Journal of the American Heart Association

EDITORIAL

Important Questions Deserve Rigorous Analysis: A Cautionary Note About Selection Bias

Lucia C. Petito, PhD; Louisa H. Smith, PhD

In this issue of the *Journal of the American Heart Association (JAHA)*, the article by Millenaar et al. titled "Sex Differences in Cardiovascular Research: A Scientometric Analysis," tackles a pervasive issue in cardiovascular research: the sex gap in publishing.[1] We commend the authors (henceforth referred to as "researchers" to avoid confusion with the authors they study) for tackling this difficult topic, adding to a growing body of literature documenting both historical underrepresentation of women in and recent upwards trends in number of publications by female authors in published cardiovascular research.[2–4] As female researchers ourselves, we agree with the researchers' conclusions that structural changes—such as parity in pay and mentorship programs—are important steps toward reducing disparities between men and women in research opportunities. However, this important issue deserves to be held to rigorous standards, both for study design and statistical approach.

See Article by Millenaar et al.

The researchers assert in their discussion that over the last decade there has been "an overall increase in the number of cardiovascular research articles, driven by a relative increase in female first and last authorship position."[1] They also draw conclusions about the overall research quality of publications written by female versus male first and senior authors by comparing impact factors and H-Indexes thereof. Other researchers have pointed out limitations in inferring gender or sex in bibliometric analyses, chiefly that biological sex assigned at birth does not necessarily reflect the spectrum of gender identity, and names do not necessarily reflect either biological sex or gender identity.[5–7] That limitation notwithstanding, we have chosen to focus our comments here on the fact that author sex was unable to be ascertained for over 30% of articles in the sample—a number that is compatible with that found in other bibliometric analyses on this topic.[1,4] Accordingly, our criticism pertains to any analysis, observational or experimental, with a substantial amount of missing data.

The researchers initially retrieved 387 463 articles from the Web-of-Science Core Collection.[8] After processing them through the Science Performance Evaluation web application and using a Python library (https://pypi.org/project/SexMachine/) to assign author sex, they excluded 117 636 articles—more than 30% of their sample—owing to their inability to assign the sex of the first author.[1,9] The researchers present some data about the articles with first authors of unknown sex in the Supplemental Material but otherwise do not address this issue in their analysis. However, the characteristics of the excluded articles differ in almost all respects from those of the included articles. Missing data such as these can introduce substantial *selection bias*.

Key Words: cardiovascular research ■ Editorials ■ missing data ■ selection bias

Correspondence to: Lucia C. Petito, PhD, Northwestern University Feinberg School of Medicine, 680 N Lake Shore Dr, Chicago, IL 60611.
E-mail: lucia.petito@northwestern.edu
For Disclosures, see page 3.

JAHA is available at: www.ahajournals.org/journal/jaha

For example, in the results section the researchers place emphasis on global findings, namely that "female cardiovascular publications were distributed unevenly around the globe," with the majority of female first-author articles from European and North American countries.[1] However, likely because Asian names are more difficult to classify via algorithm, a substantial number of articles written by first authors from Asia were excluded.[10] In their initial data pull, 102 733 papers were from Asian countries, but the researchers' analytic sample included only 44 194 articles from Asian countries and excluded 57 913 articles—56% of those from the initial pull—owing to unknown sex of the first author.[1] The researchers found that Asian countries had the lowest ratio of female to male first authors, at 0.4. To understand the possible selection bias, we can compute what percentage of the missing authors would have to be female for the researchers' conclusions to change (Figure). If just 41% of those authors with unknown sex were truly female, the female to male ratio would be 0.55, the same as was observed in North America (12 640+(0.41×57 913))/(31 554+(0.59×57 913)). If nonclassifiable Asian names were equally likely to be female as male, then the ratio would be 0.69, which would imply that the ratio observed in Asia was more favorable toward female authors than in North America. Moreover, if 66% or greater of nonclassifiable publications were written by female authors, the ratio would be >1.0, implying more publications by female first authors than male in Asian countries. Understanding how the missing data would need to be distributed to invalidate the findings from an analysis can shed light on the plausibility of those distributions.

