Application of Technology ■

CliniWeb: Managing Clinical Information on the World Wide Web

WILLIAM R. HERSH, MD, KEVIN E. BROWN, BS, LARRY C. DONOHOE, MLIS, EMILY M. CAMPBELL, MS, ASHLEY E. HORACEK, MD

Abstract The World Wide Web is a powerful new way to deliver on-line clinical information, but several problems limit its value to health care professionals: content is highly distributed and difficult to find, clinical information is not separated from non-clinical information, and the current Web technology is unable to support some advanced retrieval capabilities. A system called CliniWeb has been developed to address these problems. CliniWeb is an index to clinical information on the World Wide Web, providing a browsing and searching interface to clinical content at the level of the health care student or provider. Its database contains a list of clinical information resources on the Web that are indexed by terms from the Medical Subject Headings disease tree and retrieved with the assistance of SAPHIRE. Limitations of the processes used to build the database are discussed, together with directions for future research.

■ **JAMIA.** 1996;3:273–280.

The goal of this paper is to describe an approach to partitioning and managing clinical information on the World Wide Web (WWW).[1] Although the WWW is a remarkable technological and cooperative human endeavor, it suffers from significant limitations as a quick-access clinical information resource. For example, the distributed nature of its content makes it difficult to find specific information unless one knows where to look. Furthermore, it is difficult to separate clinical from non-clinical information as well as practitioner-oriented from consumer-oriented information. This paper describes CliniWeb, a searchable database of clinical, practitioner-oriented information on the WWW. It presents the processes through which sites were identified for inclusion in the database and an index was created. The limitations of these approaches and directions for future research are also discussed.

Affiliation of the authors: Biomedical Information Communication Center, Oregon Health Sciences University, Portland, OR.

Initial funding for CliniWeb has been provided in part by Grant LM05307 from the National Library of Medicine and Grant DE-FG06-94ER61918 from the Department of Energy.

Correspondence and reprints: William Hersh, MD, Oregon Health Sciences University BICC, 3181 SW Sam Jackson Park Road, Portland, OR 97201. e-mail: hersh@ohsu.edu

Received for publication: 12/19/95; accepted for publication: 3/5/96.

Background

Although there is great enthusiasm for the WWW as a front-end to networked clinical information,[2,3] many of its advantages are also its limitations. For example, the WWW is highly distributed and lacks an overall index or table of contents for all its information. This is in stark contrast to the traditional information retrieval (IR) world, where providers such as the National Library of Medicine (NLM), other on-line vendors, and a host of CD-ROM publishers supply highly organized access to their information.

Another problem with the chaotic growth of the WWW is the lack of separation between clinical and non-clinical information and between practitioner-oriented and consumer-oriented information. This can be problematic in a domain like health care, where the source and quality of information are important. Whereas traditional database vendors focus on providing high-quality commercial information, the WWW allows anyone to post anything. While this may be an advantage in political and other spheres,

it can be a disadvantage in health care because practitioners and educators base decisions and education, respectively, on the highest-quality scientific information. Health care professionals are likely to prefer information associated with a respected institution, such as a health sciences university or health-related government agency.

Some isolated collections of information on the WWW have been organized into discrete databases for searching, but the distributed and constantly changing nature of the WWW makes such organization impossible on a large-scale basis. The most comprehensive approach to indexing on the WWW has been the advent of so-called Web walkers, which "walk" the WWW by finding a Uniform Resource Locator (URL), indexing the individual words there, and following its links to find additional URLs.[4] Among the best known are Lycos (http://www.lycos.com/), Alta Vista (http://altavista.digital.com), WebCrawler (http://www.webcrawler.com/), and Infoseek (http://www.infoseek.com/).

Other services have attempted to provide controlled vocabulary indexing to the WWW. The pioneer in this approach is Yahoo (http://www.yahoo.com/), which manually assigns subjects to WWW pages or collections. Of course, the problem with this approach is its labor intensiveness and the inconsistency of human indexing.[5] In addition, these systems do not filter information appropriate for the level of health care providers and consumers.

