

Reproducibility Failure in Biomedical Research: Problems and Solutions

Tamarinde L. Haven¹ and John P.A. Ioannidis^{2,3}

¹Department of Methodology and Statistics, School of Social and Behavioural Sciences, Tilburg University, Tilburg, The Netherlands

²Departments of Medicine, Epidemiology and Population Health, and Biomedical Data Science, Stanford University, Stanford, California, USA; email: jioannid@stanford.edu

³Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA

 ANNUAL
REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Annu. Rev. Med. 2026. 77:537–49

First published as a Review in Advance on
October 3, 2025

The *Annual Review of Medicine* is online at
med.annualreviews.org

<https://doi.org/10.1146/annurev-med-050124-050859>

Copyright © 2026 by the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See credit lines of images or other third-party material in this article for license information.



Keywords

reproducibility, bias, peer review, publication, replication

Abstract

Reproducibility concerns in biomedical research have persisted for more than a decade, with large-scale assessments revealing significant challenges in replicating findings. Despite widespread acknowledgment of these issues, responses remain inconsistent, and proposed solutions often lack rigorous evaluation. This review examines the factors that contribute to irreproducibility in conducting, reporting, and reviewing research and assesses the effectiveness and desirability of interventions aimed at improving reproducibility. It highlights the need for balanced scientific reforms that strengthen reproducibility without stifling innovation or introducing unintended consequences. A critical appraisal of the role of meta-research is essential to ensure sustainable improvements in research quality.

INTRODUCTION

Concerns about reproducibility in biomedical research have been long-standing and reached a wide audience about 15 years ago (1). Low rates of success at replicating studies led many to believe there was a crisis (2). Large-scale reproducibility projects have continued to find high rates of lack of reproducibility and difficulties even in assessing reproducibility, as a large share of biomedical studies do not share sufficiently detailed information to be repeated (3). In response, many journals, funders, patient organizations, policymakers, and grassroots communities have put reproducibility more prominently on their agendas (4). Multiple actions and measures are discussed or even endorsed and implemented to prevent or address reproducibility problems. However, the response has been uneven. Recent evidence suggests that while most scientists recognize the crisis narrative, less than half indicate it to be grave (5). There is also large variability in perspectives about what needs to be done, and often the evidence behind specific proposed or implemented solutions is limited.

This review takes stock of the state of research into the factors and issues associated with reproducibility and examines the evidence for interventions to increase reproducibility. It pursues the questions of whether improving reproducibility in biomedical research is possible, and if so, how this goal might be achieved. It also tries to address whether improving reproducibility through the various proposed measures is desirable. Desirability should consider both the likely benefits of these measures as well as the potential harms, costs, and other collateral damage.

REPRODUCIBILITY DEFINED

Reproducibility has been defined in different ways by different authors. This review uses the definition of Goodman et al. (6) because it is most frequently applied to biomedicine. It distinguishes three forms of reproducibility: methods reproducibility [i.e., a study is described in sufficient detail such that it can be fully repeated (e.g., one can retrieve, reprocess, and reanalyze the data with the relevant code or algorithms used, or one can establish new, similar experiments, procedures, and/or analytical pipelines)]; results reproducibility (i.e., one or more independent studies are performed using the same methods as the original study and reveal findings that corroborate the evidence from the original study), and inferential reproducibility (i.e., the inferences that different experts and other evaluators make of the results of one or multiple studies are consistent, or they disagree even though they look at the same results).

REPRODUCIBILITY IN RESEARCH: STATUS QUO OF PROBLEMS

Studies looking into factors and issues associated with reproducibility can be divided into those looking at design and conduct, reporting, or reviewing of research. Although often cross-sectional in nature, these studies aim to unravel the causes of irreproducibility.

If research design and conduct are suboptimal, results may be more likely to be wrong due to either chance or bias. Various factors are interrelated, yet they have been investigated with different methods, and as such, they provide different assessments of the scope of the problem. For example, especially when the associated data are not openly available, certain questionable research practices may be hard to detect. This is why studies investigating falsification, fabrication, and questionable research practices use surveys where authors are asked to anonymously indicate their or others' engagement with these practices. When openness is enhanced, more problems may be revealed. For instance, while publications of clinical trials rarely have signs of statistical malpractice based on their presented information, up to 40% of clinical trials have been classified as zombie trials when their raw data can be scrutinized (7). In addition, the relationship between various factors and reproducibility may not be linear but have a plateau or even a

U shape. For example, irreproducibility has been tied to both the design and conduct being too lax and too rigid. When too lax (e.g., no blinding, no sample size calculations, no random allocations), the internal validity decreases. However, once key design and conduct issues are addressed, further restrictions may compromise external validity and eventually translational potential and utility.

