Searching for the Right Search — Reaching the Medical Literature

Robert Steinbrook, M.D.

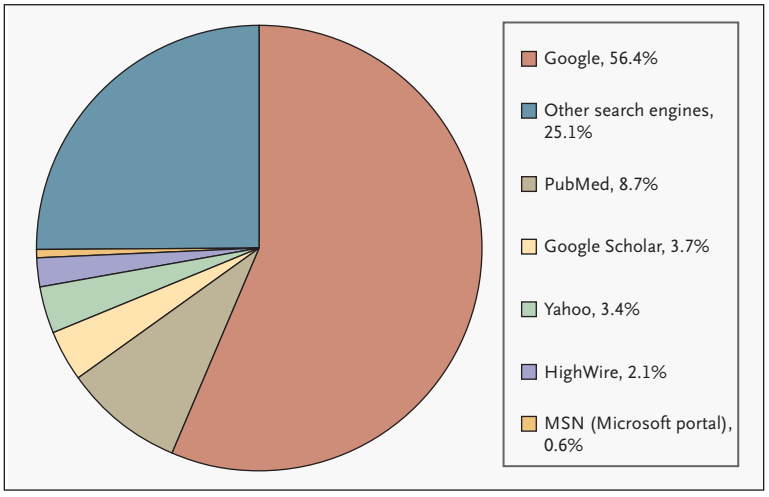
Web-based search engines are transforming our use of the medical literature. Although we continue to read the print issues of journals and to browse current issues online, we are now using links from Google — the flagship search engine of the Mountain View, California, company of the same name — and other search engines, as well as citation links in other articles, to gain direct access to the articles we want. For example, by quickly searching by the title of an article, an author, or a specific topic, we can often link to a bibliographic citation, the abstract, or the online version of the article itself. In addition, an increasing number of biomedical libraries and medical institutions can link from search results to their online subscriptions to journals, their electronic catalogue of print holdings, publishers' Web sites, and other full-text digital archives.

The ongoing changes are part of a broader trend in society. According to a May 2005 report from the Pew Internet and American Life Project, 8 in 10 Internet users have looked for health information online, showing increasing interest in diet, fitness, drugs, health insurance, experimental treatments, and particular doctors and hospitals.[1] About three quarters of scholarly journals are now online. The Web sites of most biomedical journals continue to see "marked increases in usage, with no end in sight," according to John Sack, the director of HighWire Press, a division of the Stanford University Libraries (highwire.stanford.edu). HighWire hosts the Web sites of about 918 scholarly publications (including that of the *Journal*) and maintains a large archive of full-text articles. Articles found by searching the Web (see box) are a logical evolution from the photocopies of articles that medical students and residents have pulled out of their backpacks for decades.[2]

"What readers see now are articles," Sack said recently. "They don't see articles bound in the context of issues or in the context of well-known journals. This has been happening for a while, but it has been greatly accelerated by the Internet and by Google and other search engines that are indexing everything that is out there."

There are many search engines and many ways to gain access to the online medical literature. At the moment, however, Google is the most widely used. Other widely used search tools are PubMed, a federal government portal that offers access to the enormous database of citations and abstracts at the National Library of Medicine; Google Scholar, which specifically searches the scholarly literature; and Yahoo, the search engine of the Sunnyvale, California, company of the same name. These search engines are available to anyone who has an Internet connection; none require registration, and searching is free of charge.

The rapid changes are illustrated by data compiled by HighWire Press. In June 2005, Google provided the majority (56.4 percent) of the referrals from search engines to articles in HighWire-hosted journals (see pie chart). PubMed accounted for 8.7 percent, Google Scholar 3.7 percent, and Yahoo 3.4 percent. Google



Referrals from Search Engines to Web Sites of 844 Journals Hosted by HighWire Press.

Data are for June 2005 and are from HighWire Press.

Scholar has been available only since late 2004, and many people remain unaware of it.

