DSI DATA SCIENCE AND INFORMETRICS
VOL.3, NO.4, NOV. 2023

RESEARCH ARTICLE

Quality control of research data during the pandemic: A case study of retracted COVID–19 papers

Ke Dong[a], Jiachun Wu[a,b*]

a. Research Institute for Data Management & Innovation, Nanjing University, Suzhou, China
b. School of Digital Economics and Management, Nanjing University, Suzhou, China

ABSTRACT

The utilization of research papers has been crucial in facilitating pandemic decision-making and management. Despite the unprecedented surge of scientific publications in response to the COVID-19 outbreak, the retraction of papers related to data has been increasingly frequent, indicating shortcomings in the quality control of current research data. In this study, we aim to examine the root causes and characteristics of data-related retractions amidst the deluge of pandemic research papers, with a particular focus on articles that contain flaws stemming from issues related to research data quality. Our findings suggest that retractions related to data quality deficiencies indicate a dearth of actors and mechanisms within the research quality management landscape. The monitoring and control of data quality cannot be left solely to the self-regulation of the academic community. Hence, our study proposes recommendations for ensuring research data quality in pandemic publications from three perspectives: the management of scientific research quality, data sharing and evaluation mechanisms; publication and dissemination scrutiny mechanisms; and academic early warning and tolerance mechanisms.

KEYWORDS

COVID-19; Pandemic; Research quality management; Retracted publication; Data retractions

1 Introduction

Public health emergencies have triggered a remarkable surge in scientific production, as evidenced by the explosive proliferation of publications related to SARS in 2003 and MERS in 2012 (Ruiz-Fresneda et al., 2022). However, the scientific output stemming from the COVID-19 outbreak in 2019 exceeded all previous benchmarks (Dinis-Oliveira, 2020; Palayew et al., 2020; Rollett et al., 2021). In fact, according to Odone et al. (2020), the number of publications related to COVID-19 exceeded 10,000 within a mere 107-day period commencing on 20 January 2020, accounting for approximately 2.3% of the world's scientific literature. In particular, the volume of COVID-19 preprints surpassed that during the period of the Ebola and Zika outbreaks, and played a more significant role in accelerating the sharing of research (Fraser et al., 2021; Glasziou et al., 2020). The emergence of scientific research outputs cannot be separated from the support of research data as the basic raw materials of activities. In

* Corresponding Author: hdu_wjc@163.com

the background of the global fight against COVID-19 pandemic, the data generated in the process of research on the pandemic is characterized by its complex and diverse sources, large scale, timeliness and inconsistent quality, which increased the challenges of data management, research analysis, publication review and peer validation. Boetto et al. (2021) have highlighted that the unsustainable risk associated with pandemic-related scientific research results could lead to severe consequences, including scientific fraud such as data falsification.

In 2018, the General Office of the State Council issued the "*Measures for the Management of Scientific Data*", which highlighted the importance of standardized and integrated management of scientific data (The State Council, PRC, 2018). The identified problems of scientific data quality are focused on research integrity management and retraction studies, such as academic misconduct data duplication and data falsification. Research integrity management serves as the cornerstone of data quality control within the realm of research activities, and is practiced through peer review, editorial review and third-party reader monitoring. As a branch of life sciences research with frequent retractions, epidemiological research is characterized by urgency, and their retractions highlight the deficiencies in the quality control of scientific data (Fang et al., 2012). Yeo-Teh et al. (2021) observations suggest that early retraction rates of COVID-19 research results have surpassed those of previous public health emergencies, with typical retraction cases involving hydroxychloroquine drug data (Robinson, 2021), clinical trial data (London & Kimmelman, 2020), and missed or falsified peer review (Bell & Green, 2020).

The COVID-19 retraction events reflect the deficiencies in the quality control and management of scientific data. In the later stages of the epidemic, how to mine research data management deficiencies in emergency management from existing articles on research data causing substantial deficiencies, and to carry out research data management actions at the infrastructure level, information disclosure, and quality management has attracted attention from the academic community.

This paper takes the COVID-19 retractions data as the research object, and answers the following questions from the three aspects of data retraction characteristics, the source of actors causing data retraction results, and the academic and social impact of data retraction:

(1) What are the dilemmas, performance and typical features of data quality control in the context of the dramatic expansion of the scientific publications during a pandemic?

(2) What has been the impact of data retractions and what is the urgent need for quality control of research data?

2 Review on quality control of research data for COVID–19 studies

The pressing exigency for scientific data quality control was thrust into the limelight when pandemics became a prominent topic of discussion, as evidenced by the retraction of two COVID-19 articles from esteemed publications, *The Lancet* and *The New England Journal of Medicine*, on account of unreliable data procured from Surgisphere Corporation. Boetto et al. (2021), Lee et al. (2020), and Soltani & Patini (2020) underscored that data quality control has emerged as a critical issue that impinges on the dependability of results, through their analysis of the two articles that were retracted owing to data concerns in the aforementioned journals. Furthermore, Paez (2021) has cast doubt on the reproducibility of the data presented in numerous preprints pertaining to COVID-19 studies.

With the abrupt COVID-19 pandemic emerging as a major challenge to modern medical

research, the quality management and control of COVID-19 scientific data has also attracted global attention. Previous studies attempted to investigate the privacy protection, storage and sharing management of COVID-19 scientific data from both theoretical and practical dimensions. Zong & Lu (2021) conducted a comprehensive analysis of the legal provisions pertaining to COVID-19 research data in Europe and the United States. They highlighted the importance of protecting personal private data in pandemic prevention and advocated for the establishment of a COVID-19 data protection mechanism at the level of national cooperation. In another study, Chu & Guo (2020) explored COVID-19 data management practices by examining the pandemic data management system with respect to the management subjects, objects, environment, and process of epidemic data open management. Garcia et al. (2020) proposed a COVID-19 Information Management Repository, which included data exchange standards, forms, and specifications.

In addition to COVID-19 scientific data source management research, several studies fur ther considered the issue of COVID-19 research data from the perspective of publication processes such as review and retraction of papers. The academic community has increasingly turned their focus towards unmasking scientific data quality discrepancies and executing scientific data quality regulation at disparate strata (Shankar et al., 2021). Frampton et al. (2021) investigated the implementation of the retraction process for 46 COVID-19 retracted articles and showed that these papers did not strictly follow the COPE guidelines. In response to the opacity of the withdrawal process for many COVID-19-related preprints which is not adequately explained, Teixeira da Silva (2021) pointed out that preprint websites should maintain data review mechanisms consistent with those of peer-reviewed journals and keep records of withdrawals. Due to the insidious nature of data problems, Haunschild & Bornmann (2021) have attempted to use Twitter data to identify early signs of research data problems in papers, with the aim of achieving early warning effects.

Recently, data retraction has become more and more serious, which has a gravely destructive effect on the academic ecology, and at the same time has an impact on the existing scientific research order. Although the application and use process of medical research data is relatively strict, data retractions are still occurring due to the lack of rigorous review mechanisms and responsible peer review at the publication stage. Failure to understand the manifestations of these retractions is detrimental to the well-being of the academic environment. The surge in submissions and rapid publication demand during the pandemic period have magnified the possible deficiencies in research data quality control efforts, and the frequent retractions have raised more and more concerns among scholars. A number of studies (Gao et al., 2020; Kaur et al., 2021) are focused on COVID-19 research data resources construction and privacy protection, with insufficient attention paid to data quality control. Several studies of COVID-19 retractions mentioned the data causes of retraction only in the retraction classification, but did not extend to scientific quality control; most of their data and extent were small and limited, failing to produce broader and representative findings, such as data issues in data retractions in terms of performance, causes and typical features. There is rare study that explores the causes, performance and characteristics of the occurrence of problems with quality control of research data from the perspective of articles that have been retracted due to significant deficiencies in the research data. Given the potentially dangerous implications of the retracted COVID-19 article for disease prevention and treatment, this study focuses on COVID-19 papers that have been retracted due to data issues and explores the flaws and weaknesses in current quality control of research data.

