole. The effect of the prevalence on finding high quality reviews on diagnosis using the search strategy "(sensitivity or specificity) and review" is shown in Table 9. When the pre-test probability that a document is a high-quality review on diagnosis is 9%, as it is in the hypertext collection, there is clear benefit to applying the search strategy. Using Bayes' Theorem, the post-test probability that a document is high quality is 48% when the pre-test probability is 9%, the sensitivity of the test is 83%, and the specificity of the test is 91%. However, as the pre-test probability decreases, the post-test probability also drops. This means that if the prevalence of high quality reviews on the WWW is low, many of the articles retrieved by these search strategies will not be of high quality. One implication of this finding is that these search strategies might prove most useful to clinicians when used in conjunction with other methods of identifying high quality reviews, such as document selection or classification.

Table 9: Pre-test and post-test probability of high quality reviews on diagnosis using the search strategy "(sensitivity or specificity) and review"

Pre-test probability (prevalence)	Post-test probability
9%	48%
0.9%	7.8%
0.09%	0.8%
0.009%	0.08%

Compared to other methods of improving retrieval from hypertext, the search strategies presented here have an important advantage: their implementation does not grow more difficult as the size of the collection being searched grows. Despite this advantage, however, the usefulness of these search strategies for clinicians is dependent on the prevalence of high quality material on the WWW. If high quality material is infrequent, these search strategies could be combined with other methods of improving WWW searching. For example, if a system of document classification by topic is adopted for WWW medical documents,[23] use of the search strategies described may refine searches by selecting high quality documents from a list of documents retrieved for a given topic area.

1. Limitations

This study has a number of important design limitations. First, the data are derived from a test collection of hypertext documents and may not be generalizable to the WWW as a whole. WWW documents not in the collection may differ in their use of words, and therefore may yield different performances with the search strategies described.

A second limitation related to the use of a test collection is the relatively small number of documents included. This may limit the power to detect differences between search strategies. This limitation is particularly important for search strategies used to detect review articles about diagnosis. Although there was no statistically significant difference between the sensitivity of the two strategies tested, the power to find this difference was only 33%. Thus, there may be a tradeoff between the higher sensitivity of the simple strategy and the higher specificity of the strategy based on methodological terms.

A third limitation of this study results from the use of the SWISH indexing/retrieval engine. Although the general functions of this engine are similar to those of commercial engines such as Altavista or Excite, the specific algorithms may differ enough to cause the search strategies outlined here to perform differently than with SWISH.

A fourth limitation is the non-systematic methods used to identify search strategies. The study used strategies adapted from Haynes et al., which had fairly high sensitivity and specificity. However, there may exist other search strategies that are more optimal than those identified in this study. Only a systematic assessment of search terms would identify definitively the optimal strategies.

A fifth limitation of this study is the use of a single evaluator to classify the documents by quality and emphasis. Although the quality rating had fairly high test-retest reliability, it is unclear to what extent these evaluations are reproducible if other raters were used.

A sixth limitation of this study is that it did not attempt to evaluate the performance of these search strategies in conjunction with topic terms such as the name of diseases. Further research is needed to investigate whether these search strategies are useful when applied in answering specific questions.

2. Future Directions

The limitations of this study suggest several future directions for research. First, the ability to classify consistently documents by type, focus, and quality could be evaluated by investigating inter-rater reliability. This measure would have less potential for bias than the test-retest reliability reported in this study because the raters would make independent judgments about the documents.

If documents could be classified with high reliability, a second area for further investigation would be to conduct a more systematic search for search strategies of high sensitivity and specificity. One potential method for identifying other additional search strategies is reported by Haynes et al.[18] In this study a list of candidate search terms was assembled by surveying expert searchers for terms used frequently in identifying articles of particular focus. All combinations of these terms were then tested to identify the best strategies. A similar systematic approach to identifying strategies could be done for WWW documents. This would give greater credibility to the recommended strategies.