All of these indexing approaches advance the utility of the WWW, but their indexing and retrieval functions are primitive compared to the IR capabilities in both the commercial and research sides of the traditional IR world. Complex full-screen interfaces like Grateful Med[6] and Knowledge Finder[7] assist users in searching conventional databases and provide the best possible access to both manually indexed Medical Subject Headings (MeSH) terms and text words. None of WWW search engines contains the advanced features seen in many commercial IR systems, such as proximity operators and thesaurus look-up capability. In the IR research sphere, breakthroughs with word-based automated systems[8] as well as those employing linguistic techniques[9,10] have been shown to improve aspects of retrieval performance, but they require complex interfaces that are not easily adapted to the Web.

The diffuse nature of WWW information also makes innovative projects like the NLM's Information Sources Map (ISM) difficult to apply.[11] The ISM provides meta-information about databases that allow users and systems to determine their content. On the WWW, however, boundaries are difficult to draw around databases, making meta-information difficult to utilize.

In spite of these limitations, there is much valuable clinical information on the WWW. It is often buried pages deep at a medical school or health-related government agency site and is difficult to find. It may be intermixed with consumer or organizational information that is distracting to a clinician or educator hoping to find more about clinical topics. A number of sites have attempted to organize clinical information on the Web by topic, such as Medical Matrix (http://www.kumc.edu/mmatrix/), MedWeb (http://www.cc.emory.edu/WHSCL/medweb.html), and GalaxyNet Medicine (http://galaxy.einet.net/galaxy.html). However, these sites are indexed into very broad categories, such as "Surgery" or "Heart Disease," and they give only a general location, rather than individual WWW pages.

Design Criteria of CliniWeb

The major goal of CliniWeb is to organize high-quality clinical resources on the WWW for health care educators, practitioners, and researchers. CliniWeb provides rapid access to diverse information, organized by specific topic, on the WWW. Unlike the Web walkers, it does not index everything in its path. Rather, CliniWeb explicitly omits non-clinical and consumer-oriented information, which we define as that which would not be of clinical value to health care students or practitioners.

Another goal of CliniWeb is to serve as a testbed for research into defining the optimal methods to build and evaluate a clinically oriented WWW resource. As noted above, although the WWW has attained great stature, some of its IR capabilities are lacking. CliniWeb should provide the foundation for a better understanding of the optimal approaches to organizing, indexing, and retrieving WWW-based information.

CliniWeb is available from the Oregon Health Sciences University (OHSU) home page (http://www.ohsu.edu/cliniweb/) and consists of the following components:

1. A database of clinically-oriented URLs
2. Indexing of URLs with terms from the MeSH vocabulary disease tree
3. An interface for accessing URLs by browsing and searching

CliniWeb Database

Because the presence of clinical information for health care education or practice is the single criterion for including WWW pages in the CliniWeb database, pages not dealing with clinical topics are deliberately not added. Although consumer or organizational information on the WWW is important, it can distract someone seeking information for education or patient care.

For the initial version, several months were spent searching the WWW for URLs that contained clinical information. We found that the richest sites for these pages were medical schools and health-related government agencies. Many of these sites also contained pointers to other locations of biomedical information. A total of 4,225 clinically pertinent pages were discovered and entered into a relational database containing the following information (Fig. 1):

1. Title—The name of the WWW page
2. Location—The institution or organization that created the page
3. URL—The URL to the page
4. Indexing term(s)—for MeSH terms to be added in the indexing process.

A major challenge for CliniWeb is the choice of granularity for Web pages. Because our goal was to index CliniWeb pages on a topical basis, we had to avoid the extremes of either assigning broad terms to large groups of pages or else not breaking down coherent topic groups of pages. We generally aimed to index pages at the granularity of representing topics in the MeSH disease tree. For example, we would break down an entire curriculum on neurology or cardiovascular disease, but we would keep intact a series of pages on dementia or hypertension.

Another significant challenge for CliniWeb is the maintenance of the database. Not only do new sites become available continuously, but existing sites are often modified. Even if the content is not changed, the directory structure might be. Since URLs encode directory pathways explicitly, a reorganization of a server can render a database of URLs to that server invalid. This has actually occurred with several sites, requiring updating of the database. Fortunately, most updates just require a pathway change, which can be done rapidly in the database for all URLs that share the same path. Some URLs have been removed altogether; these are detected by a utility and flagged as "deleted" in the database. The records are maintained in the database in case the URL reappears, but is not displayed on the CliniWeb pages.

