Evaluating the Accuracy of scite, a Smart Citation Index

Caitlin Bakker, MLIS, AHIP-D[a], Nicole Theis-Mahon, MLIS, AHIP[b], Sarah Jane Brown, MSc[c]

[a]Discovery Technologies Librarian, Dr. John Archer Library, University of Regina, Regina, Saskatchewan, Canada, https://www.orcid.org/0000-0003-4154-8382, Caitlin.Bakker@uregina.ca

[b]Librarian Liaison to the School of Dentistry & Health Sciences Collections Coordinator, Health Sciences Library, University of Minnesota, Minneapolis, Minnesota, https://www.orcid.org/0000-0002-6913-5195,theis025@umn.edu

[c]Librarian Liaison to the College of Pharmacy, Health Sciences Library, University of Minnesota, Minneapolis, Minnesota, https://www.orcid.org/0000-0001-7699-4417, sjbrown@umn.edu

Abstract

Objectives: Citations do not always equate endorsement, therefore it is important to understand the context of a citation. Researchers may heavily rely on a paper they cite, they may refute it entirely, or they may mention it only in passing, so an accurate classification of a citation is valuable for researchers and users. While AI solutions have emerged to provide a more nuanced meaning, the accuracy of these tools has yet to be determined. This project seeks to assess the accuracy of scite in assessing the meaning of citations in a sample of publications.

Methods: Using a previously established sample of systematic reviews that cited retracted publications, we conducted known item searching in scite, a tool that uses machine learning to categorize the meaning of citations. scite's interpretation of the citation's meaning was recorded, as was our assessment of the citation's meaning. Citations were classified as mentioning, supporting or contrasting. Recall, precision, and f-measure were calculated to describe the accuracy of scite's assessment in comparison to human assessment.

Results: From the original sample of 324 citations, 98 citations were classified in scite. Of these, scite found that 2 were supporting and 96 were mentioning, while we determined that 42 were supporting, 39 were mentioning, and 17 were contrasting. Supporting citations had high precision and low recall, while mentioning citations had high recall and low precision. F-measures ranged between 0.0 and 0.58, representing low classification accuracy.

Conclusions: In our sample, the overall accuracy of scite's assessments was low. scite was less able to classify supporting and contrasting citations, and instead labeled them as mentioning. Although there is potential and enthusiasm for AI to make engagement with literature easier and more immediate, the results generated from AI differed significantly from the human interpretation.

Introduction

Research is cited for a variety of reasons including contextualizing one's own research, acknowledging foundational works, and critiquing or disputing previous research.[1] Despite the multitude of possible reasons for citations, they are typically viewed as an endorsement or an

indication of impact.[2] In the context of a systematic review, a citation may also have a larger meaning, indicating that the data underlying the cited report have been included in the pool of data to be analyzed.

Understanding why a paper is being cited is far more complex than determining the number of times it has been cited, and often requires readers to find and assess the citation within the full text. Artificial intelligence (AI) solutions have emerged in an attempt to provide a more nuanced understanding of a citation's context and allow readers to quickly discern where there are supportive and divergent findings. These solutions have ranged from those that aim to expedite the systematic review process, including screening, extraction and appraisal,[3] to large-scale text mining initiatives that seek to reinforce or contradict scientific claims.[4] scite is one such tool.

scite is a "smart citation index" that allows researchers to analyze the context of citations by locating a citation within the text and displaying and classifying the citation into one of three areas: supporting, mentioning, or contrasting.[5] As of March 2022, scite had over one billion "Smart Citations" in its database and has continued to ingest articles via agreements with publishers and other content providers, including PubMed, Unpaywall, and preprint servers.[6] scite is available online via scite.ai, through a Chrome browser extension, and through a Zotero plugin. scite previously had a freemium model where users could access limited information, but has more recently moved to an exclusively subscription-based model. The paid version of scite includes publication reports that detail citation information for the article, the number of publications citing the article, references, an analysis of the citation statements in the sentence including the in-text citation as well as the preceding and following sentences, and a classification of the citation statement as supporting, mentioning, or contrasting. While the possibility of having a more nuanced representation of citations is an appealing one, scite's value is predicated upon the assumption that the classification of these citations is an accurate one.

