Scientometrics
https://doi.org/10.1007/s11192-023-04827-x

Author gender and citation categorization: a study of high-impact medical journals

Paul Sebo[1] · **Amrollah Shamsi**[2]

Received: 24 June 2023 / Accepted: 29 August 2023

Abstract

The number of citations is one of the main bibliometric indicators. However, not all citations can be considered equivalent; *scite* (https://scite.ai/), a new tool based on artificial intelligence, was developed to determine whether citations are positive, negative or neutral. We assessed whether publications first/last authored by women were more often cited positively (or negatively) than those first/last authored by men. Using the 2021 *Journal Citation Reports* (*JCR)* impact factor, we selected the ten highest impact journals in nine medical disciplines. Using *Web of Science*, we extracted all research and review articles published between January 2012 and December 2021 in these journals. We used *Namsor* to determine first/last authors' gender and *scite* to categorize article citations as positive ("supporting"), negative ("contradicting"), neutral ("mentioning") and "unclassified". There were 141,921 articles in the database, of which 116,204 had unabbreviated first/last names. We found that the proportion of positive and negative citations was higher for publications whose first/last authors were women (vs. men), while the opposite was true for neutral citations. This is the first study to our knowledge to document the association between gender and citation type. Further research is needed in the future to investigate the reasons for these gender differences, and to assess whether the type of citation is also associated with the gender of the citing author.

Keywords Citation · Gender · Inequality · Publication · Research · Scite

Introduction

The number of citations is one of the indicators used to quantify an individual's scientific research output. However, not all citations are equivalent. A new artificial intelligence-based tool called *scite* (https://scite.ai/) can determine whether article citations are positive/negative/neutral (Nicholson et al., 2021).

✉ Paul Sebo
paulsebo@hotmail.com

1 University Institute for Primary Care (IuMFE), University of Geneva, Geneva, Switzerland

2 Department of Medical Library & Information Sciences, Bushehr University of Medical Sciences, Bushehr, Iran

Published online: 26 September 2023

Women often face gender inequalities in the realm of scientific publication (Sebo & Clair, 2022a, 2022b). These inequalities may partly explain why female researchers have more difficulty breaking into academia. There are differences between female/male researchers in terms of the topics they address and how they addressed them (Ashmos Plowman & Smith, 2011). Gender differences may therefore also appear in the way articles are cited.

In this study, we assessed whether publications first/last authored by women were more often cited positively (or negatively) than those first/last authored by men.

Methods

Using the 2021 *Journal Citation Reports* (*JCR*) impact factor, we selected the ten highest impact journals in nine medical disciplines (Supplementary Material #1). Using *Web of Science*, we extracted all research and review articles published between January 2012 and December 2021 in these journals. We used the *Gender Guesser* software developed by *Namsor* (https://gender-guesser.com/gender-name/) to determine first/last authors' gender and *scite* (https://scite.ai) to categorize article citations as positive ("supporting"), negative ("contradicting"), neutral ("mentioning") and "unclassified" (Lund & Shamsi 2023; Nicholson et al., 2021). The data for the study was collected on 30 September 2022.

scite uses deep learning and natural language processing (NLP) techniques to analyze citation context and content, and determines the type of citation based on annotated data from scientific articles, training machine learning models to classify citations, and applying these models to new citations for classification (Nicholson et al., 2021). The classifiers in *scite* were iteratively developed, starting with RNN (Recursive Neural Network) architectures and transitioning to SciBERT, a science-pretrained base BERT (Bidirectional Encoder Representations from Transformers) model. Additional information on how smart citations are created is available in Supplementary Material #1.

We created three additional citation variables (= proportion of positive/negative/neutral citations) dividing the number of supporting/contradicting/mentioning citations by the number of citations. We used proportions to summarize data on publications and medians and interquartile ranges (IQRs) to summarize data on citations. We stratified the results by gender, year of publication and medical discipline. We compared the data by gender using univariable and multivariable negative binomial regressions, adjusting for year of publication, type of article and intra-cluster correlations within journals, except for the proportion of positive/negative/neutral citations. For these three variables, we used a generalized linear model (logit link/binomial family/robust option). We repeated the analyses after removing all review articles. We also repeated the analyses with "≥ 70% Classification Accuracy Sample", consisting of all articles whose authors' gender was determined with ≥ 70% accuracy. The statistical significance was set at a two-sided p value ≤ 0.05. All analyses were performed with STATA 15.1 (StataCorp/College Station/TX).