3 Research design and data

In this study, we use various data sources including Web of Science (WOS), PubMed, and Retraction Watch database. The former two are employed as general and medical databases to categorize and classify retracted articles based on their respective retraction types (Kuroki & Ukawa, 2018). In contrast, Retraction Watch was deemed a credible and authoritative website for investigating research misconduct due to its extensive coverage and high popularity (Dal-Ré & Ayuso, 2019; Liu & Chen, 2021). The specific process is as follows (Figure 1).

Firstly, this study identifies four keywords related to COVID-19 ("COVID-19", "coronavirus 2019", "2019-nCoV" and "SARS-CoV-2") and retrieved data on retracted articles, retraction notices and retraction records in three databases during January 1st, 2021 and July 31st, 2022. (Al-Zaman, 2021). Secondly, we remove duplicate records of retracted articles from multiple data sources according to the uniqueness of DOI, and merge the contents of the records. Thirdly, we obtain the total number of citations through the article cited statistics in Google Scholar, and manually obtain the number of citations before and after the retraction. Fourthly, we obtain the statistical results of each Altmetric metric of the article through the API interface provided by the Altmetric website in conjunction with the DOI of the article. Finally, we aggregate and identify retraction fields including: bibliographic data, reason for retraction, time of publication, time of retraction, DOI, notice of retraction, number of citations before and after retraction, total citations fields, and Altmetric score.

In this research, we have gathered a total of 252 retracted articles related to COVID-19. To assess the quality control capability of research data, we took into account various factors such as the promptness of detecting data quality issues, the involvement of author groups as direct or indirect data processing actors in diverse writing paradigms, and the participation of the academic community in data management. These factors, when combined, form the

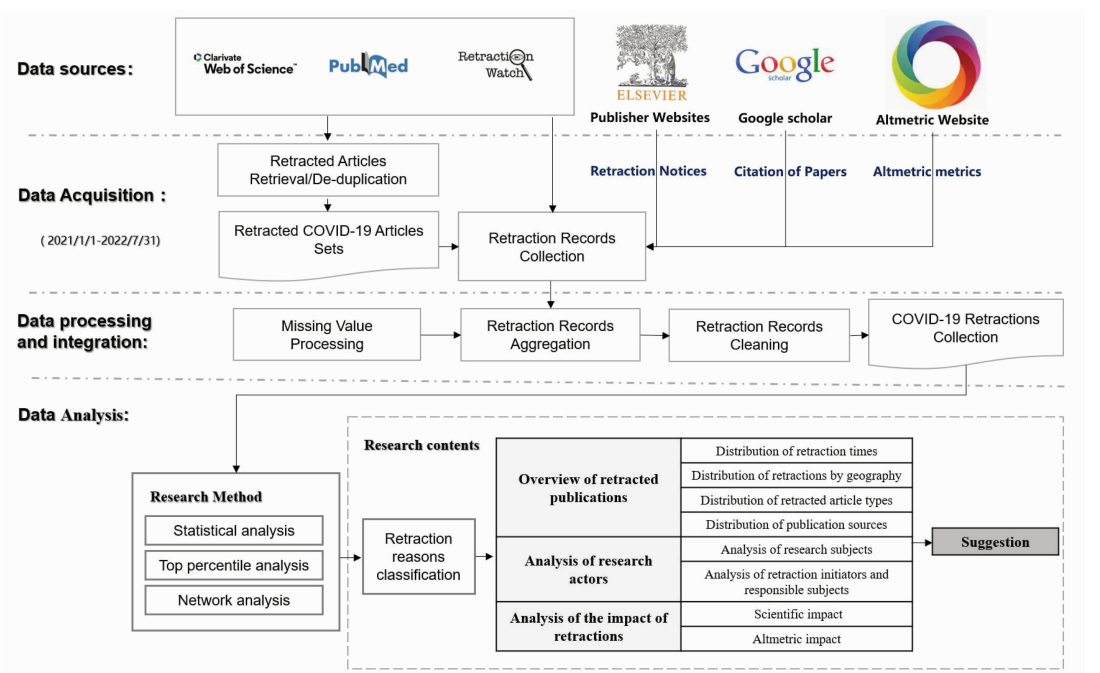


Figure 1 Process of the study on retracted COVID-19 papers

responsible subjects of research data quality control. Additionally, the impact analysis of retracted articles with data defects can unveil the efficacy of research data quality control during the dissemination process. To conduct our analysis, we employed descriptive statistical analysis, percentile analysis, social network analysis to examine the retraction characteristics, research subjects, responsible subjects, as well as the impact of the retracted articles. Our approach offers a comprehensive perspective on research data quality control.

4 Results

4.1 Retraction reasons

This study obtains 59 types of retraction reasons with reference to Retraction Watch, totaling 608 reasons; after excluding four types with unknown information, 362 records of retraction reasons are obtained. Based on the COPE 2019 retraction criteria, the "*Definition of Academic Misconduct in Scholarly Publishing Code Journals*" document released in 2019 and the classification scale in existing retraction studies, this paper summarizes the new classification scale for the retraction reasons around the scientific scenarios in which retractions occur and the subjects who have retraction problems, as shown in Table 1.

Table 1 shows that data issues accounted for 51.4% as the top issue. Previous studies conducted on data retraction rates in the basic life sciences found that although it was the top retraction reason, it has a relatively low percentage at 39.3% (Guan et al., 2021). It is evident that the surge in COVID-19 submissions has further amplified the problem of data quality control in the scientific publishing process. Meanwhile, a study showed that plagiarism ranked first in PubMed-related retractions, at 32.7%, while plagiarism and duplicate publication were lower than data problems in COVID-19 retractions, at 11.0%. With steep pressure to publish, journals and publishers are also prone to operational errors such as early publication, at 8.6%. Plagiarism is also a reflection of researchers' recklessness in the quality management of their research, while duplicate publications are contrary to scientific integrity and ethics, and not only disrupt scientific publication but also result in a waste of academic resources. In addition, the occurrence of peer review issues accounted for 2.5%. All of the above issues have had a devastating impact on the innovation and integrity of COVID-19 research.

Table 1 List of retraction reasons

No.	Retraction Reasons	Description	Occurrence (%)
1	Data issues	Actions such as falsification of data in articles, or data–related errors or problems, etc.	51.4
2	Unclear issues	Vague information about the subject of the survey, and access to article resources	13.0
3	Plagiarism or republication of articles	Plagiarism in the text of articles, etc.	11.0
4	Journal/publisher issues	Errors such as rogue editors, duplicate publications by editors/publishers, etc.	8.6
5	Authorship issues	Falsification of authorship, affiliation, etc.	6.6
6	Copyright or legal disputes	Problems such as copyright disputes or legal risks, etc.	5.0
7	Peer review issues	Issues such as false or manipulated peer reviews, etc.	2.5
8	Reference issues	Citing retracted articles, etc.	1.9

4.2 Overview of retracted publications

4.2.1 Distribution of retraction times

The retraction time lag is the time interval between publication and retraction (Shah et al., 2021), and reflects the efficiency of self-correction in academia (Elango, 2021). Figure 2 shows the time lag distribution of retractions, where the dash indicates the ratio of data retractions to total retractions at each time lag stage. We find that the average time lag for total retractions of COVID-19 articles is approximately 117 days, with a median of 77 days, and nearly 53% of articles are retracted within 3 months, and nearly 94% of retractions taking less than a year. Reviewing previous studies, Ghorbi et al. (2021) showed an average time lag of 591 days for papers retracted by Iranian authors. Li (2022) noted that the average retraction time lag in oncology worldwide was 776 days. Samp et al. (2012) counted an average time lag of 31 months for the drug literature. Elango (2021) yielded an average time lag of 2.48 years for the biomedical social review type of literature. Bhatt (2021) obtained an average time lag of 3.8 years for retractions in the PubMed database. In conclusion, COVID-19 related article retractions have a shorter time lag than general retractions.