A third area of potential research involves establishing the generalizability of the findings. Search strategies should be tested not only within a limited test collection with a particular search engine, but also on the WWW as a whole and within the subsets of medical documents identified by sites such as Medical Matrix or CliniWeb. In addition, these search strategies should be tested using a variety of search engines such as Altavista, Excite, etc. One potential method for evaluating these search strategies on the WWW would be to give subjects assigned search topics. One group of searchers would use the suggested search strategies, while the other group would search using strategies of their own devising. Clinicians could then compare the search results by evaluating the quality of materials retrieved by the two different strategies.

V. Summary and Conclusion

In summary, this study showed that search terms could be identified that selected with high sensitivity and specificity high quality review documents from a collection of hypertext documents . This study further demonstrated that a set of search strategies based on methodological terms performed better than simple strategies based on the terms therapy, etiology, prognosis, and diagnosis. In the case of the search strategies for diagnosis and therapy, those strategies based on methodological terms were equal or superior to the more simple strategies in sensitivity and were also more specific. In the case of etiology and prognosis, strategies based on methodological terms were more sensitive but less specific, than simple search strategies.

The search strategies presented in this study could prove useful to clinicians who are looking for high quality medical review documents on the WWW. However, the usefulness of these strategies depends on the prevalence of high quality documents on the WWW. For this reason these search strategies could prove most useful when combined with other methods of improving retrieval from the WWW such as indexing content using keywords.

Further research is needed to identify systematically the optimal search strategies for retrieving high quality WWW documents, to study the effect of combining these strategies with topic terms, and to determine whether these strategies will prove useful when deployed on the WWW as a whole.

References

1. Lynch C. Searching the Internet. Sci American 1997; 276:53-56.
2. Lowe H, Lomax E, Polonkey S. The world wide web: a review of an emerging Internet-based technology for the distribution of biomedical information. J Am Med Informatics Assoc 1996; 3:1-14.
3. Hersh W. Information retrieval: a health care perspective. In: Orthner H, ed. Computers and medicine. New York: Springer-Verlag New York, Inc., 1996:320.
4. Oxman A, Sackett D, Guyatt G. Users' guides to the medical literature: how to get started. JAMA 1993; 270:2093-95.
5. Guyatt G, Sackett D, Cook D. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993; 270:2596-2601.
6. Jaeschke R, Guyatt G, Sackett D. Users' guide to the medical literature. III How to use an article about diagnostic test. A. Are the results of the study valid? JAMA 1994; 271:389-391.
7. Laupacis A, Wells G, Richardson S, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. JAMA 1994; 272:234-237.
8. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature. IV. How to use an article about harm. JAMA 1994; 271:1615-1619.
9. Oxman A, Cook D, Guyatt G. Users' guides to the medical literature. VI. How to use an overview. JAMA 1994; 1994:1367-71.

19. Hersh W, Brown K, Donohoe L, Campbell E, Horacek A. CliniWeb: managing clinical information on the World Wide Web. J Am Med Informatics Assoc 1996; 3:273-280.

20. Fowler J, Maram S, Kouramajian V, Devadhar V. Automated MeSH indexing of the World-Wide Web, Nineteenth Annual Symposium on Computers in Medical Care, New Orleans, Lousiana. Hanley and Belfus, Inc. 1995: 893-897.

21. Kelsey J, Thompson W, Evans A. Methods in observational epidemiology. In: Lilienfeld A, ed. Monographs in epidemiology and biostatistics. Vol. 10. New York: Oxford University Press, 1986:366.

22. Singer A, Homan C, Stark M, Werblud M, HC Thode J, Hollander J. Comparison of types of research articles published in emergency and non-emergency medicine journals. Academic Emergency Medicine 1997; 4:1153-58.

23. Appleyard R, Malet G. A proposal for using metadata encoding techniques for health care information indexing on the WWW [abstract]. Proceedings of the 1997 AMIA annual fall symposium. Nashville, TN. Hanley and Belfus, Inc. 1997:905.

