PERSPECTIVE

Safeguarding Open Science from exploitative practices

Danny Maupin◉°, **Matt Spick**◉°, **Nophar Geifman**◉*

School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, United Kingdom

◔ These authors contributed equally to this work.
* n.geifman@surrey.ac.uk

Open research and data transparency are a bulwark against unethical activities, but can also introduce integrity risks. As with all public goods, freely available data can be exploited, and here we set out the case for the use of safeguarding practices.

The advent of Generative Artificial Intelligence (GenAI) and its ability to create fake data, images and text represents an unprecedented challenge to the integrity of the scientific literature. Data transparency through Open Science acts as a crucial safeguard against such fraudulent or unethical activity, providing an auditable trail of evidence [1]. This is in addition to other Open Science benefits, such as equitable data access and improved reproducibility of research [2]. There is, however, a fundamental and circular problem with making scientific data freely available as a defence against AI-generated fraud or other unethical behaviours: Datasets are themselves the fuel for AI engines. In other words, the very measures designed to fight fraud (open access) simultaneously power new forms of problematic activities (GenAI-assisted fast churn science). Robust evidence is emerging that open access datasets, especially in health and medicine, are being exploited in this way by paper mills and other bad actors [3,4].

High volumes of low-value or outright misleading research have many undesirable consequences including misallocation of funding, distortion of assessment metrics and reduced trust in the scientific literature. Both publishers and the wider scientific community are taking action in areas such as fake or duplicated images, the identification of tortured phrases used to avoid plagiarism checks [5], or other signals of paper mill activity; albeit this is an adversarial process and problematic research is still being published [6,7]. Nevertheless, the policies of data providers are still mostly seen through the prism of Open Science and a desire to have as much data available to as many people as possible. While some datasets operate a system of controlled access, others, e.g., the Centre for Disease Control's National Health and Nutrition

Examination Survey (NHANES), are fully open access, AI-ready and easily exploited (**Fig 1A**) [8]. Although for sensitive information (e.g., personal data that cannot be anonymized) a controlled system is clearly required, the treatment of other data types is more contested. We argue that unfettered open access is undesirable and leads to unwanted behaviours including choosing methods and data to create the illusion of statistical significance (p-hacking), hypothesising after the results are known (HARKing) and the introduction of false discoveries to the literature [9]. Conversely, heavily restricted closed systems are inequitable, and run the risk of being monopolised by users with specific research objectives or biases towards particular viewpoints on health issues.

Other resources have taken a middle path (Fig 1A), for example, the UK Biobank, which provides access to de-identified data for all eligible researchers globally—including academics, charities, and commercial entities—provided the research is health-related and in the public interest. This is managed through the Access Management System, which includes reviews by a committee to ensure proposals align with its ethical framework and public interest mandate. Once approved and a legal transfer agreement is signed, researchers are granted access to the data, typically via the secure, cloud-based UK Biobank Research Analysis Platform. A core condition of access is a prohibition on attempting to identify participants; researchers additionally pay an access fee, return their results to the Biobank for the benefit of future research, and are obliged to acknowledge their use of UK Biobank data and their application number in all published research.

This type of approach has many advantages. The UK Biobank's Access Management System and the use of application-specific registration IDs act as a form of pre-registration (as these are reported openly on the UK Biobank website), protecting against 'salami slicing' (slicing one set of results across multiple papers) and HARKing. Such requirements are, however, only as effective as their enforcement. A pilot audit of 321 research papers published between 2021 and 2025, conducted as part of a pre-registered protocol [10] using UK Biobank data targeted at topics previously identified as featured in formulaic/low-quality research templates, found that 7% did not report or appear to have a valid UK Biobank application number (Fig 1B). 25% exhibited substantial hypothesis drift compared with the disclosed research goals on the UK Biobank website. The proportion of publications without an application number varied by year but did not show any clear trend, but the proportion showing hypothesis drift increased each year, from 12% in 2021 to 28% in 2025 (Fig 1C). This may also reflect growth in post-application waivers by the UK Biobank, such as likely happened in 2021 with COVID-19-focussed research, but these are not disclosed. Whilst adding new projects to existing applications is convenient, we argue that—without public disclosure of approved variations—this reduces the effectiveness of publishing approved research questions as a form of integrity control and potentially enables fast-churn science and HARKing.

