Journal of Clinical Epidemiology 172 (2024) 111397

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Retracted papers originating from paper mills: a cross-sectional analysis of references and citations

Cristina Candal-Pedreira[a,b,c], Carla Guerra-Tort[a], Alberto Ruano-Ravina[a,b,c,*], Fabián Freijedo-Farinas[a], Julia Rey-Brandariz[a,c], Joseph S. Ross[d,e,f], Mónica Pérez-Ríos[a,b,c]

[a]*Preventive Medicine and Public Health, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain*
[b]*Health Research Institute of Santiago de Compostela (Instituto de Investigación Sanitaria de Santiago de Compostela-IDIS), Santiago de Compostela, Galicia, Spain*
[c]*Consortium for Biomedical Research in Epidemiology and Public Health (CIBER en Epidemiología y Salud Pública-CIBERESP), Madrid, Spain*
[d]*Section of General Internal Medicine and National Clinician Scholars Program, Yale School of Medicine, New Haven, CT USA*
[e]*Department of Health Policy and Management, Yale University School of Public Health, New Haven, CT, USA*
[f]*Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA*

Accepted 21 May 2024; Published online 28 May 2024

Abstract

Objectives: The aims of this study are (1) to analyze the references cited by retracted papers originated from paper mills; (2) to analyze the citations received by retracted papers originated from paper mills; and (3) to analyze the potential relationships existing between paper mill papers and their references and their citations.

Study Design and Setting: This study was a cross-sectional study. All original papers retracted in 2022 identified as having originated from paper mills and had been published at least 12 months before their retraction (hereinafter ''source-retracted papers'') were included. The Retraction Watch database was used to identify the source-retracted papers and Web of Science was used to identify the references contained within them and the citations received by them. We described the characteristics of the papers and journals. Additionally, 2 networks of source-retracted papers mutually interconnected via their citations and references were built: 1 with only retracted references and retracted citations and the other with all references and citations (retracted or unretracted).

Results: A total of 416 paper mill papers retracted in 2022 (sourced retracted papers) were identified, with a median of 1247 (interquartilic range, 907.8−1673.5) days between publication and retraction. Of all authors identified, 92.3% were affiliated with Chinese institutions. There were 14,411 references contained in the source-retracted papers and 8479 citations received by them; the median number of references and citations was 35 (29−40) and 16 (9−25), respectively. In total, 473 references and citations had also been retracted for being paper mill papers. Among the 416 sourced-retracted papers, 169 (41.9%) and 178 (42.8%) were referenced or were cited by at least another retracted paper, the majority of which also originated from paper mills. The first network analysis, which included source-retracted papers along with their retracted references and citations, found 3 clusters of 53, 48, and 44 retracted papers that were mutually interconnected. The second network analysis, with all references and citations (retracted or unretracted) identified a large cluster of 2530 interconnected papers.

Conclusion: Retracted papers originating from paper mills frequently reference and are cited by papers that are later retracted for having originated from paper mills, displaying inter-relationships. Detecting these inter-relationships can serve as an indicator for identifying potentially fraudulent publications.

Keywords: Paper mills; Research integrity; Ethics; Retraction; Scientific misconduct; Bibliometric analyses

Funding: This paper forms part of the work leading to the doctoral thesis of Cristina Candal-Pedreira, the beneficiary of a Pre-doctoral Health Research Training (*Contratos Predoctorales de Formación en Investigación en Salud)* grant (reference no. FI21/00149) from the Carlos III Institute of Health (*Instituto de Salud Carlos III/ISCIII*).

Ethics statement: Because this study used publicly available materials and did not involve human subjects, human subjects' ethics committee approval was not required.

Disclaimer: The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

* Corresponding author. Preventive Medicine and Public Health, School of Medicine, Universidade de Santiago de Compostela, San Francisco Street s/n., 15782 Santiago de Compostela, Spain.

