Quantifying new threats to health and biomedical literature integrity from rapidly scaled publications and problematic research

Running title: Implications of AI-supported analyses for Open Access and the FAIR Guiding Principles

Matt Spick [1,*], Anthony Onoja [1], Charlie Harrison [2], Stefan Stender [3], Jennifer Byrne [4,5], and Nophar Geifman [1,*]

[1] School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, United Kingdom, GU2 7XH
[2] Department of Computer Science, Aberystwyth University, Ceredigion, SY23 3DB, UK
[3] Department of Clinical Biochemistry, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[4] School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Camperdown, New South Wales, Australia
[5] NSW Health Statewide Biobank, NSW Health Pathology, Camperdown, New South Wales, Australia
[*] Correspondence: matt.spick@surrey.ac.uk or n.geifman@surrey.ac.uk

Abstract

The last three years have seen an explosion in published manuscripts analysing open-access health datasets, in many cases presenting misleading or biologically implausible findings. There is a growing evidence base to suggest that this is due in part to AI-assisted and formulaic workflows. Here we employ a top-down scientometric analysis to investigate which datasets have seen publication rates deviate from previous trends, especially where this coincides with changes to author geographical origins and increases in formulaic titles. Across 34 datasets we identify five showing hallmarks of paper mill exploitation (the FDA Adverse Event Reporting System, NHANES, UK Biobank, FinnGen and the Global Burden of Disease Study). These five datasets had, in 2024, a combined publication count of 11,554 indexed in the PubMed database. This represents an excess of around five thousand publications above the AutoRegressive Integrated Moving Average (ARIMA) forecast trend, and is a 2.8x fold change on the 4,001 publication count for these five datasets in 2021. We also identified a notable difference in the fold change for China (9.5x) versus the rest of the world (1.2x) and an increase in formulaic titles. These findings highlight potential risks to research integrity in areas such as public health and drug safety, and especially to the accessibility and interoperability principles central to Open Science and FAIR data practices. We argue that permissive open-access data policies naturally facilitate exploitative workflows, from direct API access for data dredging to large language model authoring of papers. These findings add to the case for the adoption of controlled data-access mechanisms combined with pre-registration of research protocols as safeguards, as previously used in the field of genomics, to reduce false discoveries and misleading conclusions. Such an approach would balance data availability with responsible use, preserving the goals of Open Science and mitigating against disincentives for compliance with the FAIR Guiding Principles.

Introduction

Generative AI (GenAI) and other automation tools have the potential to transform productivity in biomedical and health research. New technologies can, however, also be exploited in problematic ways, and paper mills (entities which mass-produce manuscripts for purchase) are particularly likely to benefit from such productivity gains as their business model relies on the large-scale authoring of inevitably low quality or completely fabricated manuscripts. [1,2] These issues have been identified previously. Notably, the exploitation of the National Health And Nutrition Examination Survey (NHANES) data resource has been highlighted by bottom-up analysis of individual manuscripts, [3,4] with problematic research practices including lack of false discovery correction, selective data usage / hypothesising after the results are known (also known as HARKing), and data dredging to maximise manuscript counts. Investigating individual publications is time-consuming and can be slow to react to new trends, and so a top-down approach can also be helpful when identifying problematic research practices, especially when resources are scarce. [5] As well as rapid growth in the number of publications citing NHANES, [6] concerning trends have also identified in works poorly employing two-sample Mendelian randomization applied to openly available GWAS data. [7,8] To our knowledge, however, there has not yet been a systematic top-down attempt to quantify the growth in exploitation of open access datasets. This has the potential to be a particular problem for datasets in the field of health and epidemiology, especially those that are FAIR compliant (e.g. with API-supported access), which may be vulnerable to AI-supported and pipeline-driven mass production of papers.

These issues are, of course, not new. Genomics provides a case study of a field that experienced an explosion of results, often lacking in biological plausibility or which could not be reproduced. This was addressed through a number of measures, including much more stringent genome-wide significance thresholds, [9] meta-analyses, [10] pre-registration of study designs and protocols to reduce data dredging, [11,12] and adoption of reporting standards such as STREGA, [13] all of which may be helpful in addressing AI-assisted exploitation of health data sources. Other measures employed in genomics, such as use of large-scale datasets and the promotion of open data, [14] may in contrast not work well in the AI-assisted era, as open-access naturally facilitates the type of data dredging and HARKing that has been seen in the last three years. [15,16]

In this work we attempt to establish the impact of AI-assisted formulaic templates we believe are currently in use, by examining trends in the number of publications published in journals and indexed in PubMed. We use these findings to discuss whether existing guidelines from policymakers are futureproofed to deal with these new strategies for the mass production of manuscripts, a crucial issue given the efforts to encourage Open Science and adoption of FAIR Guiding Principles. [17–20] In the worst-

case scenario, unchecked exploitation and manipulation of FAIR assets could undermine confidence in publication-based dissemination and lead to researchers turning away from open data altogether.

