

Novel Citation-Based Search Method for Scientific Literature: A Validation Study

A. Cecile J.W. Janssens,*¹ Marta Gwinn¹, J. Elaine Brockman¹, Kimberley Powell², Michael Goodman¹

¹ Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta GA, USA.

² Woodruff Health Sciences Center Library, Emory University, Atlanta GA, USA.

*Correspondence to: Professor A. Cecile J. W. Janssens, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, Georgia 30322, USA. E-mail: cecile.janssens@emory.edu, Telephone: +1 404 727 6307.

ABSTRACT

Objective: We recently developed CoCites, a citation-based search method that is designed to be more efficient than traditional keyword-based methods. The method begins with identification of one or more highly relevant publications (query articles) and consists of two searches: the co-citation search, which ranks publications on their co-citation frequency with the query articles, and the citation search, which ranks publications on frequency of all citations that cite or are cited by the query articles.

Materials and Methods: We aimed to reproduce the literature searches of published systematic reviews and meta-analyses (n=250) and assess whether CoCites retrieves all eligible articles while screening fewer titles.

Results: CoCites retrieved a median of 75% of the articles that were included in the original reviews. The percentage of retrieved articles was higher (88%) when the query articles were cited more frequently and when they had more overlap in their citations. Applying CoCites to only the highest-cited article yielded similar results. The co-citation and citation searches combined were more efficient when the review authors had screened more than 500 titles, but not when they had screened less.

Discussion: CoCites uses the expert knowledge of authors to rank related articles. The method does not depend on keyword selection and requires no special expertise to build search queries. The method is transparent and reproducible.

Conclusion: CoCites is an efficient and accurate method for finding relevant related articles.

BACKGROUND

Finding relevant related articles on a specific topic is challenging and time-consuming [1], especially when there is no uniform set of keywords to describe the topic [2]. To quickly find who else has published on the exact topic of a paper, researchers have three options: perform a new literature search, follow the “related articles” link in databases such as PubMed, Web of Science (WOS), or Scopus; or trace the citations to and from the article.

Tracking citations is an intuitive strategy that allows finding articles on the same topic as authors tend to cite papers that are directly related to their work. The reference list of the query article, the so-called “backward citations”, and the newer articles that cite the query article, the “forward citations”, might both include relevant articles [3]. While intuitive, tracking citations is considered inefficient and inaccurate, even as a complement to keyword searching [4 5]. Tracking citations can only find articles that are connected in a single citation network [6]. A review of 259 meta-analyses, in which researchers aimed to retrieve all published articles on a specific topic, showed that this occurred in less than half (46%) of cases. In 39% of the meta-analyses, the articles were in two disconnected citation networks and in 15% of the meta-analyses in three or more networks [6].

We recently developed CoCites, a new search method that finds related articles for one or more articles of interest, termed ‘query articles [7].’ CoCites is based on the principle of co-citation [8]. Co-citation relationships between articles are used in methods that visualize the similarity and clustering among, e.g., articles, authors, and research topics [9-13], and have been previously proposed for literature search methods as well. To find relevant articles for one or more query articles, Belter proposed to screen their citing, cited, co-citing, co-cited articles [14], and to rank these articles based on their number of different relations with the query articles [15]. As each article has a maximum of three relations with a query article (citing/cited, co-citing, co-cited), this ranking method works better with more query articles. Others have examined whether the use of the proximity of (co-)citations within articles efficiently and effectively retrieves relevant articles [16-18]. And again others proposed to search related articles for one or more query articles by screening all citing, cited, co-citing and co-cited articles [19 20].

CoCites consists of two searches: one based on co-citations and another based on all citations (**Figure 1**). A co-citation is an article that is cited together with the query article(s). Thus a co-citation search identifies all co-cited articles ranks them in descending order of co-citation frequency [8]. The method assumes that articles with a higher co-citation frequency are more likely to address the same specific topic as the query article [8 21-23]. By contrast, a citation search finds and ranks all articles that cite or are cited by the query articles. As the maximum rank is dictated by the number of query articles, this citation search needs a larger query set to be effective.

