Comparing test searches in PubMed and Google Scholar

Mary Shultz, MS, AHIP

See end of article for author's affiliations.

DOI: 10.3163/1536-5050.95.4.442

INTRODUCTION

Google Scholar has been met with both enthusiasm and criticism since its introduction in 2004. This search engine provides a simple way to access ''peer-reviewed papers, theses, books, abstracts, and articles from academic publishers' sites, professional societies, preprint repositories, universities and other scholarly organizations'' [1]. An obvious strength of Google Scholar is its intuitive interface, as the main search engine interface consists of a simple query box. In contrast, databases, such as PubMed, utilize search interfaces that offer a greater variety of advanced features. These additional features, while powerful, often lead to a complexity that may require a substantial investment of time to master. It has been observed that Google Scholar may allow searchers to ''find some resources they can use rather than be frustrated by a database's search screen'' [2]. Some even feel that ''Google Scholar's simplicity may eventually consume PubMed'' [3].

Along with ease of use, Google Scholar carries the familiar ''Google'' brand name. As Kennedy and Price so aptly stated, ''College students AND professors might not know that library databases exist, but they sure know Google'' [4]. The familiarity of Google may allow librarians and educators to ease students into the scholarly searching process by starting with Google Scholar and eventually moving to more complex systems. Felter noted that ''as researchers work with Google Scholar and reach limitations of searching capabilities and options, they may become more receptive to other products'' [5].

Google Scholar is also thought to provide increased access to gray literature [2], as it retrieves more than journal articles and includes preprint archives, conference proceedings, and institutional repositories [6]. Google Scholar also includes links to the online collections of some academic libraries. Including these access points in Google Scholar retrieval sets may ultimately help more users reach more of their own institution's subscriptions [7].

While its advantages are substantial, Google Scholar is not without flaws. The shortcomings of the system and its search interface have been well documented in the literature and include lack of reliable advanced search functions, lack of controlled vocabulary, and issues regarding scope of coverage and currency. Table 1 summarizes some of the reported criticisms of Google Scholar.

Vine found that while Google Scholar pulls in data from PubMed, many PubMed records are missing [20], and that Google Scholar also lacks features available in MEDLINE [12]. Others have noted that Google Scholar should not be the first or sole choice when searching for patient care information, clinical trials, or literature reviews [23, 24]. Thorough review and testing of Google Scholar, being an approach similar to that used to evaluate licensed resources, is necessary to better understand its strengths and limitations. As Jacso states, ''professional searchers must do sample test searches and correctly interpret the results to corroborate claims and get factual information about databases'' [18]. This paper compares and contrasts a variety of test searches in PubMed and Google Scholar to gain a better understanding of Google Scholar's searching capabilities.

METHODOLOGY

Ten searches were performed in PubMed using a variety of available search features. The searches were repeated in Google Scholar to approximate a user's approach to those same topics in that search engine. The searches, performed between August and September 2006, were by topic, author, title, journal name, and/or combinations of those fields (Appendix online). Topics included iron-deficiency anemia, bupropion for smoking cessation, and articles by specific authors in specific journals. The topics selected were loosely based on questions received during reference transactions or were previously developed for use during instruction.

For each search, the citations received via Google Scholar and PubMed were examined to determine a variety of characteristics including format, date, Medical Subject Headings (MeSH) where appropriate, uniqueness, duplications, and full-text availability from the author's institution.

Most searches were narrowed by date to produce sets of a reasonable size to allow comparison of unique items retrieved by each system. The search results were analyzed to determine possible reasons for the retrieval of unique items in each resource and to gather information on the general features of the Google Scholar results.

RESULTS

In eight of the ten searches, Google Scholar returned larger retrieval sets than PubMed (Table 2). Table 3 illustrates the characteristics of the items retrieved by Google Scholar, and Table 4 provides information on PubMed retrieval sets. Most items retrieved by Google Scholar were journal articles (Table 3). Items in other formats included: 9 books, 11 book reviews, 2 Web pages, 1 subject index listing, 1 thesis, 1 newsletter item, 1 bibliography, 4 author replies, 1 annual meeting abstract, and 1 draft document. These results yielded few gray literature items.

The main title link in Google Scholar citations was used to determine if full text was found. Full text was available in 46.96% (116/247) of the total citations retrieved. In most cases, it was assumed that full-text access was based on the institutional subscriptions

Supplemental Table 5 and an appendix are available with the online version of this journal.

