Use of ChatGPT to explore gender and geographic disparities in scientific peer review

Paul Sebo

Submitted to: Journal of Medical Internet Research
on: February 22, 2024

Table of Contents

Use of ChatGPT to explore gender and geographic disparities in scientific peer review

Paul Sebo[1]

[1]University of Geneva Geneva CH

Corresponding Author:
Paul Sebo
University of Geneva
Rue Michel-Servet 1
Geneva
CH

Abstract

This study used ChatGPT 4.0 to assess sentiment and politeness in 291 peer review reports across nine general medical journals. While no gender-based differences were found, notable regional disparities were observed, with articles from the Middle East, Latin America, and Africa receiving significantly lower scores. These findings underscore broader issues of inclusivity in peer review processes, particularly for researchers from regions facing systemic challenges in academic publishing.

(JMIR Preprints 22/02/2024:57667)
DOI: https://doi.org/10.2196/preprints.57667

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Short Paper

Use of ChatGPT to explore gender and geographic disparities in scientific peer review

Running headline: ChatGPT to explore disparities in peer review

Paul Sebo MD MSc[1]

[1] University Institute for primary care (IuMFE), University of Geneva, Geneva, Switzerland

Address of the corresponding author:

Dr Paul Sebo
University Institute for primary care (IuMFE), University of Geneva
1211 Geneva
Switzerland
Email address: paulsebo@hotmail.com

Word count: 1525 (abstract: 347)

Number of references: 27

Number of tables and/or figures: 2

Keywords: Africa; Artificial Intelligence; Discrimination; Peer Review; Sentiment Analysis; Women

Abstract

Background: In the realm of scientific research, peer review serves as a cornerstone for ensuring the quality and integrity of scholarly articles. Recent trends in promoting transparency and accountability led some journals to publish peer review reports alongside articles.

Objective: I used ChatGPT-4 to quantitatively assess sentiment and politeness in peer review reports from high-impact medical journals. The objective was to explore gender and geographical disparities to enhance inclusivity within the peer review process.

Methods: I identified all nine general medical journals with an impact factor>2 that publish peer review reports. I randomly selected twelve research articles per journal, all published in 2023. I collected the names of first and last authors along with the first author's country of affiliation and determined the gender for both first and last authors. For each review, I asked ChatGPT-4 to evaluate the ‘sentiment score’, ranging from -100 (negative) to 0 (neutral) to +100 (positive), and the ‘politeness score’, ranging from -100 (rude) to 0 (neutral) to +100 (polite). I repeated the measurements five times and removed the minimum and maximum values. I calculated the mean sentiment and politeness scores for each review. Statistical analyses included Wilcoxon rank-sum tests, Kruskal-Wallis rank tests, and negative binomial regressions.

Results: Analysis of 291 peer review reports corresponding to 108 articles unveiled notable regional disparities. Articles from the Middle East/Latin America/Africa exhibited lower sentiment and politeness scores compared to those from North America/Europe/Pacific and Asia (sentiment scores: 27 vs. 60 and 62 respectively; politeness scores: 43.5 vs. 67 and 65 respectively, adjusted p-values=0.02). No significant differences based on authors’

gender were observed.

Conclusion: I found notable regional disparities, with articles from the Middle East, Latin America, and Africa demonstrating significantly lower scores, while no discernible differences were observed based on authors' gender. The absence of gender-based differences suggests that gender biases may not manifest as prominently as other forms of bias within the context of peer review. The study underscores the need for targeted interventions to address regional disparities in peer review and advocates for ongoing efforts to promote equity and inclusivity in scholarly communication.

Introduction

The peer review process plays a pivotal role in validating the quality and integrity of scholarly articles**.** With an increasing emphasis on transparency and accountability, some journals adopted a practice of publishing peer review reports alongside articles.