Many evaluations have investigated how research is reported. If reported incompletely or incorrectly, methods reproducibility may be low. Designing, conducting, and reporting research are closely linked. However, the fact that a particular practice is not reported (e.g., randomized sequence generation or allocation concealment) need not necessarily indicate that it was not implemented (8, 9). For selective reporting of primary outcomes, it is sensible to suspect that if a primary outcome that was investigated is left out, the results may not have been favorable. However, poor reporting may be due to other reasons, including sloppiness. The exact dynamics of not reporting, adding, or changing outcomes can be very complex and defy easy narratives (10, 11). As with research conduct, different factors affecting reporting that may lead to irreproducibility may reinforce one another. Without the availability of a detailed protocol and statistical analysis plan, the code used to run the analyses, and the full data, detecting selective reporting and determining the trustworthiness of the results are difficult.

Other evaluations have examined the way research is reviewed. If review systems do not function properly, biased or flawed evidence may contaminate the field, and results reproducibility could sink. Editors and peer reviewers may play a role in blocking negative (i.e., statistically non-significant) research results, fueling publication bias. However, these problems are probably highly context specific and may vary across journals, publishers, and scientific fields. Some dominant narratives may favor nonsignificant results [e.g., demonstrating no harm from a contested treatment or showing noninferiority (12)].

Each of the issues listed (**Table 1**) may have deeper causes, yet one must be careful to not oversimplify. It is possible that they converge into some fundamental roots, for example, the use of perverse incentives in scientific research that allow biases to thrive and create inappropriate, inefficient, or outright fraudulent microenvironments (i.e., single researchers and their teams) and macroenvironments (i.e., institutional level, country level, or even worldwide). One may blame the incentive systems that focus on productivity (13) and the ensuing publication pressure. However, there is nothing wrong with productivity, per se, and in fact, productivity is vital and may strengthen the timely publication of scientific results. Admonitions for publishing less may even exacerbate some problems, such as selective reporting.

EVIDENCE TO INCREASE REPRODUCIBILITY: CURRENT STATE OF SOLUTIONS

Accumulating evidence about the causes of irreproducible research has fueled a line of research that proposes and evaluates interventions to strengthen reproducibility. Meta-research (also known as metascience) aims to assess and improve the way we conduct, report, and review science (54).

Most interventions have been piloted or studied in specific settings under less-than-optimal conditions (55, 56). In addition, effects are often measured using proxies instead of hard outcomes [e.g., shared data may not necessarily guarantee reproducible findings (57)]. This makes it difficult to determine true effectiveness. Some proposed reproducibility interventions may be ineffective, and others may be minimally effective or may be helpful only in specific fields or settings. Connected to this, some types of research may have a ceiling of how reproducible they can become, and some types of research may best be abandoned (as they may have exhausted their

Table 1 Overview of issues associated with irreproducibility

Issue	Description	Scope or frequency
Design and conduct		
Falsification or fabrication of data	To make up data (fabrication) or to willfully alter data (falsification)	Traditional estimates vary between 1% and 2% of scientists admitting to fraud. However, a recent preprint suggests that one in seven papers is fraudulent (14, 15).
QRPs	More subtle trespasses of ethical conduct, including failing to report critical flaws in study design and analysis or selectively citing to enhance one's findings	Recent estimates indicate that ~30% of scientists admit to one QRP in the last 3 years. Higher estimates emerge when asked about QRPs by other scientists (14, 15).
Researcher degrees of freedom	Choices in study design and analysis, such as how to measure and analyze a variable (including which factors to correct for, how to handle missing data, which software to use, and what definitions to use)	Most biomedical fields enjoy high degrees of freedom, and more complex research questions with more degrees of freedom are becoming more common (16).
P-hacking, significance chasing, fishing, and HARKing	A group of problematic statistical practices, including testing many hypotheses or running multiple statistical tests until something appears significant by chance, exploring data without a clear hypothesis, hoping to find any significant association, and HARKing (17); can be seen as QRPs; related to researcher degrees of freedom	These practices are probably almost ubiquitous in one form or another; considerations are similar to those for QRPs and researcher degrees of freedom (17–19).
Low internal validity	Research subjects not randomized, investigators not blinded, studies underpowered or sample size calculations lacking, use of outcome measures inconsistent, selection of variables to include inconsistent, validity of measures poor or lacking	Very extensive literature suggests that most published studies have one or multiple major problems; this is highly prevalent in basic, applied, and clinical research (20–22).
Lack of data management infrastructure and quality control	Lab notes not being made consistently, leading to critical details being lost or inadequately documented; no consistent measures in place for the handling of samples or storage of materials	This is not as rigorously studied but likely a common occurrence (23).
Low external validity	A diverse set of problems that limit generalizability to different or broader settings, for example, investigating only one sex or age group, heavily sterile environmental conditions, not considering immunological or microbiome factors, or investigating only one specific breed of mice or type of bacteria; in clinical trials, limiting the inclusion of patients to those that have no comorbidities and are not taking multiple medications	This applies to the vast majority of basic, applied, and clinical research (24–28).
Error	Various forms of mistakes (human and technical), including statistical errors, methodological inconsistencies, data recording flaws, and specific problems with mischaracterized cell lines, equipment artifacts, reagent inconsistencies, or other material	Some form of technical error is expected to occur in many publications, but it is unclear how many of these lead to irreproducibility. Specific technical flaws and statistical inconsistencies can now be massively detected (29, 30).
Reporting		
Selective reporting	Reporting only certain results from a study, usually the statistically significant ones, leading to a more favorable narrative, or presenting the results of secondary outcomes or nonprespecified analyses as if they represent the primary outcome	This is mostly studied in clinical trials by comparing registered protocols versus publications (31–33). Around 50% of trials had some form of selective reporting.
File drawer problem	The phenomenon that, upon having conducted their study and finding the results unfavorable, researchers disregard submitting the study in question for publication, thus keeping it in their file drawer	For trials of antidepressants where one could compare against US Food and Drug Administration licensing packages, 89% of the nonsignificant studies of drugs approved between 1987 and 2002 were not published, decreasing to 53% for drugs approved between 2008 and 2013 (34).

