The effectiveness of peer review in identifying issues leading to retractions

Xiang Zheng*, Jiajing Chen**, Alison Tollas*, Chaoqun Ni*

*xzheng246@wisc.edu; atollas@wisc.edu; chaoqun.ni@wisc.edu
0000-0002-6619-5504; N/A; 0000-0002-4130-7602
Information School, University of Wisconsin-Madison, USA

** jc12020@nyu.edu
0000-0002-2334-1401
Department of Computer Science, New York University, USA

Retractions can remove flawed research from citable literature but cannot offset the negative impact those publications have on science advances and public trust. This study analyzed the peer-review comments (from Clarivate Analytics) for a sample of retracted publications (from Retraction Watch) to investigate how the peer-review process effectively detects the areas where the retraction causes lie and whether reviewer characteristics are related to the effectiveness. We found that a small proportion of peer reviews suggested rejections during the peer review stage, while about half suggested acceptance or minor revision for those later retracted papers. The peer-review process was more effective in identifying retraction causes related to data, methods, and results than those related to text plagiarism and references. Additionally, factors such as the level of match between reviewers' expertise and the submission were significant in determining the possibility of peer reviews identifying suspicious areas in submissions.

1. Introduction

Retraction is a self-correcting mechanism for science to remove seriously flawed and published research from the citable literature (Hsiao & Schneider, 2021; Steen et al., 2013). Generally, retraction cannot thoroughly delete or hide the retracted publication from public databases. It is intended to alert readers that the published paper contains seriously flawed or erroneous contents or data that undermine its reliability (COPE Council, 2019). Nonetheless, post-publication retractions cannot offset the negative impact on the science advance and public trust (Fang & Casadevall, 2011). For example, retracted papers may be diffused on social media even after retractions and spread misinformation (Serghiou et al., 2021; Shamsi et al., 2022). Retractions may also stigmatize the authors, journals, and associated affiliations, impede the usage of correct knowledge in other papers, and damage public trust in the scientific community(Byrne, 2019; Lu et al., 2013; "The Science of Retraction," 2002; Xu & Hu, 2022). To prevent publishing problematic publications before retractions in the first place, the scientific community should increase the effectiveness of peer review to detect non-administrative errors (Azoulay et al., 2017; Bar-Ilan & Halevi, 2018; Horbach & Halffman, 2019).

However, there is a knowledge gap about whether the peer-review process is effective to detect and report problematic areas in submitted manuscripts that lead to retractions. We aim to identify the effectiveness of the peer review process in identifying suspicious areas in submissions that later lead to retraction. Understanding the factors in the peer-review process that are related to the successful identification of malicious components in the submissions is critical for evaluating the effectiveness of the peer-review process in preventing retractions.

