FEATURE / MANCHETTE

A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations

Dean Giustini and Eugene Barsky

Introduction

The *beta version* of Google Scholar (GS) has attracted worldwide attention from health professionals and librarians since its launch in November 2004 [1–4]. Though it purports to "locate scholarly literature across all disciplines in [many] formats" and to offer "the best scholarly search experience for users" [5], GS has generated considerable debate in library circles about its usefulness [6–8]. How do librarians educate users about Google's shortcomings when they (and their services) are becoming increasingly *google-ized*?

Some nagging questions about GS persist: what is "scholarly" in Google's view? how big is GS? how many databases, journals, dot.edu and dot.gov sites are indexed? how often is it updated or refreshed? In this article, we discuss what is known about GS and run simple tests of its coverage. Then, GS is compared to PubMed and its major strengths and weaknesses discussed. Scirus is also discussed (its pros and cons) as a free search alternative to GS. Based on the requirements for complex searches, we make a recommendation for using OVID MEDLINE for specific clinical queries.

Background

The Internet has helped to promote end-user searching through *freely*-accessible databases at the US National Library of Medicine (NLM). But Web search engines are also a factor in forming end-user search preferences and habits [9]. According to a 2003 Canadian Medical Association survey [10], 65% of physicians use the Web for information to support clinical practice. Many of these doctors search PubMed or tools like Google to locate information. Curiously, nearly half (46%) call themselves "novice or inexperienced" when locating reliable information.

D. Giustini.[1] University of British Columbia Biomedical Branch Library, Vancouver Hospital & Health Sciences Centre, Heather Pavilion, 700 West 10th Avenue, Vancouver, BC V5Z 1L5, Canada.
E. Barsky. Mental Health Evaluation and Community Consultation Unit, Department of Psychiatry, University of British Columbia, St. Paul's Hospital, 1081 Burrard Street, Comox Room 306C, Vancouver, BC V6Z 1Y6, Canada.

[1]Corresponding author.

Information retrieval is a challenge for users when search tools are too complex to navigate. "Clinicians and researchers conduct MEDLINE searches but lack skills to do this well", according to Haynes et al. [11]. Could GS be an efficient means to access information? Could GS be used by clinicians for specific types of questions? What types? Before listing the negative (and potentially lethal) implications of using GS in clinical decision making, let's examine why Google is so popular among our users.

First, users like Google for its simplicity, speed, and coverage; it is used more than any other Web search engine [12]. Google is the search engine of choice for more than half of all Web queries [13–15]. Users have faith in Google branding and believe high standards are applied equally to all Google products [16].

GS does index a lot of content, linking back to regular Google (and even PubMed) for optimum cross-functionality. For users not affiliated with a major university or teaching hospital, GS is seen as a welcome, free gateway to reliable scientific information. In beta version, however, GS has some serious limitations that need to be examined.

Coverage and currency: the pros and cons of Google Scholar

From its inception in late 2004, GS crawled most of PubMed–MEDLINE (1966 – present) and OLDMEDLINE (1949–1965). However, Vine noted that PubMed records in GS are a year out of date [17]. (Our tests repeatedly retrieve the same results on GS, suggesting the database is not regularly updated.)

GS indexes content from 29 of the top scholarly publishers and university presses (see Appendix A) [18]. Discussions are underway with other publishers [19]. Digital hosts at HighWire Press, MetaPress, and Ingenta are crawled by Google's bots, as are open-access journals at BioMedCentral, PubMedCentral, and document suppliers like Ingenta, societies, scholarly organizations, government agencies, and preprint–reprint servers.

What is not indexed is more difficult to determine, as Google has been vague at times about GS's content. Major health science publishers *not* crawled by Google's bots include Elsevier and Karger Press. Some major Canadian content is inadequately indexed or not indexed at all. Statistics at Sta-

tistics Canada (www.statcan.ca) or the Canadian Institute for Health Information (www.cihi.ca) are not indexed, though in-house papers are to be indexed. Library and Archives Canada's (http://www.collectionscanada.ca/) records have also started to appear.

