Prevalence of Problematic Papers in Non-Coding RNA Research

Ying Lou[1,2,+], Zhengyi Zhou[1,2,+], Zhesi Shen[1,2,*], Menghui Li[1,2]

1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China
2. Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

+ These authors contributed equally with the order determined alphabetically.

* Corresponding author: Zhesi Shen (Email: shenzhs@mail.las.ac.cn).

Abstract

This study examines the prevalence of problematic papers in the rapidly growing field of non-coding RNA (ncRNA) research. Analysis of 153,826 ncRNA papers during 2000-2023 reveals that around 1.79% have been retracted and an additional 5.68% have raised concerns on PubPeer. The number of problematic papers has steadily increased, peaking in 2019 when the concerning and retraction rates reached nearly 10.8% and 3.7%, respectively. These unreliable papers have been widely disseminated, accumulating hundreds of thousands of citations in academic literature, patents, clinical trials, and policy documents, posing a significant threat to research integrity and public health. The main issues identified include image manipulation, data falsification, fake peer reviews, and ethical lapses. The findings call for urgent, comprehensive scrutiny of ncRNA publications and broader reforms to address systemic problems driving the proliferation of problematic research.

The study underscores the prevalence of problematic non-coding RNA papers, calling for the critical necessity for heightened scrutiny to identify and address these issues.

Non-coding RNA (ncRNA) plays a crucial role as potent regulators of numerous biological functions in the detection and treatment of various diseases [1]. The study of ncRNA is thriving due to a large amount of under-investigated ncRNA [2], leading to a 'wild' growth in publications [3]. For example, the number of publications has risen from 102 in 2000 to 18,491 in 2021 (Figure 1A). Nevertheless, some under-investigated ncRNAs are being actively exploited for potentially fraudulent research [4]. For instance, over 8,700 ncRNA-related papers have been questioned on PubPeer, with more than 2,700 papers retracted, making ncRNA one of the most severely impacted areas by academic misconduct [5]. Furthermore, future research

may be misdirected by the problematic papers [4,6]. Even more concerning is that treatments based on problematic conclusions put patients at risk [6]. When many problematic papers emerge within a field, it doubts the credibility of all papers within that field. This research uncovered a significant prevalence of unreliable articles in the ncRNA field, which have been extensively circulated. Therefore, it is crucial and urgent to raise awareness about the severity of problematic papers in the ncRNA field and to advocate for stronger scrutiny of all publications.

The proportion of problematic papers is alarmingly high.

This analysis covers 153,826 papers on the topics of Micro & Long Noncoding RNA in the InCites database between 2000 and 2023 (Accessed on July 1st, 2024), considering only Articles, Reviews, and Retracted Publications. Among these papers, 2,753 papers have been retracted for various reasons, accounting for around 1.79% of the total publications. In addition, 8,739 papers have sparked concerns on PubPeer, constituting approximately 5.68% of the total. These papers have received more than 28,800 comments in PubPeer, almost all questioning their contents [7].

The number of comments on PubPeer surged from 749 in 2019 to 6,512 in 2021, consistently maintaining a high level of over 5,600 comments every year. During the same period, the number of retractions experienced a rapid increase, ultimately peaking at over 700 retractions by 2023 (Figure 1B). With the substantial increase in comments or retractions, it casts a shadow of doubt over the overall credibility of ncRNA papers.

From a historical perspective on publication years, the quantity of concerned papers and retracted papers appears intricately linked to the overall volume of publications. Specifically, the count of concerned or retracted papers ascended from 3 or 1 in 2002 to 1,751 or 600 in 2019, respectively (Figure 1C). Moreover, a consistent upward trend in the annual proportions of both concerned and retracted papers is observed. The peaks observed in 2019, with a concerning rate of 10.84% and a retraction rate of 3.71%, highlight a significant portion of papers being highly problematic (Figure 1D). These findings expose a troubling pattern of increasing proportions of questionable

research outputs over the years, highlighting the need for special attention to publications around 2019.

Unreliable knowledge has been widely disseminated, posing a threat to life and health, highlighting the critical need for urgent scrutiny.

