CASE REPORT

DOI: dx.doi.org/10.5195/jmla.2022.1117

The Citation Cloud of a biomedical article: a free, public, web-based tool enabling citation analysis

Neil R. Smalheiser; Jodi Schneider; Vetle I. Torvik; Dean P. Fragnito; Eric E. Tirk

See end of article for authors' affiliations.

Background: An article's citations are useful for finding related articles that may not be readily found by keyword searches or textual similarity. Citation analysis is also important for analyzing scientific innovation and the structure of the biomedical literature. We wanted to facilitate citation analysis for the broad community by providing a user-friendly interface for accessing and analyzing citation data for biomedical articles.

Case Presentation: We seeded the Citation Cloud dataset with over 465 million open access citations culled from six different sources: PubMed Central, Microsoft Academic Graph, ArnetMiner, Semantic Scholar, Open Citations, and the NIH iCite dataset. We implemented a free, public extension to PubMed that allows any user to visualize and analyze the entire citation cloud around any paper of interest A: the set of articles cited by A, those which cite A, those which are co-cited with A, and those which are bibliographically coupled to A.

Conclusions: Citation Cloud greatly enables the study of citations by the scientific community, including relatively advanced analyses (co-citations and bibliographic coupling) that cannot be undertaken using other available tools. The tool can be accessed by running any PubMed query on the Anne O'Tate value-added search interface and clicking on the Citations button next to any retrieved article.

Keywords: citation analysis; bibliometrics; information retrieval; evidence based medicine; science of science

See end of article for supplemental content.

BACKGROUND

Citation analysis is crucial for tracing the diffusion of knowledge across disciplines and over time, both at the micro level (individual citations) and macro level (global citation networks). For example, one may wish to follow citation chains (e.g., identifying the influence of a retracted article on later citing papers) [1, 2]. Also, Hutchins et al. employed citation patterns to predict which articles are likely to contribute to the translation of basic studies into clinical advances [3], and Boyack and Klavans employed citations to identify research frontiers [4].

Citation analysis has largely been the province of scholars in the specialties of bibliometrics, scientometrics, and innovation and policy studies, who typically carry out extensive, time-consuming analysis of proprietary citation data licensed by commercial data providers. Many members of the scientific community may not take advantage of citation analysis to find relevant articles because it can involve the use of commercial databases for which they may not have access. Recently, iCite, an extensive set of open citations in the biomedical literature, was publicly released [5] with a monthly updated dataset (https://icite.od.nih.gov/). This provides a great opportunity for biomedical investigators and other interested parties, but, to date, there is no user-friendly interface for accessing or analyzing the citation data. Here, we describe Citation Cloud, a free, public extension to PubMed that allows any user to visualize and analyze the citation cloud around any article A: the set of articles cited by A, those which cite A, those which are co-cited with A, and those which are bibliographically coupled to A.

To say that an article B is co-cited with A means that they are both cited by the same article(s) C_i [6]. Co-citation is a measure of similarity not directly based on textual or topical similarity. Note that the co-citation relationship is not fixed but can vary over time depending on how many newer articles cite both A and B. According to Small, "It appears that an interpretation of the significance of strong co-citation links must rely both on the notion of subject similarity and on the association or co-occurrence of ideas [. . .]. If it can be assumed that frequently cited papers represent the key concepts, methods, or experiments in a

DOI: dx.doi.org/10.5195/jmla.2022.1117

field, then co-citation patterns can be used to map out in great detail the relationship between these key ideas" [6].

In contrast, to say that an article B is bibliographically coupled to A means that they both cite some of the same articles C_i in their reference lists [7]. In other words, the reference lists of the two articles overlap, and the larger the number of shared references, the greater the degree of bibliographical coupling. This is also a measure of similarity that is not directly based on textual or topical similarity, although the fact that two bibliographically coupled papers share references suggests that they are also likely to share some methods, ideas, or topics. The bibliographically coupled relationship has the advantage that it can be calculated for any two articles regardless of when they were published. Also, this relationship is stable and will not change over time.