Sentiment analysis, the identification and categorization of opinions, attitudes, and emotions conveyed in text data as positive, negative, or neutral[1], finds ChatGPT standing out with unique advantages compared to 'traditional methods'[1–5]. Unlike lexicon-based approaches that may overlook nuanced expressions, ChatGPT understands natural language, capturing subtle nuances and context-specific sentiments. Unlike machine learning methods requiring labeled datasets, ChatGPT adapts to diverse domains without explicit training, reducing costs and time. Its human-like responses facilitate intuitive sentiment interpretation. Additionally, AI-driven sentiment analysis, including ChatGPT, ensures efficient, scalable (able to accommodate increasing data volumes without significant performance degradation), consistent, objective (devoid of human biases and preconceptions), and adaptable analyses. ChatGPT showed promise for sentiment analysis in several recent studies[6–10].

Verharen showed that ChatGPT was accurate in determining sentiment and politeness scores in peer reviews of scientific articles[11]. The study also showed gender inequalities, with female authors receiving less polite reviews than men. The study, limited to articles from a single journal (Nature Communications), aligns with the broader issue of gender discrimination in academic medicine[12–17]. While sentiment and politeness metrics may not directly measure bias, they serve as useful proxies for identifying potential biases in peer review. Biased reviewers may exhibit tendencies towards overly positive/negative sentiments, as well as varying levels of politeness towards authors based on factors such

as gender or geographic region.

Building upon Verharen's work, I used ChatGPT-4 to quantitatively assess sentiment and politeness in peer review reports from nine high-impact medical journals, exploring gender and geographical disparities to enhance inclusivity within the peer review process. By leveraging the capabilities of artificial intelligence, this study sheds light on the potential of AI technologies to mitigate biases and promote fairness in scholarly communication.

Methods

Selection of journals, articles, and peer review reports

I searched the Clarivate and ASAPbio websites to identify all nine general medical journals with a JCR impact factor>2 that publish peer review reports(Table 1). None of these journals employs double-blind peer review. Using simple randomization, which involves randomly selecting articles using a random number generator, I selected twelve research articles per journal, all published in 2023, and retrieved all peer review reports from the initial round for these 108 articles. This sample size was chosen to ensure robust and manageable analysis.

Data collection and gender determination

I collected the names of first/last authors along with the first author's country of affiliation and determined the gender for both first/last authors. I categorized authors' genders in two steps. I determined the gender based on names alone, classifying names as male or female accordingly. For authors whose gender could not be inferred from their names, I searched professional networks and university websites for photos or text containing gender-specific pronouns. This method enabled me to assign genders to all authors. I also used a gender detection tool (GenderAPI) to confirm my classifications. GenderAPI demonstrated high accuracy in previous studies [18,19]. Both approaches yielded similar results, with high agreement between them (first authors:percent agreement=0.9725,Cohen’s Kappa=0.9450;last authors:percent agreement=0.9691,Cohen’s Kappa=0.9295).

Sentiment and politeness scores

For each review, I asked ChatGPT-4 to evaluate the '*sentiment score*', ranging from -100(negative) to +100(positive), and the '*politeness score*', ranging from -100(rude) to +100(polite). I repeated the measurements five times and removed the minimum and maximum values. The sentiment score measures how favorable the review is, the politeness score how polite a review's language is. I used the same prompt as Verharen ("Below you will find a scientific peer review. Can you score this peer review on the sentiment, on a scale from -100(negative) to 0(neutral) to 100(positive), and politeness of language use, on a scale of -100(rude) to 0(neutral) to 100(polite) ?")[11]. All the data was collected in January/2024.

Statistical analyses

I computed the mean sentiment/politeness scores for each review, rounding to the nearest whole number, and summarized these data using the median and interquartile range(IQR), overall and by journal, gender, and affiliation. Affiliation countries were categorized into three regions(North America/Europe/Pacific, Asia, and Latin America/Middle East/Africa), following prior research[20]. Comparisons were conducted using Wilcoxon rank-sum tests (for gender) and Kruskal-Wallis rank tests (for affiliation). Negative binomial regressions were performed, adjusting for journal, affiliation, and intra-cluster correlation within articles[21,22]. To accommodate the requirement of non-negative outcome variables in negative binomial regressions, I added 100 to the sentiment/politeness scores, resulting in a scale from 0 to 200. Quadratic weighted agreement coefficients were calculated to assess the agreement between the three sentiment/politeness score measurements. Fleiss' Kappa was used instead of Cohen's Kappa due to the presence of more than two measurements[23]. All analyses were conducted using STATA15.1.