The retracted papers have garnered over 85,800 citations in academic papers, more than 600 citations in patents, been cited in clinical trials 8 times, and referenced in policy documents 11 times (Figure 1E). In addition, the papers concerned on PubPeer have accumulated over 555,100 paper citations, more than 4,700 patent citations, over 80 clinical trial citations, and over 30 policy document citations (Figure 1E). For instance, a retracted paper[1] had amassed more than 2,000 paper citations and had been referenced in patents over 130 times. Additionally, it was cited in 4 clinical trials and mentioned in 3 policy documents in Dimensions (Accessed on August 26th, 2024). Similarly, another paper, concerned on PubPeer in 2019 for irreproducible results[2], had garnered over 4,500 citations and had been referenced in patents 52 times. These findings suggest that the dissemination of unreliable knowledge has been alarmingly extensive, spanning from basic research to clinical applications and policy formulation, posing a particularly dangerous threat to life and health.

Scrutiny can start with these characteristics.

What characteristics will help us to identify problematic papers? By analyzing and compiling the retraction reasons [8], we find the primary issues are related to image, data, peer review, text, and ethics (Figure 1F). In particular, image-related problems (including image reuse and falsification/fabrication) have emerged as the most common issue which affects 1,592 papers, representing a significant 57.8% of retractions. Meanwhile, within the papers commented on PubPeer, 7,307 are linked to image-related concerns, making up 83.6% of the total cases. Data-related concerns, e.g., data reuse and data falsification/fabrication, have affected 829 papers. Fake peer review, text reuse, and lack of ethical approval are the other three leading issues. Based on the above characteristics, post-publication scrutiny can commence with the

[1] 10.1016/j.ygyno.2008.04.033
[2] https://pubpeer.com/publications/9E5F06FC1E57F825881B4FC37956D4

help of AI-based tools like ImageTwin, ImaCheck, and Proofig to detect suspicious images [9], as well as STM Integrity Hub to identify Paper Mill [10], and iThenticate to check text similarity.

Conclusion and discussion

This study examines the prevalence of problematic papers during the flourishing research in the ncRNA field. Approximately 9.2% of papers published during 2017-2020 have been commented on PubPeer, with 3.04% of papers from the same period having been retracted. Moreover, the widespread dissemination of problematic papers not only undermines the research integrity but also misdirects the subsequent studies. The issues revealed in ncRNA are just the tip of the iceberg. Problematic papers have already permeated all fields [5].

It is generally believed that the prevalence of problematic papers is often linked to the pervasive 'publish or perish' pressure, highlighting a systematic problem within academia. Maintaining scientific integrity necessitates a pause to address this challenge strategically. All stakeholders bear the responsibility of thoroughly scrutinizing publications. While it is essential to prevent the publishing of new problematic papers, it is equally vital to scrutinize those already published, which can mitigate the influence of unreliable knowledge on both the academic community and the public.

Data and Methods

About 150,000 papers categorized under the meso-level Citation Topics -- Micro & Long Noncoding RNA (ncRNA) from the InCites database were retrieved on July 1, 2024, covering the publication period from 2000 to 2023. The dataset included articles, reviews, and retracted publications.

To identify problematic papers, we collected the retraction data from the Amend platform[8] and comments on PubPeer. The dissemination of problematic papers was assessed through citation analysis. The citations from papers, patents, policy

documents and clinical trials are accessed from Dimensions. Based on the retraction notice recorded in Amend, we compiled retraction reasons and categorized issues into image related, data related, peer review related, text related, and ethical related categories. The full workflow is illustrated in Fig.2. The dateset can be accessed from https://zenodo.org/doi/10.5281/zenodo.13383979.

Acknowledgments

This study is partially supported by the National Natural Science Foundation of China (grant no. 71974017) and the LIS Outstanding Talents Introducing Program, Bureau of Development and Planning, CAS (2022).

Statement

During the preparation of this work the author(s) used LLM in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Reference

[1] Nemeth, K, Bayraktar, R, Ferracin, M et al. Non-coding RNAs in disease: from mechanisms to therapeutics. Nat Rev Genet 25, 211-232 (2024).

[2] Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol. 16, e2006643 (2018).

[3] Mattick, J.S., Amaral, P.P., Carninci, P. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 24, 430-447 (2023).

[4] Byrne JA, Grima N, Capes-Davis A, Labbé C. The Possibility of Systematic Research Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions. Biomarker Insights, 2019 14 1-12 (2019). [5] Li M. and Shen Z. Science map of academic misconduct. The Innovation 5(2), 100593 (2024).

