A call for transparency in gender assignment approaches

Elvira González-Salmón and Nicolas Robinson-Garcia

Unit for Computational Humanities And Social Sciences (U-CHASS), EC3 Research group, University of Granada, Granada (Spain)

Dear Scientometrics Editors,

The interest in studying gender differences in science has increased over the last decade in the field of bibliometrics. Calls for diversity in science (i.e., CoARA[1]), evidence demonstrating a gender gap in science (Sugimoto & Larivière, 2023), as well as the expansion of algorithmic approaches for author disambiguation (Tekles & Bornmann, 2020) and gender assignment (Mihaljevíc et al., 2019) have greatly helped explore publication and citation patterns within the scientific workforce. This qualitative leap has been made possible partly by an improvement in the quality of metadata produced by major bibliometric data providers[2]. By exploring gender inequalities in science, the scientometric community has raised awareness on this topic and explored the mechanisms producing such inequalities.

However, we have noted an important lack of rigour when reporting the use of algorithms that assign gender to names, which play a major role on the findings reached by these papers. These algorithms make important assumptions that have not been tested. First, they assume that gender can be inferred based on names (or images of faces), which is not necessarily true. Just because a name is usually associated with one gender, it does not mean that the individual carrying that name identifies with said gender. Moreover, names are not always associated with gender in the first place, which leads to the second limitation: there are many given names which are unisex (can be applied both to male and female authors), depending on the author's country of origin. Third, they consider gender as a binary variable, making invisible other identities such as non-binary or trans authors (Rasmussen et al., 2019; Lindqvist et al. 2021).

There are further limitations however, which, on many occasions, are not reported or are overlooked. Gender algorithms usually work better with Western (and English) names than with Asian names, as current methods have performed poorly in non-roman names and, overall, non-Western names (Karimi et al., 2016). Geographically unequal representations of gender in global analyses can lead to biased findings[3]. This is related with the use that gender assignment algorithms make of lists of gendered names as a fundamental component. These lists are often not reported in the studies. It is critical to understand how they are composed as some do not consider the cultural and regional variations that can exist within countries. A notable example of this limitation is evident in Slavic countries, where gender assignment based solely on given names is less efficient than focusing on both first names and surnames, since gender information can be found in surnames (Mryglod et al., 2023). Moreover, it is important to recognize that some

[1] https://coara.eu/agreement/the-agreement-full-text/

[2] Web of Science started including author full names in 2007 (they are searchable in the database since 2011) and Scopus announced their inclusion in 2022. Also now most bibliometric databases include their own researcher identifier.

[3] They may also lead to inaccurate findings (i.e. Andrea is commonly assigned to men in Italy and to women in Spain)

algorithms, such as NamSor and Gender API, do not transparently report the sources of their name-gender lists, leaving room for uncertainty regarding their origin and reliability. Therefore, advocating for increased transparency in the description of gender assignment methods in gender-related research is essential to address these limitations and promote more robust and inclusive practices within the field.

Table 1 includes a brief analysis that illustrates the extent to which transparency is needed in these studies. We retrieved journals articles published since 1981 responding to the following search query in Scopus:

TITLE-ABS-KEY (gender) OR TITLE-ABS-KEY (wom?n) OR TITLE-ABS-KEY (*male) AND (LIMIT-TO (EXACTSRCTITLE , "Scientometrics"))

Method	Geographical variable	Includes validation	Limitation (non-Western names)	Limitation (binary)	Limitation (others)	Total	%
Self-reported	0	0	0	1	0	21	9.5%
Manual (online search, pictures, etc.)	3	0	0	0	1	63	28.4%
Self-developed method	0	2	0	1	1	8	3.6%
Self-developed method (rule-based algorithm)	7	5	5	2	2	11	5%
External API	4	3	1	2	2	19	8.6%
Official data	0	0	0	0	0	36	16.2%
Name lists	0	1	2	0	2	2	1%
Not reported	0	0	0	0	0	60	27%
Methods from other papers	1	0	0	0	1	2	1%
	15	11	8	6	9	222	100%

We retrieved a total of 271 records out of which 222 used some sort of method to infer gender from their dataset. 28.4% assigned gender manually by doing online searches or based on the researchers' knowledge of gendered names. 27% of the articles analysed did not report how they got the gender information. Then, 16.2% of articles got gender information from secondary official data which already assigned gender to its subjects. In most cases this was governmental or university data. 19 articles (8.6%) used third-party algorithms (e.g., genderize.io, Gender API). The drawback of these methodologies is the lack of replicability they allow. The first three cases (manual assignment, no information and secondary data) are impossible to track back, however, the use of algorithms is no easier to examine for robustness. For instance, there is no information about where data from NamSor's Gender Guesser comes from[4]. Gender API, another commonly

[4] https://gender-guesser.com/

used service, simply states that data comes from "publicly available data, governmental data and manual additions/corrections"[5].