Methods

To identify health data sources that might be current targets of exploitation, a search of the PubMed database via API using the Entrez module from the Biopython library was conducted (version 1.85) for publications dated between 2014 and 2024, [21] related to a list of 34 health or biomedical databases (Table 1) based on the authors' collective expertise in the field. Alternative search terms for each dataset were used where appropriate for acronyms, or additional Boolean operators were used where acronyms were associated with other issues (for example the PRIDE database of proteomics mass spectrometry information) and the search strings are detailed in Supplementary Materials, Table S1. All searches were conducted for Title or Abstract, without other filters.

Table 1: Data sources investigated, ranked in alphabetic order

Dataset	Full Name	API Availability
1000 Genomes Project	1000 Genomes Project	Data can be accessed through various APIs and FTP servers
ABCD Study	Adolescent Brain Cognitive Development Study	Data access requires application; no public API available
ADNI	Alzheimer's Disease Neuroimaging Initiative	Data access requires application; no public API available
All of Us Research Program	All of Us Research Program	Data access requires application; no public API available
Biobank Japan	Biobank Japan Project	Data access requires application; no public API available
CCLE	Cancer Cell Line Encyclopedia	Data available for download; no dedicated API
CHARLS	China Health and Retirement Longitudinal Study	Data access requires application; no public API available
dbGaP	Database of Genotypes and Phenotypes	Only summary data and variable descriptions available as open access, individual data are controlled access. API available.
DHS	Demographic and Health Surveys	Data access requires registration; no public API available
eICU	eICU Collaborative Research Database	Data available for download; no dedicated API.
FAERS	FDA Adverse Event Reporting System	Accessible via the openFDA API.
FinnGen	FinnGen Study	Data access requires application; no public API available, except for GWAS which is open access and has OpenGWAS APIs
Global Burden of Disease	Global Burden of Disease Study	Data available for download; no dedicated API
gnomAD	Genome Aggregation Database	Provides API access for querying genetic variant data
GTEx	Genotype-Tissue Expression Project	Data available for download; no dedicated API
HiRID	High-Resolution Intensive Care Unit Dataset	Data available for download; no dedicated API
HRS	Health and Retirement Study	Data access requires registration; no public API available
Human Connectome Project	Human Connectome Project	Data access requires application; no public API available
Human Microbiome Project	Human Microbiome Project	Data available for download; no dedicated API
ICGC	International Cancer Genome Consortium	Provides API access for genomic data queries
MetaboLights	MetaboLights	Offers RESTful API for accessing metabolomics data
Metabolomics Workbench	Metabolomics Workbench	Data available for download; no dedicated API

MIMIC	Medical Information Mart for Intensive Care	Data access requires credentialing; no public API available
NHANES	National Health and Nutrition Examination Survey	Data available for download via CDC website downloads, API endpoints, the PIC-SURE API, and R/Python packages
OpenSAFELY	OpenSAFELY Platform	Data access requires application; no public API available.
PhysioNet	PhysioNet (Research Resource for Complex Physiologic Signals)	Offers tools for data analysis but lacks a unified API.
PRIDE	PRoteomics IDEntifications Database	Provides API access for proteomics data retrieval.
SEER	Surveillance, Epidemiology, and End Results Program	Data access requires application; no public API available.
SHARE	Survey of Health, Ageing and Retirement in Europe	Data access requires registration; no public API available
TCGA	The Cancer Genome Atlas	Data accessible via the Genomic Data Commons API
UK Biobank	UK Biobank	Data access requires application; no public API available, except for GWAS which is open access and has OpenGWAS APIs
UK10K	UK10K Project	Data available for download; no dedicated API
WHO Global Health Observatory	World Health Organization Global Health Observatory	Offers API access to a wide range of global health data
World Bank Health Data	World Bank Health Data	Data available for download; no dedicated API

The period from 2014 to 2021 was taken as a baseline undisturbed by the use of large language models (LLMs) and other forms of AI-assisted workflows, consistent with our previous work identifying an acceleration from 2022 onwards. [3] 2022 to 2024 was considered to be the period potentially affected by technology-driven productivity gains from paper mills. To quantify the deviation from trend, for example where a data source might be experiencing a natural growth trend for non-exploitative reasons, forecasts were constructed using ARIMA (AutoRegressive Integrated Moving Averages). ARIMA is a time series forecasting algorithm with three components: autoregression (using past values), differencing (to account for trends and seasonality), and moving averages (to remove noise). The excess production of manuscripts was then taken as the difference between the observed production of manuscripts and the ARIMA forecasts. The ARIMA model used parameters of autoregressive order $p = 1$, degree of differencing $d = 1$ and moving average order $r = 1$. Confidence intervals were also constructed for the forecasts. ARIMA models were implemented in Python using the statsmodels library (version 0.14.4). [22]