The statistics show that data issues are the primary reason for retraction, with a relatively long average time lag of nearly 144 days, a median of 107 days and a maximum of 666 days, which is the longest time lag for retractions overall. The distributions reflect the insidious nature of data retractions. So the quality control of post-publication research data needs to remain open for regulation, with the participation of expert peers, third parties and other subjects, and the quality assessment should include the reproducibility and accessibility of data results.

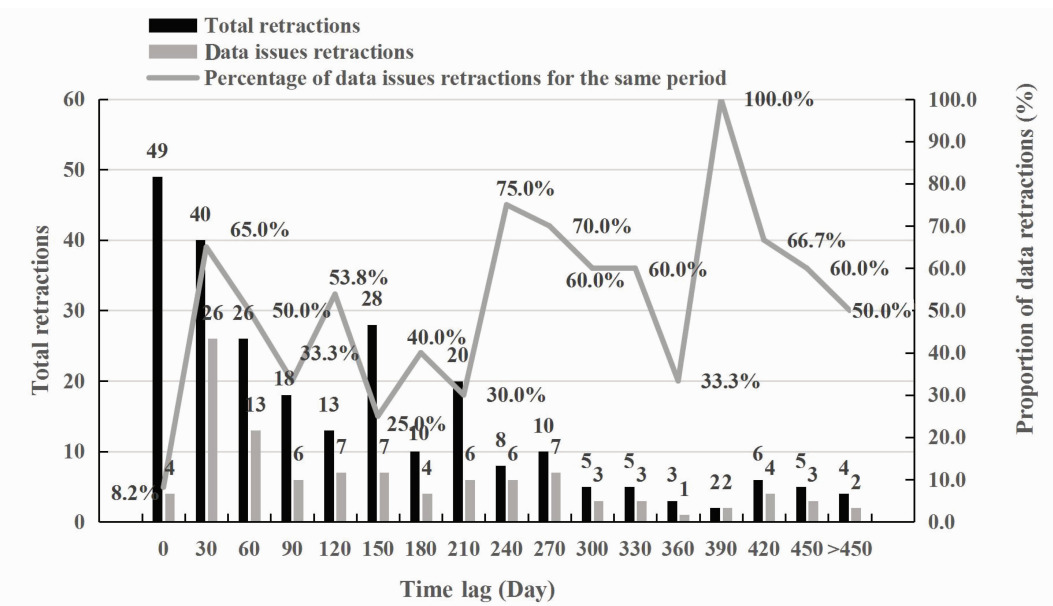


Figure 2 Time lag distribution of total retractions and data issues retractions for the COVID-19

4.2.2 Distribution of retractions by geography

A total of 56 countries or regions appear in the COVID-19 retractions, which is a relatively

wide distribution. Data retractions occur in most countries. Figure 3 shows that the US and China are at a high level in terms of the number of retracted publications, which account for 40.1% (101/252) of the total retracted publications. The UK, Italy, Canada and Germany have a higher number of publications along with a lower number of retractions. And countries such as India, Pakistan and Iran have more serious retractions, with the retraction rate of Iranian papers reaching 0.14%, much higher than the average. In addition, Malta, a special case with a total of 30 retractions, is in the third position of retractions, but it has only 243 and 301 articles published in WOS and PubMed, with a high percentage of retractions.

International collaborative research is an essential tool in the fight against global pandemics. This study's statistics show that 77.1% of data retractions were done by academics from a single country or region, with 61% of US data retractions and 66.7% of Chinese data retractions being non-international collaborations. This suggests a better standard or effectiveness of data quality control in studies related to international collaborations. Previous studies have typically used the ratio of the number of retractions to WOS publications to be able to reflect the overall picture of retractions, so this study used the following formula to calculate the proportion of COVID-19 retractions in the WOS and PubMed databases:

$$\text{COVID}-19\ \text{retractions proportion} = \frac{COVID-19\ retractions}{COVID-19\ publications\ in\ \text{database}} \times 100\% \quad (1)$$

Through calculation, we found that the withdrawal rate of COVID-19 retractions in the WOS and PubMed databases was 0.018% and 0.094%, respectively.

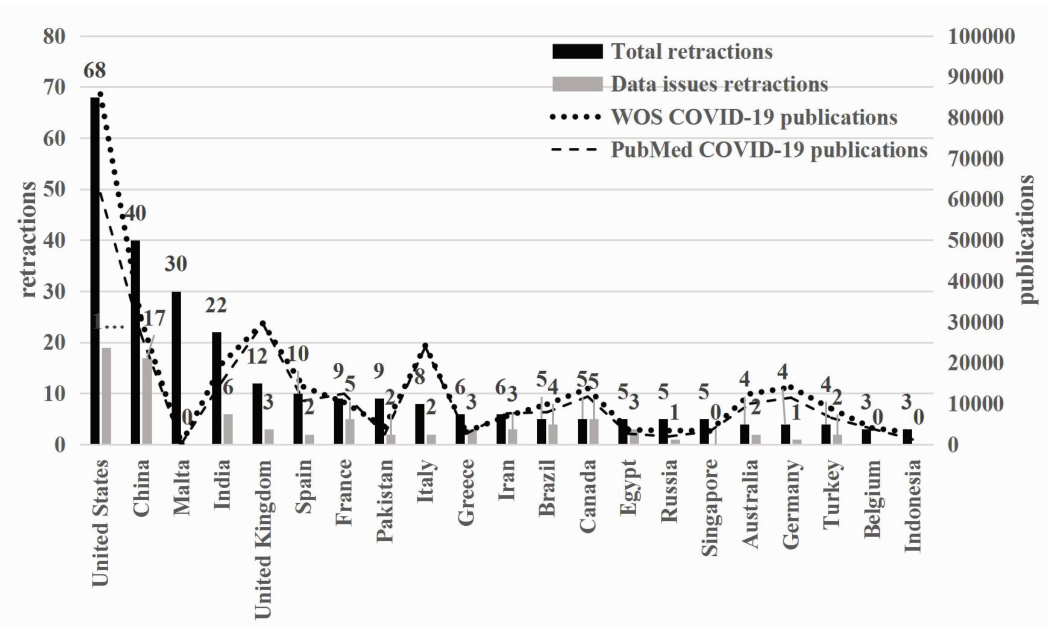


Figure 3 The top 21 countries of overall and data retractions

After further calculating the Pearson correlation coefficient between sum of all-type/data retractions counts per country and the total number of publications from that country in the two major databases WOS and PubMed, we found that the sum of all-types retractions correlated with the number of local research outputs in WOS (P<0.001, R=0.83) and PubMed

(P<0.001, R=0.79) , and the sum of data retractions correlated with the number of local research outputs in WOS (P<0.001, R=0.82) and PubMed (P<0.001, R=0.81), as shown in Figure 4. In conclusion, COVID-19 retractions are associated with local research output, and the performance of data retractions in the geographical dimension is directly related to the importance of research quality management and research data quality control objectives in policy.

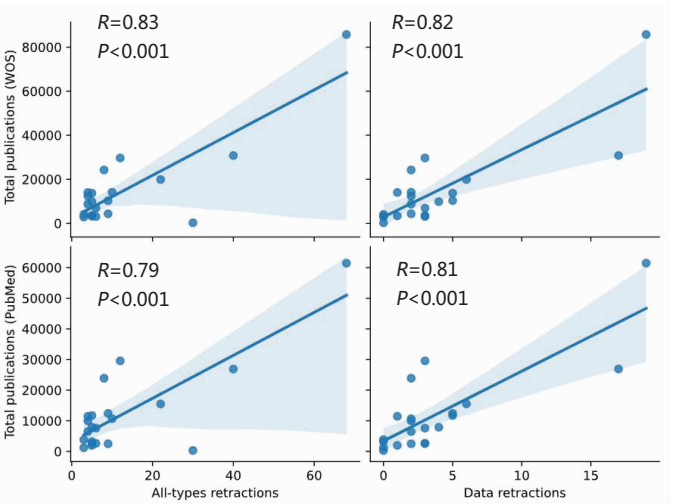


Figure 4 Pearson correlation between sum of all-type/data retractions counts per country and the total number of publications from that country in the two major databases : WOS and PubMed

4.2.3 Distribution of retracted article types

Due to the differentials of writing paradigms and publication requirements, the reasons for retractions may differ between article types. Among 252 retracted articles, 85.7% were peer-reviewed articles, and 14.3% were preprints; and according to the type of article, research articles accounted for 64.1%, reviews for 13.7% and clinical studies for 5.5%.

