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

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

Background and Objectives

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, and publishers are responding by discouraging submissions employing open-access health datasets.

Methods

Here we employ a 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.

Results

Across 36 datasets we identify nine showing hallmarks of paper mill exploitation (FAERS, NHANES, UK Biobank, FinnGen, the Global Burden of Disease Study, MIMIC, CHARLS, CDC WONDER, and TriNetX). These nine datasets had, in 2025, a combined publication count of 23,005 indexed in the OpenAlex database. This represents an excess of 11,577 publications above the AutoRegressive Integrated Moving Average (ARIMA) forecast trend, and is a 3.0x fold change on the 7,655 publication count for these nine datasets in 2022. We also identified a notable difference in the fold change for China (4.2x) versus the rest of the world (1.9x) and an increase in formulaic titles.

Conclusions

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, and that these findings add to the case for the safeguarding mechanisms to preserve the goals of Open Science

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. [1–3] These issues have been identified previously. Notably, the exploitation of the National Health And Nutrition Examination Survey (NHANES) data resource has been highlighted by targeted analysis of individual manuscripts, [4,5] 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 irrespective of the plausibility of the findings. [6] Investigating individual publications is time-consuming and can be slow to react to new trends, and so trend-analyses can also be helpful when identifying problematic research practices, especially when resources are scarce. [7] As well as rapid growth in the number of publications citing NHANES, [8] concerning trends have also been identified in works poorly employing two-sample Mendelian randomization applied to openly available GWAS data. [9,10] To our knowledge, however, there has not yet been a systematic attempt to quantify the growth in exploitation across the wider field 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, [11] meta-analyses, [12] pre-registration of study designs and protocols to reduce data dredging, [13,14] and adoption of reporting standards such as STREGA, [15] 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, [16] 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. [17,18]

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 OpenAlex. 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. [19–22] 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 OpenAlex database via API using the Entrez module from the Biopython library was conducted (version 1.85) for publications dated between 2014 and 2025, [23] related to a list of 36 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. An additional query was submitted to obtain combined publication counts for 'at risk' datasets, using OR logic across dataset terms with deduplication. As a result, totals of combinations of datasets would be expected to be lower than the arithmetic sum of per-dataset counts. All searches were conducted for Title or Abstract, filtered to show type = article, and were run on 3 January 2026.

Table 1: Data sources investigated, listed 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
CDC WONDER	Wide-ranging Online Data for Epidemiologic Research	Public-facing API using XML-based POST requests
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 III or IV	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
TriNetX	TriNetX	REST API, only accessible to those with an institutional license (universities, pharma companies, or large hospitals)
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 2022 was taken as a baseline largely 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 late-2022 onwards. [4] 2023 to 2025 was considered to be the period substantially 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 $q = 1$. Confidence intervals were also constructed for the forecasts. ARIMA models were implemented in Python using the statsmodels library (version 0.14.4). [24]

Datasets where the publication rate in 2025 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 OpenAlex 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). [25] For the geographic analysis, simple count of geographic origin was used together with fold changes of countries of origin between 2022 and 2025.

Results

The fold- changes in publications for each data source for the period between 2022 and 2025 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) and the All of Us Research Programme was excluded as its Researcher Workbench was only made available for data access in late 2020 to US researchers and for international researchers in late 2023. [26] Nine data sources (FinnGen, FAERS, NHANES, UK Biobank, Global Burden of Disease, CDC WONDER, TriNetX, CHARLS, and MIMIC III / IV) met the criterion of exceeding the 95% confidence interval for ARIMA forecasts. The median fold change was 1.2x, while the “genomics [title/abstract]” control search term produced a 2022-2025 fold change of 1.4x.