E-mail address: alberto.ruano@usc.es (A. Ruano-Ravina). @albertoruano8 (A. Ruano-Ravina)

https://doi.org/10.1016/j.jclinepi.2024.111397

Plain language summary

This study aims to analyze the potential connections between scientific papers that have been retracted for originating from paper mills. Paper mills are organizations primarily engaged in the mass production of scientific manuscripts, most of them fraudulent, which are then sold to researchers to obtain authorship. We analyzed papers retracted in 2022 for originating from paper mills (hereinafter: source-retracted papers) and their references (papers which were cited by source-retracted papers) and citations (papers which cited source-retracted papers). We found 416 retracted papers from paper mills, containing 14,411 references and receiving 8479 citations. Of these, 473 references and citations were retracted due to originating from paper mills. About 42% of the source-retracted papers were referenced or were cited by other retracted papers, many of which were from paper mills. Network analysis revealed groups of interconnected retracted papers. Retracted papers from paper mills often reference and are cited by other fraudulent papers, indicating a connection. Detecting these patterns could help identify potentially fraudulent publications.

1. Introduction

In scientific publication, the term "paper mills" refers to organizations whose purpose is to generate a high volume of scientific manuscripts, for the most part fraudulent or plagiarized, with the aim of then selling the authorship to researchers or students [1−4]. Despite being a relatively recent phenomenon, paper mills have experienced unprecedented growth [2,4−8]. The annual profit of these organizations has been estimated at a billion euros [9]. It is estimated that, in 2020, 11.0% of all papers published may have come from a paper mill organization, which would be on the order of 383,000 papers [2]. In 2022, another study concluded that 1.5%−2% of articles published could be attributed to these organizations [8]. A report by the Committee on Publication Ethics and the Scientific, Technical and Medical (STM) Integrity Hub estimated that 2 to 46% of journal article submissions may originate from paper mills [1].

Available information on the *modus operandi* of paper mill organizations is very limited, and indeed may differ between organizations. Paper mills have been identified in China, India, Russia, USA, and United Kingdom [2], and the great majority of papers identified as originating from these organizations pertain to specific fields of knowledge, particularly pharmacy, oncology, medicine, and biology [6,8]. Recently, González-Márquez et al plotted a map of retracted biomedical papers in PubMed by area of knowledge, from which it can be seen that retracted papers form groups or clusters of the same topic, suggesting the presence of paper-mill clusters [10]. This would be in line with the findings of a previous systematic review [3].

Previous studies have analyzed the problem of paper mills and the papers originating from such organizations [2,6,11]. A previous study describing paper mill papers retracted until June 2022 [6] detected a possible fraudulent citation pattern: papers retracted for being generated by paper mills received a median of 11 citations, and most of these citations were concentrated in journals ranked in the third and fourth quartiles by impact factor (IF). As a key indicator for detecting paper mill papers, Sabel et al [2] identified the practice of referencing other papers that were likewise generated by paper mills. This aligns with another known service provided by paper mill organizations: boosting the citations of the articles [2]. Hence, the analysis of the references and citations of these papers can be a valuable tool for identifying papers originated from paper mill organizations [12], and even for identifying papers coming from the same paper mill organization.

Research in this field is crucial for developing effective strategies to address the problem and safeguard the integrity of scientific publication. To date, no study has comprehensively analyzed either the references included in papers retracted for having been identified as generated from paper mills, or the citations received by these same papers. Accordingly, the aims of this study were as follows: (1) to analyze the references cited by retracted papers originated from paper mills; (2) to analyze the citations received by retracted papers that originated from paper mills; and (3) to analyze the potential relationships existing between paper mill papers, their references and their citations.

