

Viewpoint

Received: 13 Jul 2025
Accepted: 6 Aug 2025
Published: 18 Sep 2025

Declaration of Interests

The authors have no conflict of interest to declare.

Acknowledgements

Matt Spick was supported by UK Research and Innovation (UKRI1095). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding

The authors declared that this study has received no financial support.

Meeting the challenges posed by mass-produced manuscripts and click-data science

Reese Richardson¹, Matt Spick²✉

¹Center for Science of Science and Innovation, Kellogg School of Management, Northwestern University, Evanston, USA
orcid.org/0000-0002-6058-5886

²School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, United Kingdom
matt.spick@surrey.ac.uk
orcid.org/0000-0002-9417-6511



This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

Citation

Richardson R, Spick M. Meeting the challenges posed by mass-produced manuscripts and click-data science. *Eur Sci Ed.* 2025;51:e165043.

<https://doi.org/10.3897/ese.2025.e165043>

Abstract

The combination of open-access datasets, machine learning workflows, increased computing capacity, and generative artificial intelligence has effectively removed many of the rate-limiting steps in manuscript production. This has created an industry of click-data science and a flood of low-quality manuscripts based on large health datasets such as the US National Health and Nutrition Examination Survey, the UK Biobank, and the US FDA Adverse Event Reporting System. These papers often employ statistically appropriate methods and real data, but introduce misleading results and false discoveries to the literature. Here, we offer suggestions for editors on how to identify such manuscripts and reject them at the point of submission, reducing the burden on the publishing process.

Keywords:

big data, false discoveries, generative AI, integrity, paper mills

Fraudulent and low quality research is exploiting Open Science data sources

Until recently, 'big data' driven research has been hampered by a number of rate-limiting steps, including the acquisition of high-quality data, the application of complex methods, and the production of publication-ready manuscripts. The combination of easily accessible large datasets (motivated by Open Science principles),¹ robust coding libraries in languages such as R and Python, and generative artificial intelligence (GenAI) has effectively removed these rate-limiting steps from the authoring process. This represents a huge productivity gain for researchers in the form of click-data science, that is, low-effort manuscript production powered by open data and automation. The benefits, however, have accrued not only to those that wish to proceed cautiously and ethically, but also to those whose agenda is publication at all costs. The latter category includes paper mills—commercial organisations that produce manuscripts for paying clients at scale using derivative, copied, and/or fabricated text or data sets.²⁻⁵

Examples of large (often national) datasets include the US National Health and Nutrition Examination Survey (NHANES) or the UK Biobank.^{6,7} These provide wide-ranging and complex information on genomes, epidemiology or medication and provide the data on an open access basis. This often includes Application Programming Interfaces that allow data to be extracted directly into a coding environment for rapid analysis. Combined with GenAI's ability to author introductory or discussion text, this allows for end-to-end automated workflows for manuscript production. The consequences are

clearly visible in increased submission volumes and publication rates. A simple search of the PubMed database for "NHANES [tiab]" shows a four-fold increase from 1081 accepted publications in 2021 to 4060 in 2024, with further growth probable in 2025 based on the year-to-date trend. There has also been a shift towards formulaic titles reporting an association between exposure X and outcome Y for cohort Z.⁸

Similar growth has been reported for papers that use Mendelian randomisation applied to open genome-wide association study datasets,⁹ such as FinnGen.¹⁰ Another area experiencing formulaic growth is pharmacovigilance, especially exploiting the US FDA Adverse Event Reporting System (FAERS).¹¹ Often these manuscripts are simplistic: at best, they offer little insight into complex multifactorial problems, and at worst, introduce false discoveries and misleading conclusions to the literature.^{12,13} Figure 1 illustrates the acceleration in aggregated publication rates across NHANES, FAERS, the Global Burden of Disease Study, FinnGen, and the UK Biobank, but any open dataset can be exploited in this way.

The challenges posed by click-data science

The relationship between journal editors and paper mills or other unethical authors is – at its core – adversarial; adaptation is a natural outcome of such a relationship. These new, heavily automated and GenAI-assisted workflows pose different challenges from those set by manuscripts that include, for example, fraudulent Western blot images or outright p-hacking. First, for this new category of unethical submission, the

