DANISH MEDICAL JOURNAL

Letter

Correspondence on »Five major challenges for medical bibliometrics«

Myo Tha & Nilar Khin

Faculty of Medicine, MAHSA University, Selangor, Malaysia.

Dan Med J 2026;73(3):A300017. doi: 10.61409/A300017

We read with great interest the recent Danish Medical Journal article »Five major challenges for medical bibliometrics« , which provides a timely overview of limitations affecting the interpretation of bibliometric indicators in medical science. The discussion of database accuracy is particularly relevant, as the authors demonstrate that major bibliometric databases produce markedly different coverage for identical publications, raising questions about the reliability of database-derived metrics for academic evaluation and funding decisions.

The author's example, showing wide variation in reference coverage across major databases, underscores how bibliometric indicators may depend as much on database choice as on scientific output itself . However, while the authors appropriately emphasise inter-database differences in coverage, we would like to highlight a complementary source of variation: the structural characteristics of bibliographic data within databases and the preprocessing required to analyse them – particularly evident in PubMed-based bibliometrics.

PubMed remains an indispensable resource for biomedical bibliometrics owing to its free public accessibility and discipline-specific coverage . However, PubMed-based studies face two important methodological limitations: first, author affiliations are stored as unstructured free-text strings with inconsistent formatting and geographic descriptors ; and second, the preprocessing steps required to clean and interpret these strings are frequently insufficiently documented, creating opacity in analytical workflows . Consequently, for medical bibliometrics, discrepancies attributed solely to database coverage may in part reflect differences in affiliation preprocessing practices, underscoring the need for explicit reporting of data-cleaning and geographic disambiguation procedures.

Such opacity has direct consequences: bibliometric results derived from PubMed depend critically on how affiliation data are cleaned, parsed, and interpreted, yet these analytical choices remain largely invisible in published work. Discrepancies attributed to database limitations may therefore partly reflect undocumented preprocessing decisions rather than differences in database coverage

alone.

This issue also carries implications for equity in research evaluation. While subscription-based databases such as Web of Science and Scopus are readily available in well-resourced institutions, access remains limited for many researchers globally, particularly in low- and middle-income settings. Script-based preprocessing workflows further compound these barriers, as they require programming skills unevenly distributed across disciplines and regions. Database accuracy cannot therefore be separated from the accessibility and transparency of the analytical workflows applied to the data.

We suggest that discussions of database accuracy in medical bibliometrics explicitly distinguish between database coverage and preprocessing practices. Greater emphasis on reporting affiliation cleaning and geographic disambiguation as explicit methodological steps would improve reproducibility and help contextualise discrepancies between databases . Such an approach would complement the authors' call for critical awareness of bibliometric limitations and support more responsible use of bibliometric indicators in medical science.

Correspondence *Myo Tha*. E-mail: ktmyo@mahsa.edu.my

Published 10 February 2026

Conflicts of interest none. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. These are available together with the article at ugeskriftet.dk/DMJ.

Cite this as Dan Med J 2026;73(3):A300017

doi 10.61409/A300017

REFERENCES

1. Rehfeld JF. Five major challenges for medical bibliometrics. Dan Med J. 2026;73:1-6. https://doi.org/10.61409/A09250723
2. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. 2008;22:338-342. https://doi.org/10.1096/fj.07-9492LSF
3. Lee B, Brownstein JS, Kohane IS. Geoinference of author affiliations using NLP-based text classification. Sci Rep. 2024;14. https://doi.org/10.1038/s41598-024-73318-7
4. Nowakowska M. A comprehensive approach to preprocessing data for bibliometric analysis. Scientometrics. 2025. https://doi.org/10.1007/s11192-025-05415-x
5. Peng RD. Reproducible research in computational science. Science. 2011;334:1226-1227. https://doi.org/10.1126/science.1213847