



Research quality metrics as a complement to bibliometrics: an opinion paper

Harald Hannerz¹ · Kristian Schultz Hansen¹ · Jan Hyld Pejtersen² · Helene Kless Willadsen¹

Received: 26 May 2026 / Accepted: 21 September 2026
© The Author(s) 2026

Abstract

Substantial proportions of research investment may be wasted due to flawed study design, biased reporting, and publication bias, raising concerns about how research performance is evaluated. Despite this, current assessment systems rely heavily on bibliometric indicators such as publication counts, co-authorship, and citation metrics, which do not reliably reflect research quality or scientific rigor. In this opinion paper, we argue that such indicators may reward practices that undermine the accumulation of reliable knowledge, including p-hacking, selective reporting, and underpowered study designs, while undervaluing essential contributions such as replication research. We propose that bibliometric measures should be complemented by transparent and auditable research quality metrics, including indicators of reproducibility, statistical power, risk of bias, and adherence to reporting standards. We further suggest that recent advances in large language models may facilitate the scalable implementation of such metrics. Adopting quality-focused indicators could improve research evaluation and help reduce avoidable research waste.

Keywords Bibliometrics · Research waste · Research quality indicators · Large language models

Introduction

Research investments may be lost due to poor quality of studies. Chalmers and Glasziou (2009) estimated that approximately 85% of global biomedical research investment is avoidably wasted due to irrelevant research questions, flawed study design or execution, biased or unusable reporting, and publication bias. Similar concerns have been raised in other fields. Purgar et al. (2022) estimated that 82–89% of ecology research may be avoidably wasted, and a systematic review by Pejtersen et al. (2015) concluded that 95% (42 of 44) of observational studies examining the association between occupational psychosocial

✉ Harald Hannerz
hha@nrcwe.dk

¹ National Research Centre for the Working Environment, Lersø Parkallé 105, 2100 Copenhagen, Denmark

² The Danish Center for Social Science Research, Copenhagen, Denmark

factors and ischemic heart disease were effectively uninformative because of inadequate statistical power.

These estimates suggest substantial room for improving research quality and considerable potential for reducing unnecessary research expenditures. From this perspective, taxpayers and charitable donors who fund research have a legitimate interest in ensuring that the quality of the research they support is adequately monitored and assessed. However, while most research institutions and funding bodies provide bibliometric indicators intended to demonstrate the academic impact of their research output, very few report metrics that explicitly assess research quality. We argue that this imbalance reflects a widespread misconception: that bibliometric indicators—largely based on patterns of co-authorship and citation—constitute reliable proxies for research quality and researcher competence. This concern is closely aligned with the FRAME initiative (a framework for the responsible use of bibliometrics in research evaluation), which promotes responsible and context-sensitive use of bibliometric indicators in research assessment (Daraio et al., 2025).

In this commentary, we first offer brief reflections on the use of co-authorship and citation metrics in the evaluation of scientific work, highlighting the need for caution in interpreting these measures. We then outline potential approaches to complement bibliometric indicators with metrics that more directly assess research quality and, by extension, research waste.

Limitations of co-authorship as a performance indicator

The number of co-authorships is a good measure of how successful a researcher has been in the pursuit of prestige and academic credit, but it is not a good measure of the researcher's scientific competence or actual productivity.

The same scientific output could result in anything from zero to a very large number of co-authorships, depending on (i) how one interprets the Vancouver authorship guidelines, (ii) how skilled one is at claiming credit for other people's work, and (iii) how comfortable one is with taking credit for other people's work.

Example: Supervising a PhD student who writes and publishes three peer-reviewed articles can lead to:

- A. Zero co-authorships if the supervisor holds that supervision, academic sparring, and proofreading do not justify authorship
- B. Three co-authorships if the supervisor lists himself as a “co-author” on all three articles
- C. 33 co-authorships if the supervisor invites ten colleagues to list themselves as “co-authors” on the PhD student's three articles—in return for being allowed to become a “co-author” on all articles written by the colleagues' PhD students.