To assess the accuracy of scite as a classification tool, we analyzed a sample of systematic reviews that cited publications that had been retracted. This study is part of a larger research project investigating the use of retracted publications in systematic reviews in the field of pharmacy, including the ways in which authors were using these retracted publications.[7] The field of pharmacy was selected for its cross-sectional representation of research at a variety of different stages and research foci. Pharmacy research extends from bench research, such as drug development, to clinical research with individual patients, and to environmental toxicology, while also intersecting with medical specialties and other health disciplines.

Systematic reviews synthesize the totality of the evidence on a given topic. This research method is commonly placed at the pinnacle of the evidence-based pyramid. It requires authors to thoroughly assess every included report or study to identify methodological issues or concerns, and to reflect that assessment in its findings. Clinicians, students, and researchers often place a great deal of confidence in the findings of systematic reviews, and they might influence both patient care and future research. Given the importance of literature to this research method, it would stand to reason that systematic review authors would engage more thoroughly with the literature.

The uncritical incorporation of retracted publications into systematic reviews may undermine the value of the research method, as between 22% and 54% of retractions are due to methodological issues or concerns regarding data,[8-12] meaning that these reported findings

may be invalid. Previous research has found that the inclusion of retracted publications in systematic reviews can significantly alter the overall findings of the review.[13] The challenge of retracted publications in systematic reviews is of growing interest, and is one that systematic review authors are cognizant of, as evidenced in guidance available through the Cochrane Handbook and MECIR Manual.[14-16] Despite the available guidance and growing awareness, the inclusion of retracted publications in systematic reviews is an ongoing concern. AI provides an opportunity to assess how retracted publications are used in systematic reviews and holds the potential to ultimately streamline the process of appraising evidence syntheses.

Our findings regarding the methodological quality of systematic reviews that cite retracted publications are described elsewhere.[7] In completing this work, we simultaneously assessed scite's ability to classify these citations, as the potential to automate and subsequently expedite identification of retracted publications could be of benefit to the evidence synthesis community.

Methods

Using data provided by Retraction Watch from the Center for Scientific Integrity,[17] we identified a sample of retracted publications in the field of pharmacy and then conducted known item searching in Scopus and Web of Science Core Collection, including SCI-Expanded and SSCI-Expanded, to identify items which had cited these retracted publications. Two reviewers screened each citing item to limit to evidence syntheses, including systematic reviews, meta-analyses, clinical practice guidelines, scoping reviews, and rapid reviews. The initial project, which resulted in this sample, is fully described elsewhere.[7]

A data extraction form was established in Qualtrics to assess citation meaning as determined by scite and a human assessor. After piloting a subset of publications to ensure agreement, every reference in every systematic review was reviewed to assess the meaning of the citation. We conducted known item searching for every retracted publication in scite, and then reviewed the items citing the publication to identify the evidence synthesis in question. We then extracted scite's assessment of the meaning of the citation, as well as the exact text of the citation. Each reviewer recorded whether they agreed or disagreed with scite's assessment, or where they were unsure of whether the assessment was appropriate. In cases where the reviewer was in disagreement or was unsure, they entered their independent assessment into the Qualtrics form. All assessments were reviewed by the researchers collectively to ensure agreement and consistency. Citations were classified as mentioning when the citation was included in passing but was not extensively discussed. The citation was classified as supporting if the publication was one of the reports included in the systematic review results without indication of its retracted status or significant critique of the research quality. The citation was classified as contrasting if the publication were described as retracted, or if concerns about the publication or underlying research were discussed.

