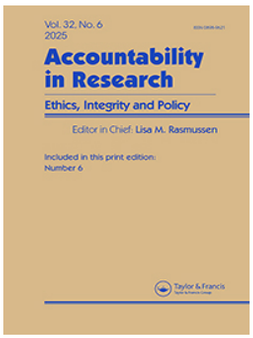

Taylor & Francis
Taylor & Francis Group

Accountability in Research
Ethics, Integrity and Policy

ISSN: 0898-9621 (Print) 1545-5815 (Online) Journal homepage: www.tandfonline.com/journals/gacr20

Analysis of scientific paper retractions due to data problems: Revealing challenges and countermeasures in data management

Wanfei Hu, Guiliang Yan, Jingyu Zhang, Zhenli Chen, Qing Qian & Sizhu Wu

To cite this article: Wanfei Hu, Guiliang Yan, Jingyu Zhang, Zhenli Chen, Qing Qian & Sizhu Wu (20 Jul 2025): Analysis of scientific paper retractions due to data problems: Revealing challenges and countermeasures in data management, Accountability in Research, DOI: 10.1080/08989621.2025.2531987

To link to this article: https://doi.org/10.1080/08989621.2025.2531987

Published online: 20 Jul 2025.

ACCOUNTABILITY IN RESEARCH, 2025
https://doi.org/10.1080/08989621.2025.2531987

Taylor & Francis
Taylor & Francis Group

Analysis of scientific paper retractions due to data problems: Revealing challenges and countermeasures in data management

Wanfei Hu, Guiliang Yan, Jingyu Zhang, Zhenli Chen, Qing Qian, and Sizhu Wu

Department of Medical Data Sharing, Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

ABSTRACT

Background: Scientific data, the cornerstone of scientific endeavors, face management challenges amid technological advances. While retractions are analyzed, a rigorous focus on data problems leading to them is missing.

Methods: This study collected 49,979 retraction records up to 17 December 2023. After screening 16,842 records were related to data problems and 19,656 were due to other reasons. Methods such as descriptive statistics, hypothesis testing, and the BERTopic (Bidirectional Encoder Representations from Transformers Topic Modelling) were applied to conduct a topic analysis of article titles.

Result: The results show that since 2000, retractions due to data problems have increased significantly ($p < 0.001$), with the percentage in 2023 exceeding 75%. Among 16,842 data-related retractions, 59.0% were in Basic Life Sciences and 40.2% in Health Sciences. Data problems involve accuracy, reliability, validity, and integrity. There are significant differences ($p < 0.001$) in subjects, journal quartiles, retraction intervals, and other characteristics between data-related and other retractions. Data-related retractions are more concentrated in high-impact journals (Q1 37.6% and Q2 43.0%).

Conclusions: Institutions, publishers, and journals should adopt image-screening tools, enforce data deposition, standardize retraction notices, provide ethics training, and strengthen peer review to address these data problems, guiding better data management and healthier scientific development.

ARTICLE HISTORY
Received 11 March 2025
Accepted 7 July 2025.

KEYWORDS
Data problems; retracted papers; retraction watch database; scientific data management

Introduction

The landscape of scientific research has been transformed by the rapid progress of information technology. The integration of data-driven and model-driven strategies, characteristic of the AI for Science (AI4S) paradigm (Wang and Miao 2023), has placed scientific data management at the

CONTACT Sizhu Wu wu.sizhu@imicams.ac.cn; Qing Qian qian.qing@imicams.ac.cn Department of Medical Data Sharing, Institute of Medical information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, 3 Yabao Road, Chaoyang District, Beijing 100020, China

forefront of global attention. Data are the cornerstone of scientific exploration, and effective, ethical data utilization is crucial (Hannigan et al. 2023; Kjelvik and Schultheis 2019). However, despite efforts by organizations like the US Office of Science and Technology Policy and UK Research and Innovation (The Office of Research Integrity 2000; UK Research and Innovation (2018), 2023) to establish policy frameworks, data-related issues continue to undermine scientific research integrity. High-profile cases such as the retraction of 31 articles by Professor Anversa due to data issues (Oransky and Marcus 2018), reports of widespread fabrication and plagiarism in neuroscience and medical research (Brainard 2023), and the 2024 suspected data fabrication in a Science retraction (Lee et al. 2024) highlight the severity of these problems. These incidents underscore the need for a comprehensive analysis of data-related issues leading to scientific paper retractions.

Previous studies on retracted papers have made contributions to understanding the characteristics of retractions. For example, research by Rubbo et al. (2019) on retraction patterns in the engineering filed found that higher impact factor journals tend to have more retractions. Candal-Pedreira et al. (2023) analyzed health sciences paper retractions in Brazil and Portugal and found that a large proportion of retracted articles were published in first and second quartile journals (ranked by impact factors) and most were in non-open access journals. Herrera-Añazco et al. (2024) studied retractions of health science articles by researchers in Latin America and the Caribbean and identified errors in procedures or data collection as the most common reason for retraction.

However, existing research has several limitations. Many studies are restricted to specific subjects or regions, resulting in small datasets that may not adequately represent the global phenomenon of retractions. For instance, Gedik, Kaya, and Kilci (2024) analyzed only 61 retracted emergency medicine articles, and Punreddy et al. (2024) focused on 77 retractions in plastic surgery and reconstruction. Moreover, most studies use descriptive statistical methods to analyze individual retraction characteristics in isolation, overlooking the correlations among different characteristics. Additionally, while some studies cover multiple levels of characteristics, they may not be comprehensive. Ferraro et al. (2023) only discussed the journal subject, and Candal-Pedreira et al. (2023) did not consider the retraction requestor, while A. Shi et al. (2024) believed that confirming author consensus is important.

Importantly, despite the importance of data in scientific research, no existing study has specifically focused on data-related problems as the core of retraction analysis. Shahraki-Mohammadi, Keikha, and Zahedi (2024) explored the relationship between retraction reasons and methodology quality in non-Cochrane retracted systematic reviews but did not focus on data issues. Retractions in the

field of oncology often involve issues with data and images (Qi et al. 2024; Yang, Sun, and Song 2024). The number of retractions in the field of molecular biology was on the rise, with data and image errors being the main reasons (Feng et al. 2024). Islam et al. (2025) found that data issues were one of the main reasons for retraction in Otolaryngology-Head & Neck Surgery literature. These studies all demonstrated the importance of data issues, but there has been no in-depth discussion on data issues.

Based on the previous research, we are intrigued by the specific manifestations and prevalence of improper scientific data use in retracted journal articles. This article seeks to address the following research questions:

RQ1: What are the trends in the number of retractions due to data problems, and how do data-related retractions change over this period?

RQ2: What are the differences between papers retracted due to data problems and those retracted for other reasons?

RQ3: What characteristics are associated with different reasons?

Material and methods

Drawing upon the relevant retraction studies conducted in recent years, we formulated our research framework, which encompasses three primary components: data collection, data filtration, data processing, and data analysis, as illustrated in Figure 1. The 9 characteristics listed in the "Analysis dimension" part of the figure were explained sequentially in the subsequent Analysis dimension section.

Data collection

We collected retraction data from the inception of Retraction Watch's database up to 17 December 2023, totaling 49,979 retraction records. For each record, we included title, journal, subjects, article type, date of retraction, reasons, number of authors, number of countries, and retraction notice of retracted papers from the Retraction Watch database (The Center for Scientific Integrity 2018). At the same time, we collected data about journal quartiles and the proportion of open-access articles for each journal from Journal Citation Reports (Clarivate 2024). Detailed descriptions of these characteristics were provided in the subsequent Analysis dimension section.

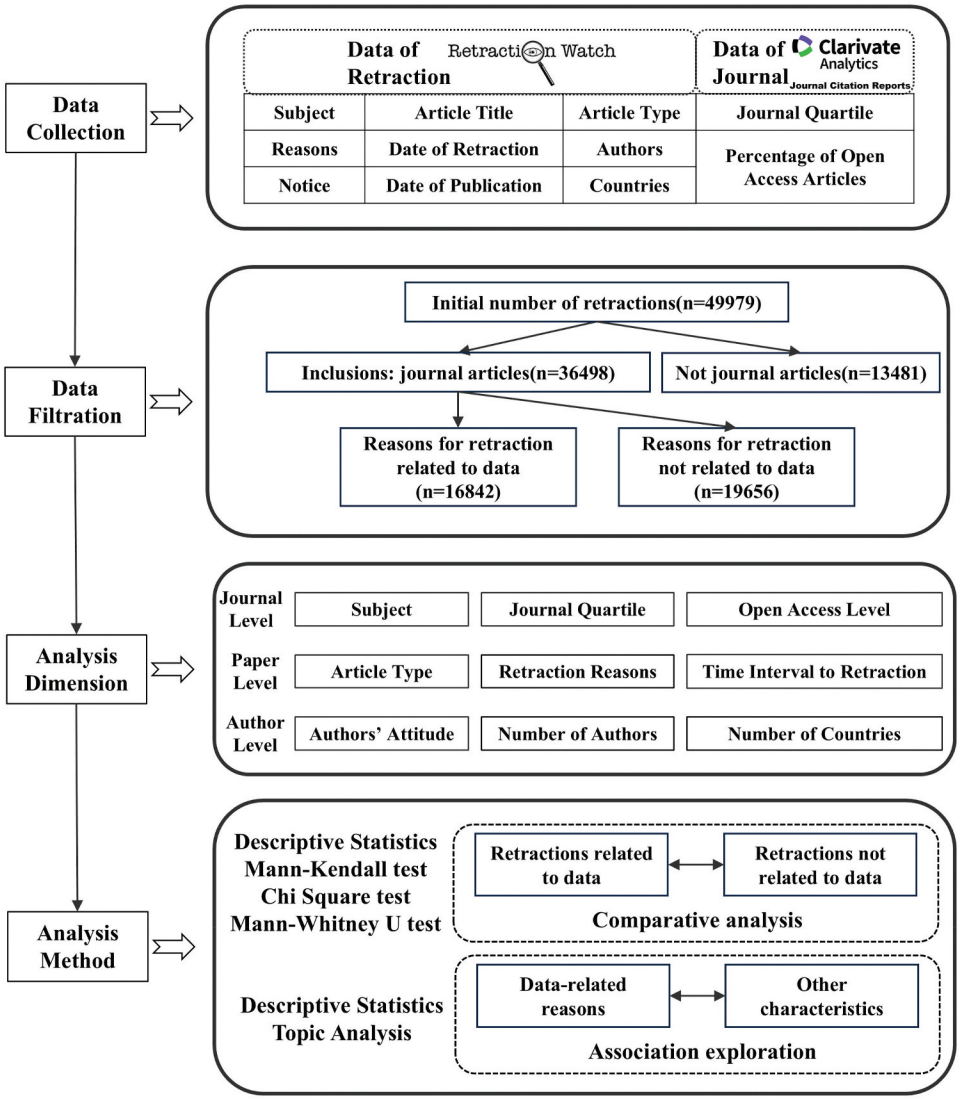

Figure 1. Flowchart of the research design in this study.