When he first saw similar data earlier in the year, Sack recalled, he was "surprised that Google had greatly surpassed PubMed and that a new product such as Google Scholar had approached half of PubMed's referrals within a few months." He added, "The data indicate that readers' habits for finding information are shifting. New readers are using the new search tools, not the old ones." In September 2005, the percentage of referrals from Google Scholar to HighWire-hosted journals surpassed the percentage from PubMed. By November 2005, there were 35.7 percent more referrals from Google Scholar than from PubMed. The reason is that although the total percentage of referrals from either Google or Google Scholar is about the same, more people are using Google Scholar. The percentage of referrals from Yahoo has also increased, but not as rapidly.

In the past, searching has often started with PubMed, which was launched in June 1996. The largest component of PubMed is the Medical Literature Analysis and Retrieval System Online, more commonly known as Medline, which covers more than 4800 biomedical journals published in the United States and 70 other countries, primarily from 1966 to the present. Medline contains

Choosing among Search Engines

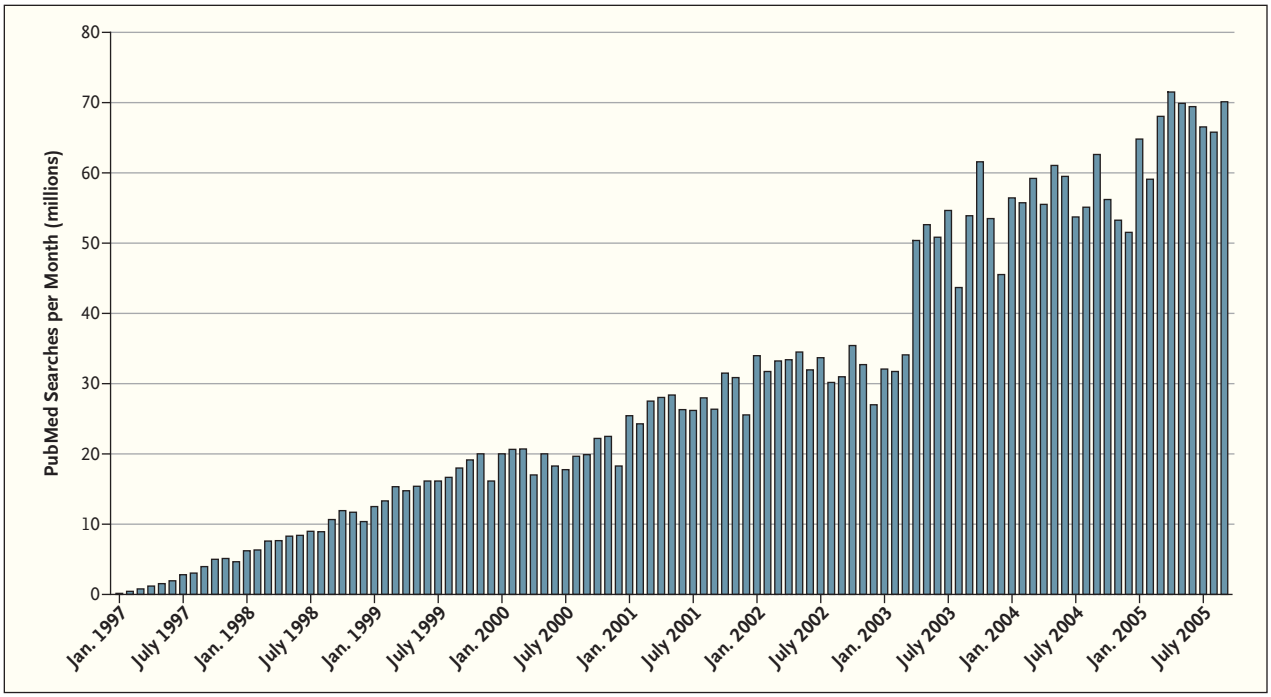
Whether one should search using PubMed (www.ncbi.nlm.nih.gov/entrez/query.fcgi), Google (www.google.com), Google Scholar (scholar.google.com/), Yahoo (www.yahoo.com), or another search engine depends on the information one desires, one's personal preference, and the search engine's strengths and weaknesses. Google, Google Scholar, and Yahoo are easy to use and provide breadth, but the results will vary widely depending on the search terms that are chosen. These search engines index the contents of PubMed as well as the online content of many scholarly publishers. Google searches have "lots of hits and lots of misses," according to Dr. Daniel Masys, chair of the Department of Biomedical Informatics at Vanderbilt University Medical Center. PubMed can provide more depth but may require more effort to use; training may help searchers to obtain the best results. In many instances, the choice will not matter, because several approaches will work. For more in-depth searching, the use of several search engines may provide complementary results.