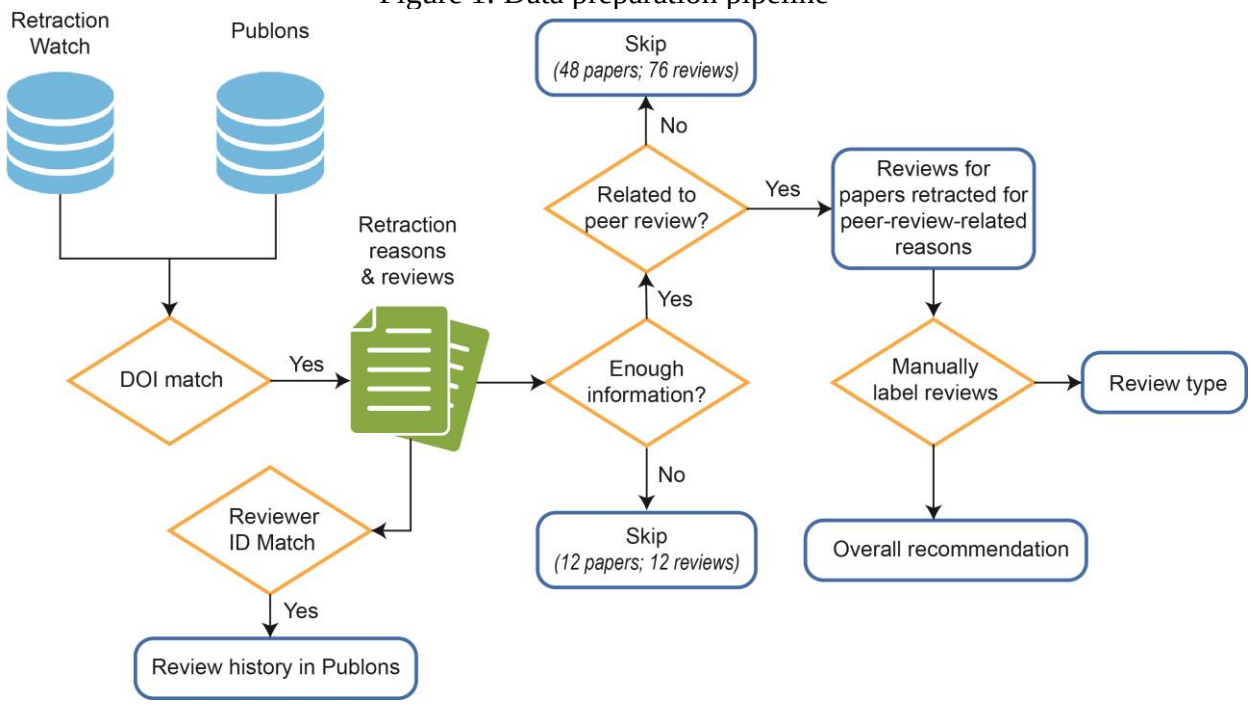
2. 2. Materials and Methods

2.1. Data sources

This study relies on two data sources: the retracted publication list by RW (The Center For Scientific Integrity, 2018) and peer review comments from Publons by Clarivate Analytics (Clarivate Analytics, 2012). RW keeps track of retracted scientific publications by documenting the critical metadata information of retracted publications. Publons documents scientists' invisible peer review contributions by tracking their peer review records. The peer review data underwent anonymization and deidentification process by Publons before they were used in this study.

We obtained peer review comments for retracted publications by matching the DOIs from the RW database and Publons. Peer review comments associated with those DOIs of retracted publications were then extracted for further analysis. We found 348 first-round peer reviews for 211 retracted papers. Among the reviews, 12 reviews (associated with 12 retracted papers) have insufficient information for analysis (e.g., "no comments") and were thus excluded for subsequent analysis. This leaves 206 retracted papers and 336 reviews for further analysis (See **Figure 1**).

Figure 1: Data preparation pipeline



The RW database records the reason(s) for each retraction. Each retracted paper is associated with one or more reasons from the 102 reasons listed by RW. We excluded 76 reviews concerning 48 retracted papers that were retracted due to administrative reasons that are unrelated to the peer review process, such as "Copyright Claims," "Objections by Third Party," and "Error by Journal/Publisher." This leaves 32 retraction reasons, covering 160 retracted papers and 260 peer-review comments in our dataset.

To increase the interpretability of the results, we aggregated the remaining 32 reasons for retraction provided by RW into seven categories: *plagiarism*, *data*, *methods & analysis*, *result*, *reference*, *author*, and *other* (COPE Council, 2019; Marcovitch, 2007; Nair et al., 2020). The number of retracted papers and the corresponding review comments by each retraction reason

category are shown in **Table 1**. It should be noted that some papers could be retracted for multiple reasons and could also have multiple peer-review comments.

Table 1. Aggregated retraction causes by the number of retracted papers and reviews

	Paper		**Corresponding review**	
Aggregated retraction causes	**Number**	**Percentage (%)**	**Number**	**Percentage (%)**
Plagiarism	58	36.25	103	39.62
Data	69	43.13	95	36.54
Method/Analysis	33	20.63	55	21.15
Result	92	57.50	135	51.92
Reference	7	4.38	8	3.08
Author	3	1.88	3	1.15
Other	2	1.25	2	0.77
Total	160	100.00	260	100.00

2.2. Coding and labeling review comments

To understand the gatekeeping role of the peer review process in identifying issues leading to retractions, we read and manually coded the peer review comments for each retracted paper. We first coded each peer review by the type of recommendation it implied in the comments. Two independent coders coded each peer review comment into one of the four recommendation categories: Reject, Major revision, Minor revision, and Accept. The two coders reached an agreement on about 88.46% of all the peer review comments, with a Cohen's Kappa value of 0.83. A third coder labeled the disagreed reviews between the two coders.