Data filtration

As shown in the data filtration section in Figure 1, the data inclusion and exclusion process of this study is outlined as follows: (1) Journal articles were included, while non-journal articles were excluded, resulting in the inclusion of 36,498 data entries. (2) Divide into two groups based on whether the reason for retraction is related to data problems 16,842 related and 19,656 unrelated. The categorization of data problems referred to the reasons provided in the Retraction Watch Database, with detailed situations presented in Table A2.

Analysis dimension

The raw data collected were processed to derive the characteristics of retracted articles. The analysis dimensions include journal level, paper level, and author level.

The characteristics at the journal level include subjects, journal quartiles, and open access level of journals. The subjects include business and technology, basic life sciences, environmental sciences, health sciences, humanities, physical sciences, and social sciences, which are derived from the original annotations in the Retraction Watch database. Journals are categorized into Q1, Q2, Q3, Q4, and No Impact Factor (No IF). Q1 to Q4 are quartile divisions by their impact factors, with Q1 (top 25%) the highest and Q4 (bottom 25%) the lowest in each subject category. The percentage of open access for each journal corresponding to each data entry was matched, ranging from 0% to 100%. To facilitate the statistical analysis, we classified the open access percentages into four categories: extremely low (0 ~ 10%), low (10%~50%), high (50%~90%), and extremely high (90%~100%).

The characteristics at the paper level include article types, reasons for retraction, and retraction time interval. Article types are categorized into one of the following four categories based on the pre-labeled results in the database: Research Article, Review Article, Clinical Research, and Other. The Retraction Watch Database labels the reasons for retraction based on the retraction notices. There are a total of 15 common reasons related to the data, detailed descriptions of these reasons can be found in Table A1. The time interval to retraction refers to the time interval between the publication of the paper and its retraction. In descriptive analysis, the time from publication to retraction is divided into five categories: within the 1st year, within the 2nd year, within the 3rd year, within the 4th year, and after the 4th year. In comparison of retraction due to data problems and other problems, time interval to retraction is converted into a continuous variable, which is the number of days from publication to retraction. This variable is more informative than categorical variables as it enables direct comparisons of the time lengths, offering a more nuanced understanding of the retraction process.

Author level characteristics include authors' attitudes, number of authors, and number of countries. Based on the description in the retraction notice text, the author's attitude toward the retraction is classified after keyword matching into six categories: voluntary retraction, agreement, disagreement, lack of consensus, no response or no statement, and not mentioned. In Table 1, they are labeled as "Request," "Agree," "Disagree," "Argue," "Not state," and "Not mention." The classification is done using Python version 3.11.5, and text segmentation is performed using the nltk package, version 3.8.1. The classification process is as follows: first, the text is segmented, then sentences containing keywords such as "Agree," "Disagree," "respond,"

Table 1. From sentence label to text label.

Sentence label			
Agree	Disagree	Not State	Text label
√	×	×	Agree
√	×	√	Agree
√	√	×	Argue
√	√	√	Argue
×	×	×	Not Mention
×	×	√	Not State
×	√	×	Disagree
×	√	√	Disagree

"reply," "accept," "retract," etc., are extracted. Each extracted key sentence is tagged with one of five sentence labels: voluntary retraction, agreement, disagreement, lack of stance/no response, not mention. The keywords used to label each sentence are listed in Table A1. These keywords were summarized during the review of the retraction notice. The matching method involves checking whether the corresponding keywords are present. After completing the matching, each retraction text generates a corresponding key sentence label sequence. The classification of each retraction text is based on whether this sequence contains the appropriate sentence label, as detailed in Table 1. A 10% random sample was drawn from all study-included data, and two independent researchers performed separate manual annotations. The annotation results showed high consistency (Cohen's Kappa = 0.91). Discrepant outcomes were adjudicated by a third researcher. The accuracy between the final manual annotations and automated labeling results was 0.95, leading to the direct adoption of the automated annotation outcomes.

Analysis method

Firstly, a descriptive analysis was performed on the retraction data. To determine the significance of the differences between the two groups (papers retracted due to data problems and those retracted for other reasons), a Chi-square test was employed for categorical variables. A p-value threshold of <0.001 was adopted to define statistical significance, aligning with prior research in retraction studies (Dal-Ré and Ayuso 2021; Lei et al. 2024). For the analysis of retraction time trends, especially the increase in retractions due to data problems since 2008, we applied the Mann–Kendall test for growth trends.

To compare the retraction time intervals (in days) of papers with data-related issues and those with non-data-related issues across various characteristics, a comparative analysis was carried out. Given the continuous nature of retraction time intervals, a normality test was first conducted, which revealed non-normality ($p < 0.001$). Consequently, the Mann-Whitney

U test was then used to determine if there were significant differences between the two groups for each characteristic. A p-value threshold of <0.001 was also adopted to define statistical significance.

To uncover prevalent themes among articles with data problems, topic analysis was performed on the article titles. The model for topic analysis is BERTopic, implemented using the Python package bertopic, version 0.16.0. The tokenization tool utilized is nltk, version 3.8.1. Bertopic represents an innovative Python library that ingeniously integrates pre-trained Transformer models with topic modeling methodologies. It has garnered widespread adoption and demonstrated commendable performance in the realm of document topic analysis (Guizzardi et al. 2023; Matsoukas et al. 2024; Raman et al. 2024).

Results

RQ1: What are the trends in the number of retractions due to data problems, and how do data-related retractions change over this period?

Retraction time trends

The first retraction due to data problems in Retraction Watch database occurred in 1967 (Retraction Watch 2014). The number of retractions before 2000 was relatively small. The annual total number of retractions and data problem retractions from 2000 to 2023 are shown in Figure 2. The overall

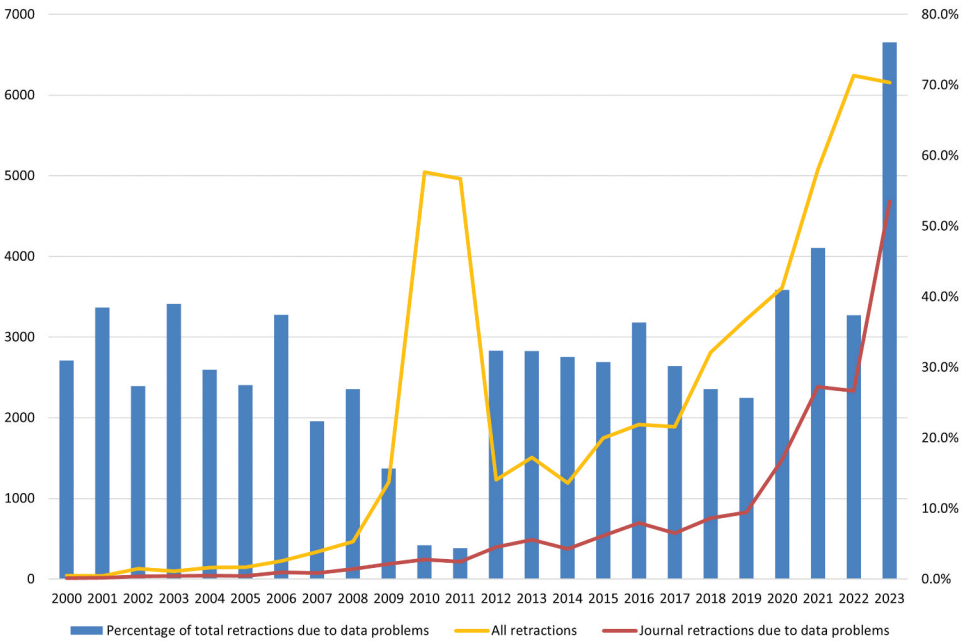

Figure 2. Changes in the number of total and data problem retractions, 2000–2023.

trend of retraction numbers peaked in 2010 and 2011, which was largely associated with the massive retractions of IEEE conference abstracts, but these retractions did not specify the reasons (McCook 2018). In addition, retractions due to data problems showed a noticeable growth trend since 2000 ($p < 0.001$), with the percentage in 2023 exceeding 75%.

Data problems

We summarized the relevant definitions in the Retraction Watch Database and the content of retraction notices and proposed data problems, corresponding reasons for retraction, amount and percentage of reasons, the description of reasons for retraction, and the main problems associated with them. The specific content is shown in Table A2.

Data problems refers to issues related to the generation, processing, presentation, or availability of data (including images, as a form of visual data) in academic publications that undermine the credibility, trustworthiness, or scientific validity of the research findings. These issues may arise from errors, intentional misconduct, or failures to adhere to scientific norms, ultimately affecting the reliability of the data as a basis for research conclusions. Specifically, in terms of retraction reasons, data problems include accuracy, reliability, validity, and integrity of the data. Table A3 provides detailed rationales for the inclusion of each reason, emphasizing that all these reasons are incorporated because they impact data accuracy, reliability, validity, or integrity to varying degrees. Accuracy refers to the degree to which data correctly reflects real-world facts, free from errors or distortions. Examples include numerical errors in data analysis or inconsistencies between images and findings. Reliability concerns the consistency and reproducibility of data, including whether sources and methods can be verified. Issues such as data duplication, falsification, or plagiarism compromise reliability. Validity measures whether data appropriately addresses the research question or objective. This includes design flaws, lack of ethical approval, or the use of randomly generated content. Integrity ensures data completeness and accessibility, such as the availability of original records. "Original Data not Provided" directly reflects integrity violations when core data cannot be retrieved or verified. The reasons description follows the definitions provided in the Retraction Watch Database, and the main problems are summarized based on the retraction notification texts.