Google and Yahoo are general search engines. PubMed and Google Scholar primarily find journal articles, the former from the life sciences and related fields, the latter from all areas of research. PubMed can sort results according to date, author, or journal but not according to the number of citations. According to Dr. Mohammad Al-Ubaydli, a visiting research fellow at the National Library of Medicine, who has studied the characteristics of various search engines, results with Google and Google Scholar can be narrowed with the use of more specific search terms — by entering "myocardial infarction thrombolysis," for instance, rather than simply "myocardial infarction." Google Scholar orders the results according to how relevant to the query it deems references to be, taking into account the full text of the article, the author, the publication in which it appeared, and how often it has been cited in other scholarly publications. Searches can be limited to an author, a publication, or a range of dates. Widely cited and important papers in a given field will often appear at the top of the results, with newer papers and others with fewer citations appearing near the end. Google Scholar also includes a "cited by" feature that links an article to the others that have cited it.

PubMed searches define databases that have extensive indexing and quality control. In addition to searching the text for particular words, it uses a controlled-vocabulary thesaurus of medical subject headings, known as MeSH. This feature permits searching with various degrees of specificity. However, "most human beings on the planet who are not librarians don't know anything about how to search with MeSH," said John Sack, the director of HighWire Press. Clinical searches through PubMed can be limited to one of four study categories (therapy, diagnosis, etiology, or prognosis), studies conducted in humans, or studies with particular types of research methods, such as systematic reviews, among other options.

Google Scholar is more difficult to focus than PubMed — but it may find papers, theses, books, preprints, abstracts, and technical reports that are not in National Library of Medicine databases. However, Google Scholar does not identify the sources that it has — or has not — indexed. Thus, there is no way to know what may be missing. Google Scholar is separate from Google Book Search, which searches the full text of books and which is related to Google's widely publicized project to digitize most of the books from several major libraries.

There are many proprietary medical resources and databases that are currently not publicly searchable by means of any Internet search engine, although subscribers may use them through the Web. Publishers have control over access to their articles. A search engine can index only the material that it identifies, "crawls," and processes. Google will index papers with access restrictions only if all users of its search tools are offered at least an abstract or an extract. The situation, however, is in flux. For example, Yahoo has a feature that searches content — with the permission of the source — that is not normally accessible. Access to the content, however, usually requires a subscription to the publication.



PubMed Searches per Month, January 1997 through September 2005.

Data are from the National Center for Biotechnology Information at the National Library of Medicine. The increase in the number of searches in the spring of 2003 reflects changes in the Web-log accounting systems; previously, the number of searches was slightly undercounted.

more than 12 million citations. PubMed contains additional citations and journals, including about 2 million citations to articles published between 1950 and 1965, and searches can extend to other databases at the National Library of Medicine, such as GenBank, PubChem, and PubMed Central (www.pubmedcentral.nih.gov), the National Institutes of Health's (NIH's) archive of biomedical and life-sciences literature.

The number of searches performed with PubMed has increased steadily to about 70 million per month (see bar graph). Yet at the same time, an increasing number of people are finding their way to citations and abstracts in PubMed through searches that begin with Google — the largest single source of referrals to PubMed — or with Google Scholar or Yahoo.

Many articles are available through Web sites maintained by journals, although there may be charges or registration requirements. Some are also available without charge through nonjournal Web sites — sometimes with the permission of the publisher, sometimes without.[3] Such sites may be personal ones established by an author or online repositories maintained by the author's institution or another institution. Archiving through nonjournal sites is incomplete, however, and it is more likely to be available for basic research articles than for clinical research articles. Some journals and publishers — as well as Web sites and Web-based links — come and go. And search engines do not store content and make it available to readers. Rather, they provide links to the actual sources of content, and they can identify only content that they have successfully indexed (see box).