We also labeled the peer review comments concerning the reasons related to the retraction. Specifically, the two coders first read the peer review comment and the retraction reasons for each retracted paper. They then coded each review comment concerning each retraction reason and "problem detection," "praise," and "solution suggestion" labels (Cho, 2008). Here, we consider "praise" to be present in a peer review comment if the comment fails to point out the retraction-related issues and uses words expressing gratitude, positivity, admiration, approval, or respect for the very area that later on led to the retraction. Two independent coders coded the 260 peer review comments for the type of comment regarding retraction reasons. The Cohen's Kappa values for "Problem Detection," "Praise," and "Solution Suggestion" were 0.83, 0.91, and 0.87, respectively. Similarly, a third coder labeled comments disagreed with by the two coders.

2.3. Reviewer characteristics

To understand the relationship between peer review and retracted science, we examined the relationship between reviewer characteristics and the likelihood of identifying (or praising and suggesting solutions to) issues leading to the retraction. For the 198 individual reviewers of these 260 reviews for 160 retracted papers, we found their review profiles and histories in Publons (anonymized). Using the accessible data, we considered the following reviewer characteristics.

- **Topic Similarity**: The topic similarity is the topical distance between a review comment and all review comments performed by the same reviewer calculated using the word2vec method

(Mikolov et al., 2013). This measures the average similarity between the peer review comment for the retracted paper and all other reviews by the same reviewer, approximating the closeness between the topic of the reviewed manuscript and the areas of expertise of the reviewer. Words were weighted by inverse document frequency to reduce common words' impact.

- **Average comment length**: This measures the average number of words in each peer review comment by the reviewer. The length of a review comment is used as a proxy of the review's quality, thoroughness, and helpfulness (Thelwall, 2022; Zong et al., 2021).
- **Acceptance rate**: This reflects the percentage of manuscripts published out of the total manuscripts reviewed by a reviewer. A high acceptance rate for a reviewer may indicate that the reviewer is less efficient in gatekeeping the manuscript quality strictly and writing high-quality peer reviews (Kurihara & Colletti, 2013; Ortega, 2017).
- **Seniority**: This measures the number of years between a reviewer's first and last peer reviews. This measure shows the length of a reviewer's review history and also indicates the reviewer's overall peer review experience from one perspective.
- **Number of reviews***:* This measures the annual number of peer reviews performed by a reviewer. This variable quantifies the commitment of a reviewer in the peer-review process and is another indicator of peer-review experience measurement.

2.4. Regression analysis

This study used logistic regression to investigate how various reviewer characteristics contribute to the probability of reviewers identifying issues leading to retractions. The outcome variables include whether the comment is labeled Problem Detection, Praise, or Solution Suggestion. The independent variables used are the reviewer characteristics mentioned above. We controlled for the disciplines of papers in regression analysis to ensure that the observed relationship is not specific to one particular field (Zhang et al., 2022). We followed the discipline classification by RW.

The regression specification is as follows.

$$logit(P) = \beta_0 + \beta_1 Seniority + \beta_2 AvgLength + \beta_3 \#reviews + \beta_4 AccRate + \beta_5 TopicSim + \sum \alpha Discipline + \epsilon \quad (1)$$

where P is the probability of detecting problems, praising, or suggesting solutions, and ϵ is the residual. We clustered the standard deviation at the paper level. We rescaled $AccRate$ and $TopicSim$ at the level of 10% in this regression to better display their coefficients in the results.

3. Results

3.1. Reviewer recommendations for retracted papers

Our coding results suggest that most reviewers failed to reject the later-retracted papers. Out of the 260 reviews associated with 160 retracted papers, 128 (49.2%) were perceived to recommend "Acceptance" (55) or "Minor revision" (73) for the manuscripts. 111 (42.7%) of the reviews recommended "Major revision" for their reviewed manuscripts. Only 21 (8.1%) were perceived to recommend "Rejection" for the manuscript. Each paper may have multiple reviews and thus could have different recommendations. In our data, 13 papers (8.1%) received consensus for a "Rejection" from their reviewers, 20 (12.5%) an "Acceptance," 30 (18.8%) a

"Minor revision," and 52 (32.5%) a "Major revision." The remaining 45 (28.1%) papers received mixed recommendations from their reviewers.