2. Study design and setting

We conducted a cross-sectional study following the Strengthening the Reporting of Observational Studies in Epidemiology guidelines. This study included original articles which have been retracted and identified as having originated from paper mills. To be eligible, papers were required to have been retracted between January 1, 2022, and December 31, 2022, and published a minimum of 12 months before retraction. This criterion was established so that the papers included could have received citations for a minimum period before retraction (minimum citation window). Corrections, expressions of concern, and communications to conferences were excluded. To identify papers that met the pre-established selection criteria, we used the Retraction Watch Database (date last accessed January 23, 2023), selecting the Paper Mill option in the "Reason(s) for Retraction" field, and "Retracted" in the "Nature of Notice" field. The way in which Retraction Watch operates has been previously described [6].

What is new?

Key findings

- This study reveals the potential interconnection of retracted paper mill papers through their references and citations.

What this adds to what was known?

- This is the first analysis of references and citations in retracted papers associated with paper mills.
- Retracted papers originating from paper mills frequently reference and are cited by papers that are later retracted for having originated from paper mills.

What is the implication what should change now?

- Further research is needed on paper mills, their characteristics, and modes of operation.
- The patterns of citation and referencing of paper mills can serve as an indicator for identifying potentially fraudulent publications.

In the case of retracted paper mill papers (hereinafter "source-retracted papers"), the following data were retrieved from Retraction Watch: title; authors; number of authors; author affiliations; authors' country or countries; type of publication; date of publication; date of retraction; journal of publication; and reason for retraction. We used the variables date of publication and date of retraction to calculate the number of days elapsed between publication and retraction. Similarly, we recorded the corresponding author's affiliation and country, based on the full text of the source-retracted paper. From Web of Science, we retrieved the number of references cited by and citations received by each of the source-retracted papers. Insofar as citations were concerned, these could come from other papers or from the source paper's own retraction notice. For analysis purposes, we excluded citations by the source paper's own retraction notice and examined only citations that came from other papers.

The following data were collected from each of the references and citations identified and included, using Web of Science: authors; number of authors; type and country of affiliation of corresponding author; title of paper; journal; journal's IF based on Journal Citation Reports (JCR); category and relative position (quartile) based on JCR; type of publication; and number of citations received. Additionally, we recorded whether the references or citations had been retracted and the reason for retraction, with the latter being retrieved from Retraction Watch. In any case where the reason for retraction was not shown in the Retraction Watch Database, the retraction note was accessed. Papers retracted for duplication/manipulation of images and/or impossibility of communicating with the author, or false peer review were categorized as ''suspected paper mill'', since these are all typical characteristics of these types of papers, as described by different sources [1,3,13].

JCR 2021 was used to ascertain the IF of the journal of publication of the source-retracted papers and of the references and citations, the category, and journal's relative position (quartile) in this category. In the case of journals included in different categories, data relating to the most favorable one were recorded (the one in which the journal was in a higher percentile). The publication modality was extracted from the journal's own website and classified as Open Access (OA) (universal and free access) or non-OA (publication by subscription, or hybrid).

2.1. Statistical analysis

We first performed a descriptive analysis of source-retracted papers identified as being generated by paper mills, with quantitative variables expressed as median and interquartilic range (IQR), and qualitative variables as relative and absolute frequencies. The number of source-retracted papers that referenced and were cited by other retracted papers was compared by quartile of journal of publication and publication modality (OA or non-OA), using the Wilcoxon signed-rank test.

Secondly, we performed a descriptive analysis of the references and citations of source-retracted papers included in this study, by reference to the variables of interest. A bivariate analysis was also performed to assess differences in characteristics, according to whether the item was a reference of or citation to the source-retracted paper. The Wilcoxon signed-rank test was used for quantitative variables and the Chi-squared test for qualitative variables. In a subsequent analysis, the sample of citations and references was restricted to those that had been retracted. The reasons for retracting references and citations were first analyzed descriptively and then compared using the Chi-squared test.

Two networks of source-retracted papers (retracted for being generated by paper mills) inter-related by their citations and references were built. Each node (dot) in the network represented 1 paper (source, reference or citation). The connections between nodes (ties) were established using the *edge betweenness* method.