In an earlier pilot study, we investigated the performance of CoCites by reproducing the literature searches of published meta-analyses. We found that the method was able to retrieve a median of 82% of the articles included in the meta-analysis. We observed that

the percentage of retrieved articles was higher when the two query articles were cited more frequently and when their topics were more similar [7].

In this article, we describe the results of a larger validation study in which we assess both completeness and efficiency of the CoCites searches. As in the pilot study, we tested CoCites' ability to reproduce the literature searches of published meta-analyses and systematic reviews. We investigated whether the method could retrieve all articles included in the reviews while screening fewer titles. We also assessed citation characteristics that impact the method's performance.

METHODS

Overview

Supplementary Figure 1 provides an overview of the project, which included several steps. We first obtained a random selection of published systematic reviews and meta-analyses (we refer to both as "reviews"). We then identified the articles that were included in the qualitative or quantitative analysis in each original review (referred to as "included articles") from which we selected the two mostly highly-cited papers, which were used as "query articles." Using a custom-designed web-based tool, we performed the co-citation search and screened the list of publications produced by that search ("screened titles") to retrieve the articles that were included in the original review ("retrieved articles"). Retrieved articles that had a co-citation frequency greater than a specified threshold (see analyses) were added to the next query set. We then performed another citation search using the updated query set and screened the new list of titles to retrieve the remaining articles included in the original review. **Supplementary Figure 2** illustrates step by step how the web tool works.

Selection of systematic reviews and meta-analyses

Systematic reviews and meta-analyses vary in rigor and quality. They may compare studies that address different research questions ("apples and oranges"), have insufficient search queries, or perform inadequate screening of articles. When evaluating performance of CoCites it is important to focus on the original reviews that meet minimum quality criteria because otherwise it is not clear if the disagreement between the two searches is attributable to the inadequacy of our method or the poor quality of the original review. Therefore, we retrieved systematic reviews and meta-analyses from WOS that cited the PRISMA or MOOSE reporting guidelines [24-26], mentioned "systematic review" or "meta-analysis" in the title, and were published in a journal with a 2015 Journal Impact Factor (Journal Citation Reports, Clarivate Analytics) of 2 or higher (**Supplementary Figure 3**). Although the last criterion is arbitrary, it allowed focusing on reviews with higher impact and presumably higher quality. We sorted the reviews on their WOS Accession number and selected the top 500 (search date: September 23, 2016).

We noticed that three journals published an exceptionally high number of reviews (Medicine, Scientific Reports, and PLoS One), which led us to limit the number of reviews per journal to a maximum of 3. We only considered reviews that had 1) evaluated the quality of the included articles; 2) reported the numbers of screened and included articles in a flowchart, and 3) reported the sample sizes of all included studies. This information was required for a sub-study investigating the impact of missing data on meta-analyses results. From the reviews that met the above criteria, we further excluded those that had inconsistencies in the references (information in main text not matching reference list), re-used the articles from an earlier, already included review, or included fewer than five articles. A total of 250 reviews were selected for analysis. All full-text files and supplementary documents were downloaded and stored.

Retrieval of included articles and selection of highest-cited articles

We downloaded bibliographic data and the reference list for each review. In WOS, the articles in the reference list are stored under a short unique identifier. We extracted the unique identifiers for all references in all reviews, removed duplicates, and downloaded bibliographic data for each article from WOS (date of download: April 25, 2017). In addition to the information on the first author, journal, and publication year, data on each article included the PubMed identification number (PMID) and the number of citations (Times Cited). PMIDs were used as an indicator of whether an article could have been found through a PubMed/Medline search or whether it was likely retrieved through other databases. Missing PMID values were hand-searched in PubMed.

For each review, we documented the end-of-search date (if pertinent) and the start date for the search period, the number of articles screened (after removal of duplicates) and the number of articles included in the qualitative or quantitative analysis in the review. We also identified the included articles in the downloaded reference lists.