The new open access citations datasets can potentially enable a broad community of scientists to utilize citations in their studies of biomedical literature—not only the simple cites and cited by relationships but also the more sophisticated co-citation and bibliographically coupled relationships. Toward this end, we created a free, public, web-based tool for PubMed articles.

CASE PRESENTATION

The Anne O'Tate tool for searching PubMed was originally described in 2008 [8], with major additions described in 2021 [9], and provides a suite of tools that help summarize, mine, and drill down the results of a PubMed query. The Citation Cloud tool is an enhancement that provides four citation analyses for a record retrieved from PubMed using the Anne O'Tate tool. The Citation Cloud can be accessed by running any query on the Anne O'Tate value-added PubMed search interface (http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi) [8, 9] and clicking on the Citations button next to any retrieved article. For example, suppose we enter the query "Retractions in the medical literature: how many patients are put at risk by flawed research?" [1] to retrieve this single article (Figure 1). We then click on the Citations button next to the article and see its citation cloud visualization in a new tab (Figure 2).

Figure 1 Screenshot of a PubMed query entered via the Anne O'Tate tool. Shown is the article retrieved using the title in the query box. The hyperlinked word Citations is displayed to the right of the article.

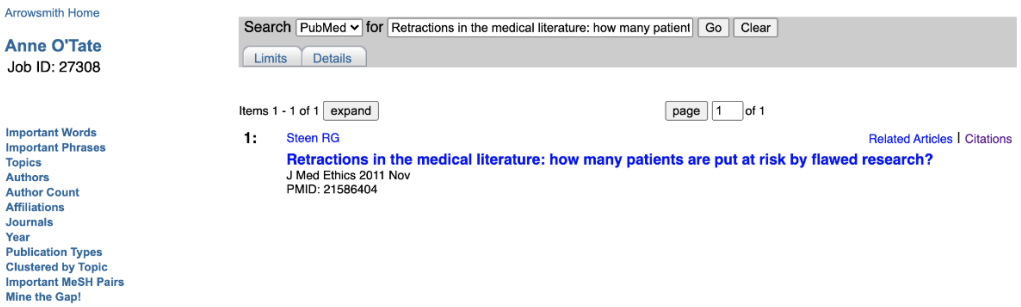
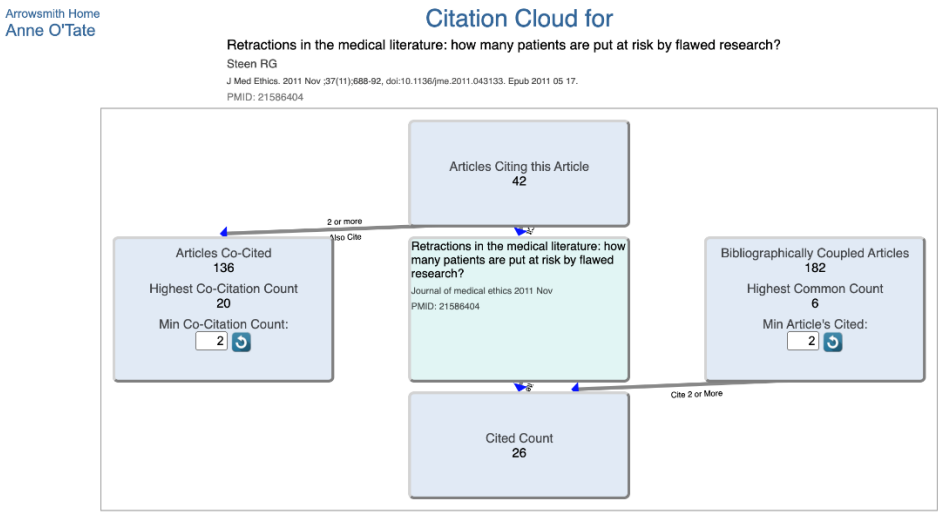


Figure 2 The citation cloud visualization for the article displayed in Figure 1.