Such calculations are a form of *sensitivity analysis*, in which assumptions about the data or analysis methods are varied in order to understand how much the results would change. Deciding whether it is possible that as many as 41% or 50% of the missing Asian authors were female requires more knowledge about the

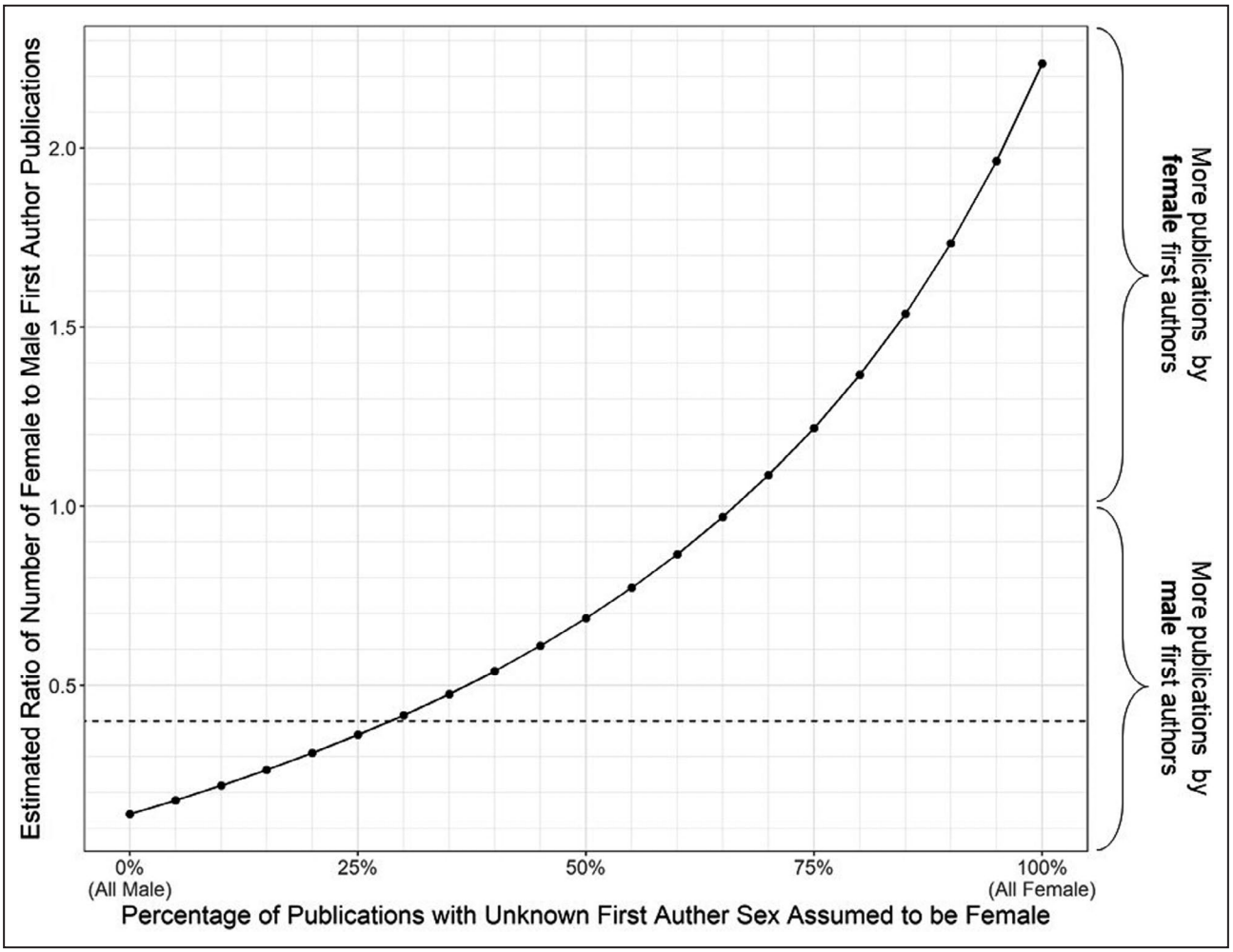

Figure 1. Estimated ratio of number of female to male first author publications under varying assumptions of what percentage of publications with unknown sex are truly written by female first authors.
The dashed line indicates the observed ratio in the nonmissing data presented in Millenaar et al.[1]

data, but we can get some hints from what the authors provided. For example, among all last authors with assigned sex, those who coauthored with first authors of unknown sex are relatively more likely to be female than those whose coauthors also had an assigned sex.[1] Patterns of sex homophily in this study and others tell us that those with unknown sex are therefore likely more often female than those with known sex.[11] A sensitivity analysis may incorporate other background knowledge as well; for example, many English-language names that were historically male are now given to girls, which may make certain names belonging to women more difficult to classify.[12]

Other forms of sensitivity analysis may involve repeatedly assigning values for sex to the missing data and repeating the analysis.[13] The assigned values would be drawn according to some probability distribution. A validation study in which the researchers hand-code sex in a small random sample of the authors with missing data could inform those probabilities. When multiple variables are missing—for example, first and last author sex—they can be drawn together from a joint probability distribution. Other sensitivity analysis methods for selection bias are applicable to specific statistical parameters, such as the odds ratio from a case-control study.[14]