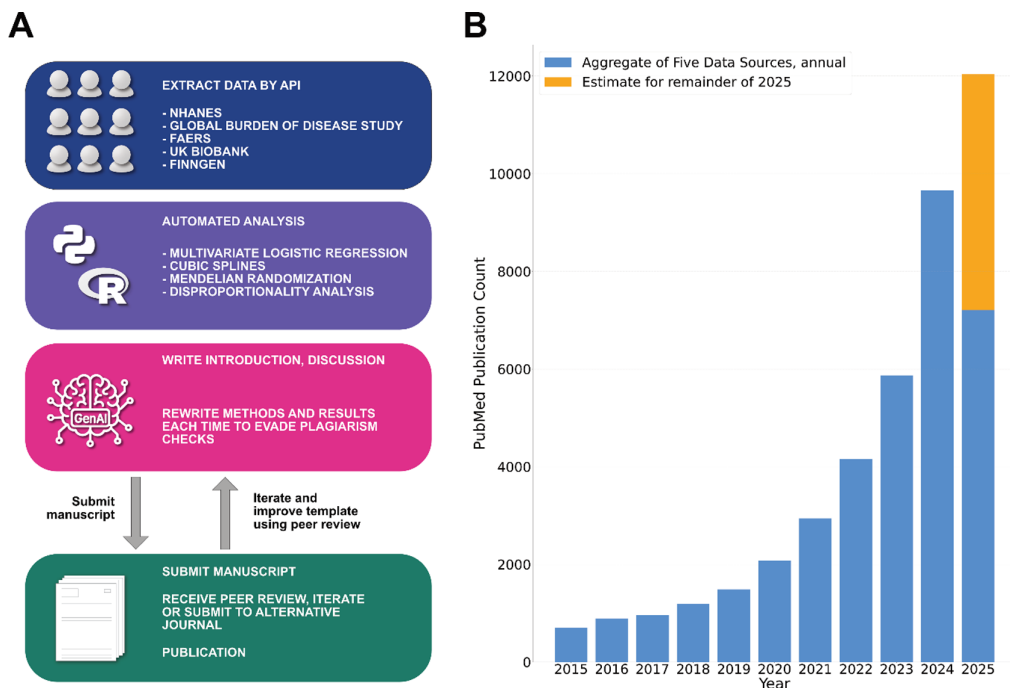


Figure 1. Artificial intelligence–assisted workflows enable rapid acceleration in publications. (A) Illustrative workflow schematic. (B) PubMed search using the following search string: “nhanes[tiab] OR uk biobank[tiab] OR finngen[tiab] OR global burden of disease study[tiab] OR faers[tiab].” The estimate for the publication count in the remainder of 2025 is derived by applying the same ratio of H2 2024 / H1 2024 publication count to the H1 2025 publication count and assumes a similar seasonal pattern of higher publication rates in the first 6 months of the year.

underlying data are not faked but are high quality, often with large numbers of participants and hence statistically well-powered cohorts. Secondly, the methods employed are sophisticated, including machine learning analyses or approaches such as Mendelian randomisation. Indeed, in the case of NHANES, we believe that mass-produced manuscripts have been in production for around 3 years and consequently have received 3 years’ worth of editor and peer reviewer feedback to iron out flaws in methodologies. Thirdly, these manuscripts often target areas of public interest such as obesity or cancer (these subjects also constitute those for which national data are most likely to be collected).

The combination of these adaptive changes by paper mills has led to manuscripts that, at least superficially, appear to address genuine issues using legitimate data and high-quality methods, with the flaws (such as selective data usage) hidden except to those willing to reproduce the studies from scratch. For editors and peer reviewers alike, this makes immediate rejection of such low-value papers more challenging. Regardless of how the publishing industry arrived at this point, the process of producing data-driven research is quickly becoming trivial. In such an environment, the question of *how* research has been conducted will decline in relative importance, and *why* the research

has been conducted will have to take on greater significance.

Practical responses to the flood of low-quality research

We have observed that each exploited dataset tends to come with its own template or formula and its own set of issues. Those concerned with genome-wide association study (GWAS) data use Mendelian randomisation and will frequently apply complex methods to derive implausible results. Examples of proposed causal relationships with no biological underpinning include semi-skimmed milk consumption protecting against depression or educational attainment increasing the risks of post-operative hernias.^{14,15} Stender et al. recommend the rejection of such papers without additional supporting evidence and have provided a template for desk rejections.⁹ Those paper mills using large-scale epidemiological data such as NHANES or the Global Burden of Disease study¹⁶ will interrogate a matrix of exposures versus outcomes and select all those that meet statistical significance (without taking account of multiple hypothesis correction that is essential when querying large datasets).^{17,18} These studies will often associate a general indicator, such as inflammation, with dozens of outcomes, reporting each as a separate manuscript. Duplicate submissions are also commonplace with slight adjustments to the cohort. For example, five essentially identical studies have been published associating infertility with the body roundness index,¹⁹ taking different subsets of the NHANES dataset each time (such as married women only, or altering the years in which the women were recruited). In our experience,

such submissions can often be immediately rejected for lack of novelty by the simple means of a literature check. More generally, as previously noted, we believe that there will need to be more emphasis in submissions on the *why* of research, rather than the *how*, given that with GenAI-assisted workflows, production of a data-driven manuscript has been reduced to a handful of clicks.