However, in recent years, research has started to focus more extensively on this methodological issue, and we found 19 articles (8.6%) that designed their own method to assign gender, either from scratch or combining previous methodological approaches. In this last group we find exemplary cases of transparent, robust and replicable reporting on gender assignment. This is the case of Ma et al. (2023), that recognizes the challenge of assigning gender to names and its binary nature, producing a method to assign gender to their dataset. Fell and König (2016) included a step-by-step validation of their initial results. El-Ouahi and Larivière (2023) dedicate an Appendix to discuss their gender assignment method. Chan and Torgler (2020) include a detailed account of the combination of methods used in their supplementary material.

These examples demonstrate that another way of conducting gender assignment is possible. Devoting time and space to explain the gender disambiguation process is not only feasible but essential to understand caveats and critically contrast findings with previous research. Thus, we make a call for transparency when reporting gender assignment. Moreover, good research needs to be replicable. Providing a clear methodology and allowing replicability is of great importance for the development of science. We encourage all researchers to apply these principles to their research.

References

Chan, H. F., & Torgler, B. (2020). Gender differences in performance of top cited scientists by field and country. *Scientometrics*, 125(3), 2421–2447. Scopus. https://doi.org/10.1007/s11192-020-03733-w

El-Ouahi, J., & Larivière, V. (2023). On the lack of women researchers in the Middle East and North Africa. *Scientometrics*, 128(8), 4321–4348. Scopus. https://doi.org/10.1007/s11192-023-04768-5

Fell, C. B., & König, C. J. (2016). Is there a gender difference in scientific collaboration? A scientometric examination of co-authorships among industrial–organizational psychologists. *Scientometrics*, 108(1), 113–141. Scopus. https://doi.org/10.1007/s11192-016-1967-5

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 53–54. https://doi.org/10.1145/2872518.2889385

Ma, Y., Teng, Y., Deng, Z., Liu, L., & Zhang, Y. (2023). Does writing style affect gender differences in the research performance of articles?: An empirical study of BERT-based textual sentiment analysis. *Scientometrics*, 128(4), 2105–2143. Scopus. https://doi.org/10.1007/s11192-023-04666-w

Mihaljević, H., Tullney, M., Santamaría, L., & Steinfeldt, C. (2019). Reflections on Gender Analyses of Bibliographic Corpora. *Frontiers in Big Data*, 2. https://www.frontiersin.org/articles/10.3389/fdata.2019.00029

Mryglod, O., Nazarovets, S., & Kozmenko, S. (2023). Peculiarities of gender disambiguation and ordering of non-English authors' names for Economic papers beyond core databases. *Journal of Data and Information Science*, 8(1), 72–89. https://doi.org/10.2478/jdis-2023-0001

[5] https://gender-api.com/en/frequently-asked-questions

Lindqvist, A., Sendén, M. G., & Renström, E. A. (2021). What is gender, anyway: A review of the options for operationalising gender. *Psychology & Sexuality*, 12(4), 332–344. https://doi.org/10.1080/19419899.2020.1729844

Rasmussen, K. C., Maier, E., Strauss, B. E., Durbin, M., Riesbeck, L., Wallach, A., Zamloot, V., & Erena, A. (2019). *The Nonbinary Fraction: Looking Towards the Future of Gender Equity in Astronomy* (arXiv:1907.04893; Version 1). arXiv. https://doi.org/10.48550/arXiv.1907.04893

Sugimoto, C. R., & Larivière, V. (2023). *Equity for Women in Science: Dismantling Systemic Barriers to Advancement*. Harvard University Press.

Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches1. *Quantitative Science Studies, 1*(4), 1510–1528. https://doi.org/10.1162/qss_a_00081