(Continued)

Downloaded from www.annualreviews.org. Guest (guest) IP: 37.132.37.129 On: Sat, 07 Feb 2026 12:49:58

Table 1 (Continued)

Issue	Description	Scope or frequency
Spin	Misrepresentation of results, usually casting the findings of a study in a more favorable light than justified, by downplaying limitations, making unwarranted generalizations, or framing nonsignificant results as clinically relevant	This is mostly investigated in clinical research, and estimates differ per field. Usually, 30–50% of abstracts contain some form of spin (35, 36).
Lack of published protocols	Experimental protocols being unavailable, leading to methodological inconsistencies and biases	Across biomedicine, only approximately 3% of papers in 2000–2020 had registered protocols linked to them, and the rate is increasing over time (37).
Shortcut citations	Suboptimal citations in methods sections, because instead of a detailed description or protocol, a citation to a procedure is given but lacks the required detail, is inaccessible, or is ultimately nonexistent (38)	In a sample of papers from neuroscience, biology, and psychiatry, 90% used shortcut citations (38).
Code not available or not readable	Unavailability of the code used to run the analyses, hampering the rerunning of the analyses and increasing the risk of errors; available but not readily usable code, potentially leading other researchers to waste valuable time and still fail to reproduce results (39)	Across biomedicine, only 1.2% of papers between 2000 and 2020 shared code, gradually increasing over time during this period (up to 3.1% in 2020) (37).
Data inaccessible or not FAIR (40)	Unavailable data, rendering other researchers unable to (independently) replicate findings, conduct meta-analyses with individual data, or perform secondary analyses (41); available data lacking relevant meta-data or otherwise being noninteroperable, rendering it difficult if not impossible to reuse (42)	There has been an increase in data sharing in recent years, reaching ~20% by 2020 (37, 43).
Reviewing		
Publication bias	Journal editors and reviewers favoring significant results over those with nonsignificant results, leading to overestimation and unwarranted confidence (44, 45)	Much of the problem may be authors or sponsors not submitting nonsignificant results (34) or using selective reporting of outcomes and analyses. Editors and reviewers may also contribute bias, but there is also evidence to the contrary regarding editors (46).
Reviewer bias	When reviewers reject manuscripts because of personal and gender biases, conflicts of interests, or personal (as opposed to professional) interests, advocacy, or ideology	Quantifying the prevalence is difficult, but reviewer bias has been documented across fields (47–50).
Paper mills	Organizations that sell fake or low-quality manuscripts, thereby contaminating the literature and distorting systematic reviews and meta-analyses (51)	Current estimates in biomedicine are around 3%, but this number is probably growing (52, 53).

Abbreviations: FAIR, findable, accessible, interoperable, and reusable; HARKing, hypothesizing after results are known; QRP, questionable research practice.

potential or be obsolete as methods and techniques evolve), rather than trying to make them more reproducible.

Reproducibility-related interventions would benefit from rigorous studies that test them explicitly and estimate risk–benefit ratios. To date, most proposed measures have no randomized trials directly testing them. Some interventions have evidence from pre–post comparisons and comparisons of journals adopting versus not adopting them, but obviously such comparisons can be confounded. Examples include studies that have shown that, among reporting guidelines, Consolidated Standards of Reporting Trials (CONSORT) adoption improved randomized controlled trial quality (58), Animal Research: Reporting of In Vivo Experiments (ARRIVE) reduced variability in animal studies (59), and Minimum Information About a Microarray Experiment (MIAME) enhanced reproducibility in microarray data (60). Among interventions that have been tested in randomized trials, including several different modes of peer review [i.e., trained

reviewers, use of statisticians, open peer review, or blinded peer review (61–63)], results may be negative or nonconclusive; even when they are favorable for the intervention (e.g., involvement of statisticians in peer review seems to improve submitted papers), it may remain unclear how to scale up the intervention across the entire relevant research enterprise.

We list the anticipated gains and contrast those with the potential harms (Table 2). Implementing measures too hastily can have detrimental consequences. In Greek mythology, King Midas is initially euphoric that everything he touches turns into gold, but when he learns that his food and loved ones will turn into gold, too, he begs the gods to undo the gift.