Results

There were 141,921 articles in the database, of which 138,393 had author names and 116,204 had unabbreviated first/last names (101,164 research articles and 15,040 reviews). The number of articles varied by year of publication from 10,178 (8.8%) in 2021 to 12,343

(10.6%) in 2013, and by discipline from 9010 (7.8%) in dermatology to 17,211 (14.8%) in neurology. As Table 1 shows, women were first/last authors of 52,295 (45.0%) and 35,828 articles (30.8%) respectively, with differences by discipline (first authors: from 32.0% in radiology to 58.9% in obstetrics/gynecology; last authors: from 19.5% in radiology to 43.9% in pediatrics).

There were 8,497,819 citations for the 116,204 articles included in the study. These citations were categorized by *scite* as follows: supporting = 317,914 (3.7%), contradicting = 50,137 (0.6%), mentioning = 7,903,822 (93.0%), and unclassified = 225,946 (2.7%). As Table 1 shows, overall the median number of citations was lower for articles first/last authored by women (vs. men). Only in anesthesiology was the median number of citations higher for articles first/last authored by women (vs. men). Overall, for first authors, $n = 3{,}344{,}939$ citations, median = 24 (IQR = 48) for women, vs. $n = 5{,}152{,}880$ citations, median = 28 (61) for men (IRR = 1.26 [95% CI 1.24–1.30], p value < 0.001). For last authors, $n = 2{,}256{,}185$ citations, median = 24 (49.5) for women, vs. $n = 6{,}241{,}634$ citations, median = 27 (58) for men (IRR = 1.23 [95% CI 1.21–1.25], p value < 0.001). These differences were present throughout the period under review (Fig. 1). Similar results were obtained with research articles only. For first authors, median = 23 (IQR = 46) for women vs. 26 (56) for men (IRR = 1.25 [95% CI 1.23–1.27], p value < 0.001); for last authors, median = 23 (46) vs. 26 (53) (IRR = 1.24 [95% CI 1.22–1.26], p value < 0.001).

Table 2 shows the unadjusted/adjusted associations between citation variables and male gender. The differences were statistically significant for all variables except for the number of negative citations in multivariable analysis. The proportion of positive/negative citations was higher for publications whose first/last authors were women (vs. men), while the opposite was true for neutral citations. The results were similar (i) when only research articles were examined and (ii) when researchers whose gender was determined with less than 70% accuracy were excluded.

The results obtained within the nine medical disciplines were for the most part similar to the overall results, although not all associations were statistically significant (Supplementary Material #2). For example, the proportion of positive citations was higher for female (vs. male) first/last authors in eight and seven disciplines respectively, but the differences were only statistically significant in seven and three disciplines.

Discussion

We found that the number of citations was lower for female than for male first/last authors. We also found that the proportion of positive/negative citations was higher for publications whose first/last authors were women (vs. men), while the opposite was true for neutral citations.

Several studies have already shown that publications authored by women are generally less cited than those authored by men (Chatterjee & Werner 2021; Sebo & Clair, 2022b), but this is the first study to our knowledge to document the association between gender and citation type. Further research is needed in the future to investigate the reasons for these gender differences, and to assess whether the type of citation is also associated with the gender of the citing author.

Table 1 Number of publications and citations, stratified by medical discipline and gender ($n=8{,}497{,}819$ citations for 116,204 publications, i.e. 101,164 research articles and 15,040 reviews)