Original research papers play the most crucial role in pandemic control, but statistics show that a high proportion of research articles are withdrawn, and nearly 60% of withdrawals are due to data quality issues. Campos-Varela & Ruano-Raviña (2019) attributed this phenomenon to the high quality requirements and large publication share of research papers compared with other article types. As can be seen in Figure 5, there were approximately 85% of data retractions in clinical study papers and nearly 50% of data retractions in case reports. Therefore, data retractions are more prone to occur in direct data-related articles like research articles, clinical studies, case reports, so continuous attention is required throughout the life cycle of research data to ensure research quality. Meanwhile, data retractions in review articles, meta-analysis and editorials reflect indirect use of data problems, with meta-analysis and editorials having a similar proportion of citation problems, and we found through the interpretation of the relevant retraction statements that most data retractions in these two types of articles were due to articles that cited problematic data, thus indirectly leading to the existence of data problems.

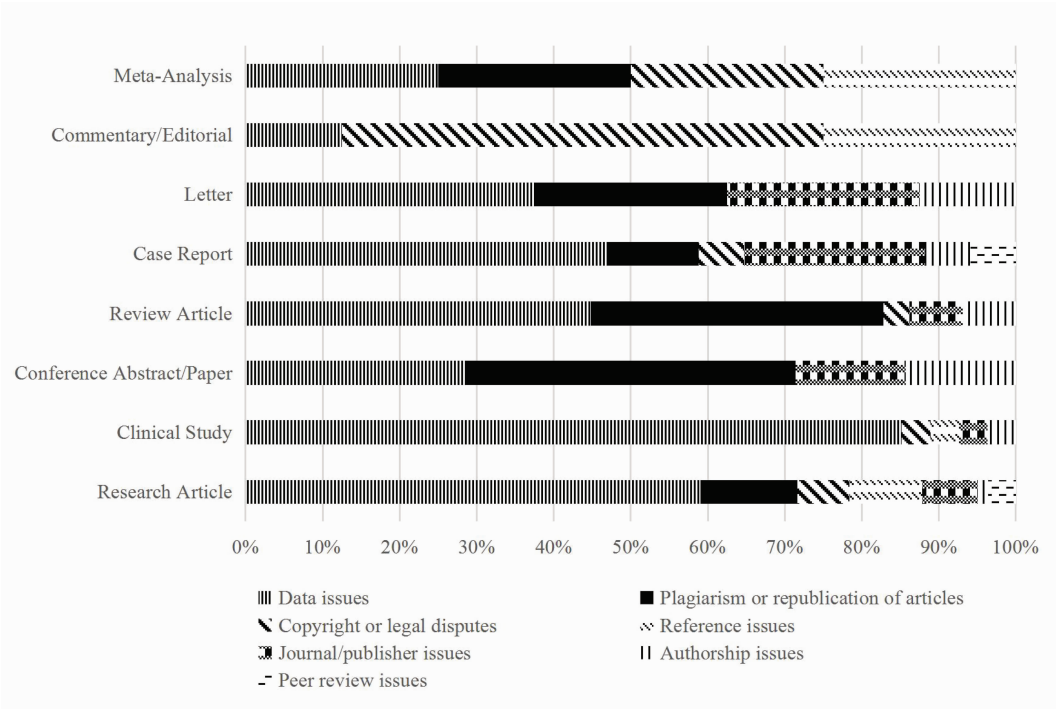


Figure 5 Distribution of retracted COVID-19 article types and retraction reasons

4.2.4 Distribution of publication sources

Retracted COVID-19 publications were from a wide variety of sources, involving 142 journals and 6 preprint platforms, with the percentage of data retractions differing widely across sources (Table 2). Ninety percent of preprint retractions are due to data quality issues, with 18,320 and 5,745 COVID-19 publications on MedRxiv and BioRxiv platforms in the same period, and the data retraction rates were 0.126% and 0.104%, respectively, both above the average data retraction rate. This indicates that while the preprint platform has been effective in promoting the timeliness of scientific publications, there are still many gaps in data quality control.

Journal publishers are considered to be a key part of scientific publishing and data quality management, and the establishment of a sound and basic data quality management system by publishers will help in data quality control at the publishing stage. Although publishers such as Elsevier and Wiley have constructed research data quality guidelines around sharing and reusability of data, as well as operations to maintain the integrity of research such as retraction and retraction of papers for problematic data, there is still a need to consider the practical operationalisation of data standards to explicitly address the need for research data quality management in the face of the proliferation of papers in pandemic. For example, as the journal with the highest number of retractions, *Early Human Development* had a total of 29 retractions, but only 4 were data retractions, and none of the rest provided reasons for the retractions. This suggested that the journal was not rigorous in its quality control of research and allowed a large number of "watered down" papers to be published. *Viruses* retracted 3 COVID-19 papers, including 2 data retractions, and we find that this journal gives a

median publication time of 2.7 days for papers in the first half of 2022 on its website. And *Cureus* gives an average time to publication of 37 days. The production cycles of the two journals listed above are substantially shorter than the time lag for most retractions, raising the question of whether such short peer review timescales serve a purpose in controlling paper quality.

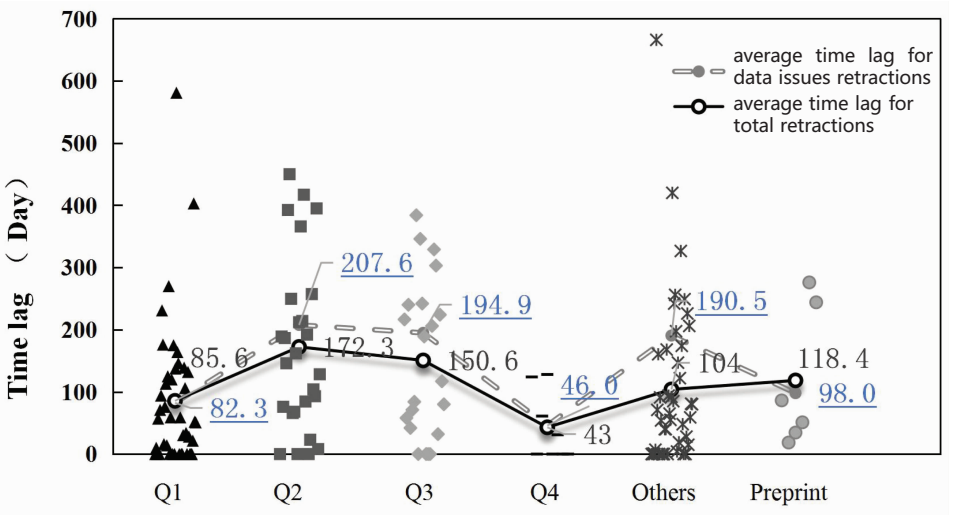
Table 2 Distribution of the main sources of retracted COVID-19 publications

Publication Sources	Publisher	Total retractions	Data issues retractions
Early Human Development	Elsevier	29	4
MedRxiv	—	24	23
Cureus	CUREUS INC	8	7
SSRN: Social Science Research Network	—	8	0
BioRxiv	—	7	6
Actas Dermo–Sifiliograficas	Elsevier	4	0
Journal of Infection	Elsevier	4	0
Scientific Reports	NATURE PORTFOLIO	4	4
Journal of Investigative Medicine	BMJ	3	0
International Journal of Clinical Practice	WILEY	3	3
Viruses	MDPI	3	2

"—" represents publication sources from preprint servers and not considered to be subject to publisher's rules

In terms of journal category quartile, Q1 journals totaled 43 categories and withdrew a total of 53 articles; Q2 journals totaled 26 categories and withdrew 65 articles; Q3 journals totaled 21 categories and withdrew 23 articles; and Q4 journals totaled 8 categories and withdrew 9 articles. The remaining journals that were not included in the SCI were mostly national or regional journals, such as the *Korean Journal of Anesthesiology*, which was included in the Korean Science Citation Index.