3.2. Problem detection for retraction reasons

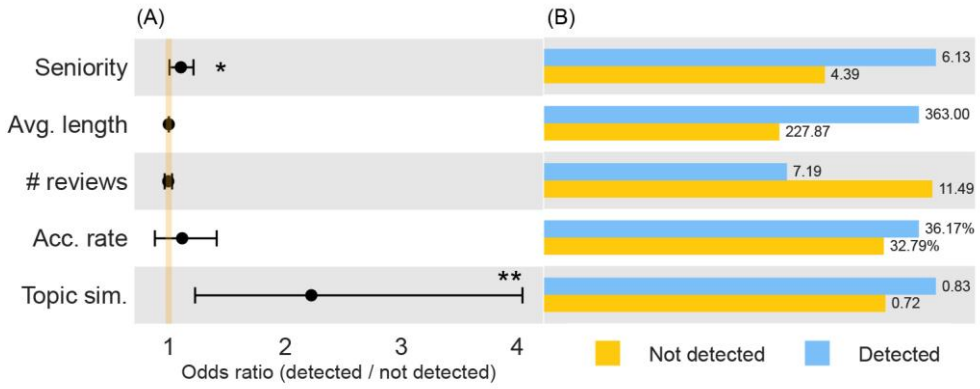
To understand the role of the peer review process in retracted science, we analyzed the effectiveness of peer reviews in problem detection. Among the 260 reviews, 192 (73.8%) failed to detect issues related to the retraction of the papers, and 68 (26.2%) detected at least one problem related to the retraction of the papers. As shown in **Table 2**, about 24.6% of the reviews identified the issues that are related to reasons for retraction, and 30.9% of papers had at least one review identifying issues related to its retraction. None of the reviews detected the problem for papers later retracted for author-related reasons. Among reasons for retractions, "result" related issues were detected by 35.6% of the peer reviews successfully, followed by "data" (33.7%) and "method/analysis" (30.9%).

Table 2. Problem detection reviews (and associated papers) by reasons for retraction. P= Paper; R=Reviews.

Reasons for retraction	**Papers (*n*=160)**		**Reviews (*n*=260)**	
	Number	**Percentage**	**Number**	**Percentage**
Plagiarism (P=58; R=103)	10	19.23%	11	11.46%
Data (P=69; R=95)	28	40.58%	32	33.68%
Method/Analysis (P=33; R=55)	14	42.42%	17	30.91%
Result (P=92; R=135)	40	43.48%	48	35.56%
Reference (P=7; R=8)	1	16.67%	1	12.50%
Author (P=3; R=3)	0	0.00%	0	0.00%
Total (P=160; R=260)	59	30.89%	68	24.55%

We further performed logistic regression analysis to investigate whether the reviewer characteristics are related to the problem-detection chances of peer review comments. Our results show that a reviewer's seniority and topic similarity are significant predictors of problem detection (see **Figure 2**). Specifically, Reviewers with higher seniorities are more likely to detect problems that later lead to the retraction of the paper (OR= 1.105, 95%CI [1.006, 1.213], p= 0.037). The average seniority of reviewers is 1.74 years longer in the problem-detected reviewer group (6.13 years) than in the not detected group (4.39 years). The topic similarity (between the current review and all reviews by a reviewer) also contributes significantly to the chance of problem detection (OR= 2.227, 95%CI [1.226, 4.043], p= 0.009).

Figure 2: Logistic regression results for problem detection of retraction reasons. **(A)** Odds ratio values of reviewer-level factors. **(B)** Mean values of reviewer-level factors. *** p < 0.001, ** P < 0.01, * P < 0.05.



When retraction reasons are aggregated into categories, the reviewer characteristics contributing significantly to the chance of problem detection vary by category (see **Table 3**). Across reasons for retraction categories, topic similarity contributes significantly to the possibility of detecting issues leading to retractions. For reviews of papers retracted due to data-related and methods and analysis-related issues, the higher the topic similarity (data: OR=2.264, 95%CI [1.092, 4.694], p=0.028; method/analysis: OR=10.284, 95%CI [1.064, 99.428], p=0.044), the more likely the peer review comment can identify issues leading to retractions. For reviews of papers retracted due to plagiarism, the acceptance rate is a significant predictor of the chance of detecting plagiarism issues. For reviews of papers retracted due to results-related issues, higher seniority and topic similarity indicates a higher probability of detecting results-related issues (seniority: OR=1.164, 95%CI [1.022, 1.327], p=0.022; topic similarity: OR=2.240, 95%CI [1.194, 4.203], p=0.012).

Table 3. Praise and solution suggestion reviews (and associated papers) by reasons for retraction. P= Paper; R=Reviews.