Coercive authorship is particularly difficult to prevent when senior researchers can use their position to demand authorship from junior colleagues. Such practices persist despite longstanding authorship guidelines, such as the ICMJE criteria, and contributor-role frameworks, such as CRediT, which generally lack independent verification and enforcement (Marcelino, 2026). It should, however, be noted that a large number of co-authors on a research paper is not necessarily indicative of questionable authorship practices. In some

cases, it may be fully justified—for example, in multi-center studies where each participating center contributes substantial research efforts and financial resources.

Limitations of citation-based metrics

The number of citations is often used as an indicator of an article's academic impact, but it is not necessarily a reliable measure of its scientific quality. In fact, a paper's visibility and citation count can be amplified through practices that do not contribute to—and may even undermine—the accumulation of reliable knowledge. Below are some examples of practices that hinder the pursuit of truth but advance the pursuit of academic recognition—and vice versa.

P-hacking, HARKing, and selective reporting

P-hacking, HARKing, and selective reporting are practices that can seriously compromise the integrity of scientific research. P-hacking refers to the manipulation of data analysis—for example, by trying multiple statistical tests, selectively excluding observations, or stopping data collection once a desired result is achieved—until statistically significant results (typically $p < 0.05$) are obtained (Simonsohn et al., 2014). HARKing (Hypothesizing After the Results are Known) involves presenting hypotheses that were formulated after inspecting the data as though they had been specified in advance (Kerr, 1998). Selective reporting further exacerbates the problem by highlighting statistically significant or theoretically convenient findings while omitting null or contradictory results (cf. Elston, 2021). These practices are quite common (Fraser et al., 2018; John et al., 2012; Schneider et al., 2024) and together they distort the scientific record. They increase the likelihood of false positives, inflate reported effect sizes, and produce biased and potentially misleading conclusions. As a result, they undermine the reliability and cumulative nature of scientific knowledge. At the same time, such practices can make findings appear more novel, coherent, and statistically convincing. This can increase the likelihood of publication in high-impact journals and lead to greater visibility, higher citation counts, and enhanced academic recognition. From the perspective of career advancement, these practices may therefore appear advantageous—despite their detrimental effects on scientific validity and long-term credibility.

Statistical testing without sufficient power

Random error is as important as systematic error when evaluating scientific evidence. The standard error (SE) reflects the expected deviation of an estimate from the true value; larger SEs imply less precise estimates and lower statistical power. Severely underpowered studies are therefore prone to producing estimates that deviate substantially from the truth.

Studies with low statistical power are less likely to detect true effects and, consequently, less likely to yield statistically significant results. This creates a selection mechanism: only those low-powered studies that, by chance, produce large and statistically significant effects are likely to be published. This phenomenon, known as publication bias, leads to a distorted literature in which reported effects are systematically exaggerated.

Statistical testing without sufficient power is therefore detrimental to the pursuit of knowledge. However, it may be advantageous in the pursuit of academic status and credentials. As noted by Forstmeier et al. (2017), “studies which report surprising (i.e. unlikely)

findings are often highly cited and thus contribute to the stature of the journal. Thus, ‘the more surprising, the better’.” Similarly, Higginson and Munafò (2016) argue that “researchers acting to maximize their fitness [publication record and career success] should spend most of their effort seeking novel results and conduct small studies that have only 10–40% statistical power.”

Well-powered replication studies

Replication studies play a crucial role in strengthening scientific knowledge by verifying whether findings are robust, reliable, and reproducible. Well-powered replications, in particular, reduce uncertainty and help distinguish genuine effects from spurious ones. However, replication studies—especially those that confirm or fail to confirm existing findings—are often perceived as less novel and therefore attract fewer citations. They may also face greater difficulty in being published in high-ranking journals. Consequently, while well-powered replication studies are essential for the advancement of knowledge, they are often undervalued within current academic incentive structures. In this sense, replication studies are beneficial for the pursuit of truth but may be disadvantageous in the pursuit of status and academic recognition.