Interestingly, Canadian health content from recognized Web sites, such as the Manitoba Association of Registered Nurses (www.crnm.mb.ca), are *not* crawled, while US institutions with a similar focus are, such as the New York Nurses Association (www.nysna.org). Canada's "grey literature" is *not* comprehensively indexed, fragmenting an already unwieldy bibliography. (Well-known government reports such as the *Romanow Report* and provincial documents such as the *Kirby Report* are increasingly available.) Health librarians should work to ensure our grey literature gets indexed on the Web by developing our own database or advocating for better coverage on standard Web tools [20,21].

Google Scholar search results: publishers and PubMed

Health librarians should show users how GS *should* and *should not* be used. Using examples to illustrate why GS is useful (or dangerous) should be a part of all librarian-led search training.

Let's start with search functionality. Do a standard search for two phrases: "common cold" and "vitamin c". Illogically, articles from the 1990s are listed first, not the most current articles. Why older articles first? GS's PageRank algorithm makes a calculated guess at what it believes is scholarly and lists articles by how relevant and popular they are — *not* how current (see Fig. 1).

Ranking of older research in a scholarly database is a big problem, compounded by a lack of re-sorting options. Filtering of results by English language, abstracts, and methodology on GS is difficult if not impossible.

Does GS compare with searching directly at publisher sites? Significant differences in recall are observed. A search at Blackwell Synergy (www.blackwell-synergy.com) yielded 456 000 citations, whereas a site search for Blackwell on GS retrieved only 80 300 citations. A site search on GS for PubMed (www.ncbi.nlm.nih.gov) citations found 1.1 million records, 14 million fewer than on PubMed itself (Fig. 2).

Searching for "heart attack" at *Nature*'s publisher site found 557 citations compared to GS's 251 (Fig. 3). Similar discrepancies were found for "electroconvulsive therapy" at Wiley (202 citations) and GS (58 citations). GS doesn't come close to what is found at publisher sites. For maximum recall, we advise searching publisher sites directly. Keyword searching in GS vis-à-vis PubMed is inadvisable, also. To maximize recall, search PubMed by keyword and MeSH simultaneously from the homepage (*click* Details).

To run simple tests of coverage and recall, Peter Jascó from the University of Hawaii has recently developed some very useful "polysearch" tools (http://www2.hawaii.edu/~jacso/scholarly/side-by-side2.htm) [22]. Polysearch runs simple queries across several sites and databases. Our testing validates Jascó's findings and conclusions. GS's coverage is incomplete, retrieving fewer unique citations than either publishers' sites or PubMed.

Fig. 1. Standard search in GS for two phrases: "common cold" and "vitamin c".

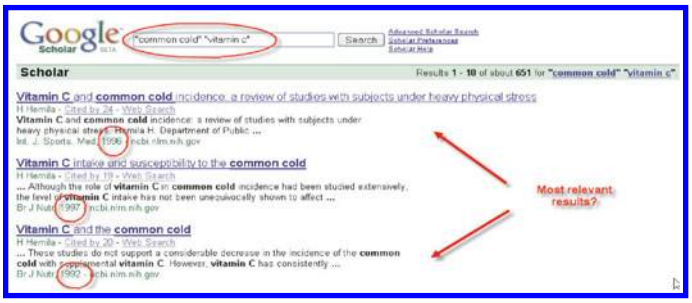
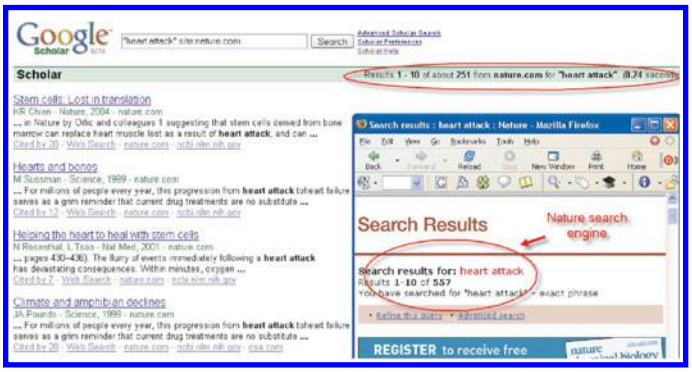


Fig. 2. Search in GS for PubMed citations.

Fig. 3. Search in GS for the phrase "heart attack" compared with search in *Nature*'s publisher site.