Table 2: Fold changes in publication numbers between 2022 and 2025 according to health data source, split between those with a change > median versus those with a change < median and ranked largest to smallest

Data sources with Fold Change > median (median = 1.2)	Fold Change	Data sources with Fold Change <= median (median = 1.2)	Fold Change
CDC WONDER	17.4	1000 Genomes Project	1.2
TriNetX	8.0	BioBank Japan	1.1
All of Us Research Program	6.9	OpenSAFELY	1.1
FAERS	5.6	DHS	1.1
FinnGen	5.5	gnomAD	1.1
Global Burden of Disease	3.2	SEER	1.1
CHARLS	2.9	GTEx	1.0
HiRID	2.5	SHARE	1.0
NHANES	2.4	Human Connectome Project	1.0
UK Biobank	2.3	HRS	1.0
MIMIC-III or MIMIC-IV	2.3	PRIDE	0.9
ABCD Study	2.1	TCGA	0.8

WHO Global Health Observatory	2.0	PhysioNet	0.7
dbGaP	1.9	CCLE	0.6
MetaboLights	1.9	Human Microbiome Project	0.6
eICU Collaborative Research Database	1.7	ICGC	0.5
Metabolomics Workbench	1.4	UK10K	0.4
ADNI	1.3	World Bank Health Data	0.0