	Odds ratio	**Std. Err**	**p-value**	**95% CI Lower**	**95% CI Upper**
Data (*n*=111)					
Seniority	1.114	0.096	0.210	0.941	1.320
Ave. length	1.001	0.001	0.642	0.998	1.003
# reviews	0.997	0.025	0.901	0.948	1.048
Acc. rate	1.152	0.229	0.477	0.780	1.699
Topic sim.	2.264	0.842	**0.028**	1.092	4.694
Method/Analysis (*n*=76)					
Seniority	1.136	0.098	0.139	0.960	1.345
Ave. length	1.002	0.002	0.267	0.998	1.006
# reviews	0.994	0.034	0.855	0.929	1.063
Acc. rate	0.748	0.287	0.448	0.352	1.586
Topic sim.	10.284	11.904	**0.044**	1.064	99.428
Plagiarism (*n*=114)					
Seniority	1.044	0.078	0.561	0.902	1.210
Ave. length	0.998	0.002	0.468	0.994	1.003
# reviews	0.937	0.083	0.464	0.788	1.115
Acc. rate	1.811	0.436	**0.014**	1.130	2.903
Topic sim.	1.460	0.728	0.448	0.549	3.881
Result (*n*=175)					
Seniority	1.164	0.078	**0.022**	1.022	1.327
Ave. length	1.000	0.001	0.801	0.998	1.002

# reviews	1.018	0.018	0.306	0.984	1.053
Acc. rate	1.073	0.147	0.604	0.821	1.403
Topic sim.	2.240	0.719	**0.012**	1.194	4.203

3.3. Praise and solution suggestion for retraction reasons

In our data, some review comments praised the areas later on that led to retractions rather than raising concerns, albeit in a small proportion (see **Table 4**). Among the 260 peer review comments for the 160 retracted papers, 27 (10.4%) reviews (for 15 papers) mentioned the retraction issue with a praising tone. Logistic regression analysis shows that the chance of praising issues leading to retractions is not related to any of the reviewer-level factors (see **Figure 3**). Given the limited number of praising peer reviews by retraction reasons, no regression analysis was performed by retraction reasons separately. In addition, 29 (11.2%) out of the 260 peer review comments provided suggestions for solving issues leading to retractions, which accounts for 17.50% (28) of the total papers in the sample.

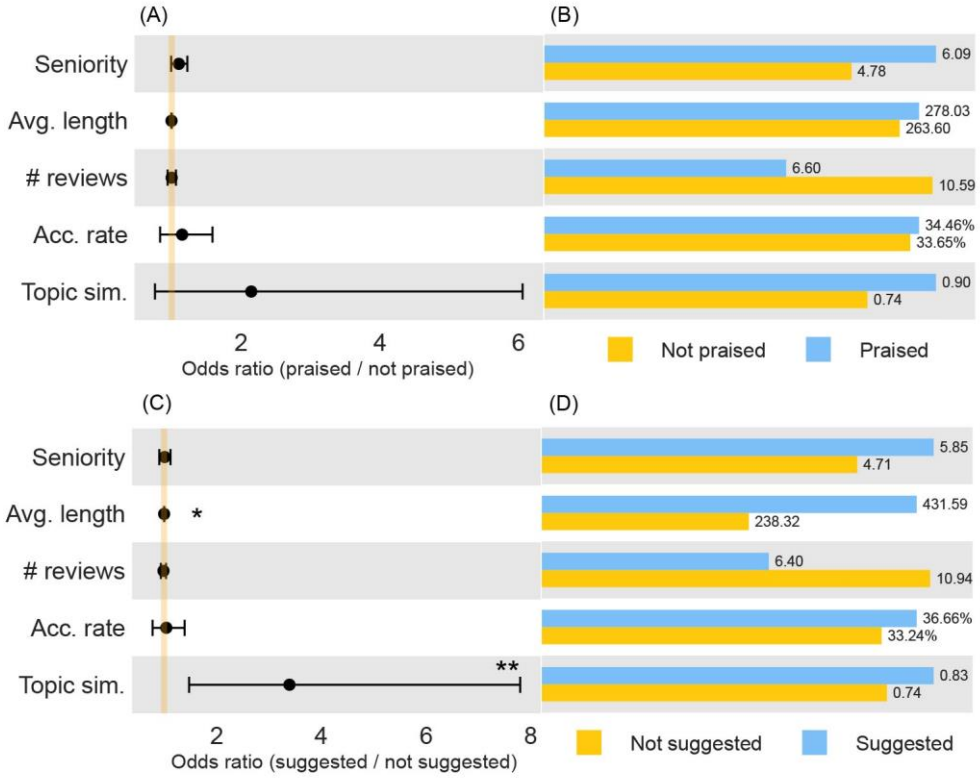
Table 4. Praise and solution suggestion reviews (and associated papers) by reasons for retraction. P= Paper; R=Reviews.