Datasets where the publication rate in 2024 exceeded the 95% confidence interval for the ARIMA forecast were identified as potentially being exploited by new technologies - and therefore worthy of further investigation. “Genomics” was used as a control term in the PubMed database, as a more mature field where extensive prior work has been done by the research industry to reduce false discoveries and implausible conclusions. Secondary analyses were conducted to identify whether manuscript titles were becoming more homogenous / formulaic, and to identify datasets where there had been a geographic shift in the affiliations of last-named authors. For the identification of potentially formulaic titles, a simple count of increased frequency of words was conducted to test

whether titles were becoming more homogenous. This was performed in Python using a count vectorizer function from the scikit-learn library (version 1.6.1). [23] For the geographic analysis, simple count of geographic origin was used together with fold changes of countries of origin between 2021 and 2024.

Results

The fold- changes in publications for each data source for the period between 2021 and 2024 are summarised in Table 2. Of the searched datasets, four were excluded for having insufficient publication counts to support later modelling steps (UK10K, World Bank Health Data, OpenSAFELY, HiRID). Six data sources (FinnGen, FAERS, NHANES, UK Biobank, Global Burden of Disease, All of US) met the criterion of exceeding the 95% confidence interval for ARIMA forecasts. The "genomics [title/abstract]" control search term produced a 2021-2024 fold change of 1.0x.

Table 2: Fold changes in publication numbers between 2021 and 2024 according to health data source, split between those with a change > 1 versus those with a change <= 1 and ranked largest to smallest

Data sources with Fold Change > 1	Fold Change	Data sources with Fold Change <= 1	Fold Change
FinnGen	14.9	Genomics (control term)	1.0
All of Us Research Program	8.4	SEER	1.0
FAERS	3.8	Human Connectome Project	0.9
NHANES	2.7	PhysioNet	0.9
CHARLS	2.6	gnomAD	0.9
MIMIC	2.6	DHS	0.9
UK Biobank	2.4	SHARE	0.9
ABCD Study	2.0	TCGA	0.9
Global Burden of Disease	2.0	MetaboLights	0.8
BioBank Japan	2.0	Human Microbiome Project	0.8
eICU Collaborative Research Database	1.8	ICGC	0.7
ADNI	1.2	1000 Genomes Project	0.7
GTEx	1.2	PRIDE	0.7
WHO Global Health Observatory	1.1	CCLE	0.7
HRS	1.1	Metabolomics Workbench	0.2
dbGaP	1.1		

The deviations from the ARIMA-estimated trend are shown in Figure 1.

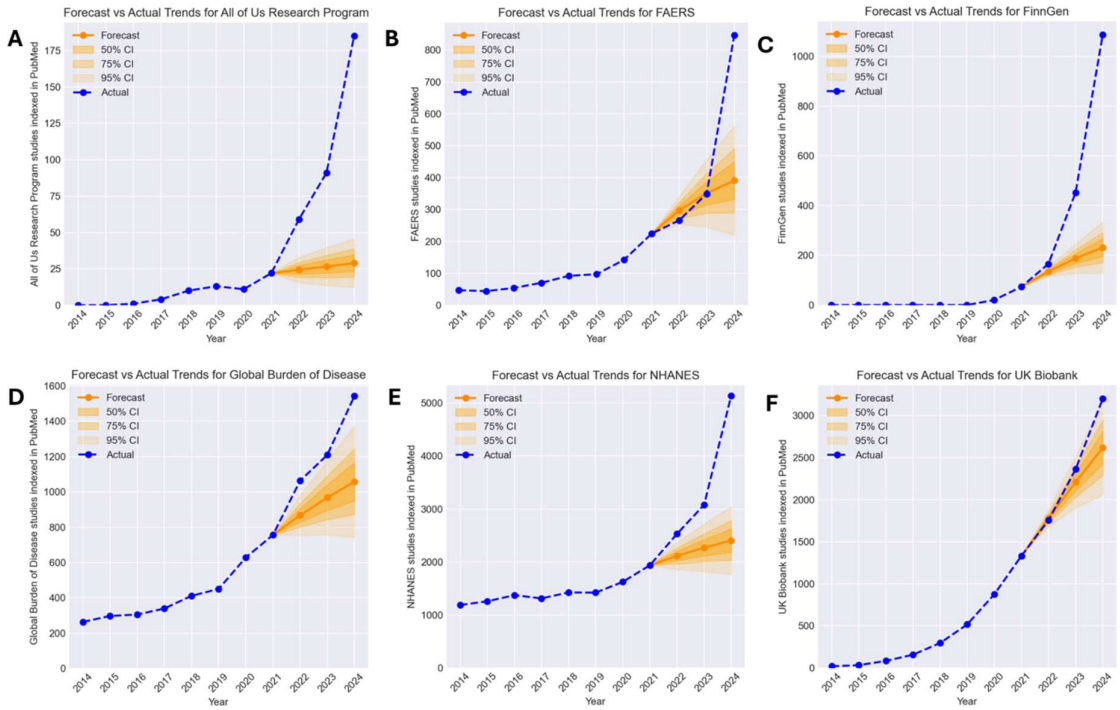

Figure 1: Actual publication counts compared with ARIMA forecasts, using 2014 to 2021 as the training period and 2022 to 2024 as the forecast period [A] All of Us Research Programme [B] FAERS [C] FinnGen [D] Global Burden of Disease Study [E] NHANES and [F] UK Biobank. Numerical data underlying these figures are included in Supplementary Materials, Table S2