DOI: dx.doi.org/10.5195/jmla.2022.1117

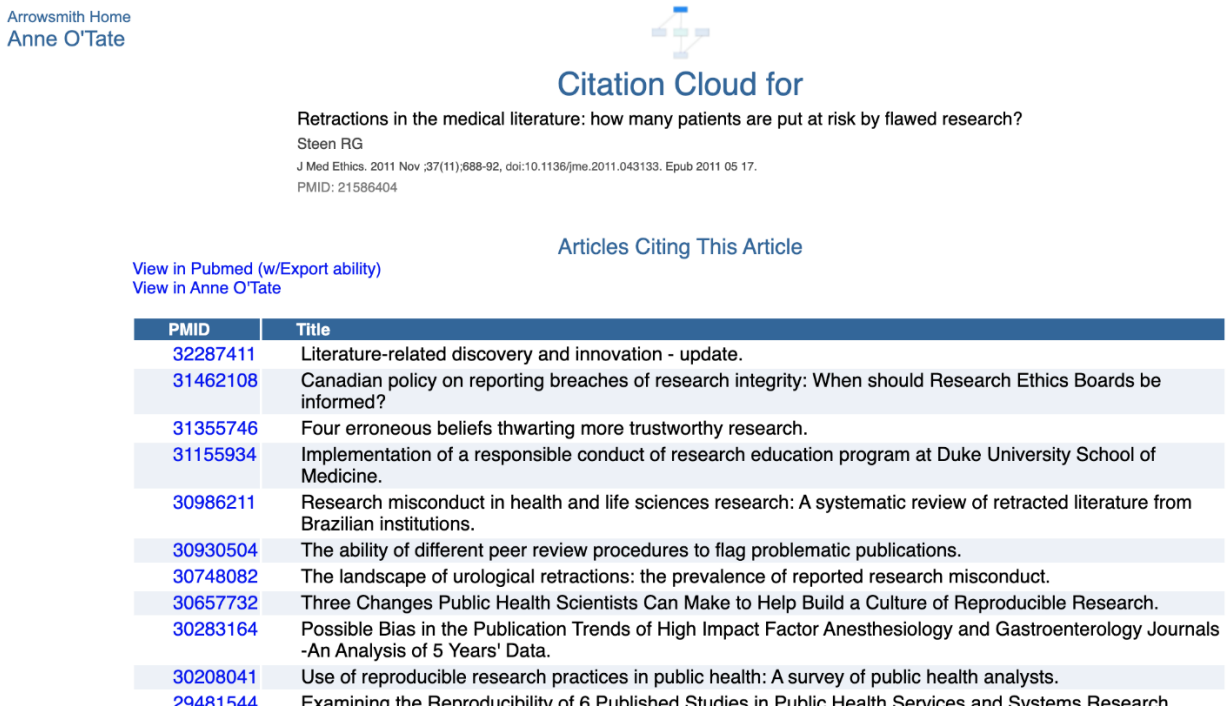
The article of interest is in the center box, which is surrounded by four other boxes interlinked by arrows that show the direction of citations. Clicking on any box opens a new tab that shows the articles in that box and has hyperlinks to allow users to export the articles to PubMed or Anne O'Tate for further mining. The "Articles Citing this Article" box consists of all articles that cite the paper of interest. In this example, there are forty-two citing articles as of the date of this query. The "Cited Count" box consists of all articles in the reference list of the paper of interest. The "Articles Co-Cited" box consists of all articles that are cited by one or more papers in the "Articles Citing this Article" box. Highly cited papers may have a very large set of co-cited articles, so we allow users to adjust the co-citation count threshold as desired. Finally, the "Bibliographically Coupled Articles" box consists of all articles that cite papers in the reference list of the paper of interest.

The upper box shows that forty-two articles have cited the article of interest; clicking on this box opens a new Results tab that lists the forty-two articles (Figure 3). Similarly, by clicking on the respective boxes, one can view and process articles that are cited by the article of interest, that are co-cited, or that are bibliographically coupled. The default option is to display a threshold of two, which means that at least two articles in the "Citing" box cited any article displayed in the "Co-Cited" box. Conversely, for the "Bibliographically Coupled" box, this means that each displayed bibliographically coupled article cited at least two references within the "Cited" box. The minimum threshold for display can be varied by the user in order to focus on the articles having the most similarity to the article of interest while minimizing the size of the list of articles displayed within the box.