Because of the limits of other online sources, central electronic repositories of journals and articles serve a critical archival function, according to Dr. David Lipman, the director of the National Center for Biotechnology Information at the National Library of Medicine, home to PubMed and PubMed Central. Within the year, PubMed Central is expected to contain between 700,000 and 800,000 reports, including many articles from back issues of about 180 journals.[4] Central repositories can also store supplemental data and may permit more detailed searches and a greater ability to retrieve and manipulate the underlying information than is possible with papers that may be archived in different formats

at different sites. "Biomedical research has changed," noted Lipman. "Every paper has more and more data. People are not just reading these papers. Researchers want to compute on the underlying data."

The NIH is seeking to expand public access to the research it sponsors and to increase the usefulness of PubMed Central. As of May 2, 2005, the NIH has asked the investigators it supports to submit voluntarily to PubMed Central an electronic copy of any scientific report, on acceptance for publication, and to specify when the article should become publicly available through the repository.[4] According to the policy, posting for public accessibility "is requested and strongly encouraged as soon as possible (and within 12 months of the publisher's official date of final publication)." However, the initial response to the voluntary policy has been slow.

With 100 percent participation, about 5500 peer-reviewed manuscripts that have been accepted but not yet published — equivalent to about 10 percent of the articles indexed monthly by PubMed — would be submitted to PubMed Central each month, according to Lipman. As of July 9, 2005, 340 such unpublished manuscripts (or about 165 per month) had been submitted — a participation rate of only 3 percent. There are no signs that the participation rate for unpublished manuscripts is increasing — in August, September, and October of 2005, it was between 2.2 and 2.7 percent. In December 2005, Senators Joseph Lieberman (D-Conn.) and Thad Cochran (R-Miss.) introduced legislation that would require the public posting of all NIH-funded peer-reviewed manuscripts at PubMed Central within six months of their publication. Failure to comply could result in the loss of public funding for federal employees or grantees.

Physicians and researchers have extremely diverse information needs. Meeting these needs requires diverse resources. Search engines and the Internet are not only changing the medical literature. They are also challenging the traditional economics of scholarly publishing and fueling heated debate about the extent to which the biomedical literature should be accessible online and available without charge to the user.[4,5] As search engines and the online medical literature itself continue to evolve, the pace of change is likely to increase.

Dr. Steinbrook is a national correspondent for the *Journal*.

1. Fox S. Health information online. Washington, D.C.: Pew Internet & American Life Project, May 17, 2005. (Accessed December 14, 2005, at http://www.pewinternet.org/PPF/r/156/report_display.asp.)
2. Sack J. HighWire Press: ten years of publisher-driven innovation. Learned Publ 2005; 18:131-42.
3. Wren JD. Open access and openly accessible: a study of scientific publications shared via the Internet. BMJ 2005;330:1128-31.
4. Steinbrook R. Public access to NIH-funded research. N Engl J Med 2005;352:1739-41.
5. Wysocki B. Scholarly journals' premier status is diluted by Web. Wall Street Journal. May 23, 2005:A1.

Is Our Behavior Written in Our Genes?

Dennis Drayna, Ph.D.

Scientists recently reached an important milestone in the understanding of genetic contributions to behavior. A new study demonstrated the role of a single gene in specifying sexual behavior in the fruit fly *Drosophila melanogaster*.[1] The findings prompt provocative thinking about the contribution of genetic factors to sexual orientation in humans, as well as about genes that might underlie a broader spectrum of human behaviors.

The investigators in the fruit-fly study, Demir and Dickson, focused on a gene called *fruitless* that has long been known to have strong effects on mating, fertility, and reproduction in fruit flies. The messenger RNA product of this gene (see figure) encodes a transcription factor that is essential for development and that can occur in any of several variously spliced forms. Two of these forms are sex-specific, one being unique to male flies and the other to female flies. Demir and Dickson used genetic manipulation to produce anatomically female flies that carried only the male form of the gene (see figure). The resulting flies exhibited courtship and mating behavior toward females that is normally engaged in only by male flies. Whereas previous studies have shown that the male form of the *fruitless* gene is necessary for male courtship, the new study shows that it is sufficient to produce this behavior, even in females — making it the first single gene to be identified as both necessary and sufficient for specifying a complex behavior in a higher-level organism.