The data sources showing trend changes above the 95% confidence intervals of forecasts were then analysed for 2021 and 2024 to identify whether there was any change in titles of manuscripts (as a simple test for formulaic manuscript production). Five datasets showed a sharp increase in certain tokens (words or phrases), as shown in Table 2. For example, for FinnGen, "Mendelian randomization" was used in 51% of titles in 2021, but 77% of titles in 2024, a change of 26 percentage points. One of the data sources – the All Of Us study - saw no increases in word frequency – all words declined in frequency, indicative of titles becoming more heterogenous rather than homogenous. A list of the top 20 token changes is included in Supplementary Material, Table S3, including words which declined in usage.

Table 3: Increased incidence of words, measured by change in % of titles where each word was used

Dataset	Title tokens	2021 % of titles featuring:	2024 % of titles featuring:	Change %
All of Us Research Program	No increased token frequencies			
FAERS	real world	12.1	34.7	+12.6
	study	13.8	32.5	+18.7
	based	7.6	25.6	+18.0
	pharmacovigilance	18.8	36.4	+17.6
	FAERS	21.9	38.9	+17.0
FinnGen	mendelian randomization	50.7	77.0	+26.3
	causal	11.0	35.4	+24.4

	bidirectional	4.1	16.9	+12.8
	study	46.6	58.6	+12.0
	cohort	18.7	27.3	+8.6
Global Burden of Disease	burden	41.0	64.2	+23.2
	global	39.0	58.3	+19.3
	disease	36.0	50.1	+14.1
	analysis	23.0	35.3	+12.3
	study	29.6	41.4	+11.8
NHANES	association	20.9	44.5	+23.6
	study	12.1	31.4	+19.3
	NHANES	19.6	34.2	+14.6
	cross-sectional	7.2	21.2	+14.0
	Index	4.2	15.9	+11.7
UK Biobank	study	30.2	39.2	+9.0
	prospective	7.1	14.5	+7.4
	mendelian randomization	15.9	22.9	+7.0
	cohort	13.2	18.5	+5.3
	association	14.9	19.2	+4.3

The six data sources were then examined for changes to geographic origin, focusing on country of affiliation for the last-named author on each publication. The growth in publications is shown as a chloropleth in Figure 2. The largest change was for publications with affiliations located in China, which increased from 19% of publications indexed in the PubMed database in 2021 to 65% in 2024, or - in absolute terms - growth of 6,972 publications. Over the same period, the share of publications originating in the United States declined from 34% to 14%, as its annual publication count increased by 182 over the three years.