Results

There were 291 reviews for the 108 articles selected for the study. Men were first/last authors of 61(56.5%) and 75(69.4%) articles respectively. The five most represented countries of affiliation were the UK(N=14), Germany(N=13), USA(N=10), China(N=9), and Italy(N=6). The three main regions of affiliation were Western Europe(N=56), Asia(N=16), and North America(N=15).

Overall, the median sentiment/politeness scores were 58(IQR=42,min=-70,max=90) and 63(IQR=28,min=-73,max=92) respectively, but there were notable variations by journal(Table 1). The three journals with the highest impact factor tended to have higher sentiment/politeness scores. There was no significant difference in scores between men/women(Table 2). The results were almost identical when using Gender API to determine gender (first authors:sentiment/politeness scores=58(35) and 65(30) for women,57.5(43) and 63(26) for men; last authors:sentiment/politeness scores=55(35) and 63(25) for women,58(45) and 63(30) for men). By contrast, articles authored by scholars from countries in the Middle East/Latin America/Africa exhibited significantly lower sentiment/politeness scores compared to those from the other two regions, with differences exceeding 30 and 20 points in absolute value, respectively(Table 2).

In light of the BMJ's notable impact factor compared to the other journals in the study, I conducted additional analyses by excluding the 51 peer-review reports for BMJ. The results remained consistent(Supplementary Material #1). The interrater agreement between the three measurements was high (sentiment scores:percent agreement=0.9958[95%CI=0.9954-0.9962],Fleiss' Kappa=0.9496[95%CI=0.9395-0.9598]; politeness scores:percent agreement=0.9962[95%CI=0.9958-0.9966],Fleiss' Kappa=0.9463[95%CI=0.9316-0.9610],all p-values<0.001).

Discussion

Main findings

I used ChatGPT to analyze sentiment/politeness in 291 peer review reports from nine general medical journals. The study unveiled notable regional disparities, with articles from the Middle East/Latin America/Africa demonstrating significantly lower scores, while no discernible differences were observed based on authors' gender.

Comparison with existing literature

The gender disparities experienced by women in academic medicine are widely recognized[12–17]. Consequently, I anticipated significant gender discrepancies in this study to the detriment of women. The absence of discernible differences between genders contrasts with Verharen's findings, where female first authors typically received less polite reviews compared to men[11]. Importantly, the study focused on a different discipline (neuroscience) and was confined to a single journal (Nature Communications).

The regional disparities highlighted in the study resonate with prior research illustrating the challenges faced by researchers, particularly those from countries in the Global South[20,24]. These findings align with those of earlier studies[25–27], and must be understood within broader socio-cultural, economic, and institutional contexts. Factors such as limited funding opportunities, language barriers, and cultural differences may contribute to biases against authors from underrepresented regions. Additionally, institutional biases within academic publishing systems may further exacerbate these discrepancies, highlighting the need for interventions to foster equity and inclusivity in scholarly discourse.

The observed disparities across geographic regions, contrasted with the absence of

significant differences based on gender, could indicate that within the context of peer review, gender biases may not manifest as prominently or uniformly as other forms of bias, such as those influenced by geographic factors. In addition, existing diversity/inclusion initiatives within the scholarly community may have been more effective in mitigating gender disparities compared to other forms of bias.