While these findings are derived from a purposive sample and cannot be extrapolated to the whole body of UK Biobank-derived research, they demonstrate the vulnerability of open-access data sources. Here, we are using the UK Biobank as a case
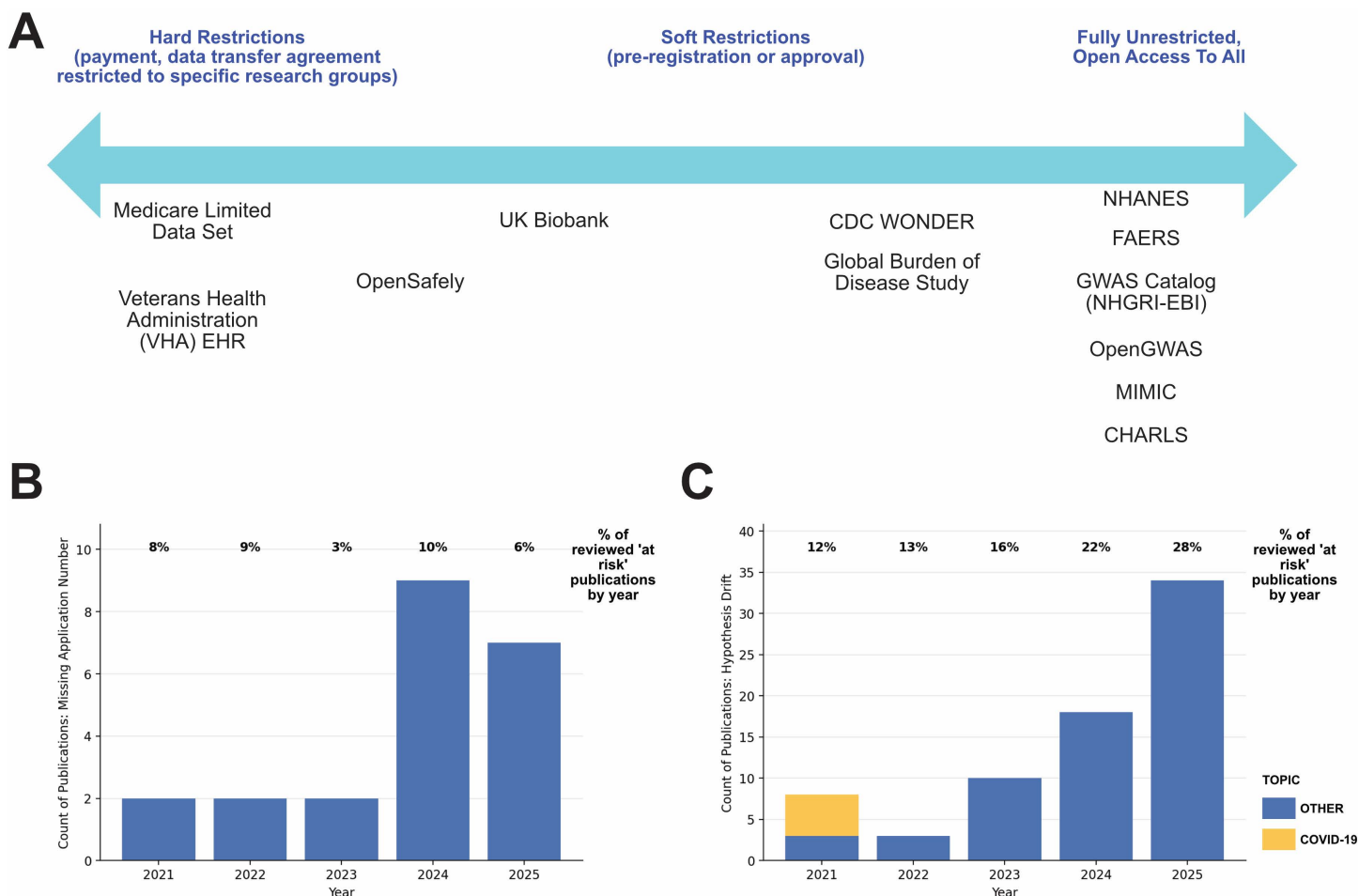
A

Hard Restrictions
(payment, data transfer agreement
restricted to specific research groups)

Soft Restrictions
(pre-registration or approval)

Fully Unrestricted,
Open Access To All

Medicare Limited
Data Set

UK Biobank

CDC WONDER

NHANES

FAERS

Global Burden of
Disease Study

GWAS Catalog
(NHGRI-EBI)