Title: aortic regurgitation
Location: CHORUS, MCW
URL: http: / / chorus. rad. mcw. edu/doc/00965.html
Indexing term(s): Aortic Valve Insufficiency

Figure 1 A CliniWeb database entry.

The initial CliniWeb database was built by a medical student who worked full time over the summer of 1995 and accumulated 4,225 URLs from 25 sites. Since then, about five hours per week have been devoted to maintaining and expanding the database, leading to the addition of 996 URLs from another 189 sites. (Most of the newer sites have far fewer clinical URLs.) Most URLs have been found by visiting the sites of academic medical centers and health-related government agencies. From those sites, links to other sites containing clinical information have been found. Additional sites to visit have been obtained through Internet publications as well as word of mouth. Once CliniWeb was made available to the public, feedback was received suggesting numerous other sites; most of them, however, did not meet our criteria of clinical content and so were not added to the database.

CliniWeb Indexing

Indexing of the CliniWeb database uses the MeSH disease classification from the C tree and F3 subtree; the latter contains psychiatric diseases. The decision to use the MeSH disease tree was based on the following considerations:

1. Indexing by diseases reflects the orientation of most clinical information on the WWW at this time
2. It has kept the indexing process manageable, requiring only about 6,000 indexing terms
3. It is rigidly hierarchical, which allows browsing capability.

Although the assignment of MeSH terms is done manually, the process is assisted by the concept-mapping engine of the SAPHIRE system.[12] This portion of SAPHIRE takes free text as its input and identifies terms from a controlled vocabulary. The vocabulary used in this instance is the MeSH disease tree, augmented by synonyms for those terms from the United Medical Language System (UMLS) Metathesaurus. An indexing assistance process has been created that takes each WWW page title and lists all MeSH disease terms that possibly match. The human indexer can

accept some or all of the suggested terms or add new ones with the assistance of an interactive version of SAPHIRE. The indexer can also submit other portions of the text to SAPHIRE to suggest additional indexing terms. All of the selected indexing terms are then added to each database record. Figure 1 shows an example.

CliniWeb Retrieval

As noted above, a design goal for CliniWeb was to allow information access either by browsing the MeSH hierarchy or by searching for specific terms. The aim for browsing was to navigate the indexing hierarchy. The goals for searching were to find MeSH terms using free text queries and then rapidly access the URLs associated with those terms.

The initial step in implementing a browsing capability was to create a browsable version of the MeSH disease classification subset. This was done using the UMLS Metathesaurus files that contain MeSH headings and their tree addresses. Each level of the MeSH tree was represented as a page in the CliniWeb WWW resource, with the term shown in color if it had further branching terms or in black if it was a leaf term. The former terms were "hot"; clicking on them would lead to their children terms in the hierarchy.

We next implemented a utility that inserted the URLs into a list directly under each term. For each entry in the CliniWeb databases, the utility took each indexing term, went to each WWW page where that term occurred in the hierarchy (going to more than one location for terms in multiple subtrees), and inserted the title as "hot" text with a link to the actual URL. For