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

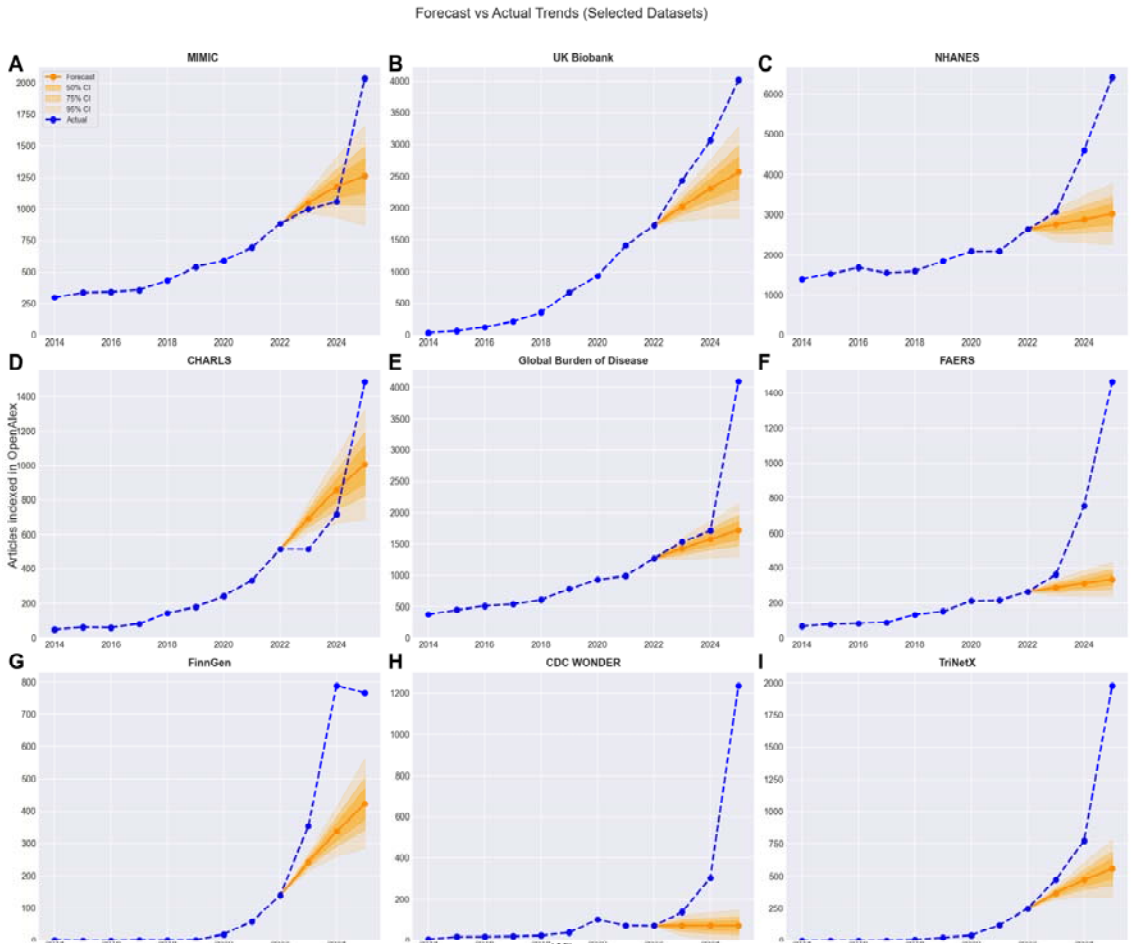

Figure 1: Actual publication counts compared with ARIMA forecasts, using 2014 to 2022 as the training period and 2023 to 2025 as the forecast period [A] MIMIC III / IV [B] UK Biobank [C] NHANES [D] CHARLS [E] Global Burden of Disease Study [F] FAERS [G] FinnGen [H] CDC WONDER, and [I] TriNetX. Numerical 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 2022 and 2025 to identify whether there was any change in titles of manuscripts (as a simple test for formulaic manuscript production). Nine datasets showed a sharp increase in certain tokens (words or phrases), as shown in Table 3, and this was

especially marked for CDC WONDER (where 69% of titles commented on 'trends' in 2025, compared with 20% of titles in 2022. Analysis of the control search term 'genomics' showed that no tokens saw a change of more than 2.5% in frequency between 2022 and 2025, in contrast to the larger changes seen in the datasets suspected of exploitation. 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 tokens, measured by change in % of titles where each token was used

Dataset	Title tokens	2022 % of titles featuring:	2025 % of titles featuring:	Change %
CDC WONDER	trends	19.7	68.8	49.1
	cdc wonder	7.0	52.7	45.6
	analysis	18.3	54.3	36.0
	mortality	52.1	88.0	35.9
	database	5.6	27.3	21.7
CHARLS	charls	9.1	29.3	20.1
	association	15.7	27.8	12.1
	index	2.3	14.3	12.0
	risk	11.1	22.5	11.5
	disease	4.5	13.1	8.6
FAERS	faers	26.0	54.8	28.9
	analysis	31.7	50.4	18.7
	database	20.2	37.5	17.2
	world	13.7	27.5	13.7
	safety	7.3	20.2	12.9
FinnGen	causal	17.1	30.7	13.5
	insights	0.0	7.8	7.8
	analysis	13.6	20.6	7.1
	potential	0.7	7.7	7.0
	relationship	2.9	9.7	6.8
Global Burden of Disease	analysis	20.6	41.8	21.2
	global burden of disease	25.1	44.7	19.6
	study	31.9	46.5	14.6
	regional	8.1	22.4	14.3
	trends	13.3	27.4	14.1
MIMIC III / IV	mimic-iv	7.9	29.1	21.2
	database	9.7	24.2	14.4
	retrospective	9.4	22.8	13.4
	mortality	18.3	31.1	12.7
	study	12.4	24.3	11.8
NHANES	nhanes	21.5	43.4	21.9
	association	26.2	41.4	15.2
	study	14.0	26.6	12.6
	index	5.5	16.9	11.4
	cross sectional	8.2	18.5	10.3
TriNetX	trinetx	4.0	22.1	18.1
	study	18.1	32.4	14.3
	cohort	10.5	23.6	13.2
	matched	4.4	16.5	12.1
	propensity	4.8	15.7	10.8
UK Biobank	risk	28.5	32.5	4.0
	plasma	0.3	3.8	3.5
	insights	1.2	4.4	3.2
	cohort	14.9	17.9	3.0
	prospective	9.1	12.1	3.0

The nine 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 27% of publications indexed in the OpenAlex database in 2022 to 45% in 2025, or - in absolute terms - growth of 8,679 publications. Over the same period, the share of publications originating in the United States declined from 25% to 19%, as its annual publication count increased by 2,408 over the three years, with nearly half of this increase (n = 1,104) in publications analysing TriNetX.