Table 1
Criticisms of Google Scholar

Criticisms	References
Advanced search functions may be unreliable	[8–10]
No ability to search controlled vocabulary or no authority control for journal names or author names	[10–13]
Some materials retrieved may not be scholarly	[14]
Secretive about how it defines ''scholarly''	[15]
Secretive about scope or coverage	[5, 9, 10, 12, 13, 16–18]
May not be current	[9–11, 14, 19]
Missing PubMed records	[11, 20]
Lack of sorting options*	[10, 11, 14]
Inclusion of duplicate citations in results	[6, 14]
Only the first 1,000 results can be viewed	[4, 19, 21]
Not as comprehensive or precise as searching native interfaces	[9, 11, 22]
Lack of limiting features	[11, 12]

* In the spring of 2006, Google Scholar introduced an option to re-sort with more current citations appearing first.

available to the author of this study. Some items retrieved might have been freely available. In 22.67% (56/247) of the results, the Google Scholar citation was simply a link out to a PubMed record. As shown in Table 4, nearly half (48.98%; 72/147) of PubMed citations provided full-text access through the author's institution.

The unique items retrieved by each interface were examined to determine why they were missed by the other system. Across all searches, Google Scholar retrieved a total of 247 citations, 125 (50.61%) of which were unique to Google Scholar. Analysis revealed the following characteristics:

- Thirty-two items (12.96%) retrieved by Google Scholar were formats other than journal articles.
- Some unique Google Scholar items (10 items, 4.05%) appeared in journals not indexed by PubMed.
- Google Scholar covered a wider date range and returned 4 items (1.62%) older than 1950 that were not in PubMed.
- Google Scholar retrieved items based on its ability to search the full text of many articles rather than solely on citation data.

PubMed retrieved a total of 147 citations across all searches, and, of these, 46 (31.29%) were unique.

DISCUSSION

Assumptions of search engine performance based purely on retrieval quantities can be misleading without closer investigation of the results. For example, Table 2 shows that many of the searches returned quantities that were close in numbers. In search #1 (dietary supplements as a treatment for iron deficiency anemia), PubMed returned twenty-five citations, while Google Scholar returned twenty-six citations. However, only four citations were common to both systems. In search #2 (Mobius syndrome), Google Scholar returned eleven citations, while PubMed found ten citations but with an overlap of only two citations retrieved by both systems.

Table 2
Number of retrieved items

Search #	PubMed results	PubMed unique items	Google Scholar results	Google Scholar unique items
1	25	21	26	20
2	10	8	11	8
3	4	2	27	24
4	10	3	52	38
5	6	0	20	10
6	11	0	20	8
7	2	1	10	8
8	13	0	18	4
9	51	7	49	4
10	15	4	14	1

Terminology was observed to be a major factor affecting retrieval and the ability of both systems to return unique items. Some unique items retrieved by Google Scholar were off topic. These ''false hits'' appear to be related to Google Scholar's full-text searching along with a lack of controlled vocabulary. For example, the purpose of search #7 was to find articles on the topic of ''wine'' that appeared in the *New England Journal of Medicine.* Google Scholar retrieved eight items where the word ''wine'' appeared in the full text but was not the main topic of the article, in one case, retrieving an article where the authors acknowledge a colleague with the surname Wine. Google Scholar also returned items that contained the search terminology but did not match the intention of the search. In the search for information about dietary supplements in the treatment of iron deficiency (search #1), Google Scholar returned some citations about high iron stores rather than deficiency (Table 5 online). Google Scholar searches for a word or sequence of letters and not the concept or meaning.

The complete citations for all unique items retrieved by PubMed were examined. One possible explanation why Google Scholar failed to retrieve the same items was that many were indexed under the appropriate MeSH term, although the search phrase might not have appeared in the title or abstract. For example, search #9 was designed to retrieve articles by Visek about the topic of ammonia. While ammonia was not searched specifically as a MeSH term, PubMed automatically mapped it to MeSH. Of the unique citations retrieved by PubMed, some were indexed under ammonia although this term did not appear in the citation (Table 5 online). While Google Scholar offers the ability to use a tilde (~) to retrieve alternative termi-

Table 3
Characteristics of Google Scholar results

	Search numbers									
Google Scholar	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Total number of results	26	11	27	52	20	20	10	18	49	14
A. Journal article citations	23	11	22	42	17	12	9	18	48	13
B. Book citations	3	0	3	2	0	0	0	0	1	0
C. Book reviews	0	0	0	5	2	4	0	0	0	0
D. Other	0	0	2	3	1	4	1	0	0	1
E. Older than 1950	0	0	0	0	0	0	0	4	0	0
F. Duplicate in Google Scholar set	0	0	0	8	5	1	1	1	1	2
G. Also in PubMed set	4	2	2	13	10	12	2	14	45	13
H. Link to a PubMed record	4	0	2	7	1	0	0	9	31	2
I. If unique, item found in PubMed directly	16	8	17	24	9	0	7	0	2	0
J. Page not found or error	2	1	1	1	0	0	0	0	0	0
K. Google Scholar item title linked to full text	13	7	18	34	6	19	5	2	3	9
Total number of unique Google Scholar items	20	8	24	38	10	8	8	4	4	1