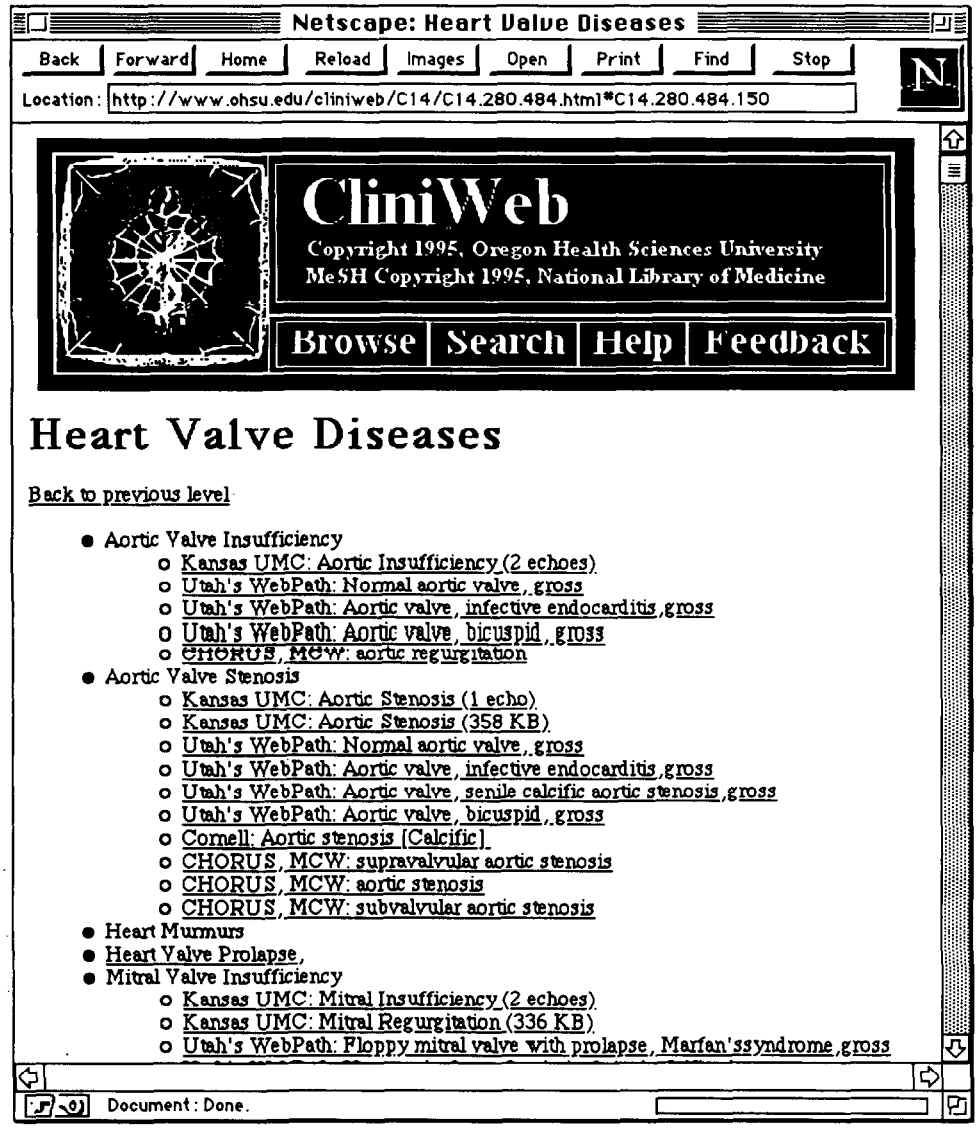

Figure 2 The MeSH disease tree with associated URLs, including the entry from Figure 1.

the example in Figure 1, the utility first went to the MeSH page containing the term "Aortic Valve Insufficiency." If the term had no URLs listed, it created a list under the term and added the URL. If terms were already there, it added the URL to the existing list. The utility made the text of the link a concatenation of the location and title, CHORUS.MCW: aortic regurgitation, with the link itself to http://chorus.rad.mcw.edu/doc/00965.html. (The HTML format for this example was <LI > <A HREF = "http://chorus.rad.mcw.edu/doc/00965.html"> CHORUS, MCW: aortic regurgitation </A>.) If other indexing terms were assigned, the same process was repeated for the other term(s). Figure 2 shows the page with the term "Aortic Valve Insufficiency." MeSH terms can be accessed by browsing from any point in the tree, starting with the top MeSH page (Fig. 3). This page is obtained at any time by selecting "Browse" from the button bar at the top of each page. Of course, users do not always want to browse through the MeSH tree to find their search terms. Therefore, a searching interface is implemented that uses the SAPHIRE concept-mapping engine to match query text to MeSH disease terms. The goal of the searching capability is to allow users to find a MeSH term and navigate directly to the list of URLs indexed for that term.