The two most highly cited articles in each review were identified based on the number of citations at the date the authors had performed their search. We programmed a web-based tool that automatically extracted the citations for each included article in each review and counted the number of citations that were published before the search date reported in the review. The two articles with the highest numbers of citations at the review search date were selected as query articles.

Application of CoCites

The strategy used to develop CoCites' co-citation and citation searches has been described previously [7] and is diagrammed in **Figure 1**. We use a custom-designed, web-based tool to perform the searches automatically (**Supplementary Figure 2**), and retrieve data from WOS through its application programming interface (API). For the co-citation search, the tool extracts the reference lists of all unique publications that cite the query articles, counts the number of times each publication appears in all reference lists and ranks them in

descending order of co-citation frequency. For the citation search, the program extracts and counts all publications that cite or are cited by the query articles and ranks them in descending order of citation frequency. The removal of duplicates in the co-citation search and the counting of frequencies is based on each article's unique identifier in the WOS database.

The WOS database includes indexed and non-indexed items. Non-indexed items are those that would not have been included in the database had they not been cited by an indexed article. Examples of non-indexed items include dissertations, reports, and articles in journals that are not covered by WOS. The non-indexed items are available in the WOS database only as cited references and include limited metadata. As their reference lists are not accessible, non-indexed articles are only retrieved when they appear frequently enough in the reference lists of the papers that cite the query articles (co-citation search) or in the reference lists of the query articles themselves (citation search). As all articles included in each review should at least be cited by the review, non-indexed articles are the ones with a missing 'Times Cited' count (see below).

Analyses

We quantified the performance of the search method for four different screening thresholds: (1) articles co-cited at least once (threshold ≥ 1 , i.e., with no exclusions); (2) articles co-cited more than once (threshold > 1); (3) articles co-cited more than once *and* found in more than 1% of the citing publications; and (4) articles that were among the top 100 of all co-cited publications. The choice of these thresholds was based on the pilot study [7], in which the '1%' threshold was investigated to reduce the number of titles needed to screen for highly-cited query articles. The total number of screened articles is the sum of items from the combined co-citation and citation searches; we were unable to reliably remove duplicate records as the database returned the results of the two searches in different formats. When reviews had a start search date of 1980 or later, we excluded earlier publications from our search results for a fair comparison of the number of screened articles.

For both the co-citation and citation searches, we calculated at each of the four thresholds 1) the percentage of articles in the original review that were retrieved using CoCites and 2) the number of titles that needed to be screened to identify eligible articles. The percentage of retrieved articles was calculated excluding the two query articles as these were known at the start of the search. We repeated these assessments for a subset of searches characterized by the high degree of similarity between the query articles. We quantified the similarity between the query articles using Simpson's similarity index [27]. This index measures the degree of co-citation between two articles as their number of co-citations divided by the number of citations of the less-cited article. For example, if two query articles are cited 10 and 20 times each, but only three times together, then the similarity score is $3/10=0.3$. A score of 0.3 means that the two query articles are co-cited in

30% of the citations of the least-cited query article. In our similarity score, the numerator was tied to the search date reported in the review, while the denominator was obtained from the bibliographic download.

RESULTS

Description of the reviews

The 250 reviews included a total of 4,761 articles. The authors of the original reviews screened from 18 to 85,714 articles (median 794; **Table 1**) and included from 5 to 85 articles (median 14). Researchers screened fewer than 200 titles in 17% of reviews, fewer than 500 in 38%, and more than 2,000 in 27%.

Searching on two highest-cited articles

The query articles had a median of 160 unique citations (inter-quartile range, IQR 80, 262). In 97% of reviews, the query articles were cited more than 20 times and in 68% more than 50 times.

The two highest-cited articles for each review had from 124 to 52,596 co-cited articles (median 5,151; **Table 1**). A median of 1,139 articles was co-cited more than once, and in 696 instances they were co-cited more than once *and* in more than in 1% of the citing articles. The co-citation and citation searches combined involved screening a median of 873 articles, which was higher than the median number of articles screened by the authors in their reviews. **Figure 2** shows that the two searches combined were less efficient when the review authors had screened fewer than 500 articles, but more efficient when they had screened more.