Strengths and limitations

This study built on the findings of Verharen, who demonstrated through several validation methods the accuracy of ChatGPT in estimating sentiment/politeness in peer review reports, surpassing that of human evaluation and traditional lexicon-based language models[11]. My study has several limitations, including its exclusive focus on high-impact general medical journals, reliance on binary gender determination without consideration for non-binary or transgender identities, and uncertainty about gender/geographic distributions of rejected papers. Future research should explore alternative approaches for gender determination such as self-identification for accurately capturing gender diversity. In addition, the sample size (108 articles and 291 peer review reports) is smaller than that of prior research by Verharen, potentially limiting the generalizability of my findings. I also acknowledge the challenges associated with using sentiment/politeness metrics to capture the nuanced biases inherent in peer review, and the dependence on algorithms like ChatGPT may introduce potential inaccuracies. Furthermore, manual scoring was not conducted for comparison. Finally, the observed association between sentiment/politeness metrics and affiliation regions suggests an alternative explanation: the possibility of higher scientific merit in papers from certain authors. My methodology lacks the capability to discern between the two hypotheses.

Conclusion

GPT-4 demonstrated effectiveness in this study by consistently evaluating sentiment and politeness in peer review reports. The study underscores the need for targeted interventions to address regional disparities in peer review and advocates for ongoing efforts to promote equity and inclusivity in scholarly communication.

Acknowledgements: None

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.

Funding source: None.

Disclosure of interest: The author alone is responsible for the content and writing of the paper.

Generative AI: Generative AI was not used in any part of the writing of the manuscript.

Data availability statement: The data associated with this article are available in the Open Science Framework (https://doi.org/10.17605/OSF.IO/WNRZU). The database is a 'dta' file for use with STATA. Supplementary Material #2 provides a detailed list and description of the variables available in the STATA file.

References

1. Mehta P, Pandya DS. A Review On Sentiment Analysis Methodologies, Practices And Applications. Int J Sci Technol Res 2020 Feb 25; Available from: https://www.semanticscholar.org/paper/A-Review-On-Sentiment-Analysis-Methodologies%2C-And-Mehta-Pandya/bbd48ed7502ece18eecf62153b23d0d7c3e08cd2 [accessed May 11, 2024]

2. Bordoloi M, Biswas SK. Sentiment analysis: A survey on design framework, applications and future scopes. Artif Intell Rev 2023 Mar 20;1–56. PMID:37362892

3. Cui J, Wang Z, Ho S-B, Cambria E. Survey on sentiment analysis: evolution of research methods and topics. Artif Intell Rev 2023 Jan 6;1–42. PMID:36628328

4. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 2022 Oct 1;55(7):5731–5780. doi: 10.1007/s10462-022-10144-1

5. Islam T, Sheakh MA, Sadik MR, Tahosin MS, Foysal MMR, Ferdush J, Begum M. Lexicon and Deep Learning-Based Approaches in Sentiment Analysis on Short Texts. J Comput Commun Scientific Research Publishing; 2024 Jan 3;12(1):11–34. doi: 10.4236/jcc.2024.121002

6. Fu Z, Hsu YC, Chan CS, Lau CM, Liu J, Yip PSF. Efficacy of ChatGPT in Cantonese Sentiment Analysis: Comparative Study. J Med Internet Res 2024 Jan 30;26:e51069. PMID:38289662

7. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, Linos E, Pereira F. A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data. JMIR Ment Health 2024 Jan 25;11:e50150. PMID:38271138

8. Tabone W, de Winter J. Using ChatGPT for human-computer interaction research: a primer. R Soc Open Sci 2023 Sep;10(9):231053. PMID:37711151

9. Kim S, Kim K, Wonjeong Jo C. Accuracy of a large language model in distinguishing anti- and pro-vaccination messages on social media: The case of human papillomavirus vaccination. Prev Med Rep 2024 Jun;42:102723. PMID:38659997

10. Wang Z, Xie Q, Feng Y, Ding Z, Yang Z, Xia R. Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. arXiv; 2024. doi: 10.48550/arXiv.2304.04339