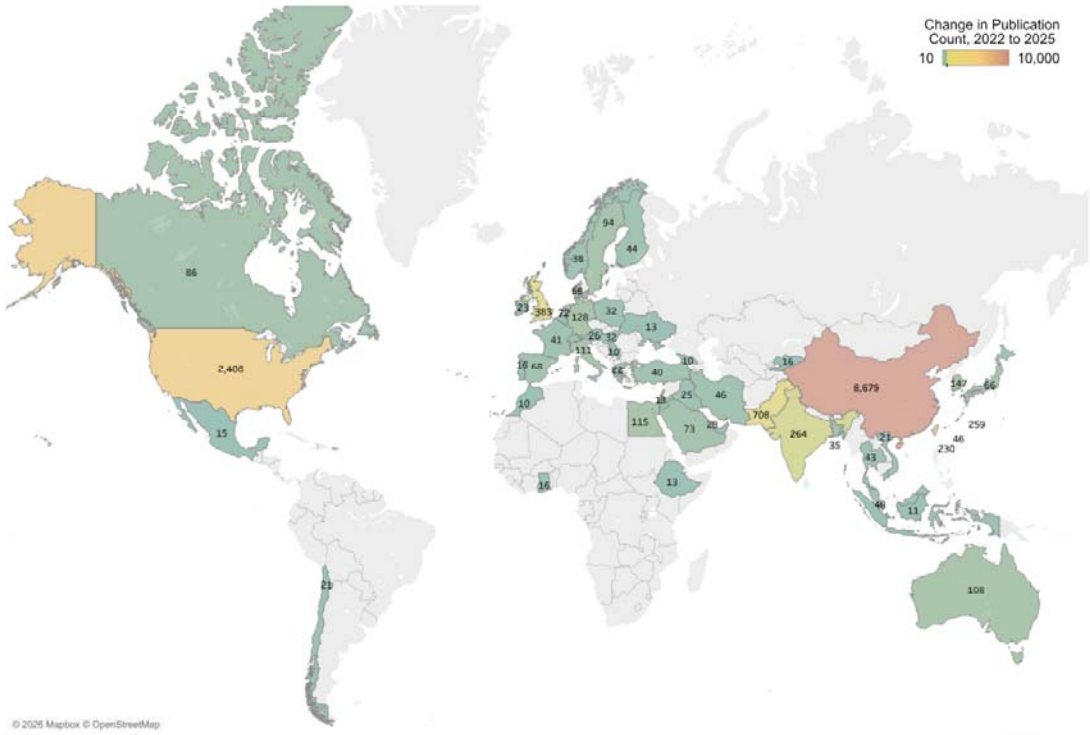

Figure 2: Chloropleth (geographical heatmap) of 2022 to 2025 change in annual publication count by country of author affiliations (fully counted), for the following nine datasets: CDC WONDER, CHARLS, FAERS, NHANES, UK Biobank, FinnGen, Global Burden of Disease Study, MIMIC III / IV, TriNetX. Filtered to only show countries with an increase in publications > 10 indexed in OpenAlex. Colour range is not linear and is centred at 200 publications. Data underlying this chloropleth are shown in Supplementary Materials, Table S4. As author affiliations are fully counted rather than fractionally, the total of affiliations does not sum to the number of publications.

Whilst the geographical split of growth in publications was very marked, it was not distributed evenly over the nine-fastest growing datasets. Seven saw a sharp increase in publications from China, based on full-attribution of affiliation (Table 4). This pattern was particularly evident for FAERS, where 66% of publications had an affiliation from China in 2025. Pakistan also showed a large change relative to its previous number, but this was focused

largely within the CDC WONDER dataset. The rapid growth in publications using TriNetX was concentrated in the US.

Table 4: Table of fold changes for last-named author affiliations for the country seeing the largest absolute growth

Dataset	Country seeing largest Fold Change	2022 affiliations from largest Fold Change country	2025 affiliations from largest Fold Change country	Fold Change
CDC WONDER	Pakistan	1	553	553.0
CHARLS	China	442	1208	2.7
FAERS	China	102	843	8.3
FinnGen	China	76	569	7.5
Global Burden of Disease Study	China	427	2389	5.6
MIMIC III / IV	China	381	1017	2.7
NHANES	China	856	4043	4.7
TriNetX	United States	205	1309	6.4
UK Biobank	China	463	1596	3.4