Of course, sensitivity analysis does not and cannot tell us what the truth is; in a world with infinite resources and time, the researchers would make a heroic effort to reach out to the corresponding authors of all publications with unknown author sex and ask for more information. In fact, prior work that has focused on a more limited number of journals has done exactly this.[2] Even then, with a larger sample size of journals, nonresponse could likely result in some remaining bias. Although selection bias can seem like an intractable problem, better data sources and study designs can help avoid it altogether, statistical methods can help correct for it, and sensitivity analysis can help us begin to understand its extent and implications in a particular situation.[15]

There is no disputing the fact that sex differences have been observed in authorship of published cardiovascular health research for decades. It is of paramount importance that women have equitable opportunities to build and thrive in careers in cardiovascular health research and that structural barriers are recognized and removed to promote their success. Although we should all celebrate that the number of female first-author cardiovascular publications has increased in the last decade, we should consider whether these findings are our own confirmation bias speaking or if they arise from unaddressed selection bias. When important questions are on the line, it is incumbent on the research community to hold the quality of evidence to a higher standard.

ARTICLE INFORMATION

Affiliations

Division of Biostatistics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL (L.C.P.); and Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA (L.H.S.).

Acknowledgments

We would like to thank Sadiya Khan for helpful comments on an earlier draft of this work.

Disclosures

Dr Petito receives research support for unrelated work from Omron Healthcare, Inc. Dr Smith has no disclosures to report.

REFERENCES

1. Millenaar D, Dillmann M, Fehlmann T, Flohr A, Mehran R, Al-Lamee R, Lauder L, Ukena C, Bohm M, Keller A, et al. Sex differences in cardiovascular research: a scientometric analysis. *J Am Heart Assoc*. 2021;10:e021522. doi: 10.1161/JAHA.121.021522
2. Asghar M, Usman MS, Aibani R, Ansari HT, Siddiqi TJ, Fatima K, Khan MS, Figueredo VM. Sex differences in authorship of academic cardiology literature over the last 2 decades. *J Am Coll Cardiol*. 2018;72:681–685. doi: 10.1016/j.jacc.2018.05.047
3. 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;352:i847. doi: 10.1136/bmj.i847
4. Lerchenmüller C, Lerchenmueller MJ, Sorenson O. Long-term analysis of sex differences in prestigious authorships in cardiovascular research supported by the National Institutes of Health. *Circulation*. 2018;137:880–882. doi: 10.1161/CIRCULATIONAHA.117.032325
5. Mihaljevic H, Tullney M, Santamaria L, Steinfeldt C. Reflections on gender analyses of bibliographic corpora. *Front Big Data*. 2019;2:29. doi: 10.3389/fdata.2019.00029
6. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. *Res Integr Peer Rev*. 2016;1:2. doi: 10.1186/s41073-016-0007-6
7. Natias JN. How to ethically and responsibly identify gender in large datasets. In: Glaser M, ed. *MediaShift*. Vol 2021. Available at: http://mediashift.org/2014/11/how-to-ethically-and-responsibly-identify-gender-in-large-datasets/. Accessed July 29, 2021.
8. Clarivate Analytics Web of science core collection: introduction. Published 2021. Available at: https://clarivate.libguides.com/woscc/basics. Accessed July 29, 2021.
9. Scholz SS, Dillmann M, Flohr A, Backes C, Fehlmann T, Millenaar D, Ukena C, Bohm M, Keller A, Mahfoud F. Contemporary scientometric analyses using a novel web application: the science performance evaluation (SciPE) approach. *Clin Res Cardiol*. 2020;109:810–818. doi: 10.1007/s00392-019-01568-x
10. Santamaria L, Mihaljevic H. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput Sci*. 2018;4:e156. doi: 10.7717/peerj-cs.156
11. Zhang C, Bu Y, Ding Y, Xu J. Understanding scientific collaboration: homophily, transitivity, and preferential attachment. *J Assoc Inf Sci Technol*. 2018;69:72–86. doi: 10.1002/asi.23916
12. Blevins C, Mullen L. Jane, John … Leslie? A historical method for algorithmic gender prediction. *Digit Humanit Q*. 2015;9:e1–e18. Available at: http://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html. Accessed July 29, 2021.
13. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer Science & Business Media; 2011. ISBN: 978-0-387-87960-4.
14. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol*. 1996;25:1107–1116. doi: 10.1093/ije/25.6.1107-a
15. Smith LH. Selection mechanisms and their consequences: understanding and addressing selection bias. *Curr Epidemiol Rep*. 2020;7:179–189. doi: 10.1007/s40471-020-00241-6