Co-citation searching retrieved a median of 57% (IQR 40, 75) of the articles at the "1%" threshold; co-citation and citation searching combined retrieved a median of 75% (IQR 50, 90; **Table 1**). Overall, 38% of retrieved articles were among the 100 top-ranked articles in the co-citation search.

Factors affecting percentage of retrieved articles

Table 2 shows that the percentage of retrieved articles was higher when the query articles were cited more frequently and when their citations overlapped more (as indicated by the higher similarity score, see methods). The combined searches retrieved a median of 83% of the articles when all articles in the review were in PubMed versus 62% when fewer than 90% were in PubMed, suggesting that the method was less likely to retrieve articles that were obtained through other databases.

Table 2 shows that in most reviews the query articles were cited more than 20 times and the similarity score between the query articles was greater than 0.2. **Table 3** shows the percentage of retrieved articles for combinations of these criteria. When the number of citing articles was higher than 20, *and* the similarity score was higher than 0.2,

CoCites retrieved a median of 80% of all included articles and 88% of included articles when all were in PubMed.

Even under these favorable conditions (**Table 3**, row 1), the method did not always work. We examined the citation characteristics of the twelve reviews in which CoCites retrieved fewer than 50% of the included articles (**Supplementary Table 1**). We observed that in 10 of 12 the reviews, co-citation searching retrieved 5 or fewer articles—too few to constitute an effective query set for the citation search. For 9 of these articles, citation searching retrieved no new publications. We explored whether using the 25 top-ranked results from the co-citation search could be effective as query articles for the citation search and found that the percentage of retrieved articles improved for all reviews. In 10 out of 12 reviews, more than 50% of included articles were retrieved, and in 6 of these, more than 75% were retrieved.

When CoCites failed to retrieve the articles that were included in the review, it nevertheless returned articles on the same topic as the query articles. **Supplementary Table 2** shows the titles of the 25 top-ranked results for each of the five reviews in which the method performed worst, retrieving only 9-13% of the included articles. It is clear from the titles that most articles are on similar topics as the query articles.

Searching on the highest-cited article

For searches with high similarity scores (Table 3, row 1), we further examined performance of the method starting only with the highest-cited article. The co-citation search retrieved a median of 63.3% (IQR 45.3, 80.0) and the combined searches retrieved 86.7% (IQR 62.6, 100), results similar to those found when using the two highest-cited articles. (66.7% and 87.5%, respectively). Using only the highest-cited query article reduced the median number of screened articles from 873 (IQR 540, 1,204) to 813 (IQR 394, 1,165).

Factors affecting retrieval of individual articles

Frequently cited articles were more likely to be retrieved by co-citation searching (Table 4); half of the infrequently-cited articles could be retrieved using citation searching. Articles that had never been cited could be retrieved using citation searching when they cited two or more articles that had already been retrieved by co-citation searching; in keeping with this finding, articles with more references were more likely to be retrieved. Finally, as expected, recently published articles were rarely retrieved because accumulating co-citations requires sufficient time since publication.

Finding the highly-cited articles

In this study, we assumed that researchers were able to retrieve the two-highest cited articles either because they are familiar with the topic or because an initial search has led them to these two articles. We explored whether these articles could be identified using

co-citation searches that started with two query articles that had been cited less frequently but at least ten times each. For this analysis, we restricted the subset of reviews with high similarity scores (**Table 3, row 1**) to those that included 10 or more articles, from which we selected the two articles with the fewest citations but at least ten citations each. We obtained the ranks of the two highest-cited articles and calculated how frequently they and other included articles appeared among the top-ranked results.

The query articles were cited by a median of 25 articles (IQR 21, 35), which was markedly lower than the median of 160 when the two highest-cited articles were used. For all but one review, at least one of the two highest-cited articles was found among the 100 top-ranked results; in 68% of the reviews, both articles were in the top 100 (**Supplementary Table 3**). The ten top-ranked results included one (72%) or both (37%) of the highest-cited articles. In all but one review, the 100 top-ranked articles retrieved multiple other articles that were cited more frequently than the two query articles. In 34 reviews, the top 50 results retrieved 5 or more articles that were cited more frequently than the query articles.