Discussion

Many approaches have been proposed for the identification of paper mill outputs or other problematic research practices, either reviewing individual papers [17] or seeking to identify concerning issues in overall publication trends (such as evidence of manipulated citation networks or retractions). [27,28] Here we use a scientometric approach to demonstrate that there has been rapid growth in numbers of publications since 2022 analysing open-source datasets, with nine datasets seeing both a break with previous trends that exceeded a 95% forecast confidence interval and signs of formulaic or templated research. This break in trend may also have been accelerated by the development of GenAI in late 2022, [29] 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 2023, but FAERS publications saw an acceleration versus the previous trend only in 2024, with TriNetX and CDC WONDER seeing a larger acceleration in 2025. 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 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. Across the datasets, there was a strong trend towards reporting the name of the dataset used, and either ‘association’, ‘trends’ or ‘mortality’. 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 or ratio for simple analysis and reporting the trend or association. Such simplistic studies fail to capture the complex and multifactorial nature of exposome-phenome associations. [30] The increased frequency of specific tokens and methodologies (for example the majority of the studies investigating CDC WONDER used Joinpoint software and the same workflow) is suggestive of a template-driven improvement in productivity and output, possibly also using Generative AI. [31]

Coincident with these changes there was also a dramatic shift in geographic origin of corresponding (last-named) author affiliations for the 9 most affected databases. The absolute increase for China between 2022 and 2025 (8679 publications) dwarfed that for the United States (2408), with the next three largest increases from Pakistan (708), the UK (383) and India (264). Authors with affiliations to Chinese hospitals have been associated with high retraction rates of papers in various fields, [32] and researchers from China are also subject to different incentives, [33] as are scientists in lower-middle income countries, [34] often due to difficulties in undertaking and publishing research and a lack of institutional support. 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. [35]

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 nine of the datasets analysed in this work. These are FAERS, NHANES, FinnGen, UK Biobank, the Global Burden of Disease Study, CHARLS, MIMIC III / IV, TriNetX, and CDC WONDER. The rapid changes over the four-year period may well be due to paper mill exploitation. Eight of the nine 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, [36] NHANES has a well developed API system, [37] FAERS research is facilitated through OpenVigil, [38] 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. FAERS itself is a valuable asset in a field that has not always had full data transparency, [39] 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. [40] Misleading publications based on such drug safety reports pose a particular risk to the public, either by amplifying unnecessary concerns or underplaying risks. TriNetX presents a different profile, with many of the publications targeted at conferences and a US-centric production model focused on medical schools. This illustrates that the demand for mass-produced manuscripts is not an issue related solely to the Global South.

Such problematic outputs present clear harms. Taken individually, the papers we find are simply not very interesting (they are formulaic and repeat the same analyses across thousands of variables), but taken together they become misleading through the introduction of false discoveries, dilution of genuine findings in the scientific literature, and reduced credibility for high quality data sources, 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 nine most problematic datasets identified here, the excess publications above trend (around twelve thousand) represented 50% of the 2025 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, [41] use of application numbers by data providers and more effective post-publication correction. [42,43] Others have also written on the need for effective tools to identify problematic manuscripts such as the recently-proposed GRABDROP checklist, [44] and we suggest that increased awareness of datasets currently experiencing worrying trends and key phrases associated with formulaic outputs will assist in this process. [7,27,45] 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. [19,46] The factors that drive compliance with these principles also make such datasets vulnerable to exploitation. To preserve the overall goal 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 safeguards against exploitative data dredging. A more radical solution given the ease of analysing simple relationships such as ‘The association of [Predictor A] with [Outcome B] in [Population C] using a cross-sectional national dataset [Open Access Dataset D]’ would be for the scientific community to deem such results as effectively already available online, and so not suitable for publication.

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”. [47] 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, the UK Biobank has made a number of new datasets available over time. Third, this analysis is retrospective, and future behaviour may change in

response to recent publisher policy changes to restrict submissions dealing with open access datasets, such as NHANES. [48,49] Given the adaptive and adversarial nature of unethical actors, we expect both scientometric analyses and targeted investigations to form part of an ongoing effort to protect research integrity and the principles of Open Science, especially given that paper mills will adapt their strategies as existing approaches are brought to light. [50,51] It should also be noted that in this work we have not integrated resources such as the Problematic Paper Screener; this may be an opportunity for future work, albeit these traditional flags (for example for the detection of tortured phrases) [52] are not typically found in this type of formulaic manuscript, possibly due to adaptive changes by paper mills. Finally, it should be noted that scientometric analysis does 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. [17,53–55] 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 mill, [18,56,57] especially when post-publication corrections / retractions are often slow, or may not happen at all. [58] Furthermore, even retracted articles can contaminate evidence syntheses, and the sheer weight of publications flagged in this work is suggestive that these negative downstream impacts are likely to increase. [59,60]