RQ2: What are the differences between papers retracted due to data problems and those retracted for other reasons?

Main characteristics

Overall 16,842 retracted papers involved data problems, while 19,656 were retracted due to other problems. There were differences in subject ($p <$

0.001), journal quartile ($p < 0.001$), open access level of journal ($p < 0.001$), time interval to retraction ($p < 0.001$), article type ($p < 0.001$), authors' attitude ($p < 0.001$) and number of authors ($p < 0.001$) between the groups (Table A4). It should be noted that the subject percentages do not sum to 100% because each retracted paper can be classified into multiple subjects, leading to overlapping counts across categories.

Basic life sciences

The Basic Life Sciences stands out with the largest number of data-related retractions and the most significant disparity in the proportion of data-related and non-data-related retractions. In the Basic Life Sciences, a comparison between data-related ($n = 9934$) and non-data-related ($n = 6087$) retraction cases reveals significant differences in various aspects (Table A4). Data-related retractions are more concentrated in high-impact journals (Q1 37.6% and Q2 43.0%), journals with higher open-access levels, and research articles. They also tend to be retracted later, involve more authors, and have a higher proportion of authors agreeing to retraction (14.4%). Non-data-related retractions, on the other hand, show different distribution patterns, with a relatively higher proportion in journals without an impact factor and Q4, and more likely to be retracted within the first year.

When comparing retractions in the Basic Life Sciences with those of all subjects, several differences emerge and all disparities shown in Figure 3 are statistically significant ($p < 0.001$, chi-square test). In the Basic Life Sciences,

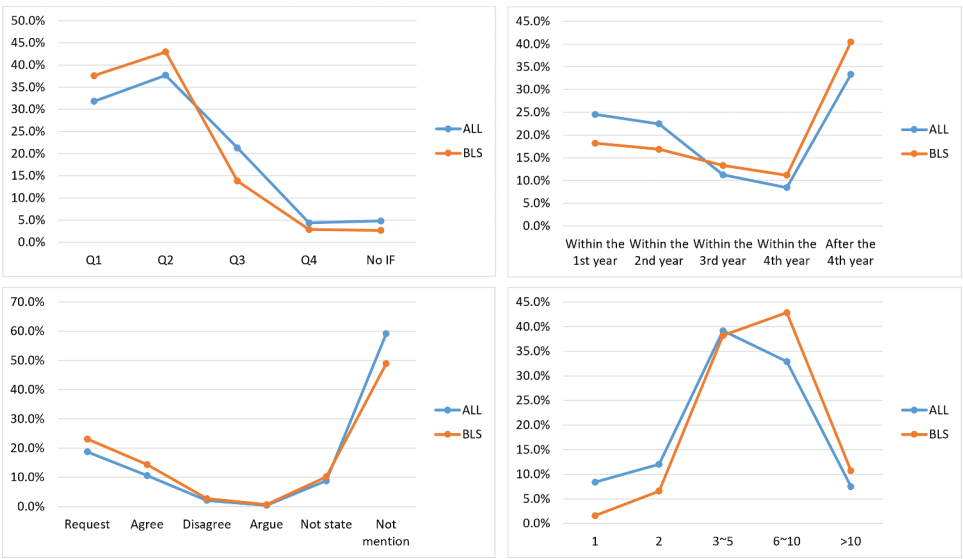

Figure 3. Comparison of retractions in the Basic Life Sciences (BLS) and all subjects (ALL) in terms of journal quartiles, time from publication to retraction, authors' attitudes, and number of authors.

the proportion of retractions in Q1 and Q2 journals is relatively higher. Retractions with a longer time interval are more common, and authors are more likely to actively request or agree to retractions. Additionally, articles with a larger number of authors are more prevalent in this discipline.

Days from publication to retraction

A comparative analysis was conducted on the time intervals to retract papers with data-related issues and those with non-data-related issues across various characteristics. The median number of days from publication to retraction and the results of the Mann–Whitney U test are presented in the following Table 2.

For characteristics like journal quartile, open-access level, authors' attitude, and the number of countries, papers with data problems generally had longer days to retraction. This implies that data-related issues may be more intricate or harder to detect in different journal contexts and papers with varying author-related factors. In Business and Technology, Environmental Sciences, Humanities, and Social Sciences, papers with data problems had shorter days to retraction, suggesting easier detection of data issues. In contrast, in Basic Life Sciences, Health Sciences, and Physical Sciences, such papers had longer intervals, indicating greater difficulty in uncovering data problems. Among article types, research articles, clinical studies, and other types with data problems had longer days to retraction. Single-author papers had shorter days to retraction, while multi-author papers had longer ones. Moreover, as the number of authors increased, the median days to retraction also grew.

RQ3: What characteristics are associated with different reasons?

Subject

In previous descriptions, it was found that data-related issues mainly concentrate on Basic Life Sciences (59%) and Health Sciences (40.2%). Figure 4 illustrates the distribution heatmap of reasons for retraction and subjects. Regarding the distribution of each reason across subjects, reasons can be classified into three categories. The first category mainly appears in Basic Life Sciences and Health Sciences, including Concerns About Data, Error in Data, Falsification of Data, Plagiarism of Data, and Plagiarism of Image. The second category is mainly found in Basic Life Sciences, covering Duplication of Data, Original Data not Provided, Unreliable Data, Concerns About Image, Duplication of Image, Error in Image, Falsification of Image, and Manipulation of Images. The third category is Randomly

Table 2. Median days from publication to retraction of papers with data-related and non-data-related problems across different characteristics.

	Days from publication to retraction		
	Data problem (*n*=16842)	Other problem (*n*=19656)	*p*-value
Subject			
Business and Technology	422	610	<0.0001**
Basic Life Sciences	1137.5	471	<0.0001**
Environmental Sciences	275	277.5	0.6962
Health Sciences	716	493	<0.0001**
Humanities	381	855	<0.0001**
Physical Sciences	632.5	450	<0.0001**
Social Sciences	430	647	<0.0001**
Journal Quartile			
Q1	1244	588	<0.0001**
Q2	815	514.5	<0.0001**
Q3	421	310	<0.0001**
Q4	730	338.5	<0.0001**
No IF	824	649	<0.0001**
Open Access Level of Journal			
Unknown	1339	623.5	<0.0001**
Extremely low	781	622	<0.0001**
Low	1393.5	385	<0.0001**
High	1575	394.5	<0.0001**
Extremely high	569	379.5	<0.0001**
Article Type			
Research Article	804	505	<0.0001**
Clinical Study	1303.5	513	<0.0001**
Review	431	756	0.000597
Other	577.5	325	0.005765
Authors' Attitude			
Request	921.5	339.5	<0.0001**
Agree	1188	439	<0.0001**
Disagree	1735	658	<0.0001**
Argue	1188	563	<0.0001**
Not state	1005	805	0.07244
Not mention	661.5	519	<0.0001**
Number of Authors			
1	383	561	<0.0001**
2	580	563	<0.0001**
3~5	833	510	<0.0001**
6~10	1069	451.5	<0.0001**
>10	1336	399	<0.0001**
Number of Countries			
1	812	531	<0.0001**
2	1036	460.5	<0.0001**
3	642	448	<0.0001**
4	563	329	<0.0001**
>4	407	293.5	0.01213

Generated Content, which is distinct from others, and retractions with this problem mainly occur in the Business and Technology discipline.

To further understand the themes focused on in retracted papers, a topic analysis was conducted in conjunction with an analysis of their

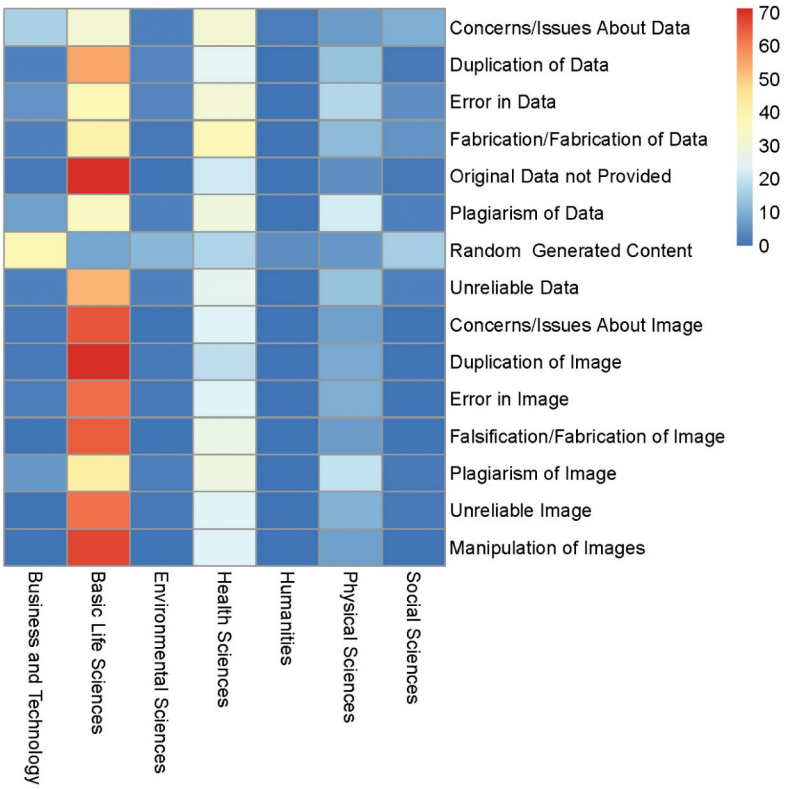

Figure 4. Heat map of the distribution of reasons for retraction and subjects.

respective disciplines. The methodology employed for thematic analysis involves conducting topic clustering on the titles of retracted papers, utilizing the BERTopic model, resulting in the creation of topic word clouds for each subject area, as illustrated in Figure A1.