Variable	Number of publications, total	Number of publications for male first authors (%)	Number of publications for female first authors (%)	Number of publications for male last authors (%)	Number of publications for female last authors (%)	Median number of citations (IQR), overall	Median number of citations for male first authors (IQR)	Median number of citations for female first authors (IQR)	Median number of citations for male last authors (IQR)	Median number of citations for female last authors (IQR)
Overall[a]	116,204	63,909 (55.00)	52,295 (45.00)	80,376 (69.17)	35,828 (30.83)	26 (55)	28 (61)	24 (48)	27 (58)	24 (49.5)
Anesthesiology	11,672	7365 (63.10)	4307 (36.90)	8897 (76.23)	2775 (23.77)	13 (25)	13 (24)	14 (26)	13 (24)	14 (26)
Dermatology	9010	4351 (48.29)	4659 (51.71)	5786 (64.22)	3224 (35.78)	19 (33)	20 (34)	19 (32)	20 (33)	19 (31)
General internal medicine	11,229	7204 (64.16)	4025 (35.84)	8199 (73.02)	3030 (26.98)	98 (218)	109.5 (234)	83 (189)	103 (225)	87 (195)
Neurology	17,211	9910 (57.58)	7301 (42.42)	12,676 (73.65)	4535 (26.35)	37 (63)	38 (66)	36 (59)	38 (64)	36 (60)
Obstetrics & Gynecology	13,841	5686 (41.08)	8155 (58.92)	8047 (58.14)	5794 (41.86)	18 (30)	19 (31)	17 (29)	19 (32)	16 (27)
Oncology	11,243	6641 (59.07)	4602 (40.93)	8108 (72.12)	3135 (27.88)	75 (132)	80 (141)	66 (115)	78 (136)	66 (118)
Pediatrics	16,158	6652 (41.17)	9506 (58.83)	9060 (56.07)	7098 (43.93)	19 (36)	19 (37)	19 (36)	20 (36)	19 (37)
Psychiatry	10,041	5352 (53.30)	4689 (46.70)	6879 (68.51)	3162 (31.49)	27 (52)	29 (55)	25 (48)	29 (54)	25 (48)
Radiology	15,799	10,748 (68.03)	5051 (31.97)	12,724 (80.54)	3075 (19.46)	20 (33)	20 (34)	18 (31)	20 (34)	19 (33)

[a]Median number of citations (IQR) for research articles only ($n=101{,}164$): overall = 25 (51), male first authors = 26 (56), female first authors = 23 (46), male last authors = 26 (53), female last authors = 23 (46)

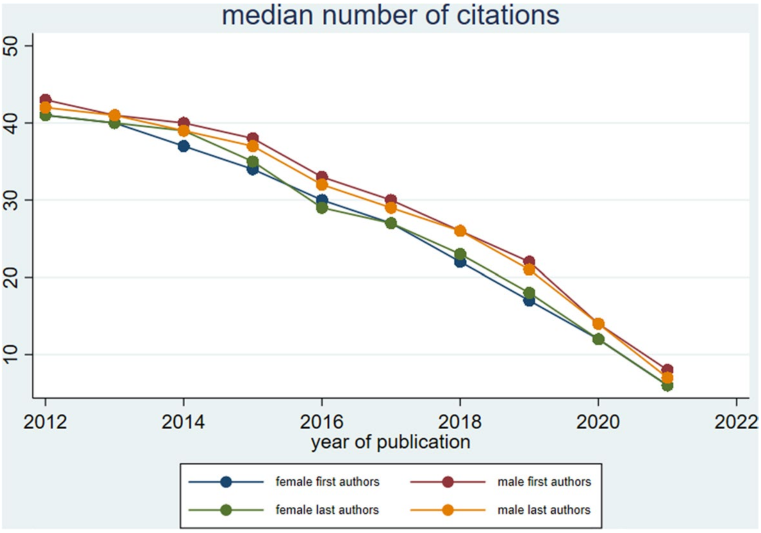

Fig. 1 Median number of citations for female and male first/last authors by year of publication