In conclusion, our scientometric analysis highlights 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: 20 tokens seeing the largest change in titles between 2022 and 2025, by ‘at risk’ dataset. Table S4: Matrix of change in publication count by country for the combined nine 'at risk' datasets

Author Contributions: Matt Spick: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing—original draft preparation, Visualization, Project administration. Anthony Onoja: Data curation. Charlie Harrison: Formal analysis, Investigation, Writing—review and editing. Stefan Stender: Methodology, Writing—review

and editing. Jennifer Byrne: Methodology, Writing—review and editing. Nophar Geifman: Conceptualization, Resources, Writing—review and editing, Supervision. 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 employed standard Python libraries without any modifications.

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 2021;595:1751–7. https://doi.org/10.1002/1873-3468.14143.
[2] Netzer NC. Artificial intelligence – the Janus-faced tool in our hands. Sleep Breath 2024;28:1861–2. https://doi.org/10.1007/s11325-024-03129-7.
[3] Richardson RAK, Hong SS, Byrne JA, Stoeger T, Amaral LAN. The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. Proceedings of the National Academy of Sciences 2025;122:e2420092122. https://doi.org/10.1073/pnas.2420092122.
[4] 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 Biology 2025;23:e3003152. https://doi.org/10.1371/journal.pbio.3003152.
[5] Byrne JA, Stender S. More science friction for less science fiction. PLOS Biology 2025;23:e3003167. https://doi.org/10.1371/journal.pbio.3003167.
[6] Grimes DR, Heathers J. “Cake causes herpes?” - promiscuous dichotomisation induces false positives. BMC Medical Research Methodology 2025;25:255. https://doi.org/10.1186/s12874-025-02712-0.
[7] 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:e3002931. https://doi.org/10.1371/journal.pbio.3002931.
[8] Mainous AG. Papermills as another challenge to research integrity and trust in science. Front Med 2025;12. https://doi.org/10.3389/fmed.2025.1557024.
[9] Stender S, Gellert-Kristensen H, Smith GD. Reclaiming mendelian randomization from the deluge of papers and misleading findings. Lipids in Health and Disease 2024;23:286. https://doi.org/10.1186/s12944-024-02284-w.
[10] Smith GD, Ebrahim S. Mendelian randomisation at 20 years: how can it avoid hubris, while achieving more? The Lancet Diabetes & Endocrinology 2024;12:14–7. https://doi.org/10.1016/S2213-8587(23)00348-0.
[11] Ioannidis JPA, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 2009;10:318–29. https://doi.org/10.1038/nrg2544.