Special features and special problems

A few special features on GS are worth mentioning. First, its overall performance is robust and comparable (or better) than other specialty health search engines (*test:* www.mammahealth.com, for example). Google's bots are capable of crawling bibliographic information from references at the end of articles, extending GS's reach beyond journal articles to books and AV materials.

Through its partnership with OCLC, links to Worldcat in the results display allow users to identify in seconds whether a local library has the book or journal needed. To expand a search, the "regular Google" link can be used to do an on-the-fly search in regular Google. Another helpful feature is linking to PubMed records. GS compensates a bit for its lack of currency by linking to PubMed records showing the URL www.ncbi.nlm.nih.gov. Users see this message after linking to PubMed: "Note: Performing your original search [in Google Scholar], '*common cold' and 'vitamin c'*, in PubMed will retrieve *150* citations."

"Cited by" is a very welcome feature [23]. By linking users to related research, GS provides for free what ISI's Web of Science (WoS) and Elsevier's Scopus provide at considerable cost. However, keep in mind that PageRank in GS is not the same as ISI's bibliometric tools, a distinction that must be iterated to users.

The article linking products Ex Libris and SFX (based on Open URL technology) are fully compatible with GS. This software allows users to see a customized display of a local library's print and electronic journal collections within GS. For users with no article linking tool, GS offers linking options under preferences, which are easily used even behind hospital firewalls.

Searching for certain medical topics is frustrating due to the lack of controlled terms and authority control. Variant titles and author names make comprehensive retrieval impossible. Fee-based document delivery through Ingenta is a problem. Users could be misled if articles are ordered for a fee — only to learn that a local library has the items. On the other hand, options for document delivery are helpful if remote users need documents and are willing to pay. Librarians should be prepared to show how to access documents, find them locally, or order them through Ingenta.

Scirus: an alternative to Google Scholar

GS is not the only choice for searching for scholarly, scientific content. Since 2001, many researchers have used Elsevier's Scirus, which claims to have the best science, technology, and medicine (STM) coverage on the Web, with more than 200 million science-specific pages indexed [23]. Unlike GS, Scirus clearly lists its content sources: ScienceDirect and BioMedCentral, Beilstein on ChemWeb, DSPACE repositories, and 13 million patents from Japan, Europe, and the United States. Elsevier is negotiating with other scientific publishers to make more content available [24].

Scirus provides focussed channel-searching by content provider and categories like "medicine" or "psychology". Improved customization and flexibility allow for more precise searching. A regular Search Engine Watch (www.searchenginewatch.com) award winner [25], Scirus gets high marks from librarians and is a good alternative to GS.

For complex searches use OVID or PubMed

Most end-users use Google because their needs are often satisfied by basic search tools [26]. However, for intermediate and advanced searchers in medicine, more functionality is needed. A pharmacist's search for the use of antibiotics, for example, introduces a number of complexities. In PubMed, a class of drugs can be searched by exploding a subject heading and its narrower terms, a feature not available in Scirus or GS. To achieve high recall, every term and antibiotic drug name would need to be keyed into GS's search box. "Explode" saves valuable time and is a feature on proprietary databases like EMBASE and CINAHL, but not on search engines like Google.

GS and Scirus are *not* able to limit searches by publication type or research methodology. This is another problem when evidence-based filters are needed to refine a search. Thus, users are forced to try wildcard and keyword combinations in GS. When age and gender are important, GS or Scirus offer no means to limit by these elements unless they are searchable as keywords in title or abstract fields.

The *gold standard* for complex searches with multiple sets is the OVID interface to MEDLINE. OVID MEDLINE offers the best functionality and flexibility for building and manipulating sets developed using PICO [27]. OVID's mapping feature makes using controlled terms easier, including explode or focus. Complex searches can be done on PubMed also, but its interface is not as intuitive or user friendly. A search history is always displayed on OVID, and easy access is provided to major limits (users do get lost in PubMed). "Clinical queries" in OVID and PubMed are synonymous (also called the Haynes filters). Both OVID and PubMed permit saved searches for later retrieval, and SDIs and e-Alerts can be sent out at regular intervals.