Note that item G will sometimes contain numbers greater than the retrieval set of PubMed. This occurred in cases in which Google Scholar returned duplicate citations that matched a single citation in PubMed. See online Appendix for search strategy details.

nology, this ability does not provide the control that subject headings do.

CONCLUSION

Performing a direct and exact comparison between searches in Google Scholar and PubMed is not possible as the systems function in very different manners. For example, PubMed searches a well-defined set of journals, while Google Scholar includes resources beyond journals and the exact scope of coverage is not extensively described. Because the systems are not searching identical data, the results are often different.

Although these two systems are difficult to compare, it is still important to explore the differences between them. Librarians should understand the strengths and weaknesses of Google Scholar and be prepared to explain them to their users [14]. It may also be wise to consider including Google Scholar in bibliographic instructional sessions and to convey how it compares to other search interfaces [11]. For example, Google Scholar does not offer the number and extent of special searching and limiting features available in PubMed. However, Google Scholar provides some advantages in that it is an easy place to begin a search to find an initial retrieval of possibly worthwhile articles. It also offers searchers the ability to find citations to older items that they would miss if they use only PubMed. Additionally, Google Scholar has the potential to provide access to the gray literature. This increased access to a part of the biomedical literature, which can be difficult to search, may have implications for the public health field [25].

One of the most advantageous features of searching PubMed is the ability to utilize the MeSH vocabulary, as Google Scholar does not currently implement controlled vocabulary searching mechanisms. MeSH provides a powerful method of narrowing results and homing in on what the searcher needs. PubMed also offers substantially more features that allow searchers to narrow their retrieval to citations from clearly identified sources, as detailed in NLM's List of Journals Indexed for MEDLINE and List of Serials Indexed for Online Users [26]. The problem faced today by searchers is not a lack of information but rather an overload of information. For a researcher conducting human studies, writing a dissertation, finding information pertinent to patient care, or conducting an in-depth literature review, Google Scholar does not appear to be a replacement for PubMed, though it may serve effectively as an adjunct resource to complement databases with more fully developed searching features. It is important to note that both PubMed and Google Scholar are often upgraded with new features or with intended improvement of existing functions. It may be worthwhile to repeat this study in one or two years to determine if further refinements have improved their performance.

ACKNOWLEDGMENTS

The author thanks the following individuals who offered invaluable advice and support: Pauline Cochrane, Robin

Table 4
Characteristics of PubMed results

	Search numbers									
PubMed	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Total number of results	25	10	4	10	6	11	2	13	51	15
Also in Google Scholar set	4	2	2	7	6	11	1	13	44	11
Item provides library link out to full text*	19	7	2	6	3	11	2	4	4	14
Total number of unique PubMed items	21	8	2	3	0	0	1	0	7	4

* Library link out icon was for author's institution. See online Appendix for search strategy details.

Beck, Sandra De Groote, AHIP, Victoria Pifalo, and Ann Carol Weller.