In previous studies, researchers have explored the relationship between journal impact factors and retraction time lags in a variety of approaches (He, 2013). Figure 6 shows the average time lag for retractions and data issues retractions for journals in each division, with longer time lags for data issues retractions in divisions Q2, Q3 and Q4 than overall, reflecting the concealment feature of data issues retractions. The shorter time lag between retraction in Q1 journals and preprints and the detection of data quality issues may be related to the strict scientific quality management of Q1 journals and the open-access scholarly communication format of preprints, where high-impact journals have the visibility and rigorous and comprehensive data review processes to respond more immediately and quickly to issues (Vuong, 2020).



Q: quartile in category.

Figure 6 Average time lags for retractions and data issues retractions for journals in different divisions

4.3 Analysis of research actors

4.3.1 Analysis of the cooperation of research subjects

The authors of the papers are the subjects directly responsible for the quality of the data. Since there is no clear definition of the number of retractions of "Repeat Offenders" (Wiedermann, 2018), this study defines repeat authors as those who have retracted twice or more, and constructs a collaborative network of repeat authors, as shown in Figure 7. The red nodes in the graph indicate that all retractions by these authors were data issues retractions, the yellow nodes indicate that none of the retractions by these authors were of data reasons, and the orange nodes indicate that some were data issues retractions.

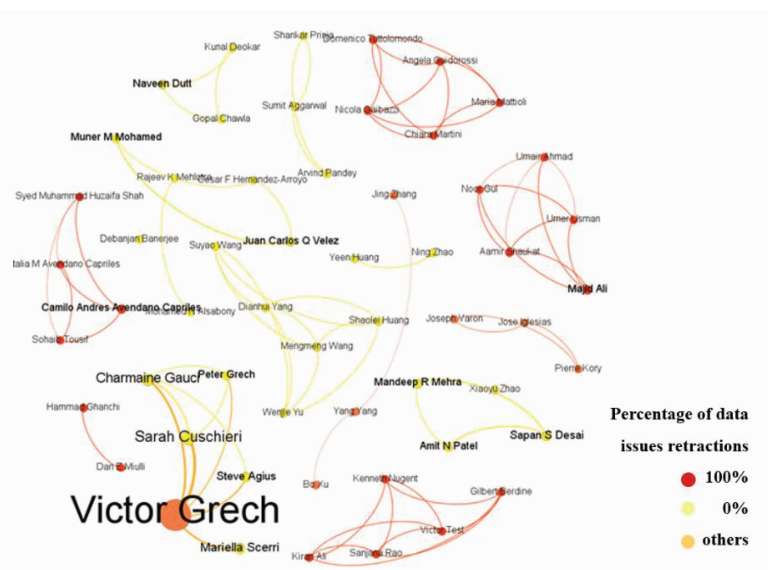


Figure 7 COVID-19 retraction repeat offenders collaborative network

From Figure 7, we find that nearly 87% of COVID-19 retractions are non-sole authors and mostly in the form of more fixed and tightly grouped collaborations of 3-5 persons, with some researchers attempting to use co-authors to escape detection of problematic articles (Foo & Tan, 2014). The size of the nodes in Figure 7 reflects the fact that there were a number of mass retractions caused by an author in the COVID-19 retractions, with the largest node weight of 29 being that of a paediatric cardiologist named Victor Grech, who was associated with 29 retracted papers, including 16 research articles and 13 review articles, which were retracted on five occasions over a three-month period. The biggest retraction occurred on 31 March 2021, with 24 retractions. 29 of the retractions were data issues retractions, except for those for which no clear reason was provided. The data credibility of such repeat offender is questionable, and more attention should be paid to monitoring the quality of data on their research.

At the institutional level, this study finds that universities and their affiliated hospitals account for the highest percentage, at 69%, and independent hospitals at 15.9%. Nearly all countries experienced some proportion of data issues retractions, with the US, China, UK and India, as countries with more retractions, having inter-institutional collaboration networks as shown in Figure 8. Data issues retractions are more likely to occur in hospitals in the UK, and mostly in universities in other countries. Non-specialist research organizations such as companies mainly provide scientific support such as data support and drug development, however, the same can lead to problems such as data and conflicts of interest, with typical cases such as Surgisphere corporation in the USA where faulty data misrepresented a significant amount of research results (Teixeira da Silva et al., 2021).

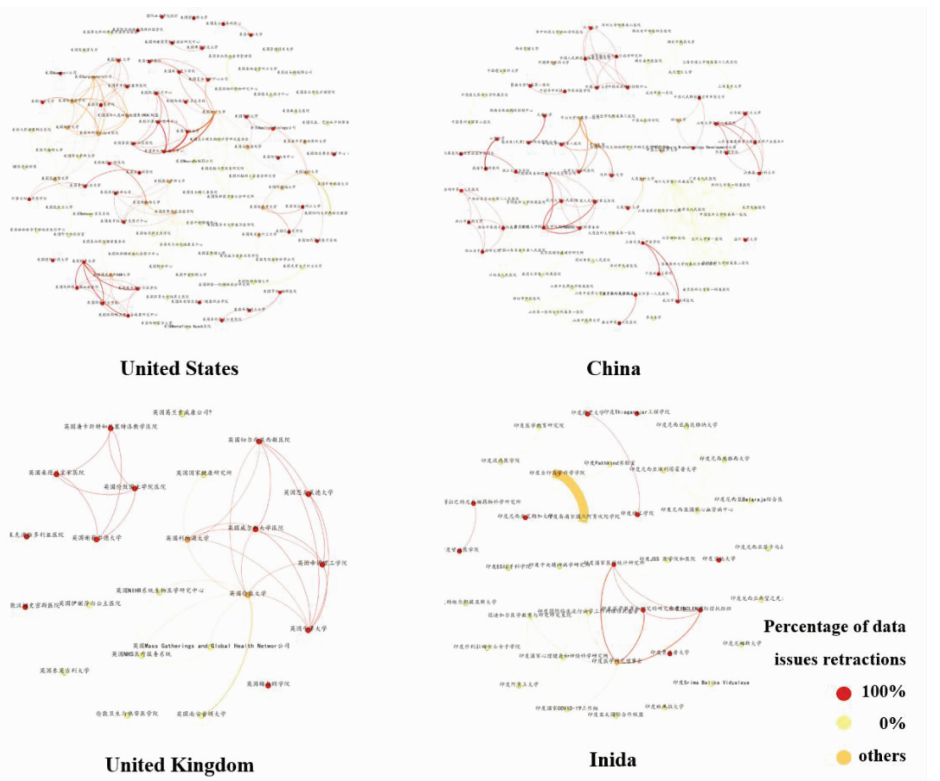


Figure 8 Collaborative networks of major national retraction agencies

4.3.2 Analysis of retraction initiators and responsible subjects

As a separate publication type, the content of a retraction notice includes the title of the retracted article, the status of the retraction, the review process, explanation of the reasons for the retraction and the response from the author of the retracted article. This study deconstructs retractions in three dimensions: the initiators of retractions, the responsible subjects, and the retraction reasons. The statistical results are shown in Table 3. The initiators include authors, editors, publishers and third parties; the responsible subjects are those directly involved in the retraction problem. Identifying the initiators and responsible subjects facilitates the analysis of the direction of data quality management from the perspective of who the data quality problem may occur.

Table 3 shows that author-initiated retractions account for a higher percentage of retractions (25.8%) and a higher percentage of data issues retractions (48.1%), in a behavior often defined by academics as an "honest mistake" (Wang, 2019). Editors and publishers, as the main initiators of retractions in the past, together accounted for the highest proportion of COVID-19 retractions as initiators at 32.9% of the total retraction and 23.1% of the data issues retractions. This difference is due to the fact that problems identified by publishers focus on duplicate publications, technical manipulation, etc., and are often not directly related to data issues retractions; journal editors, as direct processors of the scientific publishing process, need to be directly involved in the identification and quality control of problems with scientific data. In addition, the presence of third parties as initiators in data withdrawals was also high at 23.1%, indicating that third party oversight plays an active role in the identification and quality control of data problems.