Among the retractions in Business and Technology, the most prevalent themes include medical diagnosis and image processing research. In the realm of Basic Life Sciences, a significant proportion of retracted topics are related to cancer, specifically liver cancer, colorectal cancer, gastric cancer, and lung cancer, with additional emphasis on plant genetics research. Within Environmental Sciences, the dominant themes of retracted works focus on water treatment research, air quality monitoring and prediction studies, and geological hazard early warning research. In Health Sciences, the most frequently occurring themes relate to nursing quality and patient satisfaction research, obesity, and metabolism studies. As for the Humanities, the prominent topics in retracted papers encompass music education technology, and digitalization. In the Physical Sciences domain, material science emerges as the most prevalent theme, encompassing areas such as organic synthesis and catalysis, biomaterial synthesis, and hydrogen energy technologies. Additionally, quantum physics, biofuel research, and nanomedicine

studies are also notable topics. Lastly, in Social Sciences, the most common themes of retracted works involve AI-assisted teaching and machine translation.

Journal quartile

The distribution of each retraction reason across journal quartiles can be divided into four categories. As shown in Figure 5, the first category has the highest proportion of occurrences in Q1, such as Falsification of Image (51%) and Unreliable Image (55%). The second category is most prevalent in Q2, like Duplication of Image (47%) and Original Data not Provided (51%). The third category shows a similar distribution in Q1 and Q2, for example, Manipulation of Images (Q1 44% vs. Q2 44%). The fourth category has the highest proportion in Q3, with Randomly Generated Content (64%) being a typical case.

Open access level of journal

As shown in Figure 6, some retraction reasons are mainly associated with high-open-access journals, such as Duplication of Image (Extremely high 46%) and Concerns About Data (Extremely high 58%). In contrast, others are more common in low-open-access journals, like Falsification of Data (Low 40%) and Plagiarism of Data (Extremely low 38%). The open-access level of a journal may influence the visibility and accessibility of research, which in turn could affect the likelihood of detecting and reporting data-related problems.

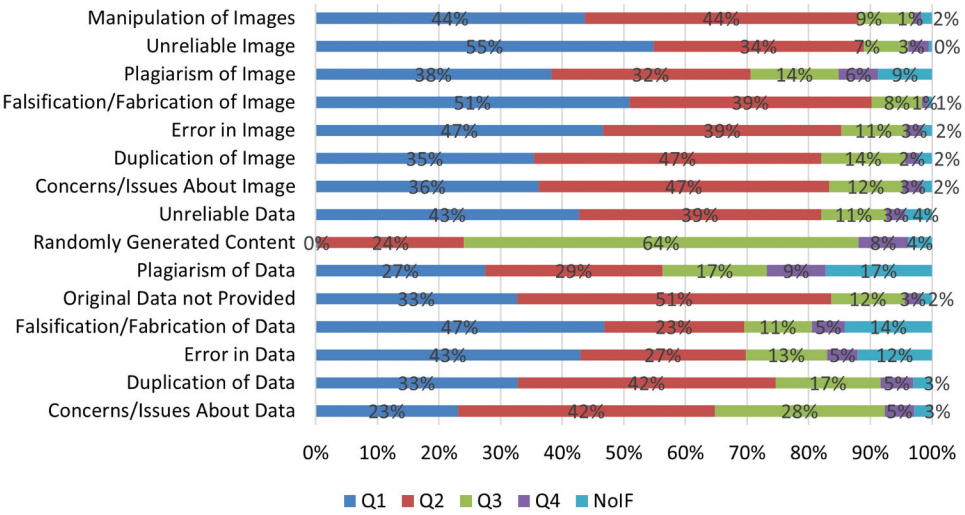

Figure 5. Distribution of reasons for retraction and journal quartiles.

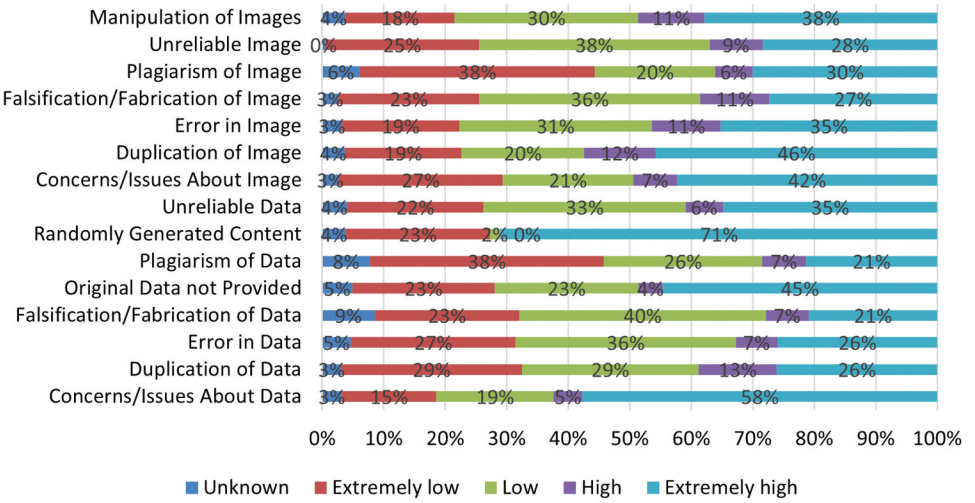

Figure 6. Distribution of reasons for retraction and journals' open access level.

Time from publication to retraction

By comparing median days from publication to retraction for different reasons, we identified the reasons associated with longer and shorter intervals (Table 3). Reasons for longer retraction time intervals mainly involve image-related issues, such as Manipulation of Images (1,790 days), Falsification of Image (1,709 days), and Unreliable Image (1,556.5 days). In contrast, reasons with shorter retraction time intervals include Plagiarism of Image (708 days), Concerns About Data (594.5 days), Error in Data (552.5 days), Plagiarism of Data (408 days), and Randomly Generated Content (371 days).

Table 3. The median days from publication to retraction for different reasons.

Reason for retraction	Median days from publication to retraction
Manipulation of Images	1790
Falsification/Fabrication of Image	1708
Falsification/Fabrication of Data	1561
Duplication of Image	1556.5
Unreliable Image	1414.5
Original Data not Provided	1278.5
Concerns/Issues About Image	1242
Duplication of Data	1227
Unreliable Data	1096
Error in Image	1050
Plagiarism of Image	708
Concerns/Issues About Data	594.5
Error in Data	552.5
Plagiarism of Data	408
Randomly Generated Content	371

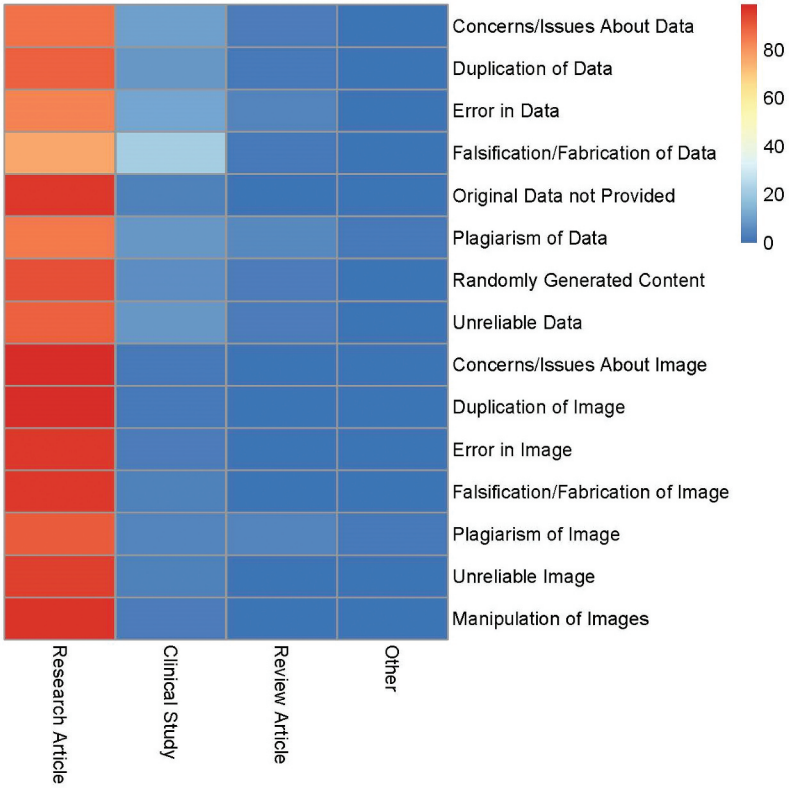

Figure 7. Heat map of the distribution of reasons for retraction and article types.

Article type

Figure 7 shows a heatmap of the distribution of article types and reasons for retraction. Research papers dominate in absolute quantity, with all reasons for retraction concentrated within this category. A comparison between the two images reveals that concerns about data and duplication of images are more prominent in research articles. In clinical studies, concerns about data and fabrication of data are highlighted, while in review papers, concerns about data and errors in data are more prominent, with errors in data standing out because when issues arise in review papers, they are categorized as errors in data.

Authors' attitude

There is a certain correlation between authors' attitudes and the reasons for retraction. Figure 8 illustrates the distribution heatmap of reasons for retraction and author attitudes. When authors retract voluntarily, the reasons with higher proportions include errors in data and origin data not provided. When authors agree to retract, the reasons are original data not provided, unreliable data, and concerns about image. For

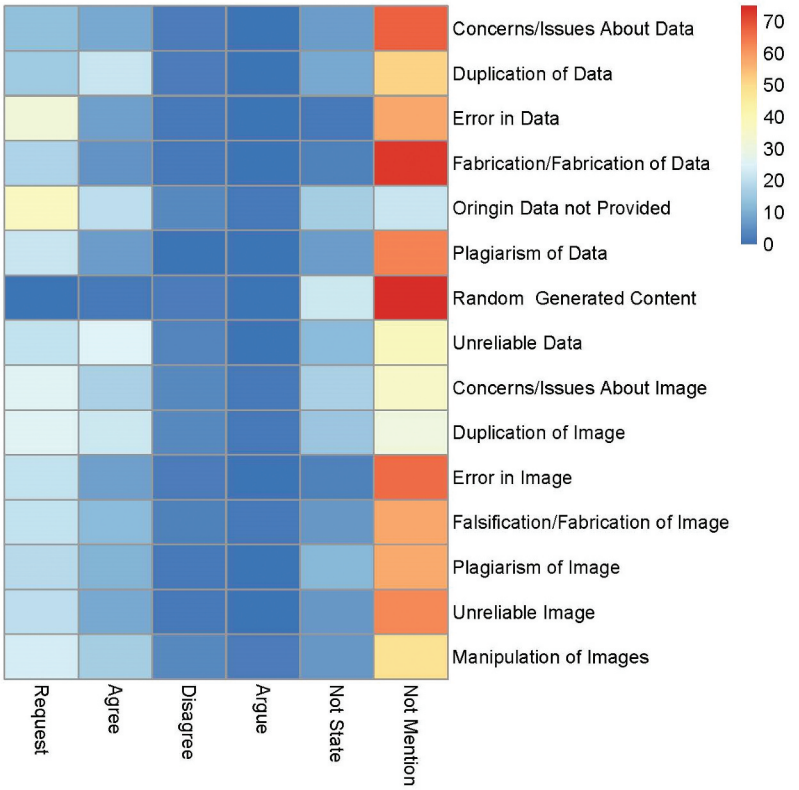

Figure 8. Heat map of the distribution of reasons for retraction and authors' attitudes.

disagreements, more frequent reasons involve origin data not being provided and concerns about image. When authors fail to reach a consensus, more frequent reasons include original data not being provided, concerns about image, and image manipulation.