DISCUSSION

In a well-defined, randomly selected sample of reviews, our citation-based search method retrieved a median of 75% of the included articles. The method performed better when the query articles were more similar and more frequently cited. CoCites' co-citation and citation searches combined retrieved 88% of included articles when all were in PubMed. In a subset of reviews with high similarity scores, the highest-cited articles could be retrieved when co-citation searching was based on less frequently cited articles.

Before discussing the implications of our findings, several methodological issues should be mentioned. First, we assumed that articles in the reviews were correctly included and excluded; however, it is possible that CoCites missed articles that should not have been included in the review and retrieved relevant articles that the authors had missed. If erroneous inclusion of articles is common, the performance of CoCites is underestimated.

And second, the authors of original reviews often utilized multiple sources, including foreign and specialty databases, conference proceedings, dissertations, and personal communications. The sources might yield articles that cannot be found through PubMed, WOS, or other major English-language literature databases. Therefore, it is not realistic to expect a 100% retrieval. When CoCites is used to find relevant articles for systematic reviews and meta-analyses, these additional sources may still need to be searched, if the topic so requires.

In line with observations from our pilot study [7], we found that co-citation searching might not retrieve articles that are infrequently cited. These articles are more likely to include abstracts, letters, articles in non-English languages, and very old articles, reports, and theses that may not be indexed in WOS or other databases. Infrequently cited

papers also include articles in WOS that were published too recently to be cited or that authors did not consider worth citing. Citation searching can retrieve such articles when they cite relevant articles in the query set. Further research is needed to assess the performance of the method in emerging and heterogeneous topics.

There is still room for improvement of the method. First, when query articles were on similar topics and cited frequently enough, CoCites retrieved fewer than 50% of the included articles in 12 reviews. The data in **Supplementary Table 1** showed that this was explained by the fact that the co-citation search did not retrieve enough articles for the citation search to be effective. When we applied the citation search to the 25 top-ranked co-citation search results instead (without screening relevant articles first), the percentage of retrieved articles increased substantially. These results warrant validation in a larger study.

Second, for highly-cited query articles, we limited the number of titles needed to screen by requiring that titles in the results needed to be cited in 1% of the articles that cite the query articles. Yet, the data in **Table 2** shows that the percentage of retrieved articles decreased when the query articles were cited more than 200 times, suggesting that the '1%' threshold may not be optimal. An alternative and reproducible strategy is to only use the, say, 100 most recent articles that cite the query articles. As the CoCites method is inefficient for highly-cited query articles, it is worth exploring alternative strategies to limit the number needed to screen.

CoCites is useful for searches that aim to find related articles on a topic. Starting with one article, a CoCites search retrieves others that can be used to repeat the search or can be added to the query set. Although the method is relevant to any systematic review, it may be especially useful when the aim is to find related articles on a niche topic or identify the key, highest-cited publications. These key publications are easily retrieved among the top-ranked articles found by a co-citation search starting with query articles with fewer citations. Repeated iterations of the co-citation search, each time with the most highly cited relevant articles, will eventually reveal the highest-cited articles on the niche topic.