The first network was built with the aim of establishing groups or clusters of retracted papers that mutually referenced/cited one another. To this end, the network was restricted to analysis of source-retracted papers and their retracted references and citations. This network shows clusters which include more than 5 inter-related elements (ie, source-retracted papers, retracted references and citations that comprise the cluster).

The second network was built by including source-retracted papers and their references and citations,

regardless of whether or not they had been retracted. In this network, references and citations were classified into 3 groups: unretracted; retracted for or being suspected of originating from a paper mill; and retracted for some other reason (authorship issues, plagiarism, error, ethical issues, fabrication/falsification, not specified). Due to the large amount of data, this network includes clusters made up of more than 45 elements (source-retracted papers, citations retracted or unretracted, and references retracted or unretracted).

Both networks in interactive format can be accessed at the following link: https://epideque.shinyapps.io/networks/. To visualize the second network, we lighten it by including a smaller number of articles. We excluded references and citations that were not retracted and that did not connect 2 other articles ($n = 1883$). The complete second network can be seen in the Supplementary Material.

Statistical significance was set at $P < .05$ for all tests conducted. All statistical analyses were performed using the Stata v.17 and R statistical software programs.

3. Results

3.1. Retracted papers originated from paper mills (source-retracted papers)

We identified 416 original papers that originated from paper mills and were retracted in 2022 (source-retracted papers), with a median of 1247 (IQR, 907.8−1673.5) days after publication. Table 1 describes their main characteristics. The source-retracted papers were published in 85 different scientific journals; 39.7% of source-retracted papers were published in journals ranked in the second quartile of their JCR category and 22.1% in journals belonging to the Medicine, Research, and Experimental category. Biomedicine & Pharmacotherapy and Molecular Medicine Reports were the journals which published the highest number of paper mill papers retracted in 2022, with 35 and 33 papers published and then retracted, respectively.

A total of 1669 unique authors were listed on the 416 source-retracted papers: two of these authors appeared in 14 of the papers, while another two were listed in 10 papers each. In 384 of the source-retracted papers (92.3%), all the authors came from the same country (China), whereas in the remaining 32 papers (7.7%), the authors came from two to five different countries. The most frequent affiliation among all the authors identified was the University of Jilin in China, which featured in 77 different sourced retracted papers (18.5%). In terms of the corresponding author's affiliation, in 389 papers (92.8%), this was a Chinese institution, and in 4 papers (1.0%), this was a Chinese institution and an institution from another country (Philippines, Russia, USA, or Italy). In 213 of the sourced retracted papers (51.2%), corresponding authors were exclusively affiliated with a hospital,in 118 (28.4%) were affiliated with

Table 1. Main characteristics of paper mill papers retracted in 2022 (source-retracted papers)

Characteristics	Total ($n = 416$)
Number of authors	
Median (IQR)	5 (4−7)
Number of affiliation countries (all authors)	
Median (IQR)	1 (1−1)
Affiliation institution of corresponding author[c]	
Hospital	213 (51.2%)
Industry	1 (0.2%)
Research center	6 (1.4%)
Research center + Hospital	1 (0.2%)
University	69 (16.6%)
University + Hospital	118 (28.4%)
University + Hospital + Research center	2 (0.5%)
University + Research center	6 (1.4%)
E-mail account[a]	
Institutional	7 (3.7%)
Private + Institutional	8 (4.2%)
Private	175 (92.1%)
Country of corresponding author	
China	386 (92.8%)
Russia	18 (4.3%)
Bangladesh	1 (0.2%)
Bulgaria	1 (0.2%)
China + Philippines	1 (0.2%)
China + Russia	1 (0.2%)
China + USA	1 (0.2%)
China + Germany	1 (0.2%)
China + Italy	1 (0.2%)
Iran	1 (0.2%)
India	1 (0.2%)
Kazakhstan	1 (0.2%)
Saudi Arabia	1 (0.2%)
USA	1 (0.2%)
IF[b]	
Median (IQR)	3.38 (2.31−4.06)
Quartile	
No IF/Not indexed	66 (15.9%)
Q1	81 (19.5%)
Q2	165 (39.7%)
Q3	86 (20.7%)
Q4	18 (4.3%)
Category	
No IF/Not indexed	66 (15.9%)
Medicine, Research & Experimental	92 (22.1%)
Oncology	75 (18.0%)
Pharmacology & Pharmacy	27 (6.5%)
Biochemistry & Molecular Biology	25 (6.0%)