Data were then extracted from Qualtrics into a comma-separated file and were summarized using R 4.1.3.[18] We assessed scite's accuracy by calculating precision, recall, and f-measure, which are well-established quantifications of the accuracy of classification systems. Precision, or positive predictive value, refers to the number of items that were classified with a meaning and truly had that meaning (i.e., the number of citations marked as supporting that truly were supporting). Recall, or sensitivity, calculates the number of items classified with a particular meaning out of all items that had that meaning (i.e., the number of citations marked as supporting out of all of the supporting citations in the dataset). As Rebala et al. note, "[e]ffectively, you want high Precision as well as high Recall; i.e., when the data is

Positive Class, the prediction is also positive, and when the prediction is positive, the actual class is also positive."[19] The f-measure is a score that includes both precision and recall and reflects overall accuracy. F-measures fall between 0 and 1, with higher scores indicating greater accuracy.

Results

Our initial sample included 1,396 retracted publications in the field of pharmacy. Of these, 312 retracted publications were cited 32,559 times. Screening to isolate evidence syntheses identified 324 references to retracted publications in 286 systematic reviews. Eleven of these referenced articles were not found in scite, resulting in a set of 313 references. In 97 cases (31%) the referenced article was in scite, but the systematic review's citation to it was not. As scite did not have access to the full text, 118 (37.7%) of the references were unclassified. Of the 98 citations that were classified, scite found that 2 (2%) were supporting, 96 (98%) were mentioning, and 0 (0%) were contrasting. These findings are described in Figure 1.