Conclusions

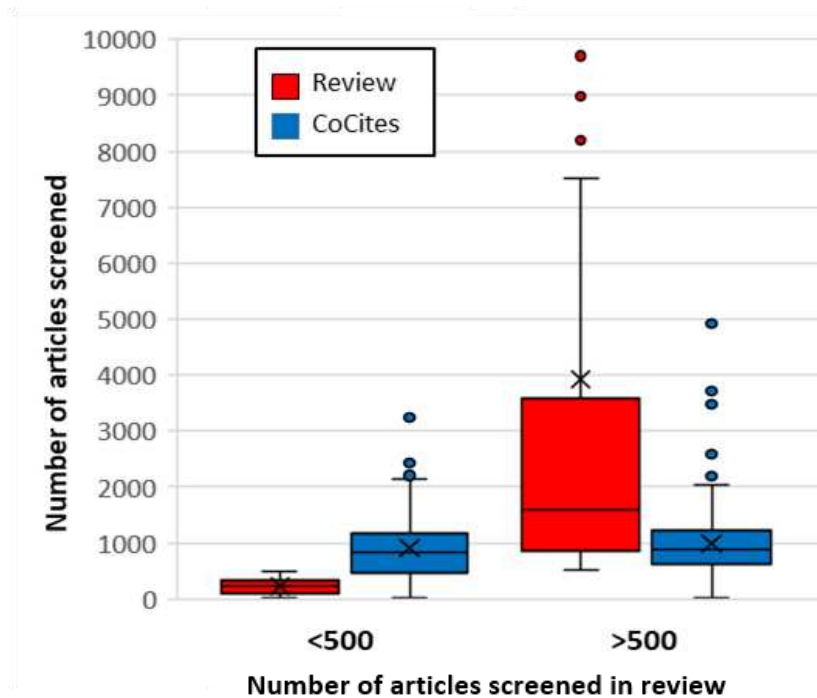
CoCites is a novel method of searching scientific literature that retrieves related articles on well defined, specific topics. The method is effective and efficient and does not require expertise in building search queries. The method is transparent and reproducible. Co-citation searching has the potential to improve the quality and reduce the time of literature searches.

Reference list

1. Bullers K, Howard AM, Hanson A, et al. It takes longer than you think: librarian time spent on systematic review tasks. *Journal of the Medical Library Association* 2018;**106**(2):198-207 doi: 10.5195/jmla.2018.323[published Online First: Epub Date]].
2. Simon M, Hausner E, Klaus SF, Dunton NE. Identifying nurse staffing research in Medline: development and testing of empirically derived search strategies with the PubMed interface. *Bmc Medical Research Methodology* 2010;**10** doi: 10.1186/1471-2288-10-76[published Online First: Epub Date]].
3. Hu X, Rousseau R, Chen J. On the definition of forward and backward citation generations. *Journal of Informetrics* 2011;**5**(1):27-36 doi: 10.1016/j.joi.2010.07.004[published Online First: Epub Date]].
4. Horsley T, Dingwall O, Sampson M. Checking reference lists to find additional studies for systematic reviews. *Cochrane Database of Systematic Reviews* 2011(8) doi: 10.1002/14651858.MR000026.pub2[published Online First: Epub Date]].
5. Wright K, Golder S, Rodriguez-Lopez R. Citation searching: a systematic review case study of multiple risk behaviour interventions. *BMC Med Res Methodol* 2014;**14**:73 doi: 10.1186/1471-2288-14-73[published Online First: Epub Date]].
6. Robinson KA, Dunn AG, Tsafnat G, Glasziou P. Citation networks of related trials are often disconnected: implications for bidirectional citation searches. *J Clin Epidemiol* 2014;**67**(7):793-9 doi: 10.1016/j.jclinepi.2013.11.015[published Online First: Epub Date]].
7. Janssens AC, Gwinn M. Novel citation-based search method for scientific literature: application to meta-analyses. *BMC Med Res Methodol* 2015;**15**:84 doi: 10.1186/s12874-015-0077-z[published Online First: Epub Date]].
8. Small H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *J. Am. Soc. Inf. Sci.* 1973;**24**(4):265-69
9. Barlow P, McKee M, Basu S, Stuckler D. The health impact of trade and investment agreements: a quantitative systematic review and network co-citation analysis. *Globalization and Health* 2017;**13**(1) doi: 10.1186/s12992-017-0240-x[published Online First: Epub Date]].
10. White HD, McCain KW. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *J. Am. Soc. Inf. Sci.* 1998;**49**(4):327-55 doi: Doi 10.1002/(Sici)1097-4571(19980401)49:4<327::Aid-Asi4>3.0.Co;2-W[published Online First: Epub Date]].
11. Jankovic MP, Kaufmann M, Kindler CH. Active research fields in anesthesia: A document co-citation analysis of the anesthetic literature. *Anesth. Analg.* 2008;**106**(5):1524-33 doi: DOI 10.1213/ane.0b013e31816d18a1[published Online First: Epub Date]].
12. van Eck NJ, Waltman L. Appropriate similarity measures for author co-citation analysis. *Journal of the American Society for Information Science and Technology* 2008;**59**:1653-61
13. Trujillo CM, Long TM. Document co-citation analysis to enhance transdisciplinary research. *Sci Adv* 2018;**4**(1) doi: ARTN e1701130. 10.1126/sciadv.1701130[published Online First: Epub Date]].
14. Belter CW. Citation analysis as a literature search method for systematic reviews. *Journal of the Association for Information Science and Technology* 2016;**67**(11):2766-77 doi: 10.1002/asi.23605[published Online First: Epub Date]].