Summarizing table

The described examples of practices that hinder the pursuit of truth but advance the pursuit of academic recognition—and vice versa are summarized in Table 1.

Here, it should be noted that, although the quest for bibliometric scores incentivizes questionable research practices, it is definitely not the sole source of the problem. Questionable research practices similar to those listed in the table existed long before the introduction of modern bibliometric measures. An example is data cooking, a term used by Charles Babbage (1830), who described it as “an art of various forms, the object of which is to give to ordinary observations the appearance and character of those of the highest degree of accuracy. One of its numerous processes is to make multitudes of observations, and out of these to select those only which agree, or very nearly agree. If a hundred observations are made, the cook must be very unlucky if he cannot pick out fifteen or twenty which will do for serving up. Another approved receipt, when the observations to be used will not come within the limit of accuracy, which it has been resolved they shall possess, is to calculate them by two different formulae. The difference in the constants employed in

Table 1 Examples of activities that hinder the pursuit of truth but advance the pursuit of academic recognition—and vice versa

Type of activity	Rating in relation to the quest for knowledge	Rating in relation to the quest for status and credentials
P-hacking	Bad	Good
Harking	Bad	Good
Selective reporting	Bad	Good
Statistical testing without sufficient power	Bad	Good
Well-powered replication studies	Good	Bad

those formulae has sometimes a most happy effect in promoting unanimity amongst discordant measures. If still greater accuracy is required, three or more formulae can be used.”

Toward metrics that directly assess research quality

A meaningful complement to bibliometric indicators would be the routine reporting of transparent and auditable quality metrics at the level of academic journals, research groups, departments, institutions, and funding bodies. Several group-level research-quality measures have been proposed, and some have already been implemented.

The Rigor and Transparency Index (RTI) is a group-level measure of scientific reporting quality, primarily designed for biomedical and biological laboratory research (<https://sciscore.com>). It is defined as the average score generated by SciScore across a set of published articles. SciScore is an AI-assisted tool that scores individual papers on a 0–10 scale according to how well they report methodological details relevant to transparency and reproducibility (cf. Menke et al., 2020, 2022).

Another example is the Transparency and Openness Promotion (TOP) Factor, a numerical scoring tool introduced by the Center for Open Science in 2020 (cf. Grant et al., 2026). It was designed to assess journal policies related to open science and transparency, thereby complementing citation-based metrics such as the Journal Impact Factor. However, TOP Factor scores were no longer added or updated after February 2025, reflecting documented limitations of journal-level scoring, including concerns about the reliability of the scores and the interpretation of journal policies.

A third example is the Research Excellence Framework (REF), the United Kingdom’s official system for assessing the quality of research conducted in higher education institutions. The REF rates research quality according to the following scale:

- 4*—Quality that is world-leading in terms of originality, significance, and rigour.
- 3*—Quality that is internationally excellent in terms of originality, significance, and rigour but which falls short of the highest standards of excellence.
- 2*—Quality that is recognised internationally in terms of originality, significance, and rigour.
- 1*—Quality that is recognised nationally in terms of originality, significance, and rigour.
- U/C—Quality that falls below the standard of nationally recognised work, or work that does not meet the published definition of research for the purposes of the assessment.

A limitation of the latter approach is that it relies substantially on expert assessment and therefore involves an element of subjective judgement. Below, we outline a set of concrete research-quality indicators that may be used to evaluate different types of empirical research without relying on expert judgement or requiring prior clinical knowledge of the subject area being investigated.

Examples of proposed research quality metrics

- Proportion of reported studies with publicly available (ideally preregistered) protocol and analysis plans.

- Proportion of reported studies with raw data and analytical algorithms publicly available within 6 months after publication of a study report.

(cf. Ioannidis et al., 2014).

- Proportion of confirmatory statistical analyses that have an acceptable statistical power or precision.

(cf. Jennions & Møller, 2003; Pejtersen et al., 2015).

- Proportion of studies with a pre-specified statistical analysis plan that were conducted without changing, adding, or omitting any primary outcome.