Each box has two hyperlinks that permit the user to export the list to PubMed (which has the ability to export the citations in various formats) or to export the list to Anne O'Tate [8, 9], where it can be further mined. For example, one can identify the most important words and phrases in the titles and abstracts of articles on the list as well as the most frequent topics, authors, or journals, [8, 9].

As a rule, for any given article, there is little overlap between its set of citing articles, co-cited articles, bibliographically coupled articles, and PubMed Similar articles. Examining the full set of relationships provides insights that complement each other. If a given article of interest has just been published or has not been cited by another article, there are no citations or co-citations to analyze, yet one may still identify related articles using the PubMed Similar Articles function and the set of bibliographically coupled articles. Conversely, follow-up studies by the same team are likely to self-cite both the original paper and others by the team, so the set of co-cited articles is likely to relate to the team's broader interests.

Figure 3 Screenshot of the contents of the "Articles Citing this Article" box. Only the top few records in the list are shown.



PMID	Title
32287411	Literature-related discovery and innovation - update.
31462108	Canadian policy on reporting breaches of research integrity: When should Research Ethics Boards be informed?
31355746	Four erroneous beliefs thwarting more trustworthy research.
31155934	Implementation of a responsible conduct of research education program at Duke University School of Medicine.
30986211	Research misconduct in health and life sciences research: A systematic review of retracted literature from Brazilian institutions.
30930504	The ability of different peer review procedures to flag problematic publications.
30748082	The landscape of urological retractions: the prevalence of reported research misconduct.
30657732	Three Changes Public Health Scientists Can Make to Help Build a Culture of Reproducible Research.
30283164	Possible Bias in the Publication Trends of High Impact Factor Anesthesiology and Gastroenterology Journals -An Analysis of 5 Years' Data.
30208041	Use of reproducible research practices in public health: A survey of public health analysts.
29481544	Examining the Reproducibility of 6 Published Studies in Public Health Services and Systems Research

DOI: dx.doi.org/10.5195/jmla.2022.1117

When an article brings together two different lines of experimentation for the first time, the citation cloud may identify a mix of related articles across both lines. For example, an article by Smalheiser et al. titled "Enoxacin elevates microRNA levels in rat frontal cortex and prevents learned helplessness" [10] brings together studies of enoxacin as a regulator of microRNA production and studies of microRNA changes during animal models of depression and stress (Figure 4). The top ten co-cited articles include four that discuss enoxacin in various diseases and two other experimental studies by the same group (Figure 5). In contrast, the top ten bibliographically coupled articles all discuss microRNA function in depression and related disorders but show no overlap with the co-cited set (Figure 6). The top ten PubMed Similar Articles set shares two articles with the top ten co-cited set and none with the top ten bibliographically coupled set (not shown).

Figure 4 Screenshot of the Smalheiser et al. citation cloud.

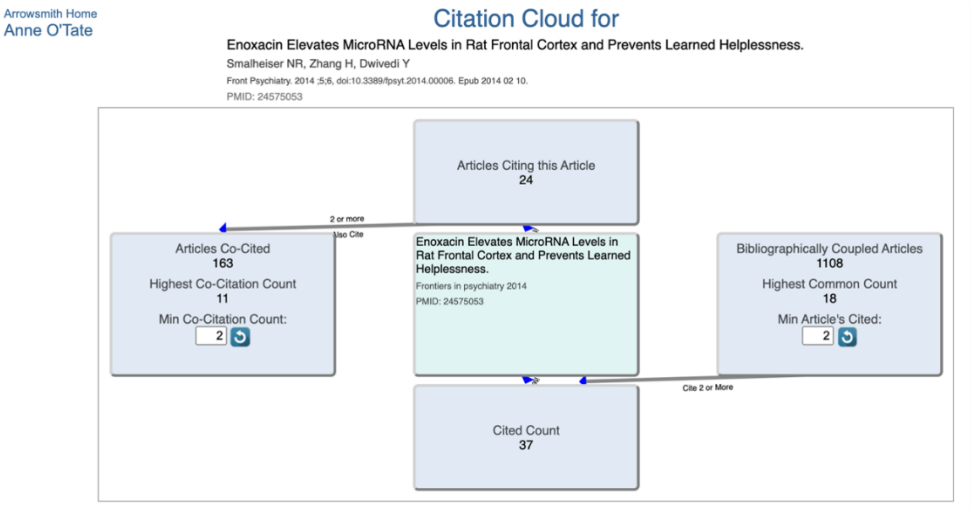


Figure 5 Top ten co-cited articles for the citation cloud shown in Figure 4.