This study had a large sample size but three limitations should be mentioned. First, the study was conducted only with high-impact medical journals. The findings are not necessarily generalizable to other journals/scientific disciplines. Second, the citation analysis was done by an artificial intelligence tool (i.e., *scite*) and not by human evaluation. The owners of *scite* published an article in which they provided data on the accuracy of their tool (precision/recall/F-score estimated on a set of 9,708 examples = 0.741/0.576/0.648 for supporting citations, 0.852/0.451/0.590 for contrasting citations, and 0.962/0.984/0.973 for mentioning citations) (Nicholson et al., 2021). However, as far as we know, there are no studies carried out by researchers with no commercial links to *scite*, and there is no accuracy data by discipline. In addition, there is a possible risk of gender bias with the use of this tool. One factor that may contribute to gender bias is the tendency for research carried out by women to favor qualitative designs and human-centered studies more often than men. This can result in greater diversity in content, and therefore less predictability and lower accuracy in citation context analysis. Third, gender was determined by *Gender Guesser* (*NamSor*), not by self-identification. However, this tool is accurate in estimating the gender of individuals from their first/last names with around 2% misclassifications (performance metrics estimated using a sample of 6131 physicians: errorCoded = 0.0202, errorCodedWithoutNA = 0.0202, and naCoded = 0) (Sebo, 2021), and the results in our study were similar using the full sample and the "≥ 70% Classification Accuracy Sample". An older study found slightly lower results for accuracy, with around 4% misclassifications in a sample of 7076 researchers, perhaps suggesting that the tool has improved over time (errorCoded = 0.1282, errorCodedWithoutNA = 0.0429, and naCoded = 0.0891) (Santamaría & Mihaljević 2018).

Table 2 Unadjusted and adjusted associations between citation variables and male gender ($n = 8{,}497{,}819$ citations[a] for 116,204 publications, i.e. 101,164 research articles and 15,040 reviews)

Variable	Unadjusted IRR or OR (95% CI) for first authors[b]	Crude *p* value[c]	Unadjusted IRR or OR (95% CI) for last authors[b]	Crude *p* value[c]	Adjusted IRR or OR (95% CI) for first authors[b]	Adjusted *p* value[d]	Adjusted IRR or OR (95% CI) for last authors[b]	Adjusted *p* value[d]
Total number of citations	1.26 (1.24–1.30)	< 0.001	1.23 (1.21–1.25)	< 0.001	1.24 (1.07–1.44)	0.004	1.22 (1.10–1.36)	< 0.001
Number of positive ("supporting") citations	1.12 (1.10–1.15)	< 0.001	1.12 (1.09–1.14)	< 0.001	1.12 (1.03–1.23)	0.01	1.10 (1.01–1.19)	0.02
Number of negative ("contradicting") citations	1.05 (1.02–1.08)	0.001	1.06 (1.03–1.10)	< 0.001	1.05 (0.95–1.15)	0.34	1.04 (0.95–1.13)	0.39
Number of neutral ("mentioning") citations	1.27 (1.25–1.29)	< 0.001	1.24 (1.22–1.26)	< 0.001	1.25 (1.07–1.45)	0.004	1.23 (1.10–1.37)	< 0.001
Number of unclassified citations	1.31 (1.27–1.35)	< 0.001	1.20 (1.16–1.24)	< 0.001	1.31 (1.12–1.52)	0.001	1.20 (1.04–1.38)	0.01
Proportion of positive ("supporting") citations								
Full sample (research articles and reviews)	0.88 (0.86–0.90)	< 0.001	0.94 (0.92–0.96)	< 0.001	0.89 (0.84–0.93)	< 0.001	0.94 (0.90–0.98)	0.002
Subsample of research articles only	0.89 (0.87–0.90)	< 0.001	0.94 (0.92–0.96)	< 0.001	0.89 (0.84–0.93)	< 0.001	0.94 (0.90–0.98)	0.004
≥ 70% classification accuracy sample[e]	0.87 (0.85–0.89)	< 0.001	0.93 (0.91–0.95)	< 0.001	0.87 (0.83–0.92	< 0.001	0.93 (0.89–0.97)	0.002
Proportion of negative ("contradicting") citations								
Full sample (research articles and reviews)	0.87 (0.84–0.91)	< 0.001	0.91 (0.87–0.95)	< 0.001	0.88 (0.82–0.94)	< 0.001	0.91 (0.85–0.97)	0.002

Table 2 (continued)