[12] Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 2013;14:379–89. https://doi.org/10.1038/nrg3472.
[13] Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proceedings of the National Academy of Sciences 2018;115:2600–6. https://doi.org/10.1073/pnas.1708274114.
[14] Chambers CD, Dienes Z, McIntosh RD, Rotshtein P, Willmes K. *Registered Reports*: Realigning incentives in scientific publishing. Cortex 2015;66:A1–2. https://doi.org/10.1016/j.cortex.2015.03.022.
[15] Little J, Higgins JPT, Ioannidis JPA, Moher D, Gagnon F, Elm E von, et al. STrengthening the REporting of Genetic Association Studies (STREGA)— An Extension of the STROBE Statement. PLOS Medicine 2009;6:e1000022. https://doi.org/10.1371/journal.pmed.1000022.
[16] Manolio TA, Fowler DM, Starita LM, Haendel MA, MacArthur DG, Biesecker LG, et al. Bedside Back to Bench: Building Bridges between Basic and Clinical Genomic Research. Cell 2017;169:6–12. https://doi.org/10.1016/j.cell.2017.03.005.
[17] Abalkina A. Publication and collaboration anomalies in academic papers originating from a paper mill: Evidence from a Russia-based paper mill. Learned Publishing 2023;36:689–702. https://doi.org/10.1002/leap.1574.
[18] Liverpool L. AI intensifies fight against ‘paper mills’ that churn out fake research. Nature 2023;618:222–3. https://doi.org/10.1038/d41586-023-01780-w.
[19] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
[20] Foster ED, Deardorff A. Open Science Framework (OSF). J Med Libr Assoc 2017;105:203–6. https://doi.org/10.5195/jmla.2017.88.
[21] McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, et al. How open science helps researchers succeed. eLife 2016;5:e16800. https://doi.org/10.7554/eLife.16800.
[22] Mirowski P. The future(s) of open science. Soc Stud Sci 2018;48:171–203. https://doi.org/10.1177/0306312718772086.
[23] Biopython · Biopython n.d. https://biopython.org/ (accessed March 26, 2025).
[24] Josef Perktold, Skipper Seabold, Kevin Sheppard, ChadFulton, Kerby Shedden, jbrockmendel, et al. statsmodels/statsmodels: Release 0.14.2 2024. https://doi.org/10.5281/ZENODO.593847.
[25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.
[26] The All of Us Research Program Investigators. The “All of Us” Research Program. N Engl J Med 2019;381:668–76. https://doi.org/10.1056/NEJMsr1809937.
[27] Candal-Pedreira C, Ross JS, Ruano-Ravina A, Egilman DS, Fernández E, Pérez-Ríos M. Retracted papers originating from paper mills: cross sectional study. BMJ 2022;379:e071517. https://doi.org/10.1136/bmj-2022-071517.
[28] Candal-Pedreira C, Guerra-Tort C, Ruano-Ravina A, Freijedo-Farinas F, Rey-Brandariz J, Ross JS, et al. Retracted papers originating from paper mills: a cross-sectional analysis of references and citations. Journal of Clinical Epidemiology 2024;172:111397. https://doi.org/10.1016/j.jclinepi.2024.111397.
[29] Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. Meta-Radiology 2023;1:100017. https://doi.org/10.1016/j.metrad.2023.100017.
[30] Patel CJ, Ioannidis JP, Manrai AK. The architecture of exposome-phenome associations 2025:2025.06.05.25329055. https://doi.org/10.1101/2025.06.05.25329055.
[31] Maupin D, Suchak T, Barnett A, Spick M. Dramatic increases in redundant publications in the Generative AI era. BMC Med 2025. https://doi.org/10.1186/s12916-025-04569-y.
[32] Van Noorden R. Exclusive: These universities have the most retracted scientific articles. Nature 2025;638:596–9. https://doi.org/10.1038/d41586-025-00455-y.
[33] Mallapaty S. China conducts first nationwide review of retractions and research misconduct. Nature 2024;626:700–1. https://doi.org/10.1038/d41586-024-00397-x.