As third parties such as public peers and readers rarely act as directly responsible subjects, this paper focuses mainly on peer reviewers in the statistics of responsible subjects. Table 3 shows that authors are responsible for 57.9% of retractions, and 96.2% for data issues retractions. Editors and peer review may also be indirectly responsible for data issues retractions, but their causes are extremely difficult to identify. In addition, the analysis of researcher responses to retraction notices finds that 55.1% mentioned that the authors were informed of the retraction, with 79.6% of the authors fully agreeing to the retraction. This indicates that the majority of researchers showed a responsible attitude towards research when they were informed.

4.4 Analysis of the impact of retractions

4.4.1 Scientific impact

This study finds that 82.1% of COVID-19 retractions were cited at least once in Google Scholar, and the average number of citations for data issues retractions was around 58, much higher than 39 citations for non-data issues retractions. The top 10 retracted papers in terms of total citations include six data issues retractions, and the two most highly cited data issues retractions were both published in high level journals such as *Lancet*, both with over 1000 citations.

This study divides the citation into pre-retraction and post-retraction citations, using the retraction time of the cited article as the cut-off point, plotted in Figure 9. We find that only 10 articles were cited only in the pre-retraction period, 95% of the articles were still cited after retraction, nearly half of the articles were cited entirely in the post-retraction period, and nearly 64.4% of the data issues retractions were cited before retraction, compared with

44.3% of the non-data issues retractions. After data issues are identified that lead to retraction, subsequent researchers often have obvious concerns regarding the quality and value of their data. 6% of data issues retractions were no longer cited after retraction, compared with 3% of non-data issues retractions.

Table 3 Deconstructing retraction notices: retraction initiators and responsible subjects

Subject(s)		Description in retractions	Percentage of total retractions	Percentage of data issues retractions
Retraction Initiator(s)	Author(s)	Appearing at the request of the author, retracted by the author, etc.	25.8%	48.1%
	Editor(s)	Appearing at the request of the editor, editor–in–chief or editorial board, etc.	21.4%	22.1%
	Publisher(s)	Appearing at the request of the publisher, the publisher expressing regret, etc.	11.5%	1.0%
	Third Party	Concerns, questions raised by public peers, readers, third party agencies, etc.	11.1%	23.1%
	Unknown	unclear or not mentioned, etc.	30.2%	5.8%
Responsible Subject(s)	Author(s)	Misconduct by the author in the course of research or scholarly publication (Fanelli et al., 2015).	57.9%	96.2%
	Editor(s)	Immature publication of journals, operational errors, etc.	0.8%	0%
	Publisher(s)	Immature publishing by publishers, duplication of publications, etc.	10.7%	1.0%
	Peer Reviewer(s)	Fake peer review, manipulation of peer review, etc.	2.8%	0%
	Unknown	unclear or not mentioned, etc.	27.8%	1.0%

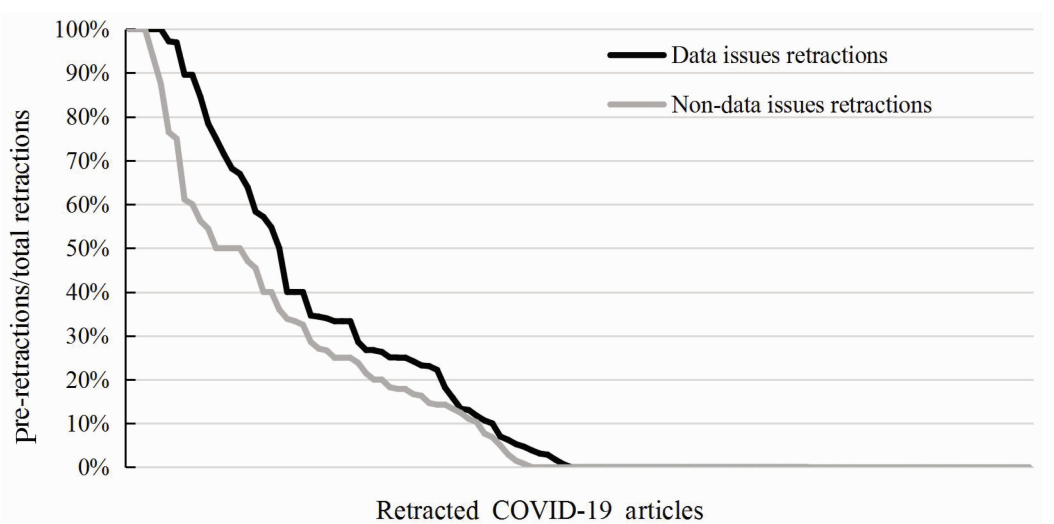


Figure 9 Distribution of citations of retracted papers before retraction

The continued citation of retracted papers will have a negative impact on research quality management, and the discipline of such behavior should be a collaborative effort of the academic community. However, through a survey of the academic community, this study finds that there are still deficiencies in the current work, for example, SSRN directly removes links to the original retracted papers, and Elsevier directly updates the retracted content on the original DOI address without carrying out more recent work. How to regulate, timely and accurately update the status of papers is also an important part of research data quality control and should be given sufficient attention.

4.4.2 Altmetric impact

Social media information is a common occurrence during outbreaks, as the public discusses new scientific research information on social media, which to some extent fills a gap in government data reporting, yet is also more likely to have widespread negative social impact when the source of the information is questionable. For example, an anti-vaccine article on Twitter received over 14,000 citations in just a few days, and was viewed over 380,000 times by the anti-vaccine community and was widely re-shared. It is obvious that the speed and breadth of information viewing and dissemination on social media is much greater than traditional channels. The reach of research results is spreading to the public through social media, and the impact of data issues retractions is being amplified.

In this study, we obtain 186 media records of COVID-19 retractions through the API interface provided by the Altmetric website, which provides data on the dissemination of articles on multiple social media such as Twitter and Mendeley, and quantify the level of exposure of papers in social media through the Altmetric score. We find that retracted articles are disseminated in short articles, images and videos on multiple social media platforms and 38.9% of retracted papers had Altmetric score greater than 20 (data collection on 8 November 2023). According to the website guidelines, the impact of these articles was significant and popular with readers. Table 4 shows data issues retractions for 9 of the top 10 retractions in the Altmetric score rankings, the highest of which was cited 35,531 times on Twitter. We show that retracted papers have a large impact on numerous social media platforms, and that quality management of scientific data in papers is critical.

5 Conclusion and discussion

Pandemics are not a passport to flawed papers, and quickly published but inaccurate results will not help in the fight. The quality control of scientific data is faced with the dilemma of journal editors' poor implementation of data review standards under overloaded submissions, effective peer review cannot be promised, authors' eagerness to publish still-improved research, and the existing data review system and its methodology are out of touch with the real and urgent situation, which will seriously hamper the response to major emergencies and the orderly development of science.