SCIENCE REFORM AND DESIRABLE REPRODUCIBILITY

The recognized reproducibility issues and the diverse proposed solutions create a rapidly evolving but also uncertain landscape of research reform. Some reform efforts have gained substantial momentum. Currently, more than 23,000 institutions worldwide, including funders and journals, have signed the San Francisco Declaration on Research Assessment (DORA) (93). DORA takes a strong stance against journal impact factors, considering them largely obsolete journal-level metrics that are widely misused and inappropriately gamed (94, 95). Connected to this is the Coalition for Advancing Research Assessment (CoARA), a collaborative effort that seeks to transform the assessment of research by emphasizing qualitative vetting of research, combined with responsible use of more diverse quantitative metrics (96).

Even for widely supported efforts, the evidence on how implementing their visions strengthens reproducibility remains scarce. Discussions surrounding reproducibility have led to policy changes at different research institutions that now use different criteria for hiring and promoting scientists, focusing more on the reproducibility of their work (97). Most institutions, however, have been slow at adopting such changes (98). Qualitative assessment is easier said than done, especially when it comes to doing it at scale (i.e., for seven million scientific papers published every year plus millions of research proposals). When and where qualitative assessment should have priority is far from clear; some have even argued that invoking qualitative assessments may be some way of handwaving at irreproducibility problems (99, 100). Relatedly, qualitative assessment could become a vehicle to justify biased decisions. A key challenge lies in finding out which reproducibility measures are worth assessing research with and how approaches can be scaled sustainably.

Some critics may question whether investing effort into strengthening reproducibility is eventually desirable and whether overemphasis on reproducibility stifles creativity and hampers scientific progress. Furthermore, certain open science practices can be gamed and may replace current suboptimal deliverables (e.g., publications and citations) with even less useful ones (e.g., bureaucratic checklists, obsessive postpublication comments, and social media impact factors). Some concepts in open science have already been hijacked for the wrong purposes, with sky-high article processing costs (the fees paid to make a publication open access) increasing profits for publishers.

Meta-research influences how efforts to promote reproducibility are perceived. The efforts of whistleblowers and integrity sleuths remain invaluable, but if hunting down scientists becomes an end in itself, we are in dangerous waters. Impassionate investigation of the extent to which certain practices are present or absent has merit. However, the field should not be perceived as a self-proclaimed police force with trial by media. Meta-researchers can be wrong, too, and evidence on reproducible research practices should be continuously appraised, as research is a living enterprise. We must be careful not to fool ourselves.

Table 2 Overview of interventions proposed to strengthen reproducibility

Intervention	Description	Adopted in areas	Anticipated gains	Potential harms, costs, and collateral damage
Design and conduct				
Preregistration	A timestamped version of the study's design and analysis plan is put in an open repository.	This is widely adopted in clinical trials (e.g., clinicaltrials.gov), with increasing adoption in observational studies (e.g., osf.io) and preclinical studies (e.g., animalstudyregistry.com). Overall, preregistration has been done for hundreds of thousands of studies to date.	Greater transparency, preventing p-hacking and data dredging, reducing publication reporting bias	Stifling research, increasing bureaucracy
Registered reports	A study with its full analysis plan is peer-reviewed and in principle accepted for publication prior to data collection; the journal will publish the study regardless of the results.	The greatest adoption is in neuroscience and human behavior studies, with hardly any adoption in clinical research (64). Overall, these have been done for hundreds of studies to date (65, 66).	Reducing publication bias and potential questionable research practices	Delayed research due to time to review, unsuitable for [some forms of (67)] exploratory research
Systematic reviews and other evidence synthesis	A structured and comprehensive synthesis of existing evidence is performed, following a predefined protocol.	These are widely used in clinical medicine, pharmacology, and epidemiology. Cochrane Reviews are considered gold-standard systematic reviews. Adoption is emerging in in vivo and in vitro research (68). Overall, there are more than 100,000 systematic reviews to date, covering a large portion of the primary literature (69).	More reliably informed next steps in research given a solid overview of what is known, showcasing biases, clarifying ambiguity	Poorly done and/or conflicted systematic reviews, perpetuating or even amplifying errors, biases, and little-value areas of research (70)
Electronic laboratory notebooks	These digital tools enable the recording, storage, and managing of laboratory data. They provide time stamps, version control, and digital sharing (71).	There are no good data on extent of adoption but these tools are probably adopted in only a minority of basic science laboratories (72).	Preventing loss of data, ensuring structured recording	Cost (although open-source versions exist), training requirements
Critical incident reporting	Error reporting mechanisms may lead to blame-free reporting (73).	This is typically integrated within ISO standards in clinical medicine. There are no good data on how many laboratories and institutions may have adopted this option, which can be integrated as part of quality control systems for in vivo and in vitro research (74).	Strengthening the self-learning and self-correcting capacity of the organization, increasing safety and efficacy	Reputation harm, time required to get a system off the ground

(Continued)

Table 2 (Continued)