A user who selects "Search" from the button bar is brought to the searching interface, shown in Figure 4. The user is prompted for disease terms. When the "Submit Query" button is pressed, the text is sent to SAPHIRE, which returns a list of potentially matching terms ranked by their weight from the SAPHIRE algorithm (see below). The terms are presented as shown in Figure 5; the user can click on any term to

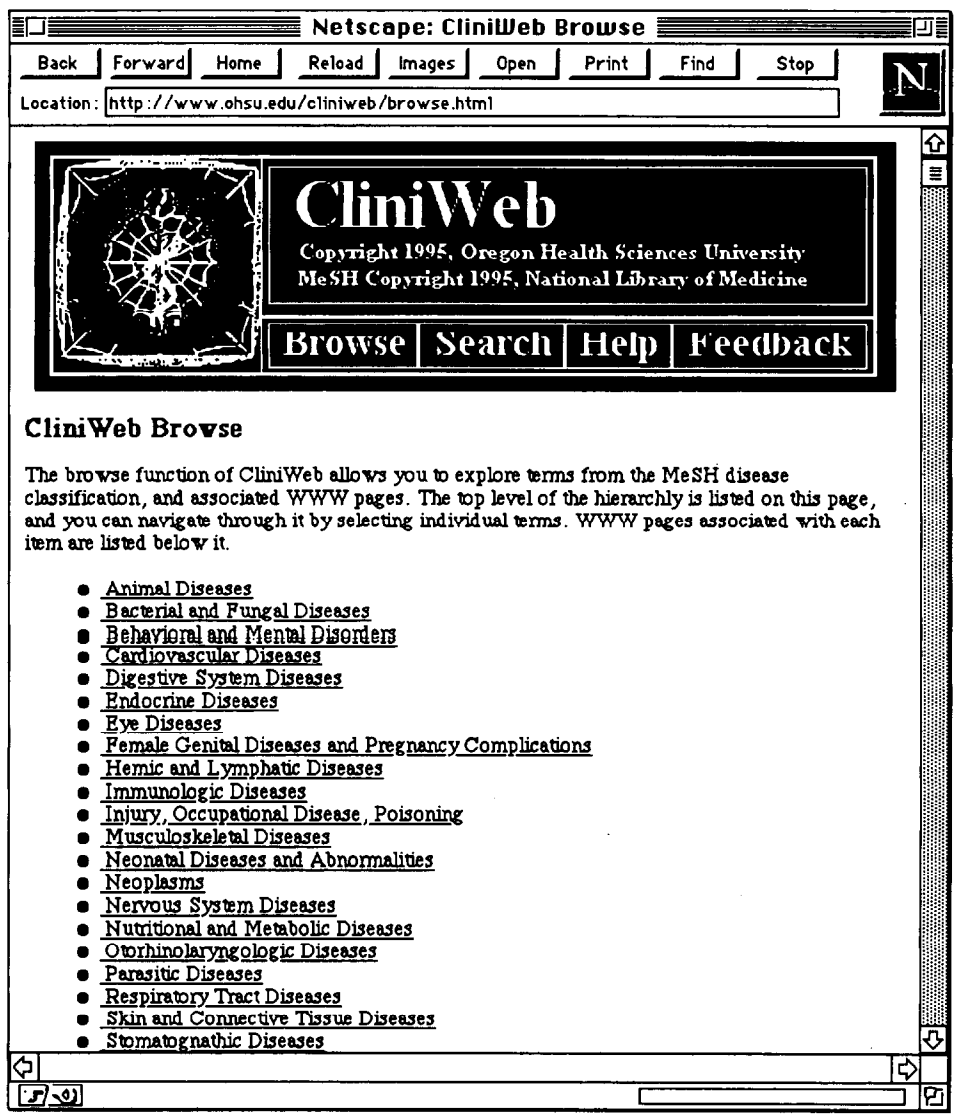

Figure 3 The top-level browsing point of CliniWeb.