Many journals and editors are aware of these problems, and resources are available that highlight how to spot and handle potential integrity issues with manuscripts, including from the Committee on Publication Ethics²⁰ and from the Collection of Open Science Integrity Guides.²¹ Domain-specific guidelines also exist; for example the Reporting of a Disproportionality Analysis for Drug Safety Signal Detection Using Individual Case Safety Reports in Pharmacovigilance (READUS-PV) checklist, originally introduced by *Drug Safety*,²² partly intended to stem the flow of low-quality pharmacovigilance papers. Albeit, as yet this has not been effective in reducing the overall rate of FAERS publications, as submissions have simply been diverted to other journals. *The Journal of Global Health* recently published a policy paper specifically on the mass manufacture of manuscripts using click-data science workflows, including a set of Guidelines for Reporting Analyses of Big Data Repositories Open to the Public (GRABDROP).²³ This tool requires authors (a) to formally disclose their previous data repository-based publications, (b) to explain the elements of study design and how the use of available datasets makes the study an original scientific contribution, (c) to list and cite all publications addressing similar research questions from the same dataset, (d) to explain how multiple hypothesis testing

was addressed, and (e) declare whether and how GenAI chatbots had been used in the development of the manuscript.

In our view, the GRABDROP guidelines offer a helpful template for editors. Some of the points, such as the disclosure of GenAI use for writing, will already be covered by most journals' policies. Others, such as the requirement to cite previous studies using the dataset, should be standard practice anyway, but requiring a checklist has the advantage of increasing the burden for exploitative, high-volume authors at little expense to authors making genuine scientific contributions. This is an important aspect given that submission handling can already be an onerous process for editors and authors.²⁴ Inevitably, the adversarial nature of paper mills means that we expect the 'next generation' to follow GRABDROP or similar guidelines superficially to avoid detection. Where responses to checklists have obviously been falsified or misrepresented, however, there will be quick and easy grounds for desk rejection.

Preservation of data as a public good

While in the short run, click-data science can be addressed by editors and peer reviewers through increased vigilance, the target-setting culture in academia makes it likely that paper mills will continue to innovate and adapt to meet customer demand. We believe that in the longer run, these trends will also raise issues around Open Science itself. There are other models of data availability than simple unfettered access, for example through pre-registration of research or requiring applications for access that state researchers' specific areas of interest to prevent data dredging.

Such steps would still meet the goals of Open Science and are largely in compliance with the FAIR Guiding Principles, which require that data be Findable, Accessible, Interoperable, and Reusable.²⁵ In this way, the value of open data as a public good can still be preserved while fighting inappropriate exploitation and misinterpretation of that data.

References

1. Foster ED, Deardorff A. Open science framework (OSF). *J Med Libr Assoc.* 2017;105(2):203-206. [\[CrossRef\]](#)
2. Byrne JA, Abalkina A, Akinduro-Aje O, et al. A call for research to address the threat of paper mills. *PLoS Biol.* 2024;22(11):e3002931. [\[CrossRef\]](#)
3. Christopher J. The raw truth about paper mills. *FEBS Lett.* 2021;595(13):1751-1757. [\[CrossRef\]](#)
4. Byrne JA, Park Y, Richardson RAK, Pathmen-dra P, Sun M, Stoeger T. Protection of the human gene research literature from contract cheating organizations known as research paper mills. *Nucleic Acids Res.* 2022;50(21):12058-12070. [\[CrossRef\]](#)
5. Richardson RAK, Hong SS, Byrne JA, Stoeger T, Amaral LAN. The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. *Proc Natl Acad Sci USA.* 2025;122(32):e2420092122. [\[CrossRef\]](#)
6. *Health Research Data for the World.* UK Biobank. 2025. Published online July 15. Available at: <https://www.ukbiobank.ac.uk/>. Accessed Aug 1, 2025.
7. Centers for Disease Control and Prevention (U.S.). National health and nutrition examination survey. National health and nutrition examination survey. 2025. Published online July 29. Available at: <https://www.cdc.gov/nchs/nhanes/index.html>. Accessed Aug 1, 2025.
8. Spick M, Onoja A, Harrison C, Stender S, Byrne J, Geifman N. Quantifying new threats to health and biomedical literature integrity from