Variable	Unadjusted IRR or OR (95% CI) for first authors[b]	Crude p value[c]	Unadjusted IRR or OR (95% CI) for last authors[b]	Crude p value[c]	Adjusted IRR or OR (95% CI) for first authors[b]	Adjusted p value[d]	Adjusted IRR or OR (95% CI) for last authors[b]	Adjusted p value[d]
Subsample of research articles only	0.88 (0.84–0.92)	< 0.001	0.92 (0.88–0.97)	< 0.001	0.88 (0.81–0.95)	< 0.001	0.92 (0.86–0.98)	0.01
≥ 70% classification accuracy sample[e]	0.87 (0.83–0.91)	< 0.001	0.89 (0.84–0.93)	< 0.001	0.87 (0.81–0.94)	< 0.001	0.89 (0.83–0.95)	0.001
Proportion of neutral ("mentioning") citations								
Full sample (research articles and reviews)	1.10 (1.08–1.12)	< 0.001	1.08 (1.06–1.10)	< 0.001	1.10 (1.05–1.15)	< 0.001	1.08 (1.04–1.12)	< 0.001
Subsample of research articles only	1.11 (1.09–1.13)	< 0.001	1.08 (1.06–1.10)	< 0.001	1.11 (1.05–1.16)	< 0.001	1.08 (1.04–1.12)	< 0.001
≥ 70% classification accuracy sample[e]	1.11 (1.09–1.13)	< 0.001	1.09 (1.07–1.11)	< 0.001	1.11 (1.06–1.17)	< 0.001	1.09 (1.05–1.13)	< 0.001

[a] Positive citations: $n = 317{,}914$; negative citations: $n = 50{,}137$; neutral citations: $n = 7{,}903{,}822$; unclassified citations: $n = 225{,}946$

[b] IRR (= incident rate ratio) for the first five variables; OR (= odds ratio) for the other variables

[c] Univariable negative binomial regression for the first five variables; generalized linear model with logit link and binomial family for the other variables

[d] Multivariable negative binomial regression for the first five variables; generalized linear model with logit link and binomial family for the other variables. Model adjusted for year of publication, type of article (research vs. review) and intra-cluster correlations within journals for "Full Sample" and "≥ 70% Classification Accuracy Sample". Model adjusted for year of publication and intra-cluster correlations within journals for "subsample of research articles only"

[e] All articles whose authors' gender was determined with ≥ 70% accuracy ($n = 102{,}150$ articles for first authors and 104,077 articles for last authors)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-023-04827-x.

Acknowledgements None.

Funding None.

Data availability statement The data underlying this article will be shared on reasonable request to the corresponding author.

Declarations

Competing interest The authors declare that they have no conflict of interest.

Ethical approval Since this study did not involve the collection of personal health-related data it did not require ethical review, according to current Swiss law.

References

Ashmos Plowman, D., & Smith, A. D. (2011). The gendering of organizational research methods: Evidence of gender patterns in qualitative research. *Qualitative Research in Organizations and Management: an International Journal, 6*(1), 64–82. https://doi.org/10.1108/17465641111129399

Chatterjee, P., & Werner, R. M. (2021). Gender disparity in citations in high-impact journal articles. *JAMA Network Open, 4*(7), e2114509. https://doi.org/10.1001/jamanetworkopen.2021.14509

Lund, B., & Shamsi, A. (2023). Examining the use of supportive and contrasting citations in different disciplines: a brief study using Scite (scite.ai) data. *Scientometrics, 128*(8), 4895–4900.

Nicholson, J. M., Mordaunt, M., Lopez, P., et al. (2021). Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies, 2*(3), 882–898. https://doi.org/10.1162/qss_a_00146

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science, 4*, e156. https://doi.org/10.7717/peerj-cs.156

Sebo, P. (2021). Performance of gender detection tools: A comparative study of name-to-gender inference services. *Journal of the Medical Library Association: JMLA, 109*(3), 414–421. https://doi.org/10.5195/jmla.2021.1185

Sebo, P., & Clair, C. (2022a). Gender gap in authorship: A study of 44,000 articles published in 100 high-impact general medical journals. *European Journal of Internal Medicine, 97*, 103–105. https://doi.org/10.1016/j.ejim.2021.09.013

Sebo, P., & Clair, C. (2022b). Gender inequalities in citations of articles published in high-impact general medical journals: A cross-sectional study. *Journal of General Internal Medicine*. https://doi.org/10.1007/s11606-022-07717-9