11. Verharen JPH. ChatGPT identifies gender disparities in scientific peer review. eLife 2023 Nov 3;12:RP90230. PMID:37922198

12. Richter KP, Clark L, Wick JA, Cruvinel E, Durham D, Shaw P, Shih GH, Befort CA, Simari RD. Women Physicians and Promotion in Academic Medicine. N Engl J Med 2020 Nov 26;383(22):2148–2157. PMID:33252871

13. Sebo P, de Lucia S, Vernaz N. Gender gap in medical research: a bibliometric study in Swiss university hospitals. Scientometrics 2020 Oct 29; doi: 10.1007/s11192-020-03741-w

14. Hart KL, Perlis RH. Trends in Proportion of Women as Authors of Medical Journal Articles, 2008-2018. JAMA Intern Med 2019 Sep;179(9):1285–1287. PMID:31135815

15. Sebo P, Clair C. Gender gap in authorship: a study of 44,000 articles published in 100 high-impact general medical journals. Eur J Intern Med 2021 Sep 28;S0953-6205(21)00313–7. PMID:34598855

16. Filardo G, da Graca B, Sass DM, Pollock BD, Smith EB, Martinez MA-M. Trends and comparison of female first authorship in high impact medical journals: observational study (1994-2014). BMJ 2016 Mar 2;352:i847. PMID:26935100

17. Sebo P, Clair C. Gender Inequalities in Citations of Articles Published in High-Impact General Medical Journals: a Cross-Sectional Study. J Gen Intern Med 2022 Jul 6; PMID:35794309

18. Gender API. Available from: https://gender-api.com [accessed Dec 31, 2023]

19. Sebo P. Performance of gender detection tools: a comparative study of name-to-gender inference services. J Med Libr Assoc JMLA 2021 Jul 1;109(3):414–421. PMID:34629970

20. Sebo P. Gender and geographical inequalities among highly cited researchers: a cross-sectional study (2014-2021). Intern Emerg Med 2023 Mar 6; PMID:36877434

21. Negative Binomial Regression | Stata Annotated Output. Available from: https://stats.idre.ucla.edu/stata/output/negative-binomial-regression/ [accessed Dec 31, 2023]

22. Negative Binomial Regression | Stata Data Analysis Examples. Available from: https://stats.idre.ucla.edu/stata/dae/negative-binomial-regression/ [accessed Dec 31, 2023]

23. Klein D. Implementing a General Framework for Assessing Interrater Agreement in Stata. Stata J SAGE Publications; 2018 Dec 1;18(4):871–901. doi: 10.1177/1536867X1801800408

24. Pouris A, Pouris A. The state of science and technology in Africa (2000–2004): A scientometric assessment. Scientometrics 2009 May 1;79(2):297–309. doi: 10.1007/s11192-009-0419-x

25. Skopec M, Issa H, Reed J, Harris M. The role of geographic bias in knowledge diffusion: a systematic review and narrative synthesis. Res Integr Peer Rev 2020 Jan 15;5(1):2. doi: 10.1186/s41073-019-0088-0

26. Lor P. Scholarly publishing and peer review in the Global South: the role of the reviewer. JLIS.it 2022 Dec 19;14(1):10–29. doi: 10.36253/jlis.it-512

27. Schneider S, Morgan C, Magill C, Hersh M, Sang K, MacIntosh R. Evidence review:

Peer review bias in the funding process: Main themes and interventions. Heriot-Watt University; 2024. doi: 10.17861/YDDH-B137

Table 1. List of high-impact general medical journals included in this cross-sectional study using ChatGPT-4 to quantitatively assess sentiment and politeness in 291 peer review reports for 108 articles published in 2023, as well as number of articles and reviews per journal, and median sentiment and politeness scores per journal