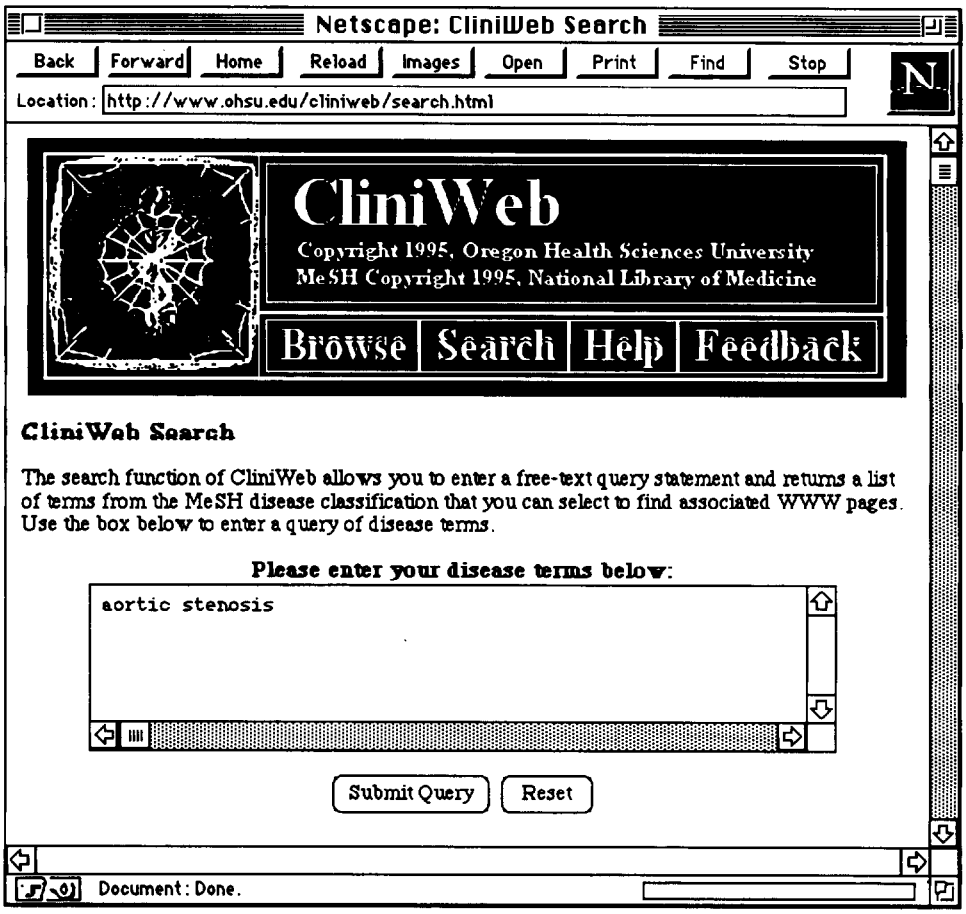

Figure 4 Entering the query *aortic stenosis* to obtain MeSH terms in CliniWeb.

go to its location in the MeSH hierarchy. This user has typed "aortic stenosis," and two matching terms are listed "Aortic Valve Stenosis" and "Aortic Subvalvular Stenosis." The WWW page in Figure 2 has the term "Aortic Valve Stenosis," which is where the user is brought if he or she selects that term. Once in a WWW MeSH page, the user can browse up and down the hierarchy by clicking on highlighted terms (such as "Heart Valve Prolapse" in Figure 2) to choose more specific terms or on "Back to Previous Level" to view more general terms.

One problem we faced in the design of CliniWeb was how to handle MeSH terms that occurred in more than one tree. The problem was minimized by adding all URLs for a given term to every instance of that term in the entire vocabulary. Thus, for the term "Lichen Planus, Oral," which occurs in the C7 (Stomatognathic Diseases) and C17 (Skin and Connective Tissues Disease) trees, both instances were given the identical and complete list of URLs for that term. An arbitrary decision had to be made regarding which tree location to bring the user to when the term was selected from the list returned by SAPHIRE. (We did not want to list multiple instances of the same term.) We chose to bring the user to the first tree address specified in the UMLS Metathesaurus file, which served as our source for MeSH terms.

In addition to an index of individual URLs, a list of WWW sites with biomedical information was compiled. As noted above, the richest sites were from medical schools and health-related government agencies. However, many other types of sites were identified, and these were organized into the following categories:

1. Medical schools
2. Government agencies
3. Medical libraries
4. Hospitals and research centers
5. Journals and professional societies

During CliniWeb's first four months of operation, it was accessed 10,689 times; 12,978 searches were performed. Although a formal evaluation of CliniWeb has not yet been undertaken, a recent search by one of the authors (WRH) shows its value. A collection of WWW material on "sinusitis" was needed. CliniWeb had three items on sinusitis: a primary care curricu-

lum from Stanford University, a grand rounds presentation from the University of Texas, and a document of complications from Vanderbilt University. An Infoseek search yielded over 100 documents, including the above three, but also many consumer-oriented publications, order forms for tonics and alternative-medicine remedies, and announcements for various symposia on the topic. The advantage for CliniWeb was to keep the search focused on clinical content. Clearly, a more comprehensive evaluation is needed to determine when each approach is most effective; this is planned for the near future.