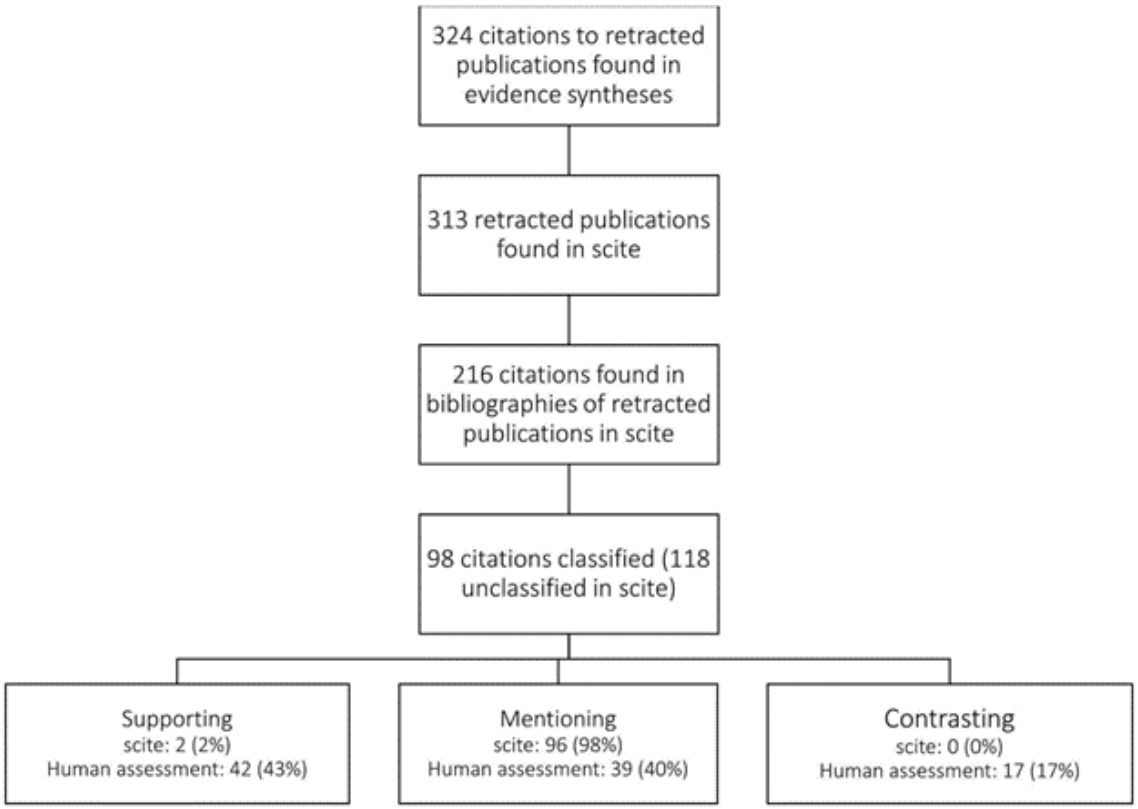

Figure 1: Results of search and assessment in scite

We manually coded the 98 references that were classified by scite to assess the accuracy of scite's classification. We agreed with two citations scite had indicated were supporting. There were an additional 40 citations that scite misclassified which we assessed as supporting. While scite did not identify any contrasting citations, we identified 17 contrasting citations. Finally, while we agreed with scite's assessment of 39 citations as mentioning, we found 57 citations were marked as mentioning when they were in fact contrasting or supporting.

We calculated the precision, recall, and f-measure of each assessment (supporting, mentioning, and contrasting), which can be seen in Table 1.

Table 1. Precision, recall and accuracy of scite

	Supporting	Contrasting	Mentioning
Precision	1.0	0.0	0.41
Recall	0.05	0.0	1.0
F-Measure	0.096	0.0	0.58

Discussion

scite shows promise in increasing the transparency of citations and has the potential to aid future researchers in understanding how authors use research, similar to the legal practice of shepardizing citations to determine how previous findings have been interpreted and applied. AI tools are rapidly developing to expedite the literature searching and appraisal process.[3, 20-22]. Although scite's solution is much less labor-intensive than reading each article to understand how it has utilized other literature, there are limitations to AI in contextualization since it may miss nuances easily discerned by human readers. In our sample of systematic reviews citing retracted publications, exclusive reliance on scite would have minimized the core problem: the supporting citation of retracted publications.

scite found that the vast majority of citations in our sample were mentioning, while the remaining were supporting. This is broadly in alignment with scite's overall assessments, as the creators indicate that "the average distribution of citation statements [is] 92.6% mentioning, 6.5% supporting, and 0.8% contrasting statements."[5] While this may be in alignment with scite's overall assessment of literature, it diverged significantly from our assessment of meaning, as we determined that, of the 96 citations scite classified as mentioning, 40 were more appropriately classified as supporting and 17 were more appropriately classified as contrasting. While scite determined that the majority of citations should be classified as mentioning, human assessors disagreed and had higher rates of supporting and contrasting citations.

In our sample, the overall accuracy of scite's assessments was low. Recall, or sensitivity, describes the number of citations classified with a particular meaning out of all citations that had that meaning. Supporting and contrasting meanings had low recall, while mentioning had high recall. This means that scite was less able to classify supporting and contrasting citations as such, instead labeling them as mentioning. Mentioning had high recall, meaning that, when citations were truly mentioning, scite recognized that they were mentioning. Precision, or positive predictive value, describes the number of items classified with a meaning that truly had that meaning. Supporting statements had high precision, meaning that the majority of citations classified as supporting were truly supporting. While scite may overlook supporting citations, it rarely misclassifies citations that are contrasting or mentioning as supporting. Conversely, mentioning had relatively low precision, indicating that users should be less confident that citations marked as mentioning are truly mentioning rather than supporting or contrasting. scite's tendency to classify the vast majority of citations as mentioning limits its overall utility. As Stacy Brody notes, "[a] high count of

mentioning citation statements may provide little more nuance or detail than a total citation count."[23]

Although scite may misclassify some citations, the larger challenge is missing content in scite's database. In our sample, less than one-third of the citations were classified in scite. In 31% of cases, the citation was not represented at all, while in 38% of cases the citation was represented but marked as unclassified due to lack of full-text access. This is a significant barrier to use. It could be argued that, even if classified inaccurately, foregrounding that citations can have different meanings is of value to users. However, if this foregrounding occurs in only a minority of cases, the pedagogical value of the tool is minimized. While scite has established relationships with several large publishers,[24] additional agreements with a broader range of publishers would enhance the utility and value of the tool.