Number of authors

Randomly Generated Content is a unique case, with a considerably high proportion of single-author papers (Figure 9). This could imply that in single-author research, there may be a higher risk of such issues, perhaps due to the lack of collaborative review and oversight that exists in multi-author projects. In multi-author papers, the distribution of retraction reasons varies with the number of authors. As the number of authors increases, the proportion of certain reasons, such as Concerns About Data and Duplication of Data, also changes. This indicates that the complexity of research collaboration may affect the likelihood and type of data-related problems that occur.

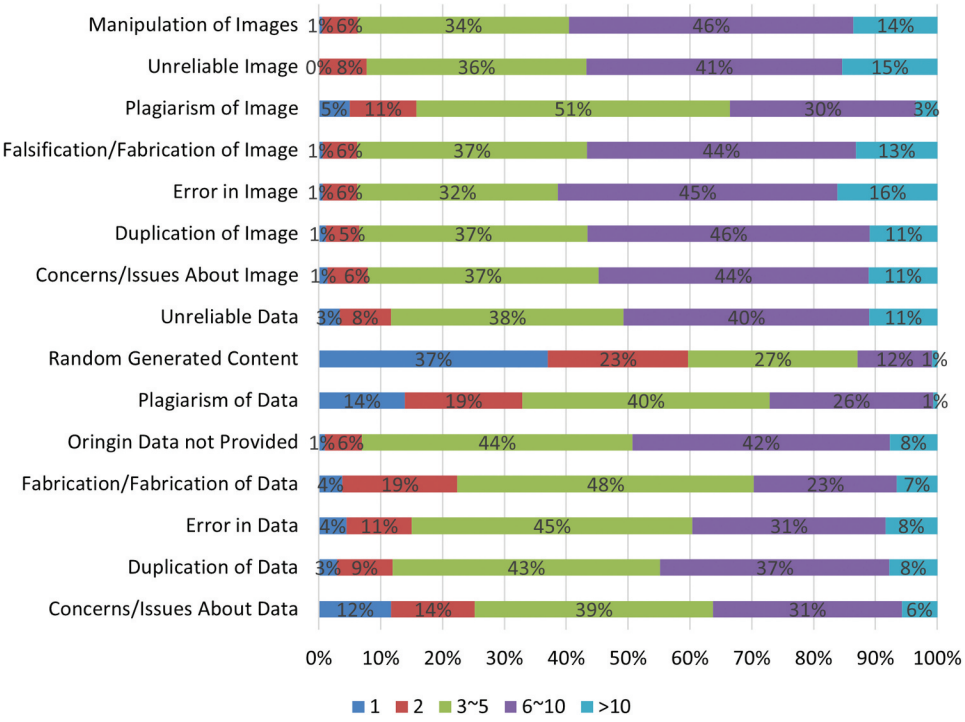

Figure 9. Distribution of reasons for retraction and number of authors.

Discussion and conclusion

Over the past decades, the landscape of scientific research has witnessed a notable increase in retractions, particularly those linked to data problems. This research analyzed scientific paper retractions caused by data issues. Since 2000, the proportion of data-related retractions has been on an upward trend ($p < 0.001$), reaching over 75% of all retractions in 2023. This upward shift may coincide with the implementation of enhanced detection measures, such as Springer Nature's Crossref Similarity Check (Springer 2024) and SAGE's research integrity team (SAGE 2022). The trend of voluntary retractions has become more prominent after 2014, and other types of author attitudes have also increased since 2017. These trends align with the implementation of targeted training and educational programs, which may reflect broader changes in research integrity practices. Data problems involve accuracy, reliability, validity, and integrity. About 47.5% of retracted papers have data-related concerns. Comparisons between data-related and other retractions show significant differences ($p < 0.001$) in various aspects. Data-related retractions concentrate on Basic Life Sciences and Health Sciences, are common in Q2 and high-open-access journals, and often occur 4 years after publication. Analysis of retraction time intervals shows that data-related retractions vary by subject, journal, article type, and author-related

characteristics. Regarding retraction reasons, they vary across different characteristics, which is the key content to be discussed next.

Manipulation of Images and Falsification of Image represent serious threats to scientific integrity. For example, Madhugiri, Nagella, and Uppar (2021) conducted a statistical analysis of retracted articles in neurosurgery and showed that the basic science category, more collaborating departments, and the H-index of the journal were associated with a longer time to retraction. Our research also indicated that papers with data issues often have long retraction time intervals. Journals need to enhance their peer-review processes by incorporating advanced image-forensic techniques. These techniques can analyze the metadata and pixel-level details of images to identify signs of manipulation (Candal-Pedreira et al. 2023). Additionally, researchers should be educated about the ethical implications of such actions. Institutions should offer training programs that emphasize the importance of image integrity in scientific research.

Falsification of Data is a fundamental violation of scientific integrity. Our results show that this problem significantly undermines the credibility of research. Previous research has pointed out that the long retraction times associated with data fabrication cases are a common challenge in uncovering this type of misconduct (Fang, Steen, and Casadevall 2012). Even papers by highly cited researchers may have instances of fake peer review (Kamali, Rahimi, and Talebi Bezmin Abadi 2022). Peer reviewers play a crucial role in detecting fabricated data. They should be trained to look for statistical anomalies, inconsistent data trends, and the absence of proper data collection protocols.

The issues of Duplication of Image and Duplication of Data, are becoming increasingly concerning. Our study shows that these problems are more prevalent in certain journal types, especially those with a high open-access level in some fields. Previous research has also pointed out the significance of this problem in scientific research. For example, in a study by Candal-Pedreira et al. (2023), it was found that in some fields, the pressure to publish and the lack of clear guidelines on data and image reuse contributed to the occurrence of such self-plagiarism. Shah et al. (2021) revealed that the likelihood of retractions in open-access articles is 62% higher than in toll-access articles, with the most prominent subjects being within the scope of basic life sciences and health sciences. Zheng, Fang, and Fu (2024) found that gold open access is advantageous in reducing the retraction time of flawed articles. This indicates that gold open access may help expedite the detection and retraction of flawed articles, ultimately promoting the practice of responsible research.

Previous studies have explored the impact of unreliable data and images in different scientific fields. For example, in a study of medical research

retractions, Khademizadeh et al. (2023) found that unreliable data was a common reason for retraction, and it often led to incorrect conclusions and potentially harmful medical decisions. Q. Shi et al. (2021), focusing on non-Cochrane systematic reviews, identified unreliable data, which meant errors in research design or data analysis, as the second most common reason for retraction. More specifically, we found papers that were retracted due to unreliable data, and images appeared more frequently in Q1, especially images. Q1 journals have a higher academic influence and therefore deserve more attention. Journals can play a part by requiring authors to provide detailed information about the data collection and image-acquisition processes, as well as any steps taken to ensure their reliability.

Our results show that lacking of original data is widespread across various fields. In basic life sciences, where experiments often need to be replicated to confirm findings, the absence of original data can prevent other researchers from validating results. This not only hinders the progress of scientific research but also undermines the trustworthiness of the entire scientific community (Fanelli et al. 2015). Research administrators should encourage and enforce the proper storage and sharing of original data. This can be achieved by providing incentives for researchers to deposit their data in publicly accessible repositories, such as offering additional research funding or recognition.

Error in Image and Error in Data, although often unintentional, can still have a significant impact on the accuracy of research. Herrera-Añazco et al. (2024) found that errors in procedures or data collection (26.5%) were the most common reason for retraction among health science articles written by individuals affiliated with academic institutions in Latin America and the Caribbean. These errors may be related to whether there is cross-border cooperation. International cooperation withdrawal shows an upward trend from 2017 to 2023 (Sharma 2024). Rossouw, Matsau, and van Zyl (2020) noted that retractions involving international collaboration were less likely to be attributed to plagiarism or data errors. Compared to other data issues, we further found that the proportion of papers that were retracted due to Error in Image was highest in terms of cross-border collaboration.

The number of retractions due to randomly generated content is on the rise. In our research, all retractions involving randomly generated content were concentrated between 2020 and 2023. This is mainly because the rapid development of generative AI, especially large language models like ChatGPT, has provided researchers with “convenience” for improper use. AI tools are being misused in paper writing. Some authors use them to generate content without proper verification, leading to papers with fabricated data, non-existent references, and inconsistent logic (Kendall and Teixeira da Silva 2024). Therefore, it is urgent to establish guidelines for the responsible and transparent use of AI tools and implement disciplinary

measures (Lei et al. 2024). A study has found that teams of 3–5 people have the highest retraction rate (42.3%) (Sharma 2021), while randomly generated content shows significant differences, with the highest proportion of single-person writing (37.0%).