[6] van der Vet, PE, Nijveen, H. Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. Research Integrity and

Peer Review 1, 3 (2016).

[7] 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(5), 655-670 (2022).

[8] Li, M., Chen, F., Tong, S., Yang, L., Shen, Z. Amend: An integrated platform of retracted papers and concerned papers. Journal of Data and Information Science, 9(2), 41-55 (2024).

[9] Jones N. How scientific journals are fighting back against a torrent of questionable images. Nature, 626 697-698 (2024).

[10] Else H. Paper-Mill detector tested in push to stamp out fake science. Nature, 612 386 (2022).

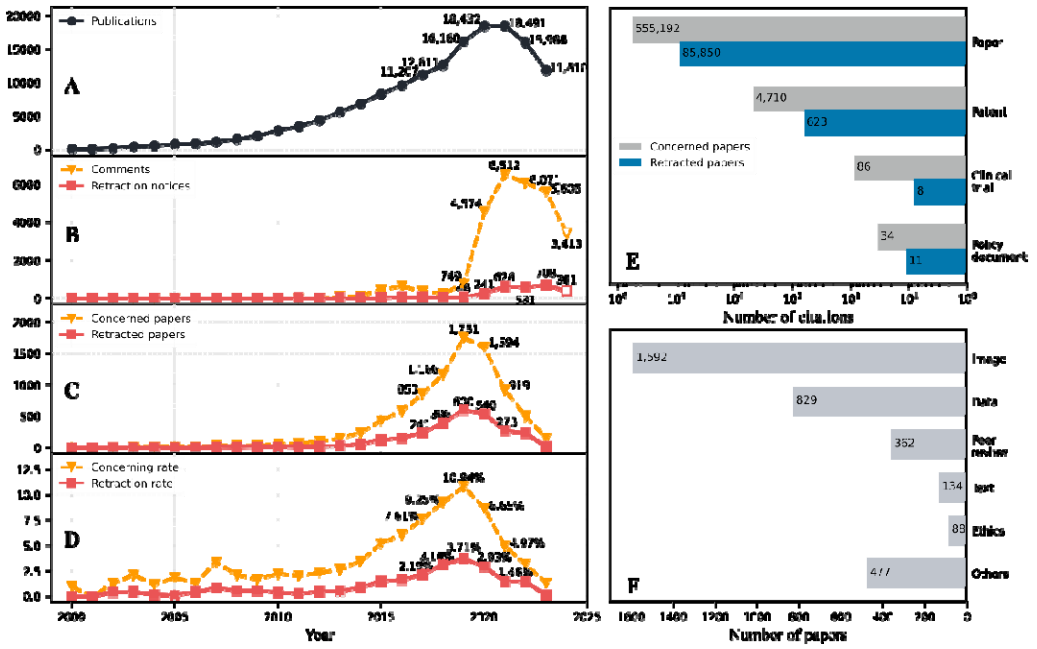


Figure 1: Characterizing the severity of problematic papers in the field of non-coding RNA. (A) The number of papers published each year. (B) The number of comments on PubPeer and retraction notices issued by journals each year. (C) The number of concerned papers and retracted papers for the year of publication. (D) The concerning rate and retraction rate for the year of publication. (E) The wide dissemination, as measured by the number of citations in papers, patents, clinical trials, and policy documents, for both the concerned papers and the retracted papers. (F) The top integrity issues, measured by the number of related retracted papers.