We recommend OVID for expert searching as it sets a high standard for commercial interfaces. PubMed is recommended for its primary strengths: currency, links to the open Web, and growing *free* content. For those without OVID, PubMed can be used to do structured literature searching also, but keeping current with changes at the site might make searching difficult for many users.

Conclusion

In summary, information professionals have no choice but to recommend Google Scholar under certain conditions and caveats. Librarians should be prepared to teach GS and PubMed side by side and answer questions about it, especially how it compares to commercial tools like OVID.

Clearly, GS provides an easy means to access the health literature. Health librarians should not dismiss it outright, especially for simple browsing, known-item searching, and linking to free materials on the open Web. Where literature reviews are required, i.e., grants, clinical trials, or systematic reviews, health librarians will continue to recommend MEDLINE, Cochrane (with Google for grey literature), and other trusted sources. Finally, clinical queries must be answered by replacing requests in context [28]. Health professionals already search Google [29] and will continue to use it (responsibly, one hopes) to satisfy their basic information needs [27].

References

1. Abram S. Google Scholar: thin edge of the wedge? *Information Outlook*. 2005;9(1):44–6.
2. Leslie M. A Google for academia. *Science*. 2004;306(5702): 1661–3.
3. Lindberg DA, Humphreys BL. 2015 — the future of medical libraries. *N Engl J Med*. 2005 Mar 17;352(11):1067–70.
4. Butler D. Science searches shift up a gear as Google starts Scholar engine. *Nature*. 2004;432(7016):423.
5. Google. Google Scholar help screen [Web page]. Mountain View, Calif.: Google Inc.; 2005 [cited 10 Aug 2005]. Available from http://scholar.google.com/scholar/help.html.
6. Lederman D. Google: friend or foe? [Web page]. Washington, D.C.: Inside Higher Ed; 2005 Apr 11 [cited 10 Aug 2005]. Available from http://www.insidehighered.com/news/2005/04/11/google.
7. Henderson J. Google Scholar: a source for clinicians. *CMAJ*. 2005;172(12):1549–50.