Intervention	Description	Adopted in areas	Anticipated gains	Potential harms, costs, and collateral damage
Multiteam and multicenter studies	Independent teams conduct the same experiment or study.	Clinical trials often use multicenter designs (75), and the concept has been piloted in preclinical research (76). Multiteam consortia are popular in epidemiology and some other fields.	Greater power, more robust results against small context- or setting-specific heterogeneity	Teams dedicated to performing the same experiment or study, as others may lack time to do their own (novel) work, thereby decreasing the creativity of a research line; unclear cost–benefit for questions that can be reliably settled with smaller studies
Reproducibility training	These educational programs, workshops, and freely available resources are intended to teach researchers about best practices to promote reproducibility, often focusing on (experimental) design, data management, open science, and transparent reporting (77).	There are many ongoing programs, both discipline agnostic and discipline specific [e.g., the UK Reproducibility Network (78), the National Institutes of Health Rigor and Reproducibility Initiative (79), and Reproducibility for Everyone, which organizes workshops at major conferences (80)].	Proper and tailored implementation, perhaps more efficient and reliable science	Resources required to hire training staff, time required to train others, lack of clarity regarding whether improvements are specifically due to the training or other concomitant experiences of the trainees, lack of clarity regarding which (if any) parts of the training make a difference
Reporting				
Reporting guidelines	These are lists of standardized items to include in a publication that are relevant for a particular (sub)field or study design. The EQUATOR platform maintains an up-to-date list of such guidelines (equator-network.org).	Many but not all medical journals publishing trials require adherence to CONSORT (81). In systematic reviews, PRISMA is widely but not ubiquitously used (82). For preclinical animal research, journals increasingly mandate ARRIVE compliance, but adopting studies remain a minority (83).	Greater transparency, crucial items included	Suboptimal if not monitored, with current adoption lacking oversight and thereby not reaching full potential; possibly erroneous or even fraudulent reporting
Open and FAIR data	Freely available data, stored in a way accessible to others, is interoperable (computer readable and system agnostic) and reusable.	There is modest adoption across diverse fields (37, 84), and it has recently been encouraged by some journals [with mixed results (85)] and key funders in both the United States and Europe, although adherence is suboptimal (86).	Ability to check research more readily or build on others' work (e.g., patient data meta-analysis)	High cloud storage capacity, cost, and resources required if all data (including low-value sets) are to be FAIR
Open access	This refers to free and unrestricted access to scientific manuscripts via journals, repositories, or preprint servers.	Open access is widely adopted in public health, genomics, epidemiology, and clinical research, and there is expanded adoption among mega journals and in some government initiatives, for example, in Europe (87–89).	Free and faster access, the ability to build on others' work without delay	Costly article publication fees, which can enrich publishers at the expense of researchers and research funders; the cost of building and maintaining institutional repositories

(Continued)

Table 2 (Continued)

Intervention	Description	Adopted in areas	Anticipated gains	Potential harms, costs, and collateral damage
Open code	This refers to the sharing of scripts, computer workflows, and algorithms alongside publications or project reports. Good annotation practices involve adding README files with instructions on how to run the code, clear variable labeling, and inline instructions.	Open code is used in tens of thousands of studies in bioinformatics, epidemiology, systems biology, and medical imaging (e.g., Bioconductor, GitHub, Code Ocean) (39).	Better ability to reuse or extend published models and techniques, easier to verify data and results	Resource development, implementation, and maintenance costs, which can vary
Open methods and protocols	This involves the full disclosing of methods and techniques through the publication or deposition of a protocol and related material on a freely accessible platform.	These are modestly adopted in preclinical research, drug development, and neuroscience (e.g., Protocols.io, <i>Nature's</i> Protocol Exchange). Journals like <i>PLOS ONE</i> and <i>BMJ Open</i> have committed to openly publishing medical protocols. Open methods and protocols are also encouraged by European funders (90).	Swifter replication, greater insight into detailed procedures	Time required from the research team to make protocols understandable to others outside the laboratory
Reviewing				
Training	Training editors and reviewers by providing structured education focused on scientific rigor and statistical analysis.	Some training programs have emerged in different settings (62), but the vast majority of biomedical scientists get no formal training in reviewing.	Potential improvement in peer review, better-quality manuscripts	Cost, time, effort, and resources required for training with uncertain benefits
Statistical reviewing	This means adding a statistician to the peer review team who focuses specifically on the quality and integrity of the statistical analyses.	Top medical journals (e.g., <i>The Lancet</i> , <i>JAMA</i> , <i>The New England Journal of Medicine</i>) regularly involve statisticians in peer review. Few other journals can do this.	Removal of statistical errors, use of better analysis techniques	Qualified statisticians and methodologists in short supply, so it may be a waste to use them to review millions of poor papers
Open peer review	This involves making the identity of the authors, reviewers, and editors open to all and publishing the review reports alongside the final publication.	Open peer review has been adopted by journals such as <i>BMJ</i> , <i>eLife</i> , and <i>PLOS</i> to enhance transparency.	Greater incentive for reviewers to review rigorously	Retaliation, reviews becoming more favorable and avoiding criticism
Postpublication peer review	Scrutinizing published manuscripts through public commenting, along with reanalysis on open platforms, such as PubPeer, may aid in correcting flawed articles (91).	Retraction Watch, PubPeer, and F1000Research facilitate postpublication review. Postpublication review options remain limited, even in top journals (92).	Correction and cleansing of the scientific literature, although often rather slow	Unwarranted reputational damage; low-quality postpublication comments; difficulty in distinguishing ethical, qualified sleuths from rogue commentators

Abbreviation: FAIR, findable, accessible, interoperable, and reusable.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The work of J.P.A.I. is supported by an unrestricted gift from Sue and Bob O'Donnell to Stanford.