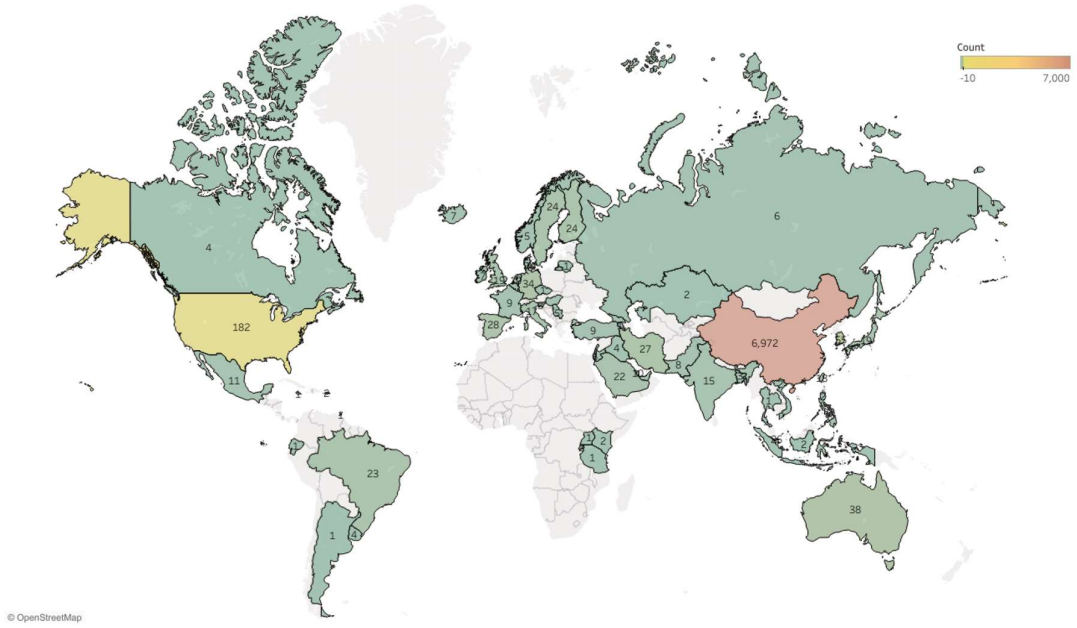

Figure 2: Chloropleth (geographical heatmap) of 2021 to 2024 change in annual publication count by country of affiliation of last-named author, for the following six datasets: FAERS, NHANES, UK Biobank, FinnGen, All of Us and Global Burden of Disease Study. Filtered to only show countries with an increase in publications indexed in PubMed. Colour range is not linear and is centred at 100 publications. Data underlying this chloropleth are shown in Supplementary Materials, Tables S4 and S5

Whilst the geographical split of growth in publications was very marked, it was not distributed evenly over the six-fastest growing datasets. Five saw a sharp increase in publications from China, based on affiliation of the last-named author (Table 3). This pattern was particularly evident for FinnGen, where 89% of publications had their primary affiliation from China in 2024. Research publications related to the All of Us study did not see the same geographic shift as the other data sources, with 21 out of 22 publications having affiliations in the United States in 2021, a pattern repeated in 2024 with 175 out of 185 in the United States. The growth was also from a much smaller base than the other datasets shown here.

Table 4: Table of fold changes for last-named author affiliations from China

Dataset	2021 China affiliations	2024 China affiliations	Fold Change (China)
All Of Us	0	0	1.0
FAERS	50	614	12.3
FinnGen	16	970	60.6
Global Burden of Disease Study	210	850	4.0
NHANES	366	3708	10.1
UK Biobank	180	1652	9.2

Discussion

Whilst many approaches have been proposed for the identification of paper mill outputs or other problematic research practices, most will either be 'bottom up' in nature, reviewing individual papers [15] or 'top down', seeking to identify concerning issues in overall publication trends (such as evidence of manipulated citation networks or retractions). [24,25] Here we use a top-down scientometric approach to demonstrate that there has been rapid growth in numbers of publications since 2021 analysing open-source datasets, with six datasets seeing a break with previous trends that exceeded a 95% forecast confidence interval. This break in trend may also have been accelerated by the development of GenAI in late 2022, [26] but it is notable that the acceleration in publication rates differs between data sources. NHANES, FinnGen and the Global Burden of Disease Study experienced a break with trend in 2022, but FAERS publications saw an acceleration versus the previous trend only in 2024. This may be suggestive of datasets being discovered at different times, albeit there are no obvious signs that datasets are being abandoned as other datasets start to be exploited

Whilst comprehensive bottom-up analysis of individual publications was beyond the scope of this work, many showed signs of titles becoming more formulaic, with 'fingerprint' methodologies, words and phrases consistent with previously described trends in the production of low quality manuscripts. A large number of papers analysing the UK Biobank and FinnGen used Mendelian randomization to propose causal relationships between different indicators and conditions, with both of these datasets seeing 'Mendelian randomization' sharply increase in frequency in titles. These works often propose implausible causal relationships, such as between the consumption of semi-skimmed milk and protection against depression [PMID 39703337], or between educational attainment and postoperative abdominal hernias [PMID 39616067], and many other hypotheses that lack biological underpinning. For NHANES the most common 'fingerprint' word was "association", which increased from being used in 21% of titles to 45% in 2024. This is consistent with our previous work which showed publications taking a well-described condition such as Type 2 Diabetes Mellitus, which is multifactorial in nature, and then analysing a single indicator (usually variations on a theme, such as body mass index, waist to hip ratio, or visceral adiposity) or ratio for simple association analysis, typically using logistic regression / cubic splines. Such simplistic studies fail to capture the complex and multifactorial nature of exposome-phenome associations. [27] The Global Burden of Disease Study saw a sharp increase in the number of publications emphasising the data source in the title, using associative approaches similar to those employed in NHANES analyses. FAERS meanwhile experienced a very large increase in 'real world' from 12% of titles to 35%, and of 'pharmacovigilance', from 19% to 36%. These papers use a general approach of taking a drug class (e.g. statins) and then data mining for a particular effect (e.g. hepatotoxicity) [PMID: 39840096], using a recurring set of disproportionality methods including Reporting Odds Ratios, Proportional Reporting Ratios and Bayesian Confidence Propagation Neural Networks. The increased frequency of specific tokens and methodologies is suggestive of a template-driven improvement in productivity and output, with the phenomenon easily observed through searches for the tokens identified here, for example using a search string of (nhanes AND study AND cross-sectional AND association) in PubMed.