Appendix

```
#       This Perl script removes absolute links from html pages.    To use it,
#       you should place to script in the directory where you want to remove
#       the absolute links.  Make sure the permissions on the script are set to +rwx.

print "\nType 1 to change .html to .htm type. ";
$answer = <STDIN>;
print "\nWhat do you want to add at the bottom of the page?\n";
$bottom = <STDIN>;

opendir(IN_DIR,'.');    # opens the current directory
@all_files =readdir(IN_DIR);   # reads in the names of all files in directory
print @all_files;
$index= 0;
foreac$index(@all_files)      # looks at each file name in turn
{

    if"$index") or (-x "$index")) {print "\n$index is not a textfile";}
    el              # ignores subdirectories and executables
    {

     en(IN_FILE, $index) or print "\ncan't open $index";
     ll_words = <IN_FILE>;          # read in words of file
     each $index2(@all_words)

     if($answer ==1)
     {

          if($index2 =~m/\.html/i)

                              # change all  .html
       {                      # to .htm

               ($begin,$end) = split(/.html/,$index2);
               $index2 = $begin.".htm".$end;
     }


 ndex2 =~m/\.org\//i)      # check  if any .org
                   # /i ignores case

$index2 =~tr/A-Z/a-z/;  # make all lowercase
print "\nindex2: $index2";
```

```
        ($junk, $good) = split(/.org\//, $index2);
                              # save after .org
        ($rescue, $junk)= split(/<a href/, $junk);
                              # recover before <a h
        $good = $rescue."<a href=\"../".$good;
        print"\nfinal good: $good";
        $index2 = $good;
}
if($index2 =~m/\.edu\//i)             # check for .ed/
{
        $index2 =~tr/A-Z/a-z/;
#        print "\nindex2: $index2";
        ($junk, $good) = split(/.edu\//, $index2);
        $good = $rescue."<a href=\"../".$good;
#        print"\nfinal good: $good";
        $index2 = $good;
}
if($index2 =~m/\.gov\//i)             # check for .gov/
{
        $index2 =~tr/A-Z/a-z/;
#        print "\nindex2: $index2";
        ($junk, $good) = split(/.gov\//, $index2);
        $good = $rescue."<a href=\"../".$good;
#        print"\nfinal good: $good";
        $index2 = $good;
}
if($index2 =~m/\.com\//)
{
        $index2 =~tr/A-Z/a-z/;
#        print "\nindex2: $index2";
        ($junk, $good) = split(/.com\//, $index2);
        $good = $rescue."<a href=\"../".$good;
#        print"\nfinal good: $good";
        $index2 = $good;
}
if($index2 =~m/\.ca\//i)              # check for .ca/
{
        $index2 =~tr/A-Z/a-z/;
#        print "\nindex2: $index2";
        ($junk, $good) = split(/.ca\//, $index2);
        $good = $rescue."<a href=\"../".$good;
#        print"\nfinal good: $good";
        $index2 = $good;
}
if($index2 =~m/\.uk/i)          #check for .uk
```

```
            {
                $index2 =~tr/A-Z/a-z/;
            #    print "\nindex2: $index2";
                ($junk, $good) = split(/.uk/, $index2);
                $good = $rescue."<a href=\"..".$good;
            #    print"\nfinal good: $good";
                $index2 = $good;
            }
            if($index2 =~m/\.fr/i)          #check for .fr
            {
                $index2 =~tr/A-Z/a-z/;
            #    print "\nindex2: $index2";
                ($junk, $good) = split(/.fr/, $index2);
                $good = $rescue."<a href=\"..".$good;
            #    print"\nfinal good: $good";
                $index2 = $good;
            }
            if($index2 =~m/<\/html>/i)
            {
                $index2 = "<Center>$bottom</Center>".$index2;
            }
        }           #end foreach
        close (IN_FILE);
        open(OUT_FILE, ">$index") or die "\ncan't write to $index";
        foreach $index3 (@all_words)          # write out new file
        {
            print OUT_FILE $index3;

        }

    }   # end of else statement
}       # end of foreach
```