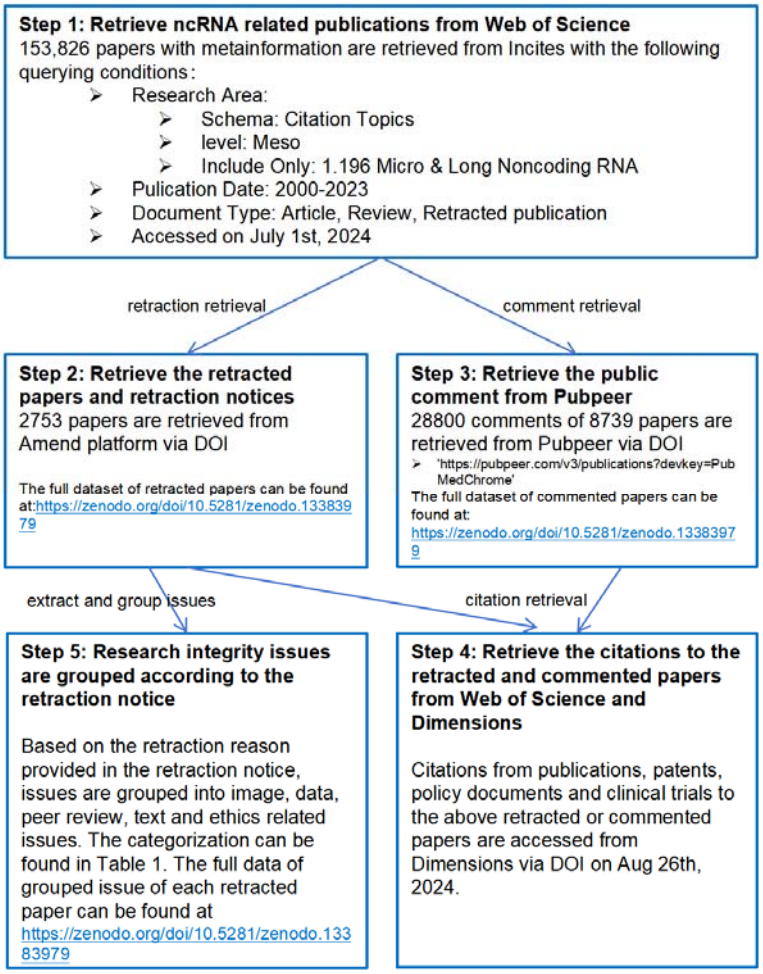
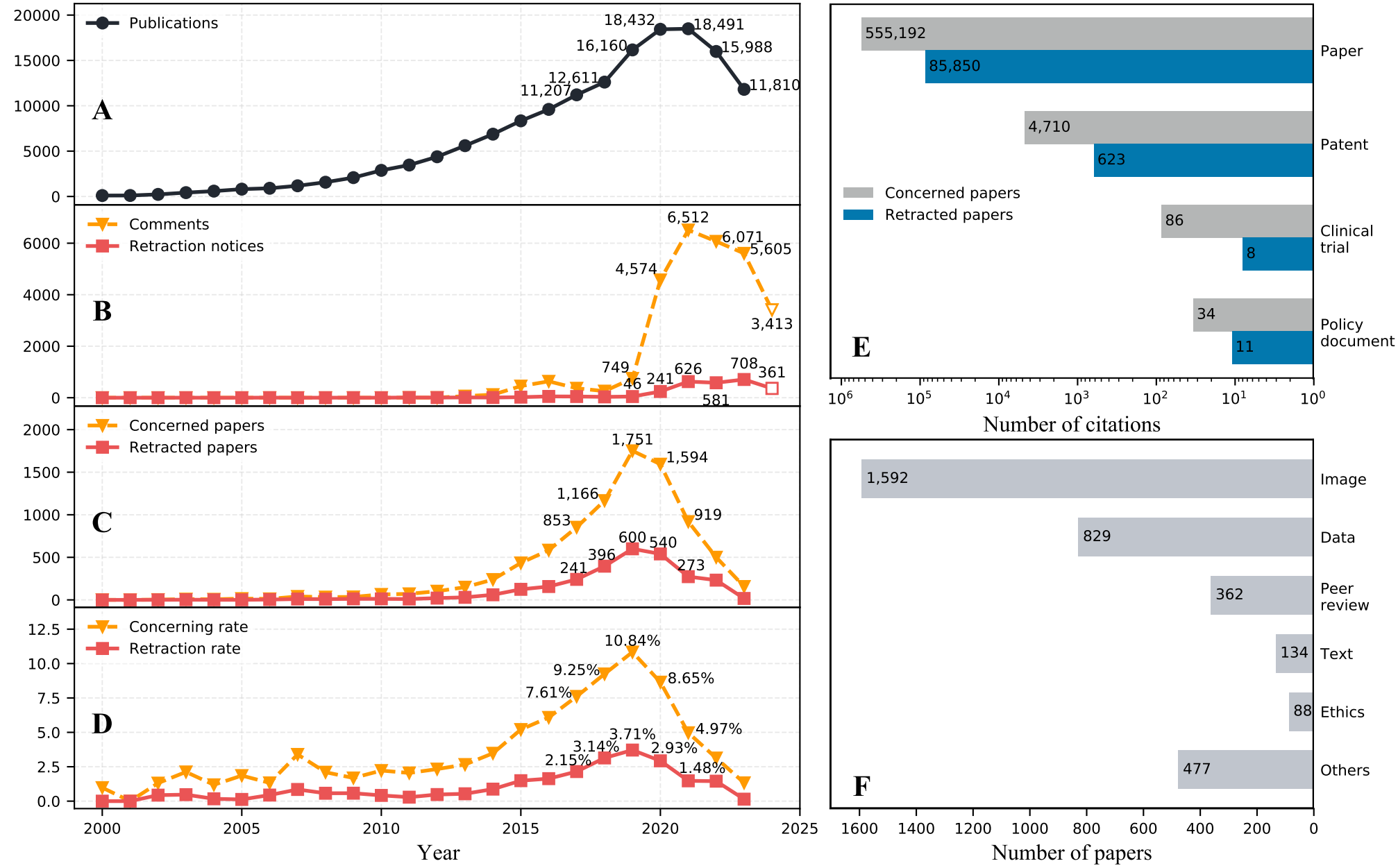