Coincident with these changes there was also a dramatic shift in geographic origin of corresponding (last-named) author affiliations. The absolute increase for China between 2021 and 2024 (6972 publications) dwarfed that for the United States (182), with the next three largest increases from Australia (38), Germany (34) and Iran (27); this distribution suggests that these changes are not due to generalised improvements in research productivity. Authors with affiliations to Chinese hospitals have been associated with high retraction rates of papers in various fields, [28] and researchers from China are also subject to different incentives, [29] as are scientists in lower-income countries, [30] often due to difficulties in undertaking and publishing research that are not always shared by higher-income

countries, as well as lack of institutional support. This may create demand for the services of paper mills. Furthermore, the Chinese supreme court has recently issued guidance on paper mills and on scientific fraud, calling for lower courts to crack down on the mass-produced academic paper industry. [31]

Overall, of the datasets showing an acceleration in publication rates, the additional evidence of formulaic titles and a geographic shift in authorship is strongly suggestive of a change in publication patterns for five of the datasets analysed in this work. These are FAERS, NHANES, FinnGen, UK Biobank and the Global Burden of Disease Study. The rapid changes over the three-year period may well be due to paper mill exploitation. These five datasets also share open access elements, often with APIs that facilitate highly productive workflows. For the UK Biobank and FinnGen, open-access GWAS data were typically used, [32] NHANES has a well developed API system, [33] FAERS research is facilitated through OpenVigil, [34] and whilst the Global Burden of Disease Study does not have an API the data are freely available for download. We have written previously on 2SMR and association templates, but drug safety is a newly exploited type of data, illustrated here with reference to FAERS for the first time. FAERS itself is a valuable asset in a field that has not always had full data transparency, [35] but is a voluntary reporting system and so cannot provide estimates of incidence, cannot analyse causality, and additionally has considerable potential for bias based on physician or public preferences. [36] Misleading publications based on such drug safety reports pose a particular risk to the public, either by amplifying unnecessary concerns or underplaying risks.

Such problematic outputs present clear harms. These include the introduction of false discoveries, dilution of genuine findings in the scientific literature, and reduced credibility for high quality data sources, This is exemplified by implausible or contradictory claims being made in the media around health issues. as was the case with reports that semi-skimmed milk causally reduces depression and anxiety. [37,38] An additional issue is that many GenAI models are trained on - and learning from - scientific literature. With the inputs for learning potentially becoming corrupted, these models will propagate false science. For the five most problematic datasets identified here, the excess publications above trend (around five thousand) represent 43% of the 2024 output, a significant proportion to be potentially misleading.

We and others have previously written on measures to mitigate against formulaic manuscript production, focusing on more frequent use of desk rejections by journals to reduce burdens on peer reviewers, dedicated statistical reviewers, [39] use of application numbers by data providers and more effective post-publication correction. [40,41] Others have also written on the need for effective tools to identify problematic manuscripts such as the GRABDROP checklist, [42] and we suggest that increased awareness of datasets currently experiencing worrying trends and key phrases associated with

formulaic outputs will assist in this process. [5,24,43] It is, however, important to stress that the threat to scientific integrity identified here presents risks to the wider goals of Open Science and compliance with the FAIR Guiding Principles, which emphasise findability, accessibility, interoperability and reusability. [17,44] The factors that drive compliance with these principles (especially accessibility and interoperability, which naturally lend themselves to low-friction options such as direct API access) also make such datasets vulnerable to exploitation. To preserve the benefits of Open Science and FAIR compliance, we believe that unrestricted open access to AI-ready data may not be the best option for all data resources. For some, approved access and / or pre-registration may offer a better way forward in balancing access for legitimate research with prevention of exploitative data dredging.

A number of limitations should be stressed in this work, the most significant of which is that this is a scientometric analysis, rather than a comprehensive review of individual papers. For example, GenAI has the potential to generate formulaic low-quality outputs, sometimes referred to in other fields as "AI slop". [45] Some of the increase seen in publications may be due to individuals independently using GenAI. It is challenging from the outside to tell the difference between deliberate, coordinated mass production by paper mills and uncoordinated mass production by individuals facilitated by large language models, and this is a limitation of this research. A second limitation is that accelerations in publication may be due to good reasons. For example, no problematic pattern was evident from analysing publications dealing with the All of Us Research Program. Whilst this does not eliminate the possibility that unethical practices are at play, the exponential growth may simply reflect its status as a new data source for the community. Third, this analysis is retrospective. Some databases that as of today are only seeing increases in publication counts below the 95% confidence interval threshold used here, may be targeted in future. We expect scientometric tools (top-down) and individual investigations (bottom-up) to form part of an ongoing effort to protect research integrity, especially given that paper mills will likely adapt their strategies as existing approaches are brought to light. Finally, it should be noted that the analyses of changes in country and publication rates targeted at specific datasets are used here to illustrate the problems, and do not substitute for existing methods of detecting problematic manuscripts, such as citation networks, evidence of image manipulation or text similarities, as well as the backstop of peer review and editorial assessment. [15,46,47] Nonetheless, the fact that this work reviews accepted manuscripts provides evidence that existing tools used by publishers may be inadequate for new challenges posed by paper mills, [16] and post-publication corrections are not always as timely as might be ideal. [48]