8. Banks M. The excitement of Google Scholar, the worry of Google Print. *Biomed Digit Libr.* 2005;2:2.
9. Pew Internet & American Life Project. Internet Health Resources [Web page]. Washington, D.C.: Pew Internet & American Life Project. 2003 Jul 16 [cited 10 Aug 2005]. Available from http://www.pewinternet.org/PPF/r/95/report_display.asp.
10. Martin S. Younger physicians, specialists use Internet more. *CMAJ.* 2004;170(12):1780.
11. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ.* 2005 May 21;330(7501):1179.
12. Sullivan D. Search Engine Watch. Searches per day [Web page]. London: Incisive Interactive Marketing LLC. 2003 Feb 25 [cited 10 August 2005]. Available from http://searchenginewatch.com/reports/article.php/2156461.
13. Sullivan D, Sherman C. Search Engine Watch. Fifth (5th) Annual Search Engine Watch Awards [Web page]. London: Incisive Interactive Marketing LLC. 2005 Mar 31 [cited 10 Aug 2005]. Available from http://searchenginewatch.com/awards/article.php/3494141#bestsearch.
14. Fox S. Dr. Google's office never closes [PowerPoint presentation]. Washington, D.C.: Pew Internet & American Life Project. 2005 Apr 14 [cited 10 Aug 2005]. Available from http://www.pewinternet.org/PPF/r/41/presentation_display.asp.
15. Boswell W. Search Engine Statistics for June 2005 [Web page]. 2005 Jul 20 [cited 10 Aug 2005]. Available from: http://websearch.about.com/b/a/186995.htm.
16. Fox S, Fallows D. Internet health resources: Health searches and email have become more commonplace, but there is room for improvement in 107 searches and overall internet access [Web page]. Washington, D.C.: Pew Internet & American Life Project. 2003 Jul 16 [cited 10 Aug 2005]. Available from http://www.pewinternet.org/PPF/r/95/report_display.asp.
17. Vine R. SiteLines. Google Scholar is a full year late indexing PubMed content [Web page]. 2005 Feb 8 [cited 10 Aug 2005]. Available from http://www.workingfaster.com/sitelines/archives/2005_02.html#000282.
18. Jasco P. Peter's Digital Reference Shelf. Google Scholar Beta [Web page]. 2004 Dec [cited 10 Aug 2005]. Available from http://www.galegroup.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=/reference/archive/200412/googlescholar.html.
19. Kennedy S, Price G. Big news: 'Google scholar' is born. *ResourceShelf* [e-newsletter]. 2004 Nov 18 [cited 10 Aug 2005]. Available from http://www.resourceshelf.com/2004/11/wow-its-google-scholar.html.
20. CABOT Database [database on the Internet]. Ottawa, Ont.: Canadian Association for Health Services and Policy Research. 2005 [cited 10 Aug 2005]. Available from http://www.cahspr.ca/cabot/.
21. Helmer D. *Health technology assessment (HTA) information resources. Chapter 10: Grey literature* [monograph on the Internet]. 2004 Aug 2 [cited 10 Aug 2005]. Available from http://www.nlm.nih.gov/nichsr/ehta/chapter10.html.
22. Jasco P. Side-by-Side, Native Search Engines vs Google Scholar [Web page]. 2005 Apr 22 [cited 10 Aug 2005]. Available from http://www2.hawaii.edu/~jacso/.
23. Felter LM. Google Scholar, Scirus, and the scholarly search revolution. *Search Medford N J.* 2005;13(2):43–8.
24. Scirus – About us [Web page]. Elsevier. 2005 [cited 10 Aug 2005]. Available from http://www.scirus.com/srsapp/aboutus/.
25. Sullivan D. 2002 Search Engine Watch Awards [Web page]. Search Engine Watch. London: Incisive Interactive Marketing LLC. 2003 Jan 28 [cited 10 Aug 2005]. Available from http://searchenginewatch.com/awards/article.php/2155921#specialty.
26. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine [report on the Internet] [cited 10 Aug 2005]. Available from http://www-db.stanford.edu/~backrub/google.html.
27. Giustini D, Barsky E. Using Google Scholar in health research: comparisons with PubMed. CHLA / ABSC Conference, Toronto, 1 June 2005 [PowerPoint presentation]. 1 June 2005 [cited 10 Aug 2005]. Available from http://www.chla-absc.ca/2005/Presentations/0601/GiustiniBarsky_CHLA2005.pdf.
28. Florance V. Information in context: integrating information specialists into practice. *J Med Libr Assoc* 2002;90(1):49–58.
29. Regazzi JJ. The battle for mindshare: a battle beyond access and retrieval. 2004 Miles Conrad Memorial Lecture, 23 February 2004. NFAIS [cited 11 Aug 2005]. Available from http://www.nfais.org/publications/mc_lecture_2004.htm.

Appendix A

Content is the vaguest part of Google Scholar. Unfortunately, Google does not explicitly disclose its sources. Google Scholar content is a follow-up to the CrossRef Search Pilot project (http://www.crossref.org/) not-for-profit network with a mandate to make reference linking throughout online scholarly literature efficient and reliable.

CrossRef Pilot was initially limited to the content of 44 member publishers and societies (see the complete list below), who collaborate to provide scholars with cross-publisher reference linking. Google Scholar's 29 publishers are apparently a subset of this list. We were able to verify nine of these sources (in bold).

Alphamed Press
American Institute of Physics
American Physical Society
American Psychiatric Publishing
American Society for Biochemistry and Molecular Biology
American Society of Civil Engineers
Annual Reviews
Ashley Publications
Association for Computing Machinery
BioMed Central
Blackwell Publishing
BMJ Publishing Group
Cambridge University Press
Cold Spring Harbor Laboratory Press
EDP Science
FASEB
IEEE
INFORMS
Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic
Institute of Physics Publishing
International Union of Crystallography
Investigative Ophthamology and Visual Science
Institute of Pure and Applied Physics (IPAP)
Journal of Clinical Oncology
S. Karger AG
Lawrence Erlbaum Associates
Mary Ann Liebert
Medicine Publishing Group
Nature Publishing Group

Oldenbourg Wissenschaftsverlag
Oxford University Press
Peeters Publishers
PNAS
RILEM Publications SARL
Royal College of Psychiatrists
Springer-Verlag
Taylor & Francis
Thieme Publishing Group
University of California Press
University of Chicago Press
Vathek Publishing
John Wiley & Sons
Wolters Kluwer International Health & Science
The World Bank