REFERENCES

1. Google Scholar help: about Google Scholar. [Web document]. Mountain View, CA: Google, 2005 [cited Aug 2006]. <http://scholar.google.com/intl/en/scholar/about.html>.
2. Kesselman M, Watstein SB. Google Scholar(tm) and libraries: point/counterpoint. Ref Serv Rev 2005; 33(4):380–7.
3. Abbasi K. Simplicity and complexity in health care: what medicine can learn from Google and iPod. J R Soc Med 2005 Sep;98(9):389.
4. Kennedy S, Price G. Big news: ''Google Scholar'' is born. [Web document]. Resource Shelf. UK: Free Pint Limited, 2004. [cited Aug 2006]. <http://www.resourceshelf.com/2004/11/18/wow-its-google-scholar/>.
5. Felter LM. Google Scholar, Scirus, and the scholarly search revolution. Searcher 2005 Feb;13(2):43–8.
6. Giles J. Science in the Web age: start your engines. Nature 2005 Dec 1;438(7068):554–5.
7. Notess GR. Scholarly Web searching: Google Scholar and Scirus. Online 2005 Jul–Aug;29(4):39–41.
8. Jacso P. Google Scholar (redux). [Web document]. Peter's Digital Reference Shelf, 2005. [cited Aug 2006]. <http://www.galegroup.com/reference/archive/200506/google>.
9. Jacso P. Google Scholar: the pros and the cons. Online Inform Rev 2005;29(2):208–14.
10. Burright M. Google Scholar: science & technology. [Web document]. Issues Sc Technol Libr 2006 Winter;45. [cited Aug 2006]. <http://www.istl.org/06-winter/databases2.html>.
11. Giustini D, Barsky E. A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations. J CHLA/J ABSC 2005;26:85–9.
12. Vine R. Google Scholar [electronic resources review]. J Med Libr Assoc 2006 Jan;94(1):97–9.
13. Myhill M. The ADVISOR reviews . . . Google Scholar. [Web document]. The Charleston ADVISOR 2005 Apr;6(4). [cited Sep 2006]. <http://www.charlestonco.com/review.cfm?id=225>.
14. Gardner S, Eng S. Gaga over Google? Scholar in the social sciences. Library Hi Tech News 2005;8:42–5.
15. Friend FJ. Google Scholar: potentially good for users of academic information. J Electronic Publishing [serial online]. 2006 Jan;9(1). [cited Sep 2006]. <http://hdl.handle.net/2027/spo.3336451.0009.105>.
16. Steinbrook R. Searching for the right search: reaching the medical literature. N Engl J Med 2006 Jan 5;354(1):4–7.
17. Wleklinski JM. Studying Google Scholar: wall to wall coverage? Online 2005 May–Jun;29(3):22–6.
18. Jacso P. As we may search: comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. Current Science 2005 Nov;89(9):1537–47.
19. Tennant R. Google, the naked emperor. Libr J 2005 Aug; 130(13):29.
20. Vine R. Google Scholar is a full year late indexing PubMed content. [Web document]. SiteLines, Feb 2005. [cited Aug 2006]. <http://www.workingfaster.com/sitelines/archives/2005_02.html#000282>.
21. Crawford W. Google: a company, not a religion. EContent 2005 Nov;28(11):42.
22. Jacso P. Side-by-side2, native search engines vs. Google Scholar. [Web document]. University of Hawai'i, 2005. [cited Aug 2006]. <http://www2.hawaii.edu/~jacso/scholarly/side-by-side2.htm>.
23. Henderson J. Google Scholar: a source for clinicians? CMAJ 2005 Jun 7;172(12):1549–50.
24. Giustini D. How Google is changing medicine. BMJ 2005 Dec 24;331(7531):1487–8.
25. Turner AM, Liddy ED, Bradley J, Wheatley JA. Modeling public health interventions for improved access to the gray literature. J Med Libr Assoc 2005 Oct;93(4):487–94.
26. National Library of Medicine. List of journals indexed for MEDLINE and list of serials indexed for online users terms and conditions. [Web document]. Bethesda, MD: The Library, 2007. [cited 29 Mar 2007]. <http://www.nlm.nih.gov/tsd/serials/terms_cond.html>.

AUTHOR'S AFFILIATION

Mary Shultz, MS, AHIP, shultz@uic.edu, Assistant Health Sciences Librarian, Library of the Health Sciences–Urbana, University of Illinois at Chicago, 102 Medical Sciences Building, 506 South Mathews Avenue, Urbana, IL 61801

Received October 2006; accepted May 2007

Characteristics of cancer blog users

Sujin Kim, PhD; Deborah S. Chung, PhD

See end of article for authors' affiliations.

DOI: 10.3163/1536-5050.95.4.445

INTRODUCTION

Blogs are a relatively new medium in computer-mediated health communication and are regarded as highly opinionated journals maintained by millions of users who read and write personal remarks on issues ranging from news stories to health care [1–3]. Of the 120 million US adults with Internet access, 7%, or 8 million people, have created blogs [4], and the increasing use of blogs has been reported in several studies [1, 4, 5]. Rainie found that the typical blogger is a young, male, Internet veteran; has a broadband connection; and is financially secure [5]. The gender of the blogger has also been a topic for research. Herring et al. found that even though women participate in blogging activities (focusing on emotional support), men are more likely to create filter blogs and k-logs (knowledge blogs) that are considered focused on information [1].

Blogs have been described as a new medium, one that shifts mainstream control of information into the hands of the audience. The potential use of blogs for cancer patients, basic scientists, clinical researchers, and practicing oncologists to discuss findings and suggestions has been envisioned in several cancer journals [6]. In addition, the use of online communication tools to share emotional support in all aspects of cancer-related issues has been frequently described [2, 6].

Figures 1 and 2 and a supplemental appendix are available with the online version of this journal.