15. Belter CW. A relevance ranking method for citation-based search results. *Scientometrics* 2017;**112**(2):731-46 doi: 10.1007/s11192-017-2406-y[published Online First: Epub Date]].
16. Gipp B, Beel J. Citation Proximity Analysis (CPA) - A New Approach for Identifying Related Work Based on Co-Citation Analysis. *Pro Int Conf Sci Inf* 2009;**2**:571-75
17. Eto M. Evaluations of context-based co-citation searching. *Scientometrics* 2013;**94**(2):651-73 doi: 10.1007/s11192-012-0756-z[published Online First: Epub Date]].
18. Callahan A, Hockema S, Eysenbach G. Contextual Cocitation: Augmenting Cocitation Analysis and its Applications. *Journal of the American Society for Information Science and Technology* 2010;**61**(6):1130-43 doi: 10.1002/asi.21313[published Online First: Epub Date]].
19. Larsen B. Exploiting citation overlaps for Information Retrieval: Generating a boomerang effect from the network of scientific papers. *Scientometrics* 2002;**54**(2):155-78 doi: 10.1023/a:1016011326300[published Online First: Epub Date]].
20. Hinde S, Spackman E. Bidirectional Citation Searching to Completion: An Exploration of Literature Searching Methods. *Pharmacoeconomics* 2015;**33**(1):5-11 doi: 10.1007/s40273-014-0205-3[published Online First: Epub Date]].
21. Ortuno FM, Rojas I, Andrade-Navarro MA, Fontaine J-F. Using cited references to improve the retrieval of related biomedical documents. *Bmc Bioinformatics* 2013;**14** doi: 10.1186/1471-2105-14-113[published Online First: Epub Date]].
22. Boyack KW, Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology* 2010;**61**(12):2389-404 doi: 10.1002/asi.21419[published Online First: Epub Date]].
23. Linder SK, Kamath GR, Pratt GF, Saraykar SS, Volk RJ. Citation searches are more sensitive than keyword searches to identify studies using specific measurement instruments. *Journal of Clinical Epidemiology* 2015;**68**(4):412-17 doi: 10.1016/j.jclinepi.2014.10.008[published Online First: Epub Date]].
24. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;**283**(15):2008-12
25. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 2009;**62**(10):1006-12 doi: 10.1016/j.jclinepi.2009.06.005[published Online First: Epub Date]].
26. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;**339**:b2700 doi: 10.1136/bmj.b2700[published Online First: Epub Date]].
27. Simpson GG. Mammals and the nature of continents. *American Journal of Science* 1943;**241**:1-31

Figure 2. Numbers of articles screened in published reviews versus numbers screened by CoCites' co-citation and citation searches



Compared for reviews in which authors screened fewer than 500 articles (left, n=95) and more than 500 articles (right, n=155).