Journal	2022 impact factor	Number of articles, n/N (%)	Number of reviews, n/N (%)	Sentiment score, median (IQR)	Politeness score, median (IQR)
BMJ	107.7	12/108 (11.1)	51/291 (17.5)	68 (25)	73 (20)
PLOS MEDICINE	15.8	12/108 (11.1)	41/291 (14.1)	60 (23)	70 (16)
BMC MEDICINE	9.3	12/108 (11.1)	29/291 (10.0)	63 (22)	73 (18)
JOURNAL OF CLINICAL MEDICINE	3.9	12/108 (11.1)	31/291 (10.7)	43 (53)	50 (42)
DIAGNOSTICS	3.6	12/108 (11.1)	28/291 (9.6)	37 (57)	57 (40)
JOURNAL OF PERSONALIZED MEDICINE	3.4	12/108 (11.1)	27/291 (9.3)	57 (53)	68 (55)
BMJ OPEN	2.9	12/108 (11.1)	27/291 (9.3)	63 (32)	60 (25)
BMC PRIMARY CARE	2.9	12/108 (11.1)	25/291 (8.6)	57 (57)	55 (40)
MEDICINA	2.6	12/108 (11.1)	32/291 (11.0)	48.5 (45)	57 (22.5)

Table 2. Associations between sentiment/politeness scores and first/last authors' gender and first authors' affiliation in this cross-sectional study using ChatGPT-4 to quantitatively assess sentiment and politeness in 291 peer review reports for 108 articles published in 2023 in 9 high-impact general medical journals

Variable	Number of articles, n/N (%)	Number of reviews, n/N (%)	Sentiment score, median (IQR)	Crude p-value[1]	Adjusted p-value[2]	Politeness score, median (IQR)	Crude p-value[1]	Adjusted p-value[2]
First authors' gender				0.49	0.48		0.37	0.68
Female	47/108 (43.5)	127/291 (43.6)	58 (39)			65 (30)		
Male	61/108 (56.5)	164/291 (56.4)	57.5 (43)			63 (24.5)		
Last authors' gender				0.52	0.88		0.74	0.86
Female	33/108 (30.6)	91/291 (31.3)	57 (35)			63 (25)		
Male	75/108 (69.4)	200/291 (68.7)	60 (44.5)			63 (30)		
First authors' affiliation				0.001	0.02[3]		<0.001	0.02[4]
North America, Europe, Pacific	82/108 (75.9)	220/291 (75.6)	60 (39)			67 (24)		
Asia	16/108 (14.8)	43/291 (14.8)	62 (30)			65 (23)		
Middle East, Latin America, Africa	10/108 (9,3)	28/291 (9.6)	27 (58)			43.5 (39.5)		

[1] Wilcoxon rank-sum test (for gender) and Kruskal-Wallis equality-of-populations rank test (for affiliation)

[2] Multivariable negative binomial regression, adjusted for journal, affiliation, and intra-cluster correlation within articles (for first/last authors' gender), and adjusted for journal and intra-cluster correlation within articles (for first authors' affiliation)

[3] IRR Asia vs. Middle East, Latin America, Africa: 1.27 (95%CI 1.06-1.51), North America, Europe, Pacific, vs. Middle East, Latin America, Africa: 1.23 (95%CI 1.02-1.47)

[4] IRR Asia vs. Middle East, Latin America, Africa: 1.30 (95%CI 1.07-1.57), North America, Europe, Pacific, vs. Middle East, Latin America, Africa: 1.27 (95%CI 1.04-1.54)

Supplementary Files

Multimedia Appendixes

Associations between sentiment/politeness scores and first/last authors' gender and first authors' affiliation in this cross-sectional study using ChatGPT 4.0 to quantitatively assess sentiment and politeness in 240 peer review reports for 96 articles published in 2023 (same data as in Table 2 but without taking into account the 51 peer-review reports for the 12 articles published in BMJ).
URL: http://asset.jmir.pub/assets/dc144172945b44a6e009fb04e507cbef.docx

List and description of variables available in the STATA file uploaded to the Open Science Framework (OSF).
URL: http://asset.jmir.pub/assets/a775c139e45cb8304cfd09d2ec737b99.docx