This research analyzes retracted COVID-19 papers and finds that data issue is the top reason for retractions. In comparison with the conventional retraction time-lag, the overall time-lag of COVID-19 retractions is shorter, but the data-issue retractions have a longer time-lag than others, because they are not easy to be detected. Article types that directly depend on data, such as original research papers and case reports, are the types of articles that need to be focused on for the implementation of research data quality control, and the review process needs to be more rigorously enforced. In terms of regional differences in data

Table 4 The top 10 COVID-19 retractions in the Altmetric score rankings

No.	Title of paper	Altmetric Score	Mendeley (Times)	Tweeters (Times)	Videos (Times)	Time lag (Day)
1	*RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID–19: A multinational registry analysis*	22,076	1,911	35,531	13	12
2	*RETRACTED: Facemasks in the COVID –19 era: A health hypothesis*	17,027	172	34,874	7	162
3	*SARS‐CoV‐2 spike impairs DNA damage repair and inhibits V(D)J recombination in vitro*	16,037	153	37,720	10	209
4	*Uncanny similarity of unique inserts in the 2019–nCoV spike protein to HIV–1 gp120 and Gag*	14,107	233	22,864	11	2
5	*RETRACTED ARTICLE: The mechanisms of action of Ivermectin against SARS–CoV–2: An evidence–based clinical review article*	9,666	39	22,576	5	104
6	*RETRACTED ARTICLE: Stay–at–home policy is a case of exception fallacy: An internet–based ecological study*	9,625	151	25,396	3	284
7	*The safety of COVID –19 vaccinations—we should rethink the policy*	8,904	195	18,879	2	8
8	*RETRACTED: Why are we vaccinating children against COVID–19?*	8,870	227	29,820	0	234
9	*WITHDRAWN: A report on myocarditis adverse events in the U.S. Vaccine Adverse Events Reporting System (VAERS) in association with COVID–19 injectable biological products*	8,216	58	19,652	1	16
10	*Effectiveness of surgical and cotton masks in blocking SARS‐CoV–2: A controlled comparison in 4 patients*	6,545	486	10,823	5	57

retractions, there is a positive correlation with the number of local research outputs, and the geographical dimensions of data retractions are related to the degree of attention paid to the construction of research quality management systems, and the level of competence in data safety and quality control, which is still insufficient in terms of concepts of research quality management, institutions, and team building in some developing countries. The flow of pandemic data retractions in academic fields will lead to the continual dissemination of problematic data and erroneous results, which in turn will cause a chain of problematic studies and retractions. The retraction study has been widely disseminated on social media, and the redacted alone cannot effectively curb the dissemination of misinformation online. In terms of social impact, the wrong treatment plan will hinder the rescue work for patients and affect the basic judgement of the public against pandemic, and the ambiguous and unpredictable public governance work will disturb the public order and cause the integrity crisis of the public towards experts and the government.

As an essential information resource for responding to pandemic, it is significant to effectively achieve quality control of research data. In view of the actual demand that quality control of scientific research data requires the collaborative co-operation of multiple academic communities, and the urgent need to improve the policy and problem-handling mechanism

of scientific research data management, the following suggestions are put forward.

First, in the dimension of research quality management actors, which includes researchers, editors, publishers, peer reviewers and third parties. For researchers, strict adherence to academic norms is still required in extraordinary times to ensure the quality of the data itself and the processing. For journal editors and publishers, they can develop and implement data quality control standards during the publication process, supplement the status operation specifications such as correction and retraction of problematic articles, enrich the additional materials at the article review stage, such as the approval of experimental data presentation, a refined division of responsibilities, and the provision of data processing log contents to ultimately achieve the purpose of aiding data peer review. For third parties such as readers, they can participate in the supervision and evaluation of scientific results through academic exchange websites such as preprints and academic communities. Quality control of scientific data needs to improve evaluation criteria in terms of both process and outcome, strengthen the sense of responsibility and self-regulation of the internal community of academics, and encourage the participation and collaborative governance of external subjects.

Second, in the dimension of policies for research data management. The policy system of scientific data management has been expanded and improved in the continuous development of science. For example, the European Union has actively explored scientific research data management by promoting the construction of the international science and technology policy database STIP Compass, and countries such as the United States, the United Kingdom and Australia attach importance to scientific research data management and have formulated policy constraints. In China, research on scientific research data management began late, but with the implementation of laws and regulations such as the *Measures for the Management of Scientific Data*, the quality of scientific research data has steadily improved. Scientific data management needs to be combined with current scientific publishing services and available resources, as well as more mature and referable foreign data management policies, to develop regulated policies and implement standards from a macro perspective to a micro level.

Third, in the dimension of data sharing and evaluation mechanism. Sources of research data in pandemic include clinical trials in hospitals, laboratory data from universities and data from non-research institutions such as drug companies. Multiple sources of data channels and interest-driven manipulation pose great risks to research data quality control. In response, research data sharing measures based on blockchain and other technologies will drive a repositioning of the quality and value of research data and become an important traceability basis for research quality control.

Fourth, in the dimension of publication and dissemination review mechanisms. Both open and rigorous scientific review facilitate research data quality management control. COVID-19 preprints on the one hand leads to over 90% of data issues retractions due to rapid publication without rigorous peer review, yet at the same time effectively shorten the time lag for data issues retractions, and their open nature facilitates research data quality monitoring. As a traditional publication channel, high-impact journals have received more attention, and their strict publication review mechanism has a proactive function in controlling the quality of research data. In the actual research quality control activities, the combination of strict internal self-censorship within the academic community and external open supervision of the review method will become a feasible path for research quality control.

Fifth, in the dimension of academic early warning and error tolerance mechanism. As the

need for scientific timeliness is particularly evident in pandemic, academic publishing needs to improve the error tolerance and correction mechanism, allowing researchers to explore new treatment options and honest errors in problematic results while insisting on accountability for data defects by researchers. In addition, the construction of an academic alert mechanism based on peer-to-peer scholarly exchange websites such as PubPeer and ResearchGate will help in the early identification of research data quality problems and encourage expert peer monitoring of the quality of research results.

Acknowledgements

The research is supported by National Social Science Foundation of China (21CTQ017).

References

Al–Zaman, Md. S. (2021). A bibliometric and co–occurrence analysis of COVID–19–related literature published between December 2019 and June 2020. *Science Editing, 8* (1), 57–63. https://doi.org/10.6087/kcse.230

Bell, K., & Green, J. (2020). Premature evaluation? Some cautionary thoughts on global pandemics and scholarly publishing. *Critical Public Health, 30* (4), 379–383. https://doi.org/10.1080/09581596.2020.1769406

Bhatt, B. (2021). A multi–perspective analysis of retractions in life sciences. *ScientometricsI, 126* (5), 4039–4054. https://doi.org/10.1007/s11192–021–03907–0

Boetto, E., Golinelli, D., Carullo, G., & Fantini, M. P. (2021). Frauds in scientific research and how to possibly overcome them. *Journal of Medical Ethics, 47* (12), e19. https://doi.org/10.1136/medethics–2020–106639

Campos–Varela, I., & Ruano–Raviña, A. (2019). Misconduct as the main cause for retraction. A descriptive study of retracted publications and their authors. *Gaceta Sanitaria, 33 (*4), 356–360. https://doi.org/10.1016/j.gaceta.2018.01.009

Chu, J. W., & Guo, C. x. (2020). Data management practice and thinking on outbreak of major infectious diseases: A case study of COVID–19. *Information studies: Theory & Application, 43* (5), 1–8. https://doi.org/10.16353/j.cnki.1000–7490.2020.05.001

Dal–Ré, R., & Ayuso, C. (2019). Reasons for and time to retraction of genetics articles published between 1970 and 2018. *Journal of Medical Genetics, 56* (11), 734–740. https://doi.org/10.1136/jmedgenet–2019–106137

Dinis–Oliveira, R. J. (2020). COVID–19 research: Pandemic *versus* "paperdemic",integrity, values and risks of the "speed science." *Forensic Sciences Research, 5* (2), 174 –187. https://doi.org/10.1080/20961790.2020.1767754

Elango, B. (2021). Retracted articles in the biomedical literature from Indian authors. *Scientometrics, 126* (5), 3965–3981. https://doi.org/10.1007/s11192–021–03895–1

Fanelli, D., Costas, R., & Larivière, V. (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., Steen, R. G., & Casadevall, A. (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

Foo, J. Y. A., & Tan, X. J. A. (2014). Analysis and implications of retraction period and coauthorship of fraudulent publications. *Accountability in Research, 21*(3), 198–210. https://doi.org/ 10.1080/08989621.2013. 848799

Frampton, G. , Woods, L. , & Scott, D. A. (2021). Inconsistent and incomplete retraction of published research: A cross–sectional study on covid–19 retractions and recommendations to mitigate risks for research, policy and practice. *PLOS ONE, 16* (10), e0258935. https://doi.org/10.1371/journal.pone.0258935

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A.(2021). The evolving role of preprints in the dissemination of COVID–19 research and their impact on the science communication landscape. *PLOS Biology, 19* (4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