To address data manipulation, falsification, and related issues, institutions, publishers, and journals should implement several key measures. First, mandate the use of image-screening software such as Crossref Similarity Check and AI detection tools to identify manipulated content or AI-generated text, particularly in high-risk fields like business and technology. Second, enforce policies that require authors to deposit original data, codes, and images in public repositories at the time of submission to ensure data integrity. Third, standardize retraction notices to include author attitudes and investigation details for transparency, and prioritize early retraction within four years to minimize post-retraction citations. Fourth, provide mandatory research ethics training on data management and responsible AI use, tailoring programs to disciplines such as life sciences for proper data preservation and social sciences for avoiding AI misuse. Finally, strengthen peer review by establishing specialized teams in high-risk fields and increasing oversight of lower-impact journals in humanities and environmental sciences.

Our study is observational in nature, as we analyzed existing retraction records without manipulating variables or controlling for external factors. This design limits our ability to establish causal relationships between data problems and the observed differences. Several confounding factors may influence our findings. Subjects like Basic Life Sciences and Health Sciences have stricter data-sharing requirements and higher scrutiny of experimental reproducibility, which may lead to more frequent detection and reporting of data problems, contributing to their overrepresentation in data-related retractions. In contrast, disciplines such as Humanities often rely on qualitative data, where "data problems" are less commonly defined or reported, resulting in lower proportions. Multi-author papers (6–10 authors or more) showed longer retraction intervals for data problems. This may reflect the complexity of coordinating data management in large teams, increasing the risk of errors or delays in detecting misconduct. Conversely, single-author papers with randomly generated content may face less internal oversight, accelerating both misconduct and its detection. Journals with high impact factors (Q1/Q2) had a higher proportion of data-related retractions, partly due to their larger readership and stricter post-publication scrutiny. These journals may also attract more high-risk research, increasing the likelihood of data irregularities. Additionally, open-access journals (especially those with extremely high open-access rates) showed more data duplication, potentially linked to pressure to publish quickly in competitive environments. Moreover, variations in retraction protocols (e.g., image-screening tools, data deposition

requirements) across publishers may affect the detection and reporting of data problems.

This study has certain limitations. Firstly, the annotations of retraction reasons are sourced from the Retraction Watch Database. While more information can be obtained from the original articles, some of these articles are not accessible, limiting the possibility of a more in-depth analysis. Secondly, a significant proportion of retraction notices fail to mention the authors' attitudes, and some notices, due to formatting issues, have not been converted into text format for data entry. More importantly, the reliability and completeness of retraction reasons reported in notices may be compromised. Consequently, journals and publishers should standardize the publication of retraction notices more systematically, so as to enable more in-depth and accurate analyses. The current study measures the degree of open access for journals based on data from the Journal Citation Reports, which can be further refined and improved upon if more authoritative standards emerge in the future. Moving forward, research can expand beyond the Retraction Watch Database by incorporating more data sources or leveraging advanced text mining and thematic analysis methods. Furthermore, the number of retractions has increased substantially in recent years. With the continuous update and iteration of AI, future data problems may be more severe. Newer AI versions can generate more sophisticated false content, making detection more difficult. Therefore, it is crucial to dynamically track the trends and changes in the characteristics of these retractions to formulate more timely response strategies.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded by the National Science and Technology Major Project for the Prevention and Treatment of Cancer, Cardiovascular, Respiratory, and Metabolic Diseases [2023ZD0509702] and Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences Program [2021-I2M-1-057].

CRediT authorship contribution statement

Wanfei Hu: Methodology, Writing-Original Draft. **Guiliang Yan**: Investigation. **Jingyu Zhang**: Data Curation, Visualization. **Zhenli Chen**: Data curation, Visualization. **Qing Qian**: Supervision, Writing- Reviewing and Editing. **Sizhu Wu**: Conceptualization, Writing- Reviewing and Editing

Data availability statement

The data is freely available via GitLab at https://gitlab.com/crossref/retraction-watch-data, which is updated daily. Users are required to cite the database in the format specified in the Retraction Watch Database User Guide (https://retractionwatch.com/retraction-watch-database-user-guide/).

References

Brainard, J. 2023. "New Tools Show Promise for Tackling Paper Mills." *Science* 380 (6645): 568–569. https://doi.org/10.1126/science.adi6513 .

Candal-Pedreira, C., A. Ruano-Ravina, J. Rey-Brandariz, N. Mourino, S. Ravara, P. Aguiar, and M. Pérez-Ríos. 2023. "Evolution and Characterization of Health Sciences Paper Retractions in Brazil and Portugal." *Accountability in Research* 30 (8): 725–742. https://doi.org/10.1080/08989621.2022.2080549 .

The Center for Scientific Integrity. 2018. *Retraction Watch Database*. http://retractiondatabase.org/ .

Clarivate. 2024. Journal Citation Reports. https://jcr.clarivate.com .

Dal-Ré, R., and C. Ayuso. 2021. "For How Long and with What Relevance Do Genetics Articles Retracted Due to Research Misconduct Remain Active in the Scientific Literature." *Accountability in Research* 28 (5): 280–296. https://doi.org/10.1080/08989621.2020.1835479 .

Fanelli, D., R. Costas, V. Larivière, and K. B. Wray. 2015. "Misconduct Policies, Academic Culture and Career Stage, Not Gender or Pressures to Publish, Affect Scientific Integrity." *PLoS One* 10 (6): e0127556. https://doi.org/10.1371/journal.pone.0127556 .

Fang, F. C., R. G. Steen, and A. Casadevall. 2012. "Misconduct Accounts for the Majority of Retracted Scientific Publications." *Proceedings of the National Academy of Sciences of the United States of America* 109 (42): 17028–17033. https://doi.org/10.1073/pnas.1212247109.

Feng, S., L. Feng, F. Han, Y. Zhang, Y. Ren, L. Wang, and J. Yuan. 2024. "Citation Network Analysis of Retractions in Molecular Biology Field." *Scientometrics* 129 (8): 4795–4817. https://doi.org/10.1007/s11192-024-05101-4 .

Ferraro, M. C., R. A. Moore, A. C. de C Williams, E. Fisher, G. Stewart, M. C. Ferguson, C. Eccleston, and N. E. O'Connell. 2023. "Characteristics of Retracted Publications Related to Pain Research: A Systematic Review." *PAIN* 164 (11): 2397. https://doi.org/10.1097/j.pain.0000000000002947 .

Gedik, M. S., E. Kaya, and A. İ. Kilci. 2024. "Evaluation of Retracted Articles in the Field of Emergency Medicine on the Web of Science Database." *The American Journal of Emergency Medicine* 82:68–74. https://doi.org/10.1016/j.ajem.2024.05.016 .

Guizzardi, S., M. T. Colangelo, P. Mirandola, and C. Galli. 2023. "Modeling New Trends in Bone regeneration, Using the BERTopic Approach." *Regenerative Medicine* 18 (9): 719–734. https://doi.org/10.2217/rme-2023-0096 .

Hannigan, A., F. Garry, C. Byrne, H. Phelan, and E. Garcia-Pelegrin. 2023. "The Role of the Arts in Enhancing Data Literacy: A Scoping Review Protocol." *PLoS One* 18 (2): e0281749. https://doi.org/10.1371/journal.pone.0281749 .

Herrera-Añazco, P., D. Fernandez-Guzman, F. Barriga-Chambi, J. K. Benites-Meza, B. Caira-Chuquineyra, and V. A. Benites-Zapata. 2024. "Retraction of Health Science Articles by Researchers in Latin America and the Caribbean: A Scoping Review." *Developing World Bioethics* 25 (1): 5–15. https://doi.org/10.1111/dewb.12439 .

Islam, A. S., E. M. Mastoloni, J. E. Fenton, and D. H. Coelho. 2025. "Article Retraction in Otolaryngology Journals: A Thirty Year Analysis." *Clinical Otolaryngology* (0): 1–7. https://doi.org/10.1111/coa.14285 .

Kamali, N., F. Rahimi, and A. Talebi Bezmin Abadi. 2022. "Learning from Retracted Papers Authored by the Highly Cited iran-Affiliated Researchers: Revisiting Research Policies and a Key Message to Clarivate Analytics." *Science and Engineering Ethics* 28 (2): 18. https://doi.org/10.1007/s11948-022-00368-3 .

Kendall, G., and J. A. Teixeira da Silva. 2024. "Risks of Abuse of Large Language Models, Like chatgpt, in Scientific Publishing: Authorship, Predatory Publishing, and Paper Mills." *Learned Publishing* 37 (1): 55–62. https://doi.org/10.1002/leap.1578 .

Khademizadeh, S., F. Danesh, S. Esmaeili, B. Lund, and K. Santos-d'Amorim. 2023. "Evolution of Retracted Publications in the Medical Sciences: Citations analysis, Bibliometrics, and Altmetrics Trends." *Accountability in Research* (0): 1–16. https://doi.org/10.1080/08989621.2023.2223996 .

Kjelvik, M. K., and E. H. Schultheis. 2019. "Getting Messy with Authentic Data: Exploring the Potential of Using Data from Scientific Research to Support Student Data Literacy." *cbe-Life Sciences Education* 18 (2): es2. https://doi.org/10.1187/cbe.18-02-0023 .

Lee, A.-H., G. S. Brandt, N. N. Iwakoshi, A. Schinzel, and L. H. Glimcher. 2024. "Retraction." *Science* 384 (6693): 280–280. https://doi.org/10.1126/science.adp1104 .

Lei, F., L. Du, M. Dong, and X. Liu. 2024. "Global Retractions Due to Randomly Generated Content: Characterization and Trends." *Scientometrics* 129 (12): 7943–7958. https://doi.org/10.1007/s11192-024-05172-3 .

Madhugiri, V. S., A. B. Nagella, and A. M. Uppar. 2021. "An Analysis of Retractions in Neurosurgery and Allied Clinical and Basic Science Specialties." *Acta Neurochirurgica* 163 (1): 19–30. https://doi.org/10.1007/s00701-020-04615-z .

Matsoukas, S., C. M. Zipser, F. Zipser-Mohammadzada, N. Kheram, A. Boraschi, Z. Jiang, L. Tetreault, M. G. Fehlings, B. M. Davies, and K. Margetis. 2024. "Scoping Review with Topic Modeling on the Diagnostic Criteria for Degenerative Cervical Myelopathy." *Global Spine Journal* 14 (7): 2155–2169. https://doi.org/10.1177/21925682241237469 .

