Variation in number of hits for complex searches in Google Scholar

Wichor Matthijs Bramer, BSc

See end of article for author's affiliation.

DOI: http://dx.doi.org/10.3163/1536-5050.104.2.009

Objective: Google Scholar is often used to search for medical literature. Numbers of results reported by Google Scholar outperform the numbers reported by traditional databases. How reliable are these numbers? Why are often not all available 1,000 references shown?

Methods: For several complex search strategies used in systematic review projects, the number of citations and the total number of versions were calculated. Several search strategies were followed over a two-year period, registering fluctuations in reported search results.

Results: Changes in numbers of reported search results varied enormously between search strategies and dates. Theories for calculations of the reported and shown number of hits were not proved.

Conclusions: The number of hits reported in Google Scholar is an unreliable measure. Therefore, its repeatability is problematic, at least when equal results are needed.

Keywords: Search Engine, Reproducibility of Results, Review Literature as Topic, Information Storage and Retrieval

Google Scholar is frequently used as a database for biomedical searching, because of its wide and still increasing coverage [1]. However, its size and coverage remain unclear [2], especially when compared to more sophisticated databases such as PubMed or Embase [3]. Google Scholar often reports a very high number of hits, compared to other databases. Because it contains more references than other databases and it indexes the full text of articles, this is not surprising. However, because Google Scholar only shows the first 1,000 hits of any search and searches cannot be compared because of the lack of a search history, it remains unclear how this number is determined and whether this is indeed a reliable number. The fact that the number of hits is never more detailed than 3 figures suggests that the numbers are an estimate.

Since May 2013, the author has regularly used Google Scholar as an additional tool for searches for systematic reviews. An option provided by Google Scholar is to "Include Citations." Unticking this option excludes from the search results those articles that are known only as a citation in an indexed paper, so that only original articles are shown. I originally unticked this option, trying to increase the number of meaningful articles in the search results. However, the references shown by Google Scholar frequently would not reach the maximum of 1,000 references. The last pages were clickable but contained no references. An example of this is shown in Figure 1. I wanted to investigate what caused this phenomenon.

My assumption was that unticking the Include Citations box would just delete the citations from the 1,000 visible references and that the number of references shown would be 1,000 minus the number of citations. Many references in Google Scholar show a number of versions [4]. Google Scholar identifies equal references and groups them into 1 search result. My hypothesis was that the number of hits reported was the sum of the number of versions reported by the first 1,000 hits. I also wanted to investigate fluctuations in reported search numbers.

METHODS

At the moment of executing searches for systematic reviews, I meticulously copied the contents of

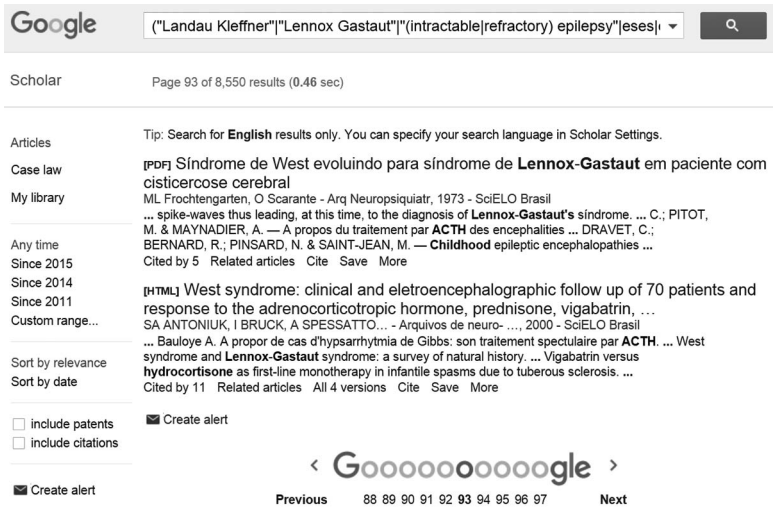

Figure 1

The last results of a Google Scholar search shown on page 93.
This page contains only 2 references; pages 94–100 are available but empty.

Google Scholar page by page into a MS Word document and performed several calculations (such as the total number of versions and the number of citations) on the contents of each document. I searched a small set of searches at least monthly for more than two years to record the development of the reported number of hits over a longer period of time.

RESULTS

At least monthly, I repeated a total of 32 separate searches that were unchanged in Google Scholar and documented the changes in the numbers of hits reported. Between June 2013 and January 2014, I observed the number of hits excluding citations (Figure 2), and between January 2014 and August 2015 those including citations (Figure 3). As shown in these figures, numbers of hits in Google Scholar for different queries sometimes showed great changes in short time periods. The first observation showed a dramatic decrease in number of hits for more than half of all citations between day 97 (September 18, 2013) and day 107 (September 28, 2013). The second observation did not show major changes until around the 300th day. Between October 17 and November 21, 2014, apparently major changes had been made to the search engine,