[34] Vasconez-Gonzalez J, Izquierdo-Condoy JS, Naranjo-Lara P, Garcia-Bereguiain MÁ, Ortiz-Prado E. Integrity at stake: confronting “publish or perish” in the developing world and emerging economies. Front Med (Lausanne) 2024;11:1405424. https://doi.org/10.3389/fmed.2024.1405424.
[35] Mallapaty S. China’s supreme court calls for crack down on paper mills. Nature 2025;639:285–6. https://doi.org/10.1038/d41586-025-00612-3.
[36] Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. Int J Epidemiol 2016;45:1717–26. https://doi.org/10.1093/ije/dyx028.
[37] Ale L, Gentleman R, Sonmez TF, Sarkar D, Endres C. nhanesA: achieving transparency and reproducibility in NHANES research. Database 2024;2024:baae028. https://doi.org/10.1093/database/baae028.
[38] OpenVigil Pharmacovigilance Search Engines n.d. https://openvigil.sourceforge.net/ (accessed May 2, 2025).
[39] Golder S, Loke YK. Is there evidence for biased reporting of published adverse effects data in pharmaceutical industry-funded studies? Br J Clin Pharmacol 2008;66:767–73. https://doi.org/10.1111/j.1365-2125.2008.03272.x.
[40] 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 2018;16:336–8. https://doi.org/10.1016/j.cgh.2017.11.025.
[41] Spick M, Higgins J, Green CL, Matsouaka R, Shin DB, Hall RP, et al. Observations from Statistical Review Editors: A Commentary. JID Innovations 2024;4:100302. https://doi.org/10.1016/j.xjidi.2024.100302.
[42] Ortega JL. Classification and analysis of PubPeer comments: How a web journal club is used. Journal of the Association for Information Science and Technology 2022;73:655–70. https://doi.org/10.1002/asi.24568.
[43] 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 2023;32. https://doi.org/10.3145/epi.2023.ene.18.
[44] 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. https://doi.org/10.7189/jogh.15.01004.
[45] Parker L, Boughton S, Bero L, Byrne JA. Paper mill challenges: past, present, and future. Journal of Clinical Epidemiology 2024;176:111549. https://doi.org/10.1016/j.jclinepi.2024.111549.
[46] Jacobsen A, De Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al. FAIR Principles: Interpretations and Implementation Considerations. Data Intellegence 2020;2:10–29. https://doi.org/10.1162/dint_r_00024.
[47] Martin A, Newell B. Synthetic Data, Synthetic Media, and Surveillance. S&S 2024;22. https://doi.org/10.24908/ss.v22i4.18334.
[48] Journals and publishers crack down on research from open health data sets n.d. https://www.science.org/content/article/journals-and-publishers-crack-down-research-open-health-data-sets (accessed October 27, 2025).
[49] Taylor L. AI: Journals are automatically rejecting public health dataset papers to combat paper mills. BMJ 2025;391:r2170. https://doi.org/10.1136/bmj.r2170.
[50] Maupin D, Spick M, Geifman N. Safeguarding Open Science from exploitative practices. PLOS Medicine 2025;22:e1004851. https://doi.org/10.1371/journal.pmed.1004851.
[51] Lumbard H, Routledge D. Open science and transparency are our strongest tools in the fight against fraudulent publishing activities. PLOS Medicine 2025;22:e1004774. https://doi.org/10.1371/journal.pmed.1004774.

[52] Cabanac G, Labbé C, Magazinov A. Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. arXivOrg 2021. https://arxiv.org/abs/2107.06751v1 (accessed October 27, 2025).
[53] van Diest RA, Seifert R, van der Heyden MAG. An extra pair of eyes: adopting innovative approaches to detect integrity issues in Naunyn–Schmiedeberg's Archives of Pharmacology. Naunyn-Schmiedeberg's Arch Pharmacol 2025;398:1–8. https://doi.org/10.1007/s00210-024-03697-1.
[54] Sanderson K. Science's fake-paper problem: high-profile effort will tackle paper mills. Nature 2024;626:17–8. https://doi.org/10.1038/d41586-024-00159-9.
[55] Byrne JA, Abalkina A, Christopher J, Soulière MF. Rethinking Peer Review Using the Swiss Cheese Model to Better Flag Problematic Manuscripts. Learned Publishing 2025;38. https://doi.org/doi/10.1002/leap.2021.
[56] Maupin D, Suchak T, Barnett A, Spick M. Dramatic increases in redundant publications in the Generative AI era 2025:2025.09.09.25335401. https://doi.org/10.1101/2025.09.09.25335401.
[57] 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;637:1047–50. https://doi.org/10.1038/d41586-025-00212-1.
[58] Naddaf M. Journal targeted by paper mill still grappling with the aftermath years later. Nature 2025. https://doi.org/10.1038/d41586-025-01010-5.
[59] Tang G, Cai H. Citation Contamination by Paper Mill Articles in Systematic Reviews of the Life Sciences. JAMA Network Open 2025;8:e2515160. https://doi.org/10.1001/jamanetworkopen.2025.15160.
[60] Gross CP, Flanagin A, Perencevich EN, Inouye SK. Mitigating the Impact of Retracted Studies in the Medical Literature. JAMA Intern Med 2025;185:621–2. https://doi.org/10.1001/jamainternmed.2025.0251.