Articles Co-Cited

View in Pubmed (w/Export ability)
View in Anne O'Tate

PMID	Title	Co-Citation Count
22427989	MicroRNA expression is down-regulated and reorganized in prefrontal cortex of depressed suicide subjects.	11
21368194	Small molecule enoxacin is a cancer-specific growth inhibitor that acts by enhancing TAR RNA-binding protein 2-mediated microRNA processing.	9
21275079	MicroRNA expression in rat brain exposed to repeated inescapable shock: differential alterations in learned helplessness vs. non-learned helplessness.	9
18641635	A small molecule enhances RNA interference and promotes microRNA processing.	9
24952960	MicroRNA 135 is essential for chronic stress resiliency, antidepressant efficacy, and intact serotonergic activity.	7
24908571	miR-1202 is a primate-specific and brain-enriched microRNA involved in major depression and antidepressant treatment.	7
26330466	Dysregulated miRNA biogenesis downstream of cellular stress and ALS-causing mutations: a new mechanism for ALS.	6
23644875	Enoxacin inhibits growth of prostate cancer cells and effectively restores microRNA processing.	5
22925464	Blood microRNA changes in depressed patients during antidepressant treatment.	5
20661255	The widespread regulation of microRNA biogenesis, function and decay.	5

DOI: dx.doi.org/10.5195/jmla.2022.1117

Figure 6 Top ten bibliographically coupled articles for the citation cloud shown in Figure 4.

Bibliographically Coupled Articles

View first 400 in Pubmed (w/Export ability)
View first 400 in Anne O'Tate

PMID	Title	Common Article Count
24733970	Emerging role of microRNAs in major depressive disorder: diagnosis and therapeutic implications.	18
33941769	MicroRNA-dependent control of neuroplasticity in affective disorders.	16
21515361	Evidence demonstrating role of microRNAs in the etiopathology of major depression.	13
24213247	The involvement of microRNAs in major depression, suicidal behavior, and related disorders: a focus on miR-185 and miR-491-3p.	12
22521503	The role of microRNAs in synaptic plasticity, major affective disorders and suicidal behavior.	12
25689819	Pathogenetic and therapeutic applications of microRNAs in major depressive disorder.	9
30071418	MicroRNAs in depression and suicide: Recent insights and future perspectives.	8
27240359	MicroRNAs: Key Regulators in the Central Nervous System and Their Implication in Neurological Diseases.	8
23642053	Pluripotent stem cell-derived somatic stem cells as tool to study the role of microRNAs in early human neural development.	8
21846569	MicroRNA function in the nervous system.	8

DISCUSSION

Although we believe this tool is easy to use, users should be aware of several limitations. The initial seed dataset of open access citations (Appendix A) is static. The iCite dataset is updated monthly, and these new citations are automatically added to the Citation Cloud dataset. Any newly added PubMed Central citations that are not included in the iCite updates will also be added. However, since not all citations are captured by these sources [5] or are openly available [11], the set of citations is far from comprehensive. Whereas we incorporated citations from over seventeen-million unique articles indexed in PubMed, including proprietary citations from Web of Science and Scopus would have given access to over twenty-one-million articles. Another limitation is that the citation cloud surrounding a single article can be quite large, especially for review articles or citation classics. Thus, it may be too cumbersome to display a citation cloud to encompass an entire list of articles. Finally, the dataset and interface focus on PubMed articles rather than articles contained in other bibliographic databases.

We expect that this new tool will augment the power of the new open access citations datasets to enable a broad community of scientists to utilize citations in their studies of biomedical literature. The Citation Cloud tool may also be useful to biomedical investigators and public users who are not carrying out citation analysis per se. Co-cited and bibliographically coupled articles represent types of similarity that are complementary to the PubMed Similar Articles ranking [12] and thus may assist in increasing recall for information retrieval [13], such as finding relevant literature for systematic reviews [14].