(cf. Chan et al., 2004).

- Proportion of research studies with loss to follow-up below a predefined negligible threshold (e.g., <5%).

(cf. Dettori, 2011).

- Average percentage of applicable reporting-guideline items adequately addressed across relevant studies.

- o CONSORT (randomized controlled trials)
- o TREND (non-randomized intervention studies)
- o STROBE (observational studies)
- o PRISMA (systematic reviews)
- o ARRIVE (animal studies)
- o COREQ [qualitative studies (indepth interviews and focus groups)]

(cf. Des Jarlais et al., 2004; Moher et al. 2001; Page et al., 2021; Percie du Sert et al., 2020; Tong et al., 2007; von Elm et al., 2007)

- The Rigor and Transparency Index (RTI) for biomedical and biological laboratory studies.

(cf. Menke et al., 2020, 2022).

- Average percentage of affirmative answers to ReMarQ items across systematic reviews.

ReMarQ [Reported methodological quality of medical systematic reviews], consisting of dichotomous items and whose application does not require subject content expertise, may be important (i) in supporting an efficient quality assessment of systematic reviews and (ii) as the basis of automated processes to support that assessment.” (Marques-Cruz et al., 2025a).

Hypothetical research quality report

To illustrate what a summary statement of an institute's research output might look like when traditional bibliometric measures are complemented by indicators of scientific rigor and transparency, we provide an account of the research output of a hypothetical OHS research institute, hereafter referred to as The Institute.

In 2028, The Institute published 200 articles in peer-reviewed journals, corresponding to 3 publications per researcher full-time equivalent (FTE), of which 132 (66%) were published in top-tier journals. Using fractional counting, whereby publication credit is divided equally among all co-authors, The Institute's contributions to these 200 publications corresponded to 20 full publications (0.3 per researcher FTE).

Of the peer-reviewed studies, 100 were either randomized trials ($n = 25$), non-randomized intervention studies ($n = 25$), observational follow-up studies ($n = 40$), or case-control studies ($n = 10$) examining the effects of exposures among human participants. In these studies, 50% of the reported analyses had a prespecified and publicly available statistical analysis plan, 10% had adequate statistical power ($> 80\%$) to detect a weak effect of the intervention or exposure under study on the primary outcomes, and 40% had less than 5% loss to follow-up. The average percentages of adequately addressed reporting-guideline items were 78% for CONSORT, 56% for TREND, and 81% for STROBE. Among analyses with a prespecified statistical analysis plan, 40% were conducted without changing or adding any primary outcomes. A weak effect was defined as an absolute Cohen's d of 0.2 for continuous outcomes (cf. Cohen, 1988) and a risk ratio of 1.2 or 0.83 for binary outcomes (cf. Monson, 1990).

Twenty of the peer-reviewed studies were systematic reviews, with an average of 80% affirmative answers to ReMarQ items and an average of 75% of PRISMA items adequately addressed.

Twenty of the peer-reviewed studies were qualitative studies using in-depth interviews or focus groups, with an average of 72% of COREQ items adequately addressed.

Thirty of the journal articles reported biological laboratory experiments. The Rigor and Transparency Index for scientific reporting in these studies was estimated at 4.8 points.

The remaining 30 papers comprised cross-sectional descriptive studies ($n = 10$), study protocols ($n = 8$), economic cost analyses ($n = 3$), narrative reviews ($n = 3$), methodology papers ($n = 3$), an ethnographic study ($n = 1$), a tutorial ($n = 1$), and an opinion paper ($n = 1$).

Raw data and analytical algorithms were publicly available within 6 months after publication in 20 out of 135 quantitative empirical studies.

Discussion

As previously mentioned, the number of citations and research publications are commonly used measures of academic impact (influence), and it is natural that academic journals, funding agencies, research departments, as well as individual researchers, seek to ensure that the research they conduct or fund is as influential as possible. However, it is important to recognize that bibliometric measures alone do not distinguish between positive and negative forms of influence (cf. the table in the section on citations). Likewise, it is often unclear whether exceptionally high academic impact reflects genuinely

high research skills and productivity or whether it is partly driven by extensive use of questionable research practices and co-authorship alliances. We therefore propose that the evaluation of research outputs could be improved by complementing bibliometric measures with metrics that assess research quality more directly.