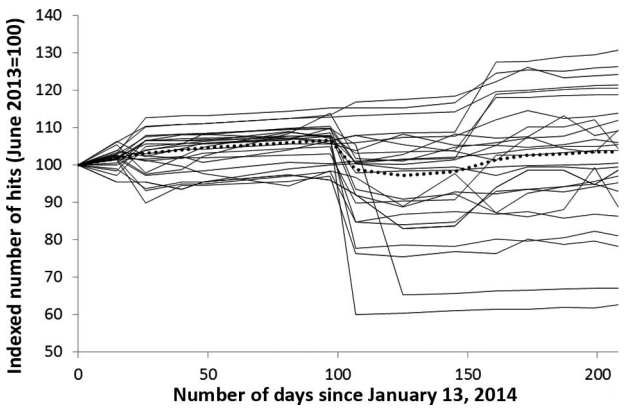

Figure 2

Changes in number of hits reported excluding citations by Google Scholar

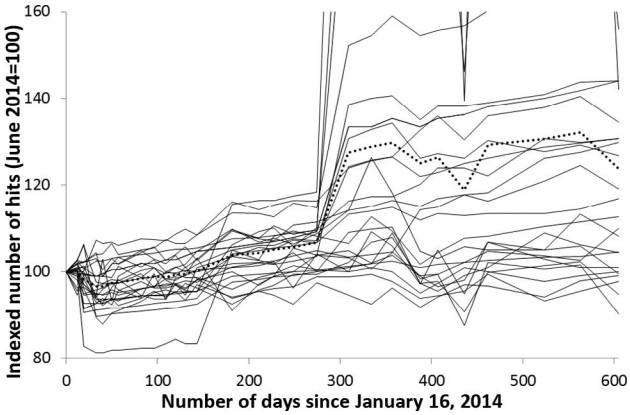

Figure 3.

Changes in number of hits reported including citations by Google Scholar

which caused some of the values to increase tremendously. One search query grew from an index of 118 to 312 (off the scale) in those 25 days, while others remained stable or even decreased.

The actual number of references shown when the option to exclude citations was chosen was registered for 98 different searches designed for systematic review projects between June 4, 2013, and February 7, 2014. For most queries, between 900 and 940 hits were viewable. No query retrieved 1,000 references, as the highest number observed was 996, but the lowest number retrieved was as low as 450.

For 68 of the aforementioned 98 queries, I registered not only the number of references excluding citations, but also the number of citations in the 1,000 references that I obtained when searching including citations was calculated. The sum of these variables ranged from 952 to 1,264, with a median of 1,014. The number of citations in the 1,000 references shown varied greatly. This ranged from 9 to 502, with a median of 127.

For 33 of the aforementioned 68 articles, the total of all reported versions of the first 1,000 references could be calculated. Though this varied greatly between the 33 observations (minimum: 1,808, maximum: 7,639, median: 4,486), the number of hits reported by Google Scholar varied even more (minimum: 631, maximum: 142,000, median: 16,500). Therefore, the ratio between the total numbers of citations varied from 0.3 to 25.8, with a median of 3.2.

DISCUSSION

The number of hits reported in Google Scholar, therefore, varied greatly, much more than it would in traditional databases. The relative changes in numbers of hits varied greatly between search strategies and seemed to be not related to an overall increase in coverage.

These changes can in part be explained by the difference in the nature of search engines and bibliographic databases. In traditional bibliographic databases, search strategies can typically be designed to retrieve all articles that meet certain criteria based on field codes. A search engine such as Google Scholar selects references matching text words, based on algorithms [5]. These algorithms change over time, often unexpectedly. Also the syntax that can be used changes from time to time; for example, the tilde (~), searching for synonyms, was recently discontinued.

Unticking the Include Citations button caused the number of shown references to drop below the regular 1,000. The missing number of hits did not equal the number of citations found in a search including citations. The total number of reported results was not equal to the total number of versions shown in the first 1,000 hits.

The observations of this study illustrate a frequently mentioned problem when searching Google Scholar for systematic searches. Repeatability with consistent results is impossible. The number of hits reported cannot be trusted as an accurate measurement.

REFERENCES

1. Harzing AW. A longitudinal study of Google Scholar coverage between 2012 and 2013. Scientometrics. 2014; 98(1):565–75.
2. Orduña-Malea E, Ayllón JM, Martín-Martín A, López-Cózar ED. About the size of Google Scholar: playing the numbers 2014. (Available from: http://arxiv.org/abs/1407.6239. [updated 5 Sep 2014; cited 13 Jan 2015].)
3. Bramer WM, Giustini D, Kramer BM, Anderson P. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. Syst Rev. 2013 Dec 23;2:115. PubMed PMID: 24360284. Pubmed Central PMCID: 3882110.
4. Pitol SP, De Groote SL. Google Scholar versions: do more versions of an article mean greater impact? Libr Hi Tech. 2014;32(4):594–611.
5. Hjørland B. Classical databases and knowledge organization: a case for Boolean retrieval and human decision-making during searches. J Assoc Inf Sci Tech. 2015;66(8):1559–75.

AUTHOR'S AFFILIATION

Wichor Matthijs Bramer, BSc, w.bramer@erasmusmc.nl, P.O. Box 2040, Erasmus MC–Erasmus University Medical Centre, 3000 CA Rotterdam, The Netherlands

Received August 2015; accepted November 2015