SAPHIRE as Used in CliniWeb

Although initially developed as a complete IR environment,[9] the SAPHIRE project has changed focus in recent years.[12] Most pertinent to CliniWeb has been a reimplementation that separates the SAPHIRE concept-matching algorithm from the rest of the retrieval system. In addition, the algorithm has been changed so that it now identifies a list of candidate matching terms instead of trying to identify the single "correct one." Although the input to the algorithm is still a string of free text, such as a query or portion of a document, the output is now a list of concepts that are ranked based on their closeness of fit to the input string, as described elsewhere.[12]

Another change in SAPHIRE has been its implementation as a server on the Internet. This allows access to SAPHIRE's functions by socket-based procedure calls. A Common Gateway Interface (CGI) script for WWW servers has been implemented that allows WWW clients to interact with the SAPHIRE server. SAPHIRE is used in CliniWeb to assist users in selecting MeSH disease terms for retrieval and to assist indexers in assigning them.

Limitations of Current Implementation

A number of limitations in the initial version of CliniWeb will be addressed in future versions. For example, we have not identified all relevant clinical information in our first pass across the WWW, as pointed out to us by electronic feedback. In addition, there is a need to develop additional procedures and tools to update and maintain the database. One tool, described above, that has already been implemented checks the validity of all URLs in the database.

The indexing process has limitations as well. Indexing assignments have been done by a medical student untrained in the methods of professional indexing. Although most assignments have been deemed adequate by the investigators in this project, a professional indexer would likely be beneficial for the long

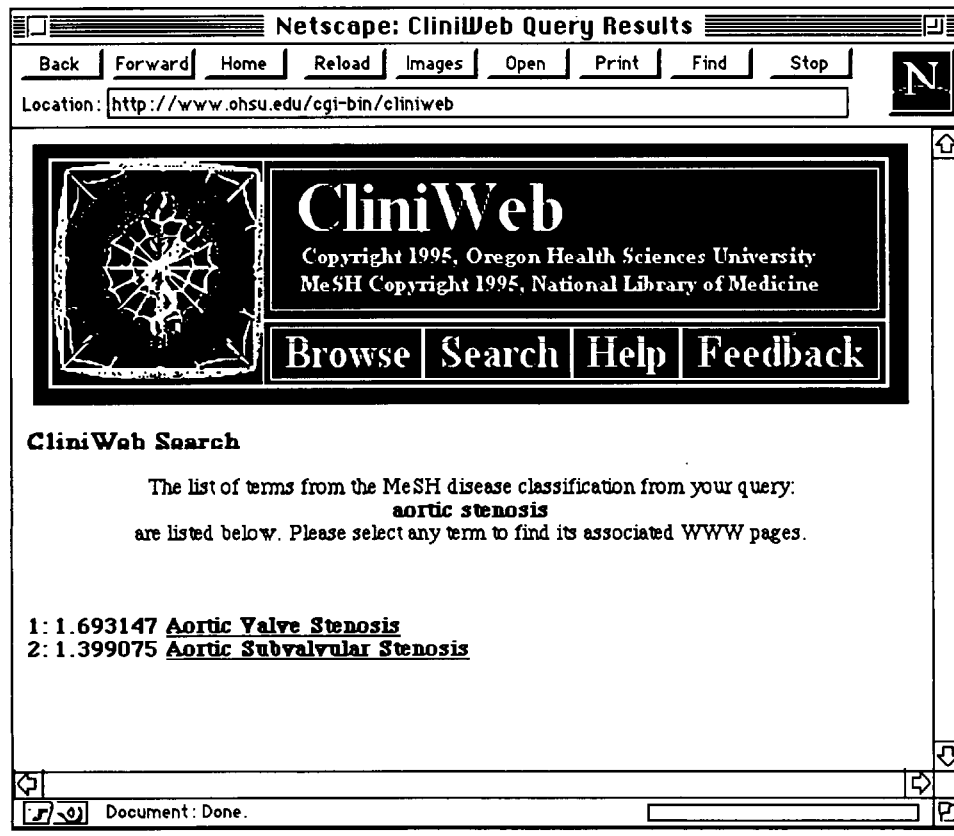

Figure 5 Retrieval of MeSH terms in CliniWeb from the query *aortic stenosis*.

run, especially as additional MeSH trees and other indexing approaches are added.

