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

Research Letter

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: 750

Number of references: 5

Number of tables and/or figures: 2

Keywords: Africa; ChatGPT; Disparity; Gender; Geographic; Global South; Inequality; Peer review; Woman

Introduction

In the evolving landscape of scientific research, 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. Verhaven showed that ChatGPT was accurate in determining sentiment and politeness scores in peer reviews of scientific articles [1]. 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 [2,3].

Building upon Verhaven's work, we used chatGPT 4.0 to quantitatively assess sentiment and politeness in peer review reports from nine high-impact medical journals, exploring geographical and gender disparities to enhance inclusivity within the peer review process.

Methods

We searched the Clarivate and ASAPbio websites to retrieve the nine general medical journals with a JCR impact factor>2 that publish peer review reports (Table 1). Using simple randomization, we selected twelve research articles per journal, published in 2023. We extracted first/last authors' names and first authors' country of affiliation. We determined first/last authors' gender by consulting websites with photos.

For each review, we asked chatGPT 4.0 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). We repeated the measurements five times and removed the minimum and maximum values. We therefore had three values for each score. The sentiment score measures how favorable the review is, the politeness score how polite a review's language is. All the data was collected in January 2024.

We calculated the mean sentiment/politeness scores for each review, rounded to the nearest whole number. We summarized these data using the median (IQR), overall and by journal/gender/affiliation. For affiliation, we grouped the countries into three regions (North America/Europe/Pacific, Asia, and Latin America/Middle East/Africa), as done elsewhere [4]. We compared the results using Wilcoxon rank-sum tests and Kruskal-Wallis rank tests. After adding 100 to the sentiment/politeness scores (minimum value=0, maximum value=200), we performed negative binomial regressions, adjusting for journal/affiliation/intra-cluster correlation within articles. Finally, we calculated quadratic weighted agreement coefficients (percent agreement and Fleiss' Kappa). All analyses were performed with STATA 15.1.

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). By contrast, articles whose authors were affiliated with countries in the Middle East/Latin America/Africa had significantly lower sentiment/politeness scores than the other two regions, with differences exceeding 30 and 20 points in absolute value, respectively (Table 2).

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, p-values<0.001).

Discussion

In summary, we used ChatGPT to analyze sentiment/politeness in 291 peer review reports corresponding to 108 articles published across 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.

These results diverge from Verhaven's study, which demonstrated that female first authors tended to receive less polite reviews(lower politeness scores), while female last authors received more favorable reviews(higher sentiment scores) compared to men [1]. In addition, the study did not find any association with authors' affiliation. Importantly, Verhaven's research focused on a different discipline(neuroscience) and was confined to a single journal(Nature Communications).

The regional disparities highlighted in our study align with previous research illustrating the challenges faced by researchers, particularly those from countries in the Global South [4,5]. Further studies are needed to confirm our findings and explore potential explanations for these disparities, considering factors such as the dominance of English in scientific discourse or the presence of ethnic bias, whether conscious or unconscious, among researchers.

Our study has limitations, including its exclusive focus on general medical journals, reliance on photos for binary gender determination without consideration for non-binary/transgender identities, and uncertainty about the gender/geographic distribution of rejected papers as all analyzed peer reports were ultimately accepted for publication.

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.

Data availability statement: The data associated with this article are available in the Open Science Framework (https://doi.org/10.17605/OSF.IO/WNRZU).

References

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

2. 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

3. 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

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

5. 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

Table 1. List of journals included in the study, 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 and politeness scores, and first/last authors' gender and first authors' affiliation (N=291 reviews for 108 articles published across nine 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)