In conclusion, our scientometric analysis highlights a substantial and rapid increase in publications analysing open-source datasets since 2021, predominantly driven by contributions from China. This surge raises significant concerns about research integrity, particularly in light of previously

documented associations between institutional pressures and increased retraction rates and the risk of paper mills employing new AI-supported workflows to produce manuscripts on an industrial scale. This growth poses a direct threat to the core principles of scientific rigour, and those of Open Science and FAIR data practices, given that the accessibility and interoperability designed to facilitate legitimate research also enables exploitation. To mitigate these risks and safeguard scientific integrity, we advocate for controlled data-access models paired with mandatory pre-registration of analyses, rather than unrestricted open access. This balanced approach is essential to preserving the intended benefits of Open Science while preventing its misuse, ensuring that scientific advancement continues to be reliable, reproducible, and trustworthy.

Supplementary Materials: Table S1: Search Strings, Table S2: Annual Data, Table S3: 2021 Country Data, Table S4: 2024 Country Data, Table S5: Token Changes,

Author Contributions: Conceptualization, M.S. and N.G.; methodology, M.S., J.B. and S.S.; software, M.S.; T.O.; formal analysis, M.S. and C.H.; investigation, M.S. and C.H.; resources, N.G.; data curation, M.S. and A.O.; writing—original draft preparation, M.S.; writing—review and editing, C.H., J.B., S.S and N.G.; visualization, M.S.; supervision, N.G.; project administration, N.G.; funding acquisition, M.S. and C.H.. All authors have read and agreed to the published version of the manuscript.

Funding: Matt Spick was supported by UK Research and Innovation (UKRI1095). Charlie Harrison was supported by the Biotechnology and Biological Sciences Research Council (BB/Y006933/1) and by UK Research and Innovation (UKRI1095). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability Statement: All data used in the preparation of this manuscript are included in Supplementary Materials. All code used in this work is available on request via GitHub as a Jupyter notebook.

Acknowledgments: The authors wish to acknowledge the wider support of the United2Act network as well as the AIBIO-UK network

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Christopher, J. The raw truth about paper mills. *FEBS Letters* **595**, 1751–1757 (2021).

2. Netzer, N. C. Artificial intelligence – the Janus-faced tool in our hands. *Sleep Breath* **28**, 1861–1862 (2024).
3. Suchak, T. *et al.* Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. *PLOS Biology* **23**, e3003152 (2025).
4. Byrne, J. A. & Stender, S. More science friction for less science fiction. *PLOS Biology* **23**, e3003167 (2025).
5. Byrne, J. A. *et al.* A call for research to address the threat of paper mills. *PLoS Biol* **22**, e3002931 (2024).
6. Mainous, A. G. Papermills as another challenge to research integrity and trust in science. *Front. Med.* **12**, (2025).
7. Stender, S., Gellert-Kristensen, H. & Smith, G. D. Reclaiming mendelian randomization from the deluge of papers and misleading findings. *Lipids in Health and Disease* **23**, 286 (2024).
8. Smith, G. D. & Ebrahim, S. Mendelian randomisation at 20 years: how can it avoid hubris, while achieving more? *The Lancet Diabetes & Endocrinology* **12**, 14–17 (2024).
9. Ioannidis, J. P. A., Thomas, G. & Daly, M. J. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* **10**, 318–329 (2009).
10. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**, 379–389 (2013).
11. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences* **115**, 2600–2606 (2018).
12. Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P. & Willmes, K. *Registered Reports*: Realigning incentives in scientific publishing. *Cortex* **66**, A1–A2 (2015).
13. Little, J. *et al.* STrengthening the REporting of Genetic Association Studies (STREGA)— An Extension of the STROBE Statement. *PLOS Medicine* **6**, e1000022 (2009).