There are several retrieval limitations with CliniWeb as well. The system currently assumes that the user is interested in just one disease. There is no capability to search on more than one disease or specific topics related to a disease, such as diagnosis or treatment. In the future, we plan to use CliniWeb as a testbed to investigate different approaches to retrieval on the Web, including the ability to search on more than one disease as well as other MeSH terms, text words, and UMLS Metathesaurus concepts. We will also assess the use of Boolean searching capabilities.

A final limitation of CliniWeb is its long-term maintenance. Like other manual indexing approaches, CliniWeb requires ongoing human indexing. Furthermore, the dynamic nature of the WWW and the lack of a table of contents for new information being posted make the discovery of new resources and their indexing labor-intensive processes. Like most large-scale medical informatics projects, an ongoing source of funding will need to be identified to ensure long-term maintenance. This is currently being addressed.

Future Directions

Based on the limitations of the current implementation, we plan the following enhancements to CliniWeb:

1. Improving the structure and maintenance of the database of URLs
2. Improving indexing term assignment
3. Broadening the scope of indexing terms used (i.e., using additional MeSH terms beyond those from the disease tree)
4. Assessing the benefit of automated indexing, based on words and UMLS Metathesaurus concepts
5. Developing alternative query interfaces, such as direct text word searching, as well as the use of Boolean operators to connect multiple text words and/or MeSH terms

We also plan to evaluate CliniWeb systematically. Although the dynamic nature of the WWW precludes conventional retrieval evaluation studies, which require fixed databases, we have implemented an online feedback form. Responses are being collated by type of suggestion in order to classify requests for changes and implement those that occur with the greatest frequency. These feedback forms have also been alerting us to content material we had not discovered.

In summary, we have created a system for managing clinical information on the WWW. It provides indexing of URLs by specific clinical topics from the MeSH disease tree and allows retrieval by searching or browsing. Further work will focus on maintaining the database in the dynamic Web environment while adding additional indexing and retrieval capabilities.

References ■

1. Berners-Lee T, Cailliau R, Luotonen A, Nielsen HF, Secret A. The World-Wide Web. Comm Assoc Comp Mach. 1994; 37:76–82.
2. Cimino JJ, Socratorus SA, Clayton PD. Internet as clinical information system: application development using the World Wide Web. J Am Med Inform Assoc. 1995;2:273–84.
3. Lowe HJ, Lomax EC, Polonkey SE. The World Wide Web: a review of an emerging Internet-based technology for the distribution of biomedical information. J Am Med Inform Assoc. 1996;3:1–14.
4. Bowman CI, Danzig PB, Manber U, Schwartz MF. Scalable Internet resource discovery: research problems and approaches. Comm Assoc Comp Mach. 1994;37:98–107.
5. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc. 1983;71:176–83.
6. Haynes RB, McKibbon KA. Grateful Med. MD Comput. 1987;4:47–57.
7. McCarberg B. Medline Knowledge Finder. JAMA. 1989;261: 1812.
8. Salton G. Developments in automatic text retrieval. Science. 1991;253:974–80.
9. Hersh WR, Hickam DH. Information retrieval in medicine: the SAPHIRE experience. J Am Soc Info Sci. 1995;46:743–7.
10. Evans DA, Hersh WR, Monarch IA, Lefferts RG, Handerson SK. Automatic indexing of abstracts via natural language processing using a simple thesaurus. Med Decis Making. 1991;11:S108–15.
11. Miller PL, Frawley SJ, Wright L, Roderer NK, Powsner SM. Lessons learned from a pilot implementation of the UMLS information sources map. J Am Med Inform Assoc. 1995;2: 102–15.
12. Hersh WR, Leone TJ. The SAPHIRE server: a new algorithm and implementation. SCAMC Proc. 1995;858–62.