McCook, A. 2018. "One Publisher, More Than 7000 Retractions." *Science* 362 (6413): 393–393. https://doi.org/10.1126/science.362.6413.393 .

The Office of Research Integrity. 2000. *Federal Research Misconduct Policy | Ori - the Office of Research Integrity*. https://ori.hhs.gov/federal-research-misconduct-policy .

Oransky, I., and A. Marcus. 2018. *Harvard and the Brigham Call for 31 Retractions of Cardiac Stem Cell research—STAT*. https://www.statnews.com/2018/10/14/harvard-brigham-retractions-stem-cell/ .

Punreddy, A., P. G. Guirguis, M. Youssef, and M. Botros. 2024. "Current Trends in Retraction of Plastic Surgery and Reconstruction Research." *Journal of Plastic, Reconstructive & Aesthetic Surgery* 93:136–139. https://doi.org/10.1016/j.bjps.2024.04.055 .

Qi, Q., J. Huang, Y. Wu, Y. Pan, J. Zhuang, and X. Yang. 2024. "Recent Trends: Retractions of Articles in the Oncology Field." *Heliyon* 10 (12): e33007. https://doi.org/10.1016/j.heliyon.2024.e33007 .

Raman, R., D. Pattnaik, H. H. Lathabai, C. Kumar, K. Govindan, and P. Nedungadi. 2024. "Green and Sustainable AI Research: An Integrated Thematic and Topic Modeling Analysis." *Journal of Big Data* 11 (1): 55. https://doi.org/10.1186/s40537-024-00920-x .

Retraction Watch. 2014. "Leading Chemist Notches Two Retractions in One Journal, Separated by 47 Years." *Retraction Watch*. February 25. https://retractionwatch.com/2014/02/25/leading-chemist-notches-two-retractions-in-one-journal-separated-by-47-years/ .

Rossouw, T. M., L. Matsau, and C. van Zyl. 2020. "An Analysis of Retracted Articles with Authors or co-Authors from the African Region: Possible Implications for Training and

Awareness Raising." *Journal of Empirical Research on Human Research Ethics* 15 (5): 478–493. https://doi.org/10.1177/1556264620955110 .

Rubbo, P., C. L. Helmann, C. Bilynkievycz dos Santos, and L. A. Pilatti. 2019. "Retractions in the Engineering Field: A Study on the Web of Science Database." *Ethics & Behavior* 29 (2): 141–155. https://doi.org/10.1080/10508422.2017.1390667 .

SAGE. 2022. *2022 World Conference on Research Integrity: SAGE's approach to research integrity and preserving trust*. https://perspectivesblog.sagepub.com/blog/author-services/2022-world-conference-on-research-integrity-sages-approach-to-research-integrity-and-preserving-trust .

Shah, T. A., S. Gul, S. Bashir, S. Ahmad, A. Huertas, A. Oliveira, F. Gulzar, A. H. Najar, and K. Chakraborty. 2021. "Influence of Accessibility (Open and Toll-Based) of Scholarly Publications on Retractions." *Scientometrics* 126 (6): 4589–4606. https://doi.org/10.1007/s11192-021-03990-3 .

Shahraki-Mohammadi, A., L. Keikha, and R. Zahedi. 2024. "Investigate the Relationship Between the Retraction Reasons and the Quality of Methodology in non-Cochrane Retracted Systematic Reviews: A Systematic Review." *Systematic Reviews* 13 (1): 24. https://doi.org/10.1186/s13643-023-02439-3 .

Sharma, K. 2021. "Team Size and Retracted Citations Reveal the Patterns of Retractions from 1981 to 2020." *Scientometrics* 126 (10): 8363–8374. https://doi.org/10.1007/s11192-021-04125-4 .

Sharma, K. 2024. "Over Two Decades of Scientific Misconduct in India: Retraction Reasons and Journal Quality Among Inter-Country and Intra-Country Institutional Collaboration." *Scientometrics* 129 (12): 7735–7757. https://doi.org/10.1007/s11192-024-05192-z .

Shi, A., B. Bier, C. Price, L. Schwartz, D. Wainright, A. Whithaus, A. Abritis, I. Oransky, and M. Angrist. 2024. "Taking it Back: A Pilot Study of a Rubric Measuring Retraction Notice Quality." *Accountability in Research* 1–12. https://doi.org/10.1080/08989621.2024.2366281 .

Shi, Q., Z. Wang, Q. Zhou, R. Hou, X. Gao, S. He, S. Zhao, Y. Ma, X. Zhang, Q. Guan, et al. 2021. "An Overview of Retraction Status and Reasons of non-Cochrane Systematic Reviews in Medicine." *Journal of Clinical Epidemiology* 139:57–67. https://doi.org/10.1016/j.jclinepi.2021.06.020 .

Springer. 2024. *Plagiarism Prevention with CrossCheck*. https://www.springer.com/gp/authors-editors/editors/plagiarism-prevention-with-crosscheck/4238 .

UK Research and Innovation. 2018. *ESRC Research Data Policy*. https://www.ukri.org/publications/esrc-research-data-policy/ .

UK Research and Innovation. 2023. *Policy Framework on Research Data*. https://www.ukri.org/who-we-are/epsrc/our-policies-and-standards/policy-framework-on-research-data/ .

Wang, F., and Q. Miao. 2023. "Novel Paradigm of ai-Driven Scientific Research: From AI4S to Intelligent Science." *Bulletin of Chinese Academy of Sciences* 38 (4): 536–540. No. https://doi.org/10.16418/j.issn.1000-3045.20230406002 .

Yang, W., N. Sun, and H. Song. 2024. "Analysis of the Retraction Papers in Oncology Field from Chinese Scholars from 2013 to 2022." *Journal of Cancer Research and Therapeutics* 20 (2): 592. Q4. https://doi.org/10.4103/jcrt.jcrt_1627_23 .

Zheng, E.-T., Z. Fang, and H.-Z. Fu. 2024. "Is Gold Open Access Helpful for Academic Purification? A Causal Inference Analysis Based on Retracted Articles in Biochemistry." *Information Processing & Management* 61 (3): 103640. https://doi.org/10.1016/j.ipm.2023.103640.

Appendix

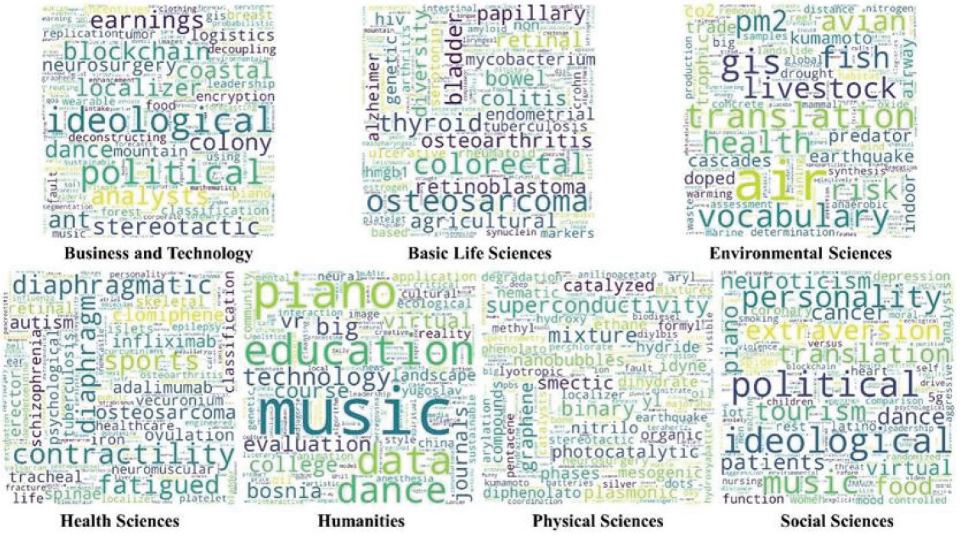

Figure A1. Topic word clouds for each subject.

Table A1. The keywords used to label each sentence.

Sentence label	Keywords
voluntary retraction	'authors retract,''author retract,''author withdraw,''author withdrew,' 'authors remove,' 'author remove,' 'authors hereby retract,' 'retract our article'
agreement	'all authors agree,' 'all the authors agree,' 'all of the authors agree,' 'all of authors agree'
disagreement	'did not agree,' "didn't agree," 'whether agree,' 'never agree,' 'not in agreement'
lack of stance/no response	'no response,' 'no reply,' 'not replied,' 'not responded,''not respond,' 'not reply,' 'failed to respond,' 'lack of response,''no acknowledgment was received,''not be reached,''not responsive'
not mention	'neither agreed nor disagreed,' 'neither agreement nor disagreement,' 'not agreed or disagreed,' 'whether they agree or disagree,' 'not confirm agreement or disagreement,''not comment on the retraction'

Table A2. Description of reasons for retraction and main problems.

Data Problem	Reason	*N*=16842	Reason description	Main problems
Accuracy	Error in Data	1826 (10.8%)	A mistake made in the data, either in data entry, gathering, or identification	Data contradicted further findings, Wrong experimental procedures, and Errors in data analysis.
	Error in Image	1209 (7.2%)	A mistake made in the preparation or printing of an image	Contradiction between images, Errors in images caused by incorrect data, Logical inconsistencies in images
	Unreliable Data	1406 (8.3%)	The accuracy or validity of the data is questionable	Unreproducible results, Ambiguous description, Inconsistence between data and results, Unreliable original data
	Unreliable Image	208 (1.2%)	The accuracy or validity of the image is questionable	Irregular images, Blurry key information in images, and Images contradicted further findings
	Manipulation of Images	1180 (7.0%)	The changing of the presentation of an image by reversal, rotation, or similar action	Images stretched and cropped, Inappropriate image manipulation, Duplication, and inappropriate presentation of images
Reliability	Duplication of Data	647 (3.8%)	Also known as "self-plagiarism." Used when the all or part of the data from an item written by one or all authors of the original article, are repeated in the original article without appropriate citation.	High similarity of data in the paper, Inappropriate reuse of previously published data
	Falsification/ Fabrication of Data	1558 (9.3%)	Intentional changes to data so that it is not representative of the actual finding	Irrational data, Contradiction between data, Fake data, Fake results
	Plagiarism of Data	295 (1.8%)	Used when the all or part of the data from an item not written by one or all authors of the original article, are repeated in the original article without appropriate citation.	Plagiarism of previously published data, Unauthorized use of unpublished data belonging to others, Intentional modification of data ownership
	Duplication of Image	3854 (22.9%)	Also known as "self-plagiarism." Used when an image from an item written by one or all authors of the original article is repeated in the original article without appropriate citation.	Repeated appearance of the same image, Inappropriate reuse of previously published images
	Falsification/ Fabrication of Image	501 (3.0%)	Intentional changes to an image so that it is not representative of the actual data	Irrational images, Contradiction between images, Fake images
	Plagiarism of Image	343 (2.0%)	Used when an image from an item not written by one or all authors of the original article is repeated in the original article without appropriate citation.	Plagiarism of previously published images, Unauthorized use of unpublished images belonging to others, Intentional modification of image ownership

(*Continued*)

Table A2. (Continued).