This proposal is consistent with broader initiatives promoting responsible research assessment. The San Francisco Declaration on Research Assessment (DORA) cautions against using journal-based metrics, particularly the Journal Impact Factor, as proxies for the quality of individual research contributions, and emphasizes assessment of the content of research rather than its publication venue (Cagan, 2013). Similarly, the Leiden Manifesto recommends that quantitative indicators support rather than replace qualitative expert assessment and stresses that indicators should be transparent, context-sensitive, and used with awareness of the incentives they create (Hicks et al., 2015). Empirical findings from The Metric Tide likewise showed that individual bibliometric and altmetric indicators could not provide a like-for-like replacement for peer review in the UK Research Excellence Framework (Wilsdon, 2015). These perspectives support our argument for complementing conventional bibliometrics with more direct measures of research quality, while also indicating that such measures should inform rather than mechanically determine research assessment.

In the above section, we listed some examples of research quality indicators that have been suggested in the literature and provided an illustration of what a summary statement of the research output in a hypothetical research institute might look like if traditional bibliometric measures were complemented by indicators of scientific rigor and transparency. Theoretically, such a statement could be achieved through manual post-publication peer review. However, manually reviewing hundreds of papers would be very costly and time-consuming. Hence, to make such quality indicators fully operational and practical to apply, two steps are required. First, indicator-specific inclusion criteria must be defined (i.e., whether a given study or analysis is relevant and should be included in the denominator of a particular quality measure). Second, it is necessary to find or develop a tool that can (i) process a list of references with links to the corresponding papers, (ii) examine these papers and identify the studies or analyses that meet the inclusion criteria for a given quality indicator, and (iii) assess the extent to which the corresponding success criteria have been satisfied.

We strongly believe that such a task could be accomplished by Large Language Models (LLMs) (cf. Khan et al., 2025; Mahmoudi et al., 2025) and that, accordingly, tool development would primarily involve the formulation and systematic testing of chatbot prompts. It should be noted that a validated LLM-based tool already exists for the evaluation of the quality of systematic reviews, with 20 yes/no questions for reviews without meta-analyses and six additional questions for reviews with a meta-analysis (Marques-Cruz et al., 2025b).

There have also been less successful attempts to use LLMs to evaluate research quality. For example, Thelwall (2025) discussed two attempts to estimate research quality using an LLM. The first quality indicator was based on LLM predictions of future citations, while the second involved an LLM-based rating of the Research Excellence Framework. Thelwall concluded that such LLM-based scores are less transparent than citation counts. The latter approach is a typical example of a research quality rating that relies heavily on expert assessment. In contrast, the quality indicators proposed in the present commentary typically consist of dichotomous items that can be assessed without subject-matter expertise.

Finally, it should be noted that (i) the potential quality indicators listed in the present commentary are limited to the evaluation of empirical research, (ii) the list is not

exhaustive, and (iii) the quality indicators apply only to published and peer-reviewed work and do not reflect research waste resulting from unpublished studies.

We hope that the present commentary will inspire the development and dissemination of practical tools for evaluating research quality. We further hope that research quality indicators will be widely adopted as a complement to bibliometrics in future evaluations of research outputs, as this may incentivize scientific rigor and, in turn, reduce research waste resulting from flawed study design or execution, biased or unusable reporting, and publication bias.