LITERATURE CITED

1. Begley CG, Ellis LM. 2012. Raise standards for preclinical cancer research. *Nature* 483(7391):531–33
2. Baker M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604):452–54
3. Errington TM, Mathur M, Soderberg CK, et al. 2021. Investigating the replicability of preclinical cancer biology. *eLife* 10:e71601
4. Munafò MR, Nosek BA, Bishop DVM, et al. 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* 1(1):0021
5. Cobey KD, Ebrahimzadeh S, Page MJ, et al. 2024. Biomedical researchers' perspectives on the reproducibility of research. *PLOS Biol.* 22(11):e3002870
6. Goodman SN, Fanelli D, Ioannidis JPA. 2016. What does research reproducibility mean? *Sci. Transl. Med.* 8(341):341ps12
7. Ioannidis JPA. 2021. Hundreds of thousands of zombie randomised trials circulate among us. *Anaesthesia* 76(4):444–47
8. Hill CL, LaValley MP, Felson DT. 2002. Discrepancy between published report and actual conduct of randomized clinical trials. *J. Clin. Epidemiol.* 55(8):783–86
9. Pildal J, Chan A-W, Hróbjartsson A, et al. 2005. Comparison of descriptions of allocation concealment in trial protocols and the published reports: cohort study. *BMJ* 330(7499):1049
10. Goldacre B, Drysdale H, Dale A, et al. 2019. COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials* 20(1):118
11. Goldacre B, Drysdale H, Marston C, et al. 2019. COMPare: qualitative analysis of researchers' responses to critical correspondence on a cohort of 58 misreported trials. *Trials* 20(1):124
12. Ioannidis JPA. 2024. Inverse publication reporting bias favouring null, negative results. *BMJ Evid. Based Med.* 29(1):6–9
13. Trueblood JS, Allison DB, Field SM, et al. 2025. The misalignment of incentives in academic publishing and implications for journal reform. *PNAS* 122(5):e2401231121
14. Xie Y, Wang K, Kong Y. 2021. Prevalence of research misconduct and questionable research practices: a systematic review and meta-analysis. *Sci. Eng. Ethics* 27(4):41
15. Heathers J. 2025. Approximately 1 in 7 scientific papers are fake. Preprint, MetaROR. <https://metaror.org/article/approximately-1-in-7-scientific-papers-are-fake/>
16. Ioannidis JPA, Greenland S, Hlatky MA, et al. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 383(9912):166–75
17. Kerr NL. 1998. HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* 2(3):196–217
18. Head ML, Holman L, Lanfear R, et al. 2015. The extent and consequences of p-hacking in science. *PLOS Biol.* 13(3):e1002106
19. Ware JJ, Munafò MR. 2015. Significance chasing in research practice: causes, consequences and possible solutions. *Addiction* 110(1):4–8
20. Wang Y, Parpia S, Couban R, et al. 2024. Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors. *J. Clin. Epidemiol.* 165:111211
21. Gluud LL. 2006. Bias in clinical intervention research. *Am. J. Epidemiol.* 163(6):493–501
22. Vinkers CH, Lamberink HJ, Tijdink JK, et al. 2021. The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. *PLOS Biol.* 19(4):e3001162