Table 1 Articles screened and retrieved in replicating the results of literature searches in 250 published reviews

	Articles screened, number	Articles retrieved, percentage
In published review	794 (273, 2,132)	100
<i>Co-citation search</i>		
All co-cited articles	5,151 (2,709, 10,490)	75.0 (58.2, 87.5)
Co-cited > 1	1,119 (544, 2,509)	60.0 (45.2, 78.3)
Co-cited > 1%*	696 (461, 978)	56.1 (40.0, 75.0)
100 Top-ranked**	109 (103, 123)	37.5 (22.5, 50.0)
<i>Citation search</i>		
Citing or cited by > 1	83 (38, 176)	50.0 (17.9, 75.8)
Total	873 (540, 1,204) ***	75.0 (50.0, 90.1)

All values are median and inter-quartile range (IQR). *Co-cited more than once and in more than 1 percent of the citing articles. The articles retrieved from this search were used to run the citation search. ** Median is higher than 100 because we included all articles that had the same co-citation frequency as the 100th article. *** Sum of results in the co-citation 'co-cited>1%' and citation searches combined, without removing duplicates. See details in methods.

Table 2 Factors that influence the percentage of articles included in each review that were retrieved by co-citation and citation searches combined.

	Number of reviews	Percentage of articles retrieved Median (IQR)
Number of citing articles		
<20	7	37.5 (28.6, 60.0)
20-50	26	63.3 (39.4, 85.7)
50-100	48	65.0 (50.0, 87.5)
100-200	74	80.0 (50.0, 90.1)
>200	95	77.6 (47.4, 93.8)
Similarity index*		
<0.1	19	34.4 (15.4, 56.3)
0.1-0.2	26	56.8 (36.1, 72.3)
0.2-0.5	109	75.0 (51.7, 92.3)
>=0.5	96	83.3 (60.4, 94.4)
Percentage of articles in PubMed		
<90	64	62.4 (41.3, 80.3)
90-100	64	67.1 (49.4, 84.5)
100	122	82.8 (59.6, 100.0)

* Similarity index = number of co-citations between query articles / number of citations of the less-cited query article. IQR = inter-quartile range.

Table 3 Percentage of retrieved articles by the number of citing articles and similarity index

Number of citing articles	Similarity index*	Percentage of articles in PubMed	Number	Percentage retrieved, median
>20	>0.2	100	101	87.5 (68.0, 100.0)
>100	>0.2	100	70	90.8 (77.6, 100.0)
>20	>0.5	100	52	87.5 (77.4, 100.0)
>100	>0.5	100	32	91.9 (82.2, 100.0)
>20	>0.2	>90	150	83.3 (64.1, 96.5)
>100	>0.2	>90	111	85.7 (66.7, 96.7)
>20	>0.5	>90	71	87.5 (69.2, 95.2)
>100	>0.5	>90	47	87.5 (76.9, 95.2)
>20	>0.2	All	200	80.0 (60.0, 94.1)
>100	>0.2	All	139	81.2 (64.3, 95.2)
>20	>0.5	All	94	84.5 (61.5, 94.4)
>100	>0.5	All	59	86.8 (66.7, 94.4)

* Similarity index = number of co-citations between query articles / number of citations of the less-cited query article. IQR = inter-quartile range

Table 4 Factors that affect retrieval of individual articles

	Total	Co-citation search		Co-citation + citation searches	
		Number retrieved	Percentage*	Number retrieved	Percentage*
Overall	4261	1938	45.5	2674	62.8
Times cited					
Not indexed	591	157	26.6	181	30.6
0	365	1	0.3	176	48.2
1-5	787	104	13.2	385	48.9
6-9	411	150	36.5	237	57.7
10-19	658	406	61.7	480	72.9
20-49	862	616	71.5	691	80.2
>50	587	504	85.9	524	89.3
Number of references					
Not indexed	591	157	26.6	181	30.6
<5	97	44	45.4	45	46.4
5-9	108	65	60.2	75	69.4
>=10	3465	1672	48.3	2373	68.5
Years since publication					
0-1	544	44	8.1	285	52.4
1-2	487	118	24.2	243	49.9
2-5	1081	402	37.2	606	56.1
5-10	994	585	58.9	686	69.0
>10	1155	789	68.3	854	73.9

* Percentage of articles included in reviews that were retrieved, by category. Query articles for all 250 reviews (n=500) were removed from the dataset.