Author contributions The authors are listed in alphabetical order. The first draft of the manuscript was written by HKW and HH. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The writing of this opinion paper was funded by the institutions of the authors. There are no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Babbage C. Reflections on the decline of science in England, and some of its causes, 1830. <https://www.gutenberg.org/files/1216/1216-h/1216-h.htm>.
- Cagan, R. (2013). The San Francisco declaration on research assessment. *Disease Models & Mechanisms*, 6(4), 869–870. <https://doi.org/10.1242/dmm.012955>
- Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Lancet*, 374(9683), 86–89. [https://doi.org/10.1016/S0140-6736\(09\)60329-9](https://doi.org/10.1016/S0140-6736(09)60329-9)
- Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, 291(20), 2457–2465. <https://doi.org/10.1001/jama.291.20.2457>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.
- Daraio, C., Glänzel W., & Gorraiz, J. (2025). Advancing Responsible Bibliometric Practices in Research Assessment, An Introduction to the ISSI 2025 Special Track (FRAME), Proceedings of the 20th International Conference on Scientometrics & Informetrics (ISSI 2025) edited by Sargsyan, Glänzel and Abramo, Yerevan (Armenia), 23–27 June 2025, ISBN 978-9939-1-2086-7, ISSN 2175-1935, Vol I, pp. 3–5. https://doi.org/10.51408/issi2025_159.
- Des Jarlais, D. C., Lyles, C., Crepaz, N., TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health*, 94(3), 361–366. <https://doi.org/10.2105/ajph.94.3.361>
- Dettori, J. R. (2011). Loss to follow-up. *Evidence-Based Spine-Care Journal*, 2(1), 7–10. <https://doi.org/10.1055/s-0030-1267080>
- Elston, D. M. (2021). Cherry picking, HARKing, and P-hacking. *Journal of the American Academy of Dermatology*. <https://doi.org/10.1016/j.jaad.2021.06.844>

- Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings—a practical guide. *Biological Reviews of the Cambridge Philosophical Society*, 92(4), 1941–1968. <https://doi.org/10.1111/brv.12315>
- Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS ONE*, 13(7), Article e0200303. <https://doi.org/10.1371/journal.pone.0200303>
- Grant, S., Corker, K.S., Mellor, D., Stewart, S.L.K., Mayo-Wilson, E., Cashin, A.G., Lagisz, M., Moher, D., Umpierre, D., Barbour, V., Buck, S., Collins, G.S., Hazlett, H.F., Hrynaskiewicz, I., Lee, C.J., Parker, T.H., Rethlefsen, M.L., Toomey, E., Nosek, B.A. (2026). TOP Advisory Board. TOP 2025: An update to the Transparency and Openness Promotion Guidelines. *Research Integrity and Peer Review*, 11(1):40. doi: <https://doi.org/10.1186/s41073-026-00223-0>. PMID: 42557573; PMCID: PMC13439764.
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature*, 520(7548), 429–431. <https://doi.org/10.1038/520429a>
- Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology*, 14(11), Article e2000995. <https://doi.org/10.1371/journal.pbio.2000995>
- Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., & Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*, 383(9912), 166–175. [https://doi.org/10.1016/S0140-6736\(13\)62227-8](https://doi.org/10.1016/S0140-6736(13)62227-8)
- Jennions, M. D., & Möller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*, 14(3), 438–445. <https://doi.org/10.1093/beheco/14.3.438>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Khan, M. A., Ayub, U., Naqvi, S. A. A., Khakwani, K. Z. R., Sipra, Z. B. R., Raina, A., Zhou, S., He, H., Saeidi, A., Hasan, B., Rumble, R. B., Bitterman, D. S., Warner, J. L., Zou, J., Tevaarwerk, A. J., Leventakos, K., Kehl, K. L., Palmer, J. M., Murad, M. H., ... Riaz, I. B. (2025). Collaborative large language models for automated data extraction in living systematic reviews. *Journal of the American Medical Informatics Association*, 32(4), 638–647. <https://doi.org/10.1093/jamia/ocae325>
- Mahmoudi, H., Chang, D., Lee, H., Ghaffarzadegan, N., & Jalali, M. S. (2025). Critical assessment of large language models' (ChatGPT) performance in data extraction for systematic reviews: Exploratory study. *JMIR AI*, 4, e68097. <https://doi.org/10.2196/68097>
- Marcelino, R. (2026). Beyond genuine collaboration: The rise of strategic co-authorship in contemporary academic publishing. *Research Integrity and Peer Review*, 11(1), Article 12. <https://doi.org/10.1186/s41073-026-00197-z>
- Marques-Cruz, M., Pinto, F., Vieira, R. J., Bognanni, A., Perestrelo, P., Gil-Mata, S., Duarte, V. H., Barbosa, J. P., Cardoso-Fernandes, A., Martinho-Dias, D., Franco-Pego, F., Germini, F., Arienti, C., Chu, A. W. L., Riera-Serra, P., Jemiole, P., Rodrigues, P. P., Fonseca, J. A., Azevedo, L. F., ... Sousa-Pinto, B. (2025b). Use of artificial intelligence to support the assessment of the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, 187, Article 111944. <https://doi.org/10.1016/j.jclinepi.2025.111944>
- Marques-Cruz, M., Vieira, R. J., Martinho-Dias, D., Barbosa, J. P., Cardoso-Fernandes, A., Franco-Pêgo, F., Perestrelo, P., Gil-Mata, S., Taveira-Gomes, T., Pêgo, J. M., Fonseca, J. A., Azevedo, L. F., & Sousa-Pinto, B. (2025a). Reported methodological quality of medical systematic reviews: Development of an assessment tool (ReMarQ) and meta-research study. *Research Synthesis Methods*, 16(1), 175–193. <https://doi.org/10.1017/rsm.2024.14>
- Menke, J., Eckmann, P., Ozyurt, I. B., Roelandse, M., Anderson, N., Grethe, J., Gamst, A., & Bandrowski, A. (2022). Establishing institutional scores with the rigor and transparency index: Large-scale analysis of scientific reporting quality. *Journal of Medical Internet Research*, 24(6), Article e37324. <https://doi.org/10.2196/37324>
- Menke, J., Roelandse, M., Ozyurt, B., Martone, M., & Bandrowski, A. (2020). The rigor and transparency index quality metric for assessing biological and medical science methods. *iScience*, 23(11), Article 101698. <https://doi.org/10.1016/j.isci.2020.101698>
- Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*, 357(9263), 1191–1194.
- Monson, R. (1990). *Occupational epidemiology* (2nd edition). CRC Press Inc.

- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>
- Pejtersen, J. H., Burr, H., Hannerz, H., Fishta, A., & Hurwitz, E. N. (2015). Update on work-related psychosocial factors and the development of ischemic heart disease: A systematic review. *Cardiology in Review*, 23(2), 94–98. <https://doi.org/10.1097/CRD.0000000000000033>
- Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., ... Würbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biology*, 18(7), Article e3000410. <https://doi.org/10.1371/journal.pbio.3000410>
- Purgar, M., Klanjscek, T., & Culina, A. (2022). Quantifying research waste in ecology. *Nature Ecology & Evolution*, 6, 1390–1397. <https://doi.org/10.1038/s41559-022-01820-0>
- Schneider, J. W., Allum, N., Andersen, J. P., Petersen, M. B., Madsen, E. B., Mejlgaard, N., & Zachariae, R. (2024). Is something rotten in the state of Denmark? Cross-national evidence for widespread involvement but not systematic use of questionable research practices across all fields of research. *PLoS ONE*, 19(8), Article e0304342. <https://doi.org/10.1371/journal.pone.0304342>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547. <https://doi.org/10.1037/a0033242>
- Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6):349–357. doi: <https://doi.org/10.1093/intqhc/mzm042>. Epub 2007 Sep 14. PMID: 17872937
- Thelwall, M. (2025). Research quality evaluation by AI in the era of large language models: Advantages, disadvantages, and systemic effects—an opinion paper. *Scientometrics*, 130, 5309–5321. <https://doi.org/10.1007/s11192-025-05361-8>
- von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet*, 370(9596), 1453–1457. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X)
- Wilsdon, J. (2015). *The metric tide: Independent review of the role of metrics in research assessment and management*. SAGE Publications Ltd.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.