While this project focused on citations occurring within systematic reviews in the field of pharmacy, it occurs in the broader context of AI in systematic reviews and critical appraisal. Almost every aspect of information seeking has an associated AI solution. AI tools and solutions have been developed to facilitate more efficient searching,[25] to select relevant publications and disregard irrelevant publications,[26] to extract study findings and characteristics,[27] and to assess the quality of the publication.[28] These advances have tremendous potential to expedite aspects of information seeking and allow time to be reallocated to critical engagement with and application of the literature. However, despite the potential of these tools, our findings highlight one context in which reliance upon AI would have led to skewed and inaccurate findings.

Limitations

Our research has several notable limitations. First, we focus on a sample of publications both within a discipline and using a specific study design. While scite is not programmed to perform differently based on study design or discipline, it is possible that studies using different publications would reveal different levels of accuracy. We also chose to focus on systematic reviews that had cited at least one retracted publication. Retraction is a relatively rare and extreme publication state. As such, the specific citations included in our sample are, by nature of their retracted status, not indicative of the majority of publications. Although we chose to focus on these outliers under the assumption that these publications would warrant stronger critique, this does limit the generalizability of our findings. Finally, this work was done in conjunction with a larger research project, rather than as a standalone project. As such, the sample was derived through this larger work, rather than being selected for the sole purpose of assessing scite. Other sampling methods may lead to different results.

Conclusions

Students, clinicians and researchers are increasingly seeking new technologies to enable rapid engagement with the scholarly literature. While AI holds tremendous promise, both in the completion of systematic reviews and in the critical appraisal of literature, reliance on AI in this context may be premature. This research emphasizes the

potential challenge of limited data sources and relatively low accuracy, both of which may cause challenges for adoption.

References

1. Garfield E. Can Citation Indexing be Automated? Stat Assoc Methods Mech Doc Symp Proc. 1965;269:189–92. doi:10.1038/227669a0

2. White HD. Citation Analysis and Discourse Analysis Revisited. Appl Linguist. 2004 Mar 1;25(1):89–116. doi:10.1093/applin/25.1.89

3. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. 2022 May;13(3):353–62. doi:10.1002/jrsm.1553

4. Opscidia - Plateforme pour accélérer la veille technologique [Internet]. 2022 [cited 2023 Feb 10]. Available from: https://www.opscidia.com/

5. Nicholson JM, Mordaunt M, Lopez P, Uppala A, Rosati D, Rodrigues NP, et al. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. Quant Sci Stud. 2021 Nov 5;2(3):882–98. doi:10.1162/qss$_{a0}$0146

6. scite. Where do you get your articles from? [Internet]. scite. [cited 2022 Jun 28]. Available from: https://help.scite.ai/en-us/article/where-do-you-get-your-articles-from-1vglydm/

7. Brown SJ, Bakker CJ, Theis-Mahon NR. Retracted publications in pharmacy systematic reviews. J Med Libr Assoc. 2022 Feb 11;110(1):47–55. doi:10.5195/jmla.2022.1280

8. Moylan EC, Kowalczuk MK. Why articles are retracted: a retrospective cross-sectional study of retraction notices at BioMed Central. BMJ Open. 2016;6(11):e012047. Published 2016 Nov 23. doi:10.1136/bmjopen-2016-012047

9. Wager E, Williams P. Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. J Med Ethics. 2011;37(9):567-570. doi:10.1136/jme.2010.040964