Gao, F., Tao, L., Huang, Y., & Shu, Z. (2020). Management and data sharing of COVID–19 pandemic informa–

tion. *Biopreservation and Biobanking, 18* (6), 570–580. https://doi.org/10.1089/bio.2020.0134

Garcia, M., Lipskiy, N., Tyson, J., Watkins, R., Esser, E. S., & Kinley, T. (2020). Centers for disease control and prevention 2019 novel coronavirus disease (COVID-19) information management: Addressing national health-care and public health needs for standardized data definitions and codified vocabulary for data exchange. *Journal of the American Medical Informatics Association, 27* (9), 1476–1487. https://doi.org/10.1093/jamia/ocaa141

Ghorbi, A., Fazeli-Varzaneh, M., Ghaderi-Azad, E., Ausloos, M., & Kozak, M. (2021). Retracted papers by Iranian authors: Causes, journals, time lags, affiliations, collaborations. *Scientometrics, 126* (9), 7351-7371. https://doi.org/10.1007/s11192-021-04104-9

Glasziou, P. P., Sanders, S., & Hoffmann, T. (2020). Waste in covid-19 research. *BMJ, 369,* m1847. https://doi.org/10.1136/bmj.m1847

Guan, Q., Dong, K., & Xia, Y. K. (2021). Research on data management problems and countermeasures based on life science retracted papers. *Library & Information, 3*, 47-56. https://doi.org/10.11968/tsyqb.1003-6938.2021041

Haunschild, R., & Bornmann, L. (2021). Can tweets be used to detect problems early with scientific papers? A case study of three retracted COVID-19/SARS-CoV-2 papers. *Scientometrics, 126* (6), 5181–5199. https://doi.org/10.1007/s11192-021-03962-7

He, T. (2013). Retraction of global scientific publications from 2001 to 2010. *Scientometrics, 96* (2), 555-561. https://doi.org/10.1007/s11192-012-0906-3

Kaur, J., Kaur, J., Dhama, A. S., Kumar, V., & Singh, H.(2021). Management of COVID-19 pandemic data in India: Challenges faced and lessons learnt. *Frontiers in Big Data, 4,* 790158. https://doi.org/10.3389/fdata.2021.790158

Kuroki, T., & Ukawa, A. (2018). Repeating probability of authors with retracted scientific publications. *Accountability in Research, 25* (4), 212–219. https://doi.org/10.1080/08989621.2018.1449651

Lee, K. H., Kim, J. S., Hong, S. H., et al. (2020). Risk factors of COVID-19 mortality: A systematic review of current literature and lessons from recent retracted articles. *European Review for Medical and Pharmacological Sciences, 24* (24), 13089–13097. https://doi.org/10.26355/eurrev_202012_24216

Li, J. Q. (2022). Analysis of the time to retraction of global papers in oncology in 2011-2020 and the influencing factors. *Chinese Journal of Scientific and Technical Periodicals, 33* (5), 561-565. https://doi.org/10.11946/cjstp/202107290592

Liu, X., & Chen, X. (2021). Authors' noninstitutional emails and their correlation with retraction. *Journal of the Association for Information Science and Technology, 72* (4), 473–477. https://doi.org/10.1002/asi.24419

London, A. J., & Kimmelman, J. (2020). Against pandemic research exceptionalism. *Science, 368* (6490), 476–477. https://doi.org/10.1126/science.abc1731

Odone, A., Salvati, S., Bellini, L., Bucci, D., Capraro, M., Gaetti, G., Amerio, A., & Signorelli, C.(2020). The runaway science: A bibliometric analysis of the COVID-19 scientific literature: How COVID-19 has changed academic publishing. *Acta Bio Medica Atenei Parmensis, 91*(9-S), 34–39. https://doi.org/10.23750/abm.v91i9-S.10121

Paez, A. (2021). Reproducibility of research during COVID-19: Examining the case of population density and the basic reproductive rate from the perspective of spatial analysis. *Geographical Analysis,54* (4),860-880. https://doi.org/10.1111/gean.12307

Palayew, A., Norgaard, O., Safreed-Harmon, K., et al. (2020). Pandemic publishing poses a new COVID-19 challenge. *Nature Human Behaviour, 4* (7), 666–669. https://doi.org/10.1038/s41562-020-0911-0

Robinson, K. (2021). A false promise of COVID-19 'big' health data? Health data integrity and the ethics and realities of Australia's health information management practice. *Health Information Management Journal, 50* (1-2), 9-12. https://doi.org/10.1177/1833358320941190

Rollett, R., Collins, M., Tamimy, M. S., et al.(2021). COVID-19 and the tsunami of information. *Journal of Plastic, Reconstructive & Aesthetic Surgery, 74* (1), 199–202. https://doi.org/10.1016/j.bjps.2020.08.112

Ruiz-Fresneda, M. A., Jiménez-Contreras, E., Ruiz-Fresneda, C., Ruiz-pérez R. (2022). Bibliometric analysis of international scientific production on pharmacologic treatments for SARS-CoV-2/COVID-19 during 2020.

Frontiers in Public Health, 9, 778203. https://doi.org/10.3389/fpubh.2021.778203

Samp, J. C., Schumock, G. T., & Pickard, A. S. (2012). Retracted publications in the drug literature. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 32*(7), 586–595. https://doi.org/10.1002/j.1875–9114.2012.01100.x

Shah, T. A., Gul, S., Bashir, S., Ahmad, S., Huertas, A., Oliveira, A., Gulzar, F., Najar, A. H., & Chakraborty, K. (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

Shankar, K., Jeng, W., Thomer, A., Weber, N., & Yoon, A. (2021). Data curation as collective action during COVID–19. *Journal of the Association for Information Science and Technology, 72* (3), 280–284. https://doi.org/10.1002/asi.24406

Soltani, P., & Patini, R. (2020). Retracted COVID–19 articles: A side–effect of the hot race to publication. *Scientometrics, 125* (1), 819–822. https://doi.org/10.1007/s11192–020–03661–9

Teixeira da Silva, J. A. (2021). Silently withdrawn or retracted preprints related to Covid–19 are a scholarly threat and a potential public health risk: Theoretical arguments and suggested recommendations. *Online Information Review, 45* (4), 751–757. https://doi.org/10.1108/OIR–08–2020–0371

Teixeira da Silva, J. A., Bornemann–Cimenti, H., & Tsigaris, P. (2021). Optimizing peer review to minimize the risk of retracting COVID–19–related literature. *Medicine, Health Care and Philosophy, 24* (1), 21–26. https://doi.org/10.1007/s11019–020–09990–z

The State Council, PRC. (2018). Notification of the General Office of the State Council on the Issuance of *Measures for the Management of Scientific Data*. http://www.gov.cn/zhengce/content/2018 –04/02/content_5279272.htm.

Vuong, Q. (2020). The limitations of retraction notices and the heroic acts of authors who correct the scholarly record: An analysis of retractions of papers published from 1975 to 2019. *Learned Publishing, 33* (2), 119–130. https://doi.org/10.1002/leap.1282

Wang, F. C. (2019). Survey on the status of retraction in China. *Chinese Journal of Scientific and Technical Periodicals, 30* (16), 1360– 1365. https://doi.org/10.11946/cjstp.201908180575

Wiedermann, C. J. (2018). Inaction over retractions of identified fraudulent publications: Ongoing weakness in the system of scientific self–correction. *Accountability in Research, 25* (4), 239–253. https://doi.org/10.1080/08989621.2018.1450143

Yeo–Teh, N. S. L., & Tang, B. L. (2021). An alarming retraction rate for scientific publications on coronavirus disease 2019 (COVID–19). *Accountability in Research, 28* (1), 47–53. https://doi.org / 10.1080/08989621.2020.1782203

Zong, J., & Lu, J. Q. (2021). An exploration into data governance and the boundaries of artificial intelligence application amid the COVID–19 crisis. *Science and Technology Management Research, 41* (17), 162–169. https://doi.org/10.3969/j.issn.1000–7695.2021.17.020