Data Problem	Reason	*N*=16842	Reason description	Main problems
Validity	Concerns/ Issues About Data	8002 (47.5%)	Any question, controversy, or dispute over the validity of the data	Error in data caused by defective research design, Data without relevant permissions, Without appropriate ethics approval, Unavailable informed consent documents
	Concerns/ Issues About Image	1879 (11.2%)	Any question, controversy, or dispute over the validity of the image	Error in images caused by defective research design, Images without relevant permissions
	Randomly Generated Content	1502 (8.9%)	Text or data that was created via a randomizing algorithm such as Mathgen or Scigen	Incoherent and nonstandard wording, Unusual or tortured phrases, Nonsensical language and excessive citation of work, Abuse of the Large Language Model
Integrity	Original Data not Provided	1646 (9.8%)	The original data or images for the published study are no longer available or are not given to the editorial staff.	Loss of original data, Incomplete original data, Issues with original data accessibility

Table A3. Description of reasons for retraction and rationale for inclusion.

Data Problem	Reason	Rationale for Inclusion
Accuracy	Error in Data	It refers to a mistake made in data entry, gathering, or identification, which directly affects the accuracy of data, making it fail to truly reflect the actual situation, thus belonging to data problems.
	Error in Image	It involves a mistake in the preparation or printing of an image, leading the image to fail to accurately present the original data and undermining the accuracy of visual information, hence falling into the category of data problems.
	Unreliable Data	The accuracy or validity of the data is questionable (e.g., uncorrected sampling bias or measurement errors), indicating that the data cannot be reliably trusted, which is a data problem related to accuracy.
	Unreliable Image	The accuracy or validity of the image is in doubt (e.g., blurriness or missing key information), making it unable to truly reflect research phenomena and affecting data accuracy, so it is a data problem.
	Manipulation of Images	It means changing the presentation of an image by reversal, rotation, or similar actions, which may mislead the interpretation of image data and damage the accuracy of data presentation, thus being a data problem.
Reliability	Duplication of Data	Also known as "self-plagiarism," it occurs when all or part of the data from an item written by one or all authors of the original article is repeated in the original article without appropriate citation. This undermines the uniqueness and originality of the data, affecting its reliability and is a data problem.
	Falsification/Fabrication of Data	It involves intentional changes to data so that it is not representative of the actual finding, seriously violating academic norms and making the data unable to stably support research conclusions, thus belonging to data problems related to reliability.
	Plagiarism of Data	It refers to the repetition of all or part of the data from an item not written by one or all authors of the original article in the original article without appropriate citation. This infringes intellectual property rights and impairs the reliability of data sources, hence being a data problem.
	Duplication of Image	Also called "self-plagiarism," it happens when an image from an item written by one or all authors of the original article is repeated in the original article without appropriate citation. Similar to data duplication, it weakens the originality and reliability of research data, so it is a data problem.
	Falsification/Fabrication of Image	It is the intentional changes to an image so that it is not representative of the actual data, directly destroying the reliability of the image as research evidence and thus being a data problem.
	Plagiarism of Image	It means the use of an image from an item not written by one or all authors of the original article in the original article without appropriate citation. Like data plagiarism, it harms the reliability of data sources and is a data problem.

(*Continued*)

Table A3. (Continued).

Data Problem	Reason	Rationale for Inclusion
Validity	Concerns/Issues About Data	It covers any question, controversy, or dispute over the validity of the data, directly challenging the rationality of data as research evidence and belonging to data problems related to validity.
	Concerns/Issues About Image	It involves any question, controversy, or dispute over the validity of the image, indicating that the image cannot effectively support research findings and affecting data validity, so it is a data problem.
	Randomly Generated Content	It refers to text or data created via a randomizing algorithm such as Mathgen or Scigen. Such data is irrelevant to the actual research and completely lacks scientific validity, thus being a data problem related to validity.
Integrity	Original Data not Provided	It means the original data or images for the published study are no longer available or are not given to the editorial staff, making the research results unable to be independently verified and damaging the traceability and academic integrity of data, hence belonging to data problems related to integrity.

Table A4. Comparison of characteristics between data-related and non-data-related retractions.

	Data problems (*n*=16842)	Other problems (*n*=19656)	*p*-value
Subject			<0.0001**
Business and Technology	2442(14.5%)	5070(25.8%)	
Basic Life Sciences	9934(59.0%)	6087(31.0%)	
Environmental Sciences	671(4.0%)	1190(6.1%)	
Health Sciences	6773(40.2%)	7397(37.6%)	
Humanities	290(1.7%)	747(3.8%)	
Physical Sciences	2174(12.9%)	4312(21.9%)	
Social Sciences	1404(8.3%)	3295(16.8%)	
Journal Quartile			<0.0001**
Q1	5357(31.8%)	5365(27.3%)	
Q2	6350(37.7%)	4850(24.7%)	
Q3	3585(21.3%)	3003(15.3%)	
Q4	737(4.4%)	1082(5.5%)	
No IF	813(4.8%)	5356(27.2%)	
Open Access Level of Journal			0.0001**
Unknown	715(4.2%)	4836(24.6%)	
Extremely low	3577(21.2%)	6569(33.4%)	
Low	4176(24.8%)	4189(21.3%)	
High	1093(6.5%)	656(3.3%)	
Extremely high	7281(43.2%)	3406(17.3%)	
Time Interval to Retraction			<0.0001**
Within the 1st year	4132(24.5%)	8020(40.8%)	
Within the 2nd year	3781(22.4%)	3851(19.6%)	
Within the 3rd year	1894(11.2%)	2314(11.8%)	
Within the 4th year	1421(8.4%)	1610(8.2%)	
After the 4th year	5614(33.3%)	3861(19.6%)	
Article Type			0.0001**
Research Article	14888(88.4%)	15683(79.8%)	
Clinical Study	1522(9.0%)	1398(7.1%)	
Review	370(2.2%)	2072(10.5%)	
Other	62(0.4%)	503(2.6%)	
Authors' Attitude			<0.0001**
Request	3160(18.8%)	3588(18.3%)	
Agree	1793(10.6%)	1234(6.3%)	
Disagree	355(2.1%)	510(2.6%)	
Argue	74(0.4%)	37(0.2%)	
Not state	1494(8.9%)	1265(6.4%)	
Not mention	9966(59.2%)	13022(66.2%)	
Number of Authors			0.0001**
1	1417(8.4%)	3987(20.3%)	
2	2023(12.0%)	3895(19.8%)	
3~5	6593(39.1%)	7707(39.2%)	
6~10	5541(32.9%)	3456(17.6%)	
>10	1268(7.5%)	611(3.1%)	
Number of Countries			0.1512
1	14186(84.2%)	16433(83.6%)	
2	1985(11.8%)	2340(11.9%)	
3	427(2.5%)	544(2.8%)	
4	145(0.9%)	199(1.0%)	
>4	99(0.6%)	140(0.7%)	

Table A5. Comparison of characteristics between data-related and non-data-related retractions in the Basic Life Sciences.

Basic Life Sciences	Data problem (*n*=9934)	Other problem (*n*=6087)	*p*-value
Journal Quartile			<0.0001**
Q1	3737(37.6%)	1710(28.1%)	
Q2	4268(43.0%)	2027(33.3%)	
Q3	1374(13.8%)	944(15.5%)	
Q4	289(2.9%)	453(7.4%)	
No IF	266(2.7%)	953(15.7%)	
Open Access Level of Journal			0.0001**
Unknown	407(4.1%)	958(15.7%)	
Extremely low	1906(19.2%)	1495(24.6%)	
Low	2710(27.3%)	1694(27.8%)	
High	926(9.3%)	350(5.7%)	
Extremely high	3985(40.1%)	1590(26.1%)	
Time Interval to Retraction			<0.0001**
Within the 1st year	1810(18.2%)	2625(43.1%)	
Within the 2nd year	1675(16.9%)	1231(20.2%)	
Within the 3rd year	1321(13.3%)	654(10.7%)	
Within the 4th year	1109(11.2%)	520(8.5%)	
After the 4th year	4019(40.5%)	1057(17.4%)	
Article Type			0.0001**
Research Article	9379(94.4%)	4953(81.4%)	
Clinical Study	417(4.2%)	529(8.7%)	
Review	112(1.1%)	493(8.1%)	
Other	26(0.3%)	112(1.8%)	
Authors' Attitude			<0.0001**
Request	2299(23.1%)	1412(23.2%)	
Agree	1430(14.4%)	15(0.2%)	
Disagree	268(2.7%)	166(2.7%)	
Argue	65(0.7%)	401(6.6%)	
Not state	1012(10.2%)	177(2.9%)	
Not mention	4860(48.9%)	3916(64.3%)	
Number of Authors			0.0001**
1	160(1.6%)	505(8.3%)	
2	654(6.6%)	714(11.7%)	
3~5	3797(38.2%)	2565(42.1%)	
6~10	4256(42.8%)	1899(31.2%)	
>10	1067(10.7%)	404(6.6%)	
Number of Countries			<0.0001**
1	8294(83.5%)	5015(82.4%)	
2	1266(12.7%)	751(12.3%)	
3	249(2.5%)	157(2.6%)	
4	81(0.8%)	95(1.6%)	
>4	44(0.4%)	69(1.1%)	