10. Nair S, Yean C, Yoo J, Leff J, Delphin E, Adams DC. Reasons for article retraction in anesthesiology: a comprehensive analysis.*Raisons justifiant la rétractation d'un article en anesthésiologie: une analyse exhaustive.* Can J Anaesth. 2020;67(1):57-63. doi:10.1007/s12630-019-01508-3

11. Bozzo A, Bali K, Evaniew N, Ghert M. Retractions in cancer research: a systematic survey. Res Integr Peer Rev. 2017;2:5. Published 2017 May 12. doi:10.1186/s41073-017-0031-1

12. Chauvin A, De Villelongue C, Pateron D, Yordanov Y. A systematic review of retracted publications in emergency medicine. Eur J Emerg Med. 2019;26(1):19-23. doi:10.1097/MEJ.0000000000000491

13. Garmendia CA, Nassar Gorra L, Rodriguez AL, Trepka MJ, Veledar E, Madhivanan P. Evaluation of the Inclusion of Studies Identified by the FDA as Having Falsified Data in the Results of Meta-analyses: The Example of the Apixaban Trials [published correction appears in JAMA Intern Med. 2021 Mar 1;181(3):409]. JAMA Intern Med. 2019;179(4):582-584. doi:10.1001/jamainternmed.2018.7661

14. Cochrane Library. Managing potentially problematic studies [Internet]. Cochrane Database of Systematic Reviews: editorial policies. [cited 2023 Feb 10]. Available from: https://www.cochranelibrary.com/cdsr/editorial-policiesproblematic-studies

15.Higgins JPT, Lasserson T, Chandler J, Tovey D, Thomas J, Flemyng E, et al. MECIR Manual [Internet]. London, England: Cochrane; 2022 Feb [cited 2023 Feb 10]. Available from: https://community.cochrane.org/mecir-manual

16. Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Marshall C, et al. 4.S1 Technical Supplement to Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions Version 63 [Internet]. Chichester, UK: John Wiley Sons, Ltd; 2022 [cited 2023 Feb 10]. Available from: https://training.cochrane.org/handbook/current/chapter-04-technical-supplement-searching-and-selecting-studies

17. Center for Scientific Integrity. Retraction Watch Database [Internet]. [cited 2022 Jul 21]. Available from: http://retractiondatabase.org/RetractionSearch.aspx?

18. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/

19. Rebala G, Ravi A, Churiwala S. Classification. In: An Introduction to Machine Learning [Internet]. 2019 [cited 2022 Jun 28]. p. 57–67. Available from: https://doi.org/10.1007/978-3-030-15729-6

20. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021 Feb;3(2):125–33. doi:10.1038/s42256-020-00287-7

21. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Jul 11;8(1):163. doi:10.1186/s13643-019-1074-9

22. Zhang Y, Liang S, Feng Y, Wang Q, Sun F, Chen S, et al. Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. Syst Rev. 2022 Jan 15;11(1):11.

doi:10.1186/s13643-021-01881-5

23. Brody S. Scite. J Med Libr Assoc. 2021 Nov 22;109(4):707–10. doi:10.5195/jmla.2021.1331

24. scite for Publishers [Internet]. scite.ai. [cited 2022 Nov 12]. Available from: https://scite.ai

25. Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smalheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. J Am Med Inform Assoc. 2017 Nov 1;24(6):1165–8. doi:10.1093/jamia/ocx053

26. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015 Jan 14;4(1):5. doi:10.1186/2046-4053-4-5

27. Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak. 2010 Sep 28;10(1):56. doi:10.1186/1472-6947-10-56

28. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc. 2016 Jan 1;23(1):193–201. doi:10.1093/jamia/ocv044

Author Contributions

Caitlin Bakker: Conceptualization, Data curation, Investigation, Writing - original draft, Writing - review & editing;

Nicole Theis-Mahon: Conceptualization, Data curation, Investigation, Writing - original draft, Writing - review & editing;

Sarah Jane Brown: Conceptualization, Data curation, Investigation, Writing - original draft, Writing - review & editing