	Praise				**Solution Suggestion**			
	Paper		**Review**		**Paper**		**Review**	
	Num.	Percent.	Num.	Percent.	Num.	Percent.	Num.	Percent.
Author (P=3; R=3)	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Data (P=69; R=95)	8	11.59%	8	8.42%	13	18.84%	14	14.74%
Method/Analysis (P=33; R=55)	5	15.15%	5	9.09%	10	30.30%	10	18.18%
Plagiarism(P=58; R=103)	5	8.62%	5	4.85%	6	10.34%	6	5.83%
Reference (P=7; R=8)	0	0.00%	0	0.00%	1	14.29%	1	12.50%
Result (P=92; R=135)	9	9.78%	9	6.67%	19	20.65%	19	14.07%
Total (P=160; R=260)	15	9.38%	27	10.38%	28	17.50%	29	11.15%

Our results show that average comment length and topic similarity contribute significantly to the chance of providing suggestions to issues leading to retractions (see **Figure 3**). Specifically, the longer the reviews written by a reviewer (OR=1.002, 95%CI [1.000, 1.003], p=0.020), the higher the topic similarity (OR=3.395, 95%CI [1.478, 7.800], p=0.004), the more likely a peer review comment can provide solution suggestions to issues leading to retractions. Given the limited number of solution suggestion reviews by retraction reasons, no regression analysis was performed by retraction reasons separately.

Figure 3: Logistic regression results for Praise and solution suggestion. (A) Odds ratio values of reviewer-level factors for praise comments. (B) Mean values of reviewer-level factors for praise comments. (C) Odds ratio values of reviewer-level factors for solution suggestion

comments. (D) Mean values of reviewer-level factors for solution suggestion comments. *** p < 0.001, ** P < 0.01, * P < 0.05.



4. Discussion

This study evaluated the effectiveness of peer-review comments in preventing retractions by analyzing a sample of peer-review comments and comparing them to the reasons for retraction. By manually coding the peer review comments, the study found that only 42.7% of the peer reviews suggested "major revision," and 8.1% suggested "rejection" for papers that were later retracted. These findings suggest that while some peer reviews did raise issues and suggest solutions that were later cited as reasons for retraction, the papers still slipped through the editorial peer-review system. We also found that the effectiveness of the peer-review process in identifying problematic areas varies depending on the type of issue leading to retraction. The study found that issues leading to retractions due to data, methods/analysis, and results were detected by peer reviews at a higher rate than issues leading to plagiarism, author, and reference-related retractions. Finally, our study found that except for the reason of plagiarism, the higher the topic similarity between the current review and all reviews by the same reviewer, the more likely the current peer-review comment will detect potential retraction issues.

In conclusion, preventing retractions requires intricate and multidimensional efforts involving authors, peer academic institutions, funders, journals, publishers, peer reviewers, and others in the scientific community. The peer review process does seem to detect issues that later lead to retractions and further suggest solutions or recommend rejections to the manuscript. However, its effect is limited. We suggest that editors should pay close attention to peer review comments and perform additional inspections to trace clues of potential issues.

Open science practices

The Publons data are not publicly available. The Retraction Watch data can be accessed at http://retractiondatabase.org/. The aggregated anonymized underlying dataset is available upon request.

Acknowledgments
We gratefully acknowledge the teams of Retraction Watch for providing the retraction records data, and Clarivate Analytics for the anonymized peer review data.

Author contributions
Xiang Zheng: Investigation, Formal analysis, Validation, Writing - Original Draft, Writing - Review & Editing. Jiajing Chen: Software, Data Curation, Formal analysis. Alison Tollas: Writing - Original Draft. Chaoqun Ni: Conceptualization, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing, Visualization.

Competing interests
The authors declare no competing interests.