- rapidly scaled publications and problematic research. *medRxiv*. [Preprint]. July 9, 2025 [accessed July 16, 2025]. Available from: <https://doi.org/10.1101/2025.07.07.25331008>.
9. Stender S, Gellert-Kristensen H, Smith GD. Reclaiming Mendelian randomization from the deluge of papers and misleading findings. *Lipids Health Dis*. 2024;23(1):286. [\[CrossRef\]](#)
10. FinnGen: an expedition into genomics and medicine | FinnGen. Available at: <https://www.finn-gen.fi/en>. Accessed Aug 1, 2025.
11. Research C for de and. FDA Adverse Event Reporting System (FAERS) database. *FDA* 2024. Published online Oct 16. Available at: <https://www.fda.gov/drugs/drug-approvals-and-databases/fda-adverse-event-reporting-system-faers-database>. Accessed July 31, 2025.
12. Suchak T, Aliu AE, Harrison C, Zwigelaar R, Geifman N, Spick M. Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. *PLoS Biol*. 2025;23(5):e3003152. [\[CrossRef\]](#)
13. Byrne JA, Stender S. More science friction for less science fiction. *PLoS Biol*. 2025;23(5):e3003167. [\[CrossRef\]](#)
14. Wu C, Liu Y, Lai Y, et al. Association of different types of milk with depression and anxiety: a prospective cohort study and Mendelian randomization analysis. *Front Nutr*. 2024;11:1435435. [\[CrossRef\]](#)
15. Li Z, Gong B, Xia L, Li X. Effect of educational attainment on the incidence of postoperative abdominal hernia: A two-step multivariable Mendelian randomization study. *Asian J Surg*. 2024 Nov 29:S1015-9584(24)02766-0. [\[CrossRef\]](#) PMID: 39616067.
16. Global burden of disease (GBD). Available at: <https://www.healthdata.org/research-analysis/gbd>. Accessed Aug 1, 2025.
17. Noble WS. How does multiple testing correction work? *Nat Biotechnol*. 2009;27(12):1135-1137. [\[CrossRef\]](#)
18. Patel CJ, Ioannidis JPA, Manrai AK. The architecture of exposome-phenome associations. *medRxiv* [Preprint]. June 6, 2025:[accessed July 10, 2025]. Available from: <https://doi.org/10.1101/2025.06.05.25329055>.
19. Infertility and BRI papers derived from NHANES. PubMed. Available at: <https://pubmed.ncbi.nlm.nih.gov/?term=nhanes%5Btiab%5D+A+ND+body+roundness+index%5Btiab%5D+AND+in+fertility>. Accessed July 9, 2025.
20. About COPE. COPE: Committee on Publication Ethics. Available at: <https://publicationethics.org/about/our-organisation>. Accessed Nov 14, 2024.
21. Richardson R. The Collection of Open Science Integrity Guides (COSIG): expanding participation in post-publication peer review. 2025. Published online June 4. Available at: <https://zenodo.org/records/15588204>. Accessed July 11, 2025.
22. Fusaroli M, Salvo F, Begaud B, et al. The reporting of a disproportionality analysis for drug safety signal detection using individual case safety reports in pharmacovigilance (READUS-PV): development and statement. *Drug Saf*. 2024;47(6):575-584. [\[CrossRef\]](#)
23. Rudan I, Song P, Adeloye D, Campbell H. Journal of Global Health's Guidelines for Reporting Analyses of Big Data Repositories Open to the Public (GRABDROP): preventing 'paper mills', duplicate publications, misuse of statistical inference, and inappropriate use of artificial intelligence. *J Glob Health*. 2025;15:01004. [\[CrossRef\]](#)

24. Bowman N, Spence P, Open HL. Organized, and onerous: understanding and recognizing the labors of open science. *J Assoc Commun Admin.* 2023;40:61-70. Available at: <https://stars.library.ucf.edu/jaca/vol40/iss1/4>.

25. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018. [CrossRef]

ease / publications

ese / European Science Editing

European Science Editing is an official publication of EASE. It is an open access peer-reviewed journal that publishes original research, review and commentary on all aspects of scientific, scholarly editing and publishing.

<https://ese.arphahub.com/>
<https://www.ease.org.uk>
<https://www.linkedin.com/company/easeeditors/>
<https://bsky.app/profile/easeeditors.bsky.social>
<https://www.facebook.com/EASEeditors/>
<https://mstdn.science/@EASE>
<https://www.instagram.com/easeeditors/>
<https://www.youtube.com/easeeditors>



© 2025 the authors. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.