Veterans Health
Administration
(VHA) EHR

OpenSafely

OpenGWAS

MIMIC

CHARLS

B

C

Fig 1. **The safeguarding spectrum of public health datasets. (A)** Datasets by access type. **(B)** Publications using UK Biobank data without an application number. **(C)** Publications using UK Biobank data with hypothesis drift vs. original research goals on UK Biobank website. % for 2021 excludes COVID-19 related publications.

study only, and would not expect it to be more compromised than any other data source; indeed, our contention is that all open-access data sources are public goods that will be exploited by unethical actors. Those arguing for open and equitable access to research data make their case based on strong ethical arguments. In addition, in experimental research, open access supports reproducibility and has so far been less susceptible to 'fast churn science'. Once a dataset is compromised and credibility has been damaged, however, it becomes more challenging for researchers to publish research. NHANES provides an illustration of this phenomenon, with some publishers no longer accepting submissions based on open-access public health datasets [11]. In other words, there is a real cost to not exerting some measure of preventive control over data usage.

Of course, data providers are not enforcement agencies for good science. Historically, a trust-based system has worked well, but GenAI allows unethical actors to achieve higher productivity than ethical researchers and in our view, has enabled the explosion in formulaic manuscript production since 2023. We suggest that unrestricted open access will continue to compromise trust in research using exploited assets, and that safeguarding measures will only be truly effective with better disclosure processes (for example, disclosing variations to approved research questions) and publication

checks. The latter could include de-registering applications that show signs of significant hypothesis drift and an increased focus on whether data have been obtained legitimately, in line with Committee on Publishing Ethics guidance on authorised data use [12]. These safeguards would not be incompatible with Open Science, but would protect open practices and maintain the ethical and scientific benefits; without such changes to reassert the balance between Open Science and research integrity, our expectation is that the credibility of research based on open access data sources will continue to decline.

References

1. Lumbard H, Routledge D. Open science and transparency are our strongest tools in the fight against fraudulent publishing activities. PLoS Med. 2025;22(9):e1004774. https://doi.org/10.1371/journal.pmed.1004774 PMID: 41004541

2. Verschraegen S, Schiltz M. Knowledge as a global public good: the role and importance of open access. Soc Without Border. 2007;2:157–74. https://doi.org/10.1163/187219107X203540

3. 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 U S A. 2025;122(32):e2420092122. https://doi.org/10.1073/pnas.2420092122 PMID: 40758886

4. Richardson R, Spick M. Meeting the challenges posed by mass-produced manuscripts and click-data science. Europe Sci Edit. 2025;51. https://doi.org/10.3897/ese.2025.e165043

5. Cabanac G, Labbé C, Magazinov A. Tortured phrases: a dubious writing style emerging in science. Evidence of critical issues affecting established journals. arXiv. 2021. https://doi.org/10.48550/arXiv.2107.06751

6. Byrne JA, Abalkina A, Akinduro-Aje O, Christopher J, Eaton SE, Joshi N, et al. A call for research to address the threat of paper mills. PLoS Biol. 2024;22(11):e3002931. https://doi.org/10.1371/journal.pbio.3002931 PMID: 39576835

7. Abalkina A, Aquarius R, Bik E, Bimler D, Bishop D, Byrne J, et al. "Stamp out paper mills"—science sleuths on how to fight fake research. Nature. 2025;637(8048):1047–50. https://doi.org/10.1038/d41586-025-00212-1 PMID: 39870791

8. Ale L, Gentleman R, Sonmez TF, Sarkar D, Endres C. nhanesA: achieving transparency and reproducibility in NHANES research. Database (Oxford). 2024;2024:baae028. https://doi.org/10.1093/database/baae028 PMID: 38625809

9. Suchak T, Aliu AE, Harrison C, Zwiggelaar 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. https://doi.org/10.1371/journal.pbio.3003152 PMID: 40338847

10. Spick M, Geifman N, Maupin D. Assessing publication compliance with UK Biobank disclosure requirements: an evaluation of research disclosures. OSF Registries; 2025. https://doi.org/10.17605/OSF.IO/K4YMW

11. Journals and publishers crack down on research from open health data sets. [cited 27 Oct 2025]. Available from: https://www.science.org/content/article/journals-and-publishers-crack-down-research-open-health-data-sets

12. Barbour V, Kleinert S, Wager E, Yentis S. Guidelines for retracting articles. Committee on Publication Ethics; 2009. https://doi.org/10.24318/cope.2019.1.4