Figure 2. Workflow of data collection and processing.

Table 1. Grouped issues based on retraction reasons.

Issue Group	Retraction Reasons
Image	Image Falsification; Image Plagiarism; Image Fabrication; Image Duplication; Image Error
Data	Data Falsification; Data Plagiarism; Data Fabrication; Data Duplication; Data Error
Peer Review	Fake Peer Review
Text	Text Plagiarism; Text Duplication; Text Error
Ethics	Lack of IRB/IACUC Approval; Lack of Informed/Patient Consent; Privacy leak; Lack of trial registration
Others	Other reasons except the above reasons


A
Publications
18,432
18,491
16,160
15,988
12,611
11,207
11,810
B
Comments
Retraction notices
6,512
6,071
5,605
4,574
3,413
749
46
241
626
708
361
581
C
Concerned papers
Retracted papers
1,751
1,594
1,166
919
853
600
540
396
241
273
D
Concerning rate
Retraction rate
10.84%
9.25%
8.65%
7.61%
4.97%
3.71%
3.14%
2.93%
2.15%
1.48%
Year
E
555,192
85,850
Paper
4,710
623
Patent
86
8
Clinical trial
34
11
Policy document
Concerned papers
Retracted papers
Number of citations
F
1,592
Image
829
Data
362
Peer review
134
Text
88
Ethics
477
Others
Number of papers

Step 1: Retrieve ncRNA related publications from Web of Science

153,826 papers with metainformation are retrieved from Incites with the following querying conditions:

- Research Area:
  - Schema: Citation Topics
  - level: Meso
  - Include Only: 1.196 Micro & Long Noncoding RNA
- Pulication Date: 2000-2023
- Document Type: Article, Review, Retracted publication
- Accessed on July 1st, 2024

retraction retrieval

comment retrieval

Step 2: Retrieve the retracted papers and retraction notices

2753 papers are retrieved from Amend platform via DOI

The full dataset of retracted papers can be found at:https://zenodo.org/doi/10.5281/zenodo.13383979

Step 3: Retrieve the public comment from Pubpeer

28800 comments of 8739 papers are retrieved from Pubpeer via DOI

- 'https://pubpeer.com/v3/publications?devkey=PubMedChrome'

The full dataset of commented papers can be found at: https://zenodo.org/doi/10.5281/zenodo.13383979

extract and group issues

citation retrieval

Step 5: Research integrity issues are grouped according to the retraction notice

Based on the retraction reason provided in the retraction notice, issues are grouped into image, data, peer review, text and ethics related issues. The categorization can be found in Table 1. The full data of grouped issue of each retracted paper can be found at https://zenodo.org/doi/10.5281/zenodo.13383979

Step 4: Retrieve the citations to the retracted and commented papers from Web of Science and Dimensions

Citations from publications, patents, policy documents and clinical trials to the above retracted or commented papers are accessed from Dimensions via DOI on Aug 26th, 2024.