23. Soundararajan S, Mishra S. 2023. Data management: the first step in reproducible research. *Indian J. Occup. Environ. Med.* 27(4):359–63
24. Pound P, Ritskes-Hoitinga M. 2018. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *J. Transl. Med.* 16(1):304
25. Dirnagl U, Bannach-Brown A, McCann S. 2022. External validity in translational biomedicine: understanding the conditions enabling the cause to have an effect. *EMBO Mol. Med.* 14(2):e14334
26. Clayton JA, Collins FS. 2014. Policy: NIH to balance sex in cell and animal studies. *Nature* 509(7500):282–83
27. Van Norman GA. 2019. Limitations of animal studies for predicting toxicity in clinical trials: Is it time to rethink our current approach? *JACC Basic Transl. Sci.* 4(7):845–54
28. Rothwell PM. 2006. Factors that can affect the external validity of randomised controlled trials. *PLOS Clin. Trials* 1(1):e9
29. Hunter P. 2021. Technical bias and the reproducibility crisis. *EMBO Rep.* 22(2):e52327
30. Nuijten MB, Polanin JR. 2020. “*statscheck*”: automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Res. Synth. Methods* 11(5):574–79
31. Chan A-W, Hróbjartsson A, Haahr MT, et al. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 291(20):2457–65
32. Anderson ML, Chiswell K, Peterson ED, et al. 2015. Compliance with results reporting at ClinicalTrials.gov. *N. Engl. J. Med.* 372(11):1031–39
33. DeVito NJ, Morley J, Smith JA, et al. 2024. Availability of results of clinical trials registered on EU Clinical Trials Register: cross sectional audit study. *BMJ Med.* 3(1):e000738
34. Turner EH, Cipriani A, Furukawa TA, et al. 2022. Selective publication of antidepressant trials and its influence on apparent efficacy: updated comparisons and meta-analyses of newer versus older trials. *PLOS Med.* 19(1):e1003886
35. Boutron I, Dutton S, Ravaud P, Altman DG. 2010. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 303(20):2058–64
36. Khan MS, Lateef N, Siddiqi TJ, et al. 2019. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: a systematic review. *JAMA Netw. Open* 2(5):e192622
37. Serghiou S, Contopoulos-Ioannidis DG, Boyack KW, et al. 2021. Assessment of transparency indicators across the biomedical literature: How open is open? *PLOS Biol.* 19(3):e3001107
38. Standvoss K, Kazezian V, Lewke BR, et al. 2024. Shortcut citations in the methods section: frequency, problems, and strategies for responsible reuse. *PLOS Biol.* 22(4):e3002562
39. Hamilton DG, Hong K, Fraser H, et al. 2023. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. *BMJ* 382:e075767
40. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3(1):160018
41. Hardwicke TE, Ioannidis JPA. 2018. Populating the Data Ark: an attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE* 13(8):e0201856
42. Tedersoo L, Küngas R, Oras E, et al. 2021. Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* 8(1):192
43. Chan A-W, Song F, Vickers A, et al. 2014. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 383(9913):257–66
44. Dalton JE, Bolen SD, Mascha EJ. 2016. Publication bias: the elephant in the review. *Anesth. Analg.* 123(4):812–13
45. Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR. 1991. Publication bias in clinical research. *Lancet* 337(8746):867–72
46. Olson CM, Rennie D, Cook D, et al. 2002. Publication bias in editorial decision making. *JAMA* 287(21):2825–28
47. Aczel B, Barwich A-S, Diekman AB, et al. 2025. The present and future of peer review: ideas, interventions, and evidence. *PNAS* 122(5):e2401232121

48. Mahoney MJ. 1977. Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cogn. Ther. Res.* 1(2):161–75
49. Tomkins A, Zhang M, Heavlin WD. 2017. Reviewer bias in single- versus double-blind peer review. *PNAS* 114(48):12708–13
50. Helmer M, Schottdorf M, Neef A, Battaglia D. 2017. Gender bias in scholarly peer review. *eLife* 6:e21718
51. Candal-Pedreira C, Ross JS, Ruano-Ravina A, et al. 2022. Retracted papers originating from paper mills: cross sectional study. *BMJ* 379:e071517
52. Abalkina A, Aquarius R, Bik E, et al. 2025. ‘Stamp out paper mills’ — science sleuths on how to fight fake research. *Nature* 637(8048):1047–50
53. Van Noorden R. 2023. How big is science’s fake-paper problem? *Nature* 623:466–67
54. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. 2015. Meta-research: evaluation and improvement of research methods and practices. *PLOS Biol.* 13(10):e1002264
55. Dudda L, Kormann E, Kozula M, et al. 2025. Open science interventions to improve reproducibility and replicability of research: a scoping review. *R. Soc. Open Sci.* 12:242057
56. Hair K, Smith S, Buljan I, et al. 2023. *A protocol for a systematic online living evidence summary of the interventions to improve reproducibility (iRISE-SOLES), Open Science Framework*, iRISE Consortium, accessed Feb. 18, 2025. <https://camarades.shinyapps.io/irise-soles/>
57. Weissgerber TL, Gazda MA, Nilsonne G, et al. 2024. Understanding the provenance and quality of methods is essential for responsible reuse of FAIR data. *Nat. Med.* 30(5):1220–21
58. Kane RL, Wang J, Garrard J. 2007. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. *J. Clin. Epidemiol.* 60(3):241–49
59. Percie du Sert N, Hurst V, Ahluwalia A, et al. 2020. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *J. Cereb. Blood Flow Metab.* 40(9):1769–77
60. Brazma A, Hingamp P, Quackenbush J, et al. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* 29(4):365–71
61. Jefferson T, Rudin M, Brodney Folse S, Davidoff F. 2007. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst. Rev.* 2007:MR000016
62. Bruce R, Chauvin A, Trinquart L, et al. 2016. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. *BMC Med.* 14(1):85
63. Ross-Hellauer T, Horbach SPJM. 2024. Additional experiments required: a scoping review of recent evidence on key aspects of Open Peer Review. *Res. Eval.* 33:rvae004
64. Anthony N, Tisseaux A, Naudet F. 2023. Published registered reports are rare, limited to one journal group, and inadequate for randomized controlled trials in the clinical field. *J. Clin. Epidemiol.* 160:61–70
65. Chambers CD, Tzavella L. 2021. The past, present and future of Registered Reports. *Nat. Hum. Behav.* 6(1):29–42
66. Soderberg CK, Errington TM, Schiavone SR, et al. 2021. Initial evidence of research quality of registered reports compared with the standard publishing model. *Nat. Hum. Behav.* 5(8):990–97
67. Dirnagl U. 2020. Preregistration of exploratory research: learning from the golden age of discovery. *PLOS Biol.* 18(3):e3000690
68. Ioannidis JPA. 2023. Systematic reviews for basic scientists: a different beast. *Physiol. Rev.* 103(1):1–5
69. Niforatos JD, Weaver M, Johansen ME. 1995. Assessment of publication trends of systematic reviews and randomized clinical trials, 1995 to 2017. *JAMA Intern. Med.* 179(11):1593–94
70. Ioannidis JPA. 2016. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q.* 94(3):485–514
71. Vandendorpe J, Adam B, Wilbrandt J, et al. 2024. Ten simple rules for implementing electronic lab notebooks (ELNs). *PLOS Comput. Biol.* 20(6):e1012170
72. Higgins SG, Nogiwa-Valdez AA, Stevens MM. 2022. Considerations for implementing electronic laboratory notebooks in an academic research environment. *Nat. Protoc.* 17(2):179–89
73. Dirnagl U, Przesdzing I, Kurreck C, Major S. 2016. A laboratory critical incident and error reporting system for experimental biomedicine. *PLOS Biol.* 14(12):e2000705
74. Bernalov A, Bernard R, Gilis A, et al. 2021. Introduction to the EQUIPD quality system. *eLife* 10:e63294
75. Senn SJ, Lewis RJ. 2019. Treatment effects in multicenter randomized clinical trials. *JAMA* 321(12):1211–12