14. Manolio, T. A. *et al.* Bedside Back to Bench: Building Bridges between Basic and Clinical Genomic Research. *Cell* **169**, 6–12 (2017).
15. Abalkina, A. Publication and collaboration anomalies in academic papers originating from a paper mill: Evidence from a Russia-based paper mill. *Learned Publishing* **36**, 689–702 (2023).
16. Liverpool, L. AI intensifies fight against 'paper mills' that churn out fake research. *Nature* **618**, 222–223 (2023).
17. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
18. Foster, E. D. & Deardorff, A. Open Science Framework (OSF). *J Med Libr Assoc* **105**, 203–206 (2017).
19. McKiernan, E. C. *et al.* How open science helps researchers succeed. *eLife* **5**, e16800 (2016).
20. Mirowski, P. The future(s) of open science. *Soc Stud Sci* **48**, 171–203 (2018).
21. Biopython · Biopython. https://biopython.org/.
22. Josef Perktold *et al.* statsmodels/statsmodels: Release 0.14.2. Zenodo https://doi.org/10.5281/ZENODO.593847 (2024).
23. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
24. Candal-Pedreira, C. *et al.* Retracted papers originating from paper mills: cross sectional study. *BMJ* **379**, e071517 (2022).
25. Candal-Pedreira, C. *et al.* Retracted papers originating from paper mills: a cross-sectional analysis of references and citations. *Journal of Clinical Epidemiology* **172**, 111397 (2024).
26. Liu, Y. *et al.* Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology* **1**, 100017 (2023).
27. Patel, C. J., Ioannidis, J. P. & Manrai, A. K. The architecture of exposome-phenome associations. 2025.06.05.25329055 Preprint at https://doi.org/10.1101/2025.06.05.25329055 (2025).

28. Van Noorden, R. Exclusive: These universities have the most retracted scientific articles. *Nature* **638**, 596–599 (2025).
29. Mallapaty, S. China conducts first nationwide review of retractions and research misconduct. *Nature* **626**, 700–701 (2024).
30. Vasconez-Gonzalez, J., Izquierdo-Condoy, J. S., Naranjo-Lara, P., Garcia-Bereguiain, M. Á. & Ortiz-Prado, E. Integrity at stake: confronting 'publish or perish' in the developing world and emerging economies. *Front Med (Lausanne)* **11**, 1405424 (2024).
31. Mallapaty, S. China's supreme court calls for crack down on paper mills. *Nature* **639**, 285–286 (2025).
32. Hartwig, F. P., Davies, N. M., Hemani, G. & Davey Smith, G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol* **45**, 1717–1726 (2016).
33. Ale, L., Gentleman, R., Sonmez, T. F., Sarkar, D. & Endres, C. nhanesA: achieving transparency and reproducibility in NHANES research. *Database* **2024**, baae028 (2024).
34. OpenVigil Pharmacovigilance Search Engines. https://openvigil.sourceforge.net/.
35. Golder, S. & Loke, Y. K. Is there evidence for biased reporting of published adverse effects data in pharmaceutical industry-funded studies? *Br J Clin Pharmacol* **66**, 767–773 (2008).
36. Chedid, V., Vijayvargiya, P. & Camilleri, M. Advantages and Limitations of the Federal Adverse Events Reporting System in Assessing Adverse Event Reporting for Eluxadoline. *Clinical Gastroenterology and Hepatology* **16**, 336–338 (2018).
37. Drinking plant-based milk 'increases risk of depression', says study. *Women's Health* https://www.womenshealthmag.com/uk/food/a63357490/plant-based-milk-depression/ (2025).
38. Depression and anxiety reduced by semi-skimmed milk. *Diabetes* https://www.diabetes.co.uk/news/2025/jan/depression-and-anxiety-reduced-by-semi-skimmed-milk.html (2025).

39. Spick, M. *et al.* Observations from Statistical Review Editors: A Commentary. *JID Innovations* **4**, 100302 (2024).

40. Ortega, J. L. Classification and analysis of PubPeer comments: How a web journal club is used. *Journal of the Association for Information Science and Technology* **73**, 655–670 (2022).

41. Ortega, J.-L. & Delgado-Quirós, L. How do journals deal with problematic articles. Editorial response of journals to articles commented in PubPeer. *Profesional de la información* **32**, (2023).

42. 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* **15**, 01004 (2025).

43. Parker, L., Boughton, S., Bero, L. & Byrne, J. A. Paper mill challenges: past, present, and future. *Journal of Clinical Epidemiology* **176**, 111549 (2024).

44. Jacobsen, A. *et al.* FAIR Principles: Interpretations and Implementation Considerations. *Data Intellegence* **2**, 10–29 (2020).

45. Martin, A. & Newell, B. Synthetic Data, Synthetic Media, and Surveillance. *S&S* **22**, (2024).

46. van Diest, R. A., Seifert, R. & van der Heyden, M. A. G. An extra pair of eyes: adopting innovative approaches to detect integrity issues in Naunyn–Schmiedeberg's Archives of Pharmacology. *Naunyn-Schmiedeberg's Arch Pharmacol* **398**, 1–8 (2025).

47. Sanderson, K. Science's fake-paper problem: high-profile effort will tackle paper mills. *Nature* **626**, 17–18 (2024).

48. Naddaf, M. Journal targeted by paper mill still grappling with the aftermath years later. *Nature* (2025) doi:10.1038/d41586-025-01010-5.