ACKNOWLEDGMENTS

This research was supported by NIH Grants R01LM010817 and P01AG039347. The study sponsor had no role in study design; in the collection, analysis, and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

DATA AVAILABILITY STATEMENT

Information regarding the open citation dataset and the EAV architecture of our web tool are presented in Appendix A.

REFERENCES

1. Steen RG. Retractions in the medical literature: how many patients are put at risk by flawed research? J Med Ethics. 2011 Nov;37(11):688–92. DOI: https://doi.org/10.1136/jme.2011.043133.
2. Van der Vet, PE, Nijveen H. Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. Res Integr Peer Rev. 2016;1(3). DOI: https://doi.org/10.1186/s41073-016-0008-5.
3. Hutchins BI, Davis MT, Meseroll RA, Santangelo GM. Predicting translational progress in biomedical research. PLoS Biol. 2019 Oct 10;17(10):e3000416. DOI: https://doi.org/10.1371/journal.pbio.3000416.
4. Boyack KW, Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? J Assoc Inf Sci Technol. 2010 Dec;61(12):2389–404.
5. Hutchins BI, Baker KL, Davis MT, Diwersy MA, Haque E, Harriman RM, Hoppe TA, Leicht SA, Meyer P, Santangelo GM. The NIH Open Citation Collection: a public access, broad coverage resource. PLoS Biol. 2019 Oct 10;17(10):e3000385. DOI: https://doi.org/10.1371/journal.pbio.3000385.
6. Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inf Sci. 1973 Jul;24(4):265–9.
7. Kessler MM. Bibliographic coupling between scientific papers. American Documentation. 1963 Jan;14(1):10–25.

DOI: dx.doi.org/10.5195/jmla.2022.1117

8. Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: a tool to support user-driven summarization, drill-down and browsing of PubMed search results. collaboration Biomed Discov Collab. 2008 Dec 1;3(1):2. DOI: https://doi.org/10.1186/1747-5333-3-2.

9. Smalheiser NR, Fragnito DP, Tirk EE. Anne O'Tate: Value-added PubMed search engine for analysis and text mining. PLoS One. 2021 Mar 8;16(3):e0248335. DOI: https://doi.org/10.1371/journal.pone.0248335.

10. Smalheiser NR, Zhang H, Dwivedi Y. Enoxacin elevates microRNA levels in rat frontal cortex and prevents learned helplessness. Front Psychiatry. 2014 Feb 10;5:6. DOI: https://doi.org/10.3389/fpsyt.2014.00006.

11. Shotton D. Funders should mandate open citations. Nature. 2018 Jan 11;553(7687):129.

12. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007 Dec 1;8(1):423.

13. Glänzel W. Bibliometrics-aided retrieval: where information retrieval meets scientometrics. Scientometrics. 2015 Mar 1;102(3):2215–22.

14. Belter CW. Citation analysis as a literature search method for systematic reviews. J Assoc Inf Sci Technol. 2016 Nov;67(11):2766–77.

SUPPLEMENTAL FILES

- **Appendix A**

AUTHORS' AFFILIATIONS

Neil R. Smalheiser, neils@uic.edu, http://orcid.org/0000-0003-1079-3406, Professor in Psychiatry, University of Illinois at Chicago, Chicago, IL

Jodi Schneider, jodi@illinois.edu, http://orcid.org/0000-0002-5098-5667, Assistant Professor in iSchool, University of Illinois at Urbana-Champaign, Champaign, IL

Vetle I. Torvik, vtorvik@illinois.edu, http://orcid.org/0000-0002-0035-1850, Associate Professor in the iSchool, University of Illinois at Urbana-Champaign, Champaign, IL

Dean P. Fragnito, dean@xornet.com, Principal, Xornet, Inc.

Eric E. Tirk, etirk@xornet.com, Coprincipal, Xornet, Inc.

Received August 2020; accepted August 2021

Articles in this journal are licensed under a Creative Commons Attribution 4.0 International License.

This journal is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.

ISSN 1558-9439 (Online)