76. Hunniford VT, Grudniewicz A, Fergusson DA, et al. 2023. A systematic assessment of preclinical multilaboratory studies and a comparison to single laboratory studies. *eLife* 12:e76300
77. Toelch U, Ostwald D. 2018. Digital open science—teaching digital tools for reproducible and transparent research. *PLOS Biol.* 16(7):e2006022
78. UK Reproducibility Network Steering Committee. 2021. From grassroots to global: a blueprint for building a reproducibility network. *PLOS Biol.* 19(11):e3001461
79. Collins FS, Tabak LA. 2014. Policy: NIH plans to enhance reproducibility. *Nature* 505(7485):612–13
80. Auer S, Haeltermann NA, Weissgerber TL, et al. 2021. Science Forum: a community-led initiative for training in reproducible research. *eLife* 10:e64719
81. Altman DG. 2005. Endorsement of the CONSORT statement by high impact medical journals: survey of instructions for authors. *BMJ* 330(7499):1056–57
82. Page MJ, Moher D. 2017. Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions: a scoping review. *Syst. Rev.* 6(1):263
83. Hair K, Macleod MR, Sena ES, et al. 2019. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARUS). *Res. Integr. Peer Rev.* 4(1):12
84. Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. 2017. Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ* 5:e3208
85. Rowhani-Farid A, Aldcroft A, Barnett AG. 2020. Did awarding badges increase data sharing in *BMJ Open*? A randomized controlled trial. *R. Soc. Open Sci.* 7(3):191818
86. Couture JL, Blake RE, McDonald G, Ward CL. 2018. A funder-imposed data publication requirement seldom inspired data sharing. *PLOS ONE* 13(7):e0199789
87. Robinson-Garcia N, Costas R, van Leeuwen TN. 2020. Open Access uptake by universities worldwide. *PeerJ* 8:e9410
88. Piwowar H, Priem J, Larivière V, et al. 2018. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* 6:e4375
89. Else H. 2021. A guide to Plan S: the open-access initiative shaking up science publishing. *Nature*, Apr. 8. <https://www.nature.com/articles/d41586-021-00883-6>
90. Batista Leite S, Brooke M, Carusi A, et al. 2024. Promoting reusable and open methods and protocols (PRO-MaP) can improve methodological reporting in the life sciences. *PLOS Biol.* 22(9):e3002835
91. Bordignon F. 2020. Self-correction of science: a comparative study of negative citations and post-publication peer review. *Scientometrics* 124(2):1225–39
92. Hardwicke TE, Thibault RT, Kosie JE, et al. 2022. Post-publication critique at top-ranked journals across scientific disciplines: a cross-sectional assessment of policies and practice. *R. Soc. Open Sci.* 9(8):220139
93. DORA. 2025. San Francisco Declaration on Research Assessment. DORA. <https://sfdora.org/read/>
94. Seglen PO. 1997. Why the impact factor of journals should not be used for evaluating research. *BMJ* 314(7079):498–502
95. *PLOS Medicine* Editors. 2006. The impact factor game. *PLOS Med.* 3(6):e291
96. CoARA. 2025. The agreement. CoARA. <https://coara.org/agreement/the-agreement-full-text/>
97. Moher D, Bouter L, Kleinert S, et al. 2020. The Hong Kong Principles for assessing researchers: fostering research integrity. *PLOS Biol.* 18(7):e3000737
98. Rice DB, Raffoul H, Ioannidis JPA, Moher D. 2020. Academic criteria for promotion and tenure in biomedical sciences faculties: cross sectional analysis of international sample of universities. *BMJ* 369:m2081
99. Ioannidis JPA, Maniadis Z. 2023. In defense of quantitative metrics in researcher assessments. *PLOS Biol.* 21(12):e3002408
100. Ioannidis JPA, Maniadis Z. 2024. Quantitative research assessment: using metrics against gamed metrics. *Intern. Emerg. Med.* 19(1):39–47