References
Azoulay, P., Bonatti, A., & Krieger, J. L. (2017). The career effects of scandal: Evidence from scientific retractions. Research Policy, 46(9), 1552–1569. https://doi.org/10.1016/j.respol.2017.07.003

Bar-Ilan, J., & Halevi, G. (2018). Temporal characteristics of retracted articles. Scientometrics, 116(3), 1771–1783. https://doi.org/10.1007/s11192-018-2802-y

Byrne, J. (2019). We need to talk about systematic fraud. Nature, 566(7742), 9–9. https://doi.org/10.1038/d41586-019-00439-9

Cho, K. (2008). Machine classification of peer comments in physics. 192–196. Scopus.

Clarivate Analytics. (2012). Publons. Publons. http://publons.com

COPE Council. (2019). COPE Guidelines: Retraction Guidelines. https://doi.org/10.24318/cope.2019.1.4

Fang, F. C., & Casadevall, A. (2011). Retracted Science and the Retraction Index. Infection and Immunity, 79(10), 3855–3859. https://doi.org/10.1128/IAI.05661-11

Horbach, S. P. J. M., & Halffman, W. (2019). The ability of different peer review procedures to flag problematic publications. Scientometrics, 118(1), 339–373. https://doi.org/10.1007/s11192-018-2969-2

Hsiao, T.-K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. Quantitative Science Studies, 2(4), 1144–1169. https://doi.org/10.1162/qss_a_00155

Kurihara, Y., & Colletti, P. M. (2013). How Do Reviewers Affect the Final Outcome? Comparison of the Quality of Peer Review and Relative Acceptance Rates of Submitted Manuscripts. American Journal of Roentgenology, 201(3), 468–470. https://doi.org/10.2214/AJR.12.10025

Lu, S. F., Jin, G. Z., Uzzi, B., & Jones, B. (2013). The Retraction Penalty: Evidence from the Web of Science. Scientific Reports, 3(1), Article 1. https://doi.org/10.1038/srep03146

Marcovitch, H. (2007). Misconduct by researchers and authors. Gaceta Sanitaria, 21(6), 492–499. https://doi.org/10.1157/13112245

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, 3111–3119.

Nair, S., Yean, C., Yoo, J., Leff, J., Delphin, E., & Adams, D. C. (2020). Reasons for article retraction in anesthesiology: A comprehensive analysis. Canadian Journal of Anesthesia/Journal Canadien d'anesthésie, 67(1), 57–63. https://doi.org/10.1007/s12630-019-01508-3

Ortega, J. L. (2017). Are peer-review activities related to reviewer bibliometric performance? A scientometric analysis of Publons. Scientometrics, 112(2), 947–962. https://doi.org/10.1007/s11192-017-2399-6

Serghiou, S., Marton, R. M., & Ioannidis, J. P. A. (2021). Media and social media attention to retracted articles according to Altmetric. PLOS ONE, 16(5), e0248625. https://doi.org/10.1371/journal.pone.0248625

Shamsi, A., Lund, B. D., & SeyyedHosseini, S. (2022). Sharing of retracted COVID-19 articles: An altmetric study. Journal of the Medical Library Association : JMLA, 110(1), 97–102. https://doi.org/10.5195/jmla.2022.1269

Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why Has the Number of Scientific Retractions Increased? PLoS ONE, 8(7), e68397. https://doi.org/10.1371/journal.pone.0068397

The Center For Scientific Integrity. (2018). The Retraction Watch Database. Retraction Watch. http://retractiondatabase.org/

The science of retraction. (2002). Nature Neuroscience, 5(12), Article 12. https://doi.org/10.1038/nn1202-1249

Thelwall, M. (2022). Journal and disciplinary variations in academic open peer review anonymity, outcomes, and length. Journal of Librarianship and Information Science, 09610006221079345. https://doi.org/10.1177/09610006221079345

Xu, S. B., & Hu, G. (2022). Retraction Stigma and its Communication via Retraction Notices. Minerva, 60(3), 349–374. https://doi.org/10.1007/s11024-022-09465-w

Zhang, G., Xu, S., Sun, Y., Jiang, C., & Wang, X. (2022). Understanding the peer review endeavor in scientific publishing. Journal of Informetrics, 16(2), 101264. https://doi.org/10.1016/j.joi.2022.101264

Zong, Z., Schunn, C. D., & Wang, Y. (2021). Learning to improve the quality peer feedback through experience with peer feedback. Assessment & Evaluation in Higher Education, 46(6), 973–992. https://doi.org/10.1080/02602938.2020.1833179