(Continued)

Table 1. Continued

Characteristics	Total (n = 416)
Pathology	22 (5.3%)
Multidisciplinary Sciences	16 (3.8%)
Physiology	13 (3.1%)
Endocrinology & Metabolism	12 (2.9%)
Neurosciences	11 (2.6%)
Cell biology	10 (2.4%)
Other	47 (11.4%)
Publishing model	
Open Access	270 (35.1%)
Non–Open Access	146 (64.9%)

IF, impact factor; IQR, interquartilic range.
[a] Data available for 190 papers.
[b] Data available for 350 papers, as 66 papers were published in nonindexed journals.
[c] Percentages may not total 100 due to rounding.

a hospital and a university, and in 69 (16.6%) were exclusively affiliated with a university. In 7 papers, the corresponding author had provided an institutional e-mail (3.7% of those papers providing information about the e-mail account), and in 8, the corresponding author had provided both an institutional and a private e-mail (4.2%).

3.2. References and citations of source-retracted papers

The 416 source-retracted papers included a total of 15,437 references, and there were 8689 citations to them by other papers. We were able to extract metadata from 14,411 (93.4%) references and 8479 (97.6%) citations. The median number of references per source-retracted paper was 35 (IQR, 29–40), and the median number of citations received was 16 (IQR, 9–25).

Table 2 shows the main characteristics of the references and citations retrieved. Most of the corresponding authors of the references and citations were affiliated with an

Table 2. Main characteristics of references within and citations to paper mill papers retracted in 2022

	References		Citations		
	n = 14,411		n = 8479		P value
Number of authors					
Median (IQR)	6.00	(3.0–8.0)	6.00	(4.0–8.0)	0.27
Country of corresponding author[a]					<0.001
China	5834	(54.0%)	5587	(78.3%)	
USA	3111	(28.8%)	331	(4.6%)	
Japan	456	(4.2%)	88	(1.2%)	
Germany	365	(3.4%)	71	(1.0%)	
South Korea	300	(2.8%)	129	(1.8%)	
Italy	310	(2.9%)	141	(2.0%)	
Iran	113	(1.1%)	334	(4.7%)	
Other	307	(2.8%)	455	(6.4%)	
IF					
Median (IQR)	3.9	(2.5–6.0)	2.9	(1.6–4.2)	<0.001
Journal Quartile based on IF (from JCR)					<0.001
Q1	6749	(46.8%)	2149	(25.3%)	
Q2	3757	(26.1%)	2422	(28.6%)	
Q3	1523	(10.6%)	1589	(18.7%)	
Q4	674	(4.7%)	806	(9.5%)	
Not indexed	1709	(11.9%)	1513	(17.8%)	
Journal category (from JCR)					<0.001
Biochemistry & Molecular Biology	1001	(6.9%)	340	(4.0%)	
Medicine, Research & Experimental	935	(6.5%)	938	(11.1%)	
Multidisciplinary Sciences	803	(5.6%)	244	(2.9%)	
Oncology	2803	(19.5%)	1132	(13.4%)	
Pharmacology & Pharmacy	629	(4.4%)	679	(8.0%)	
Not indexed	1709	(11.9%)	1513	(17.5%)	
Other	6531	(45.3%)	3633	(42.8%)	

IF, impact factor; JCR, Journal Citation Reports; IQR, interquartilic range.
[a] 3616 references and 1343 citations without information.

Table 3. Description of retracted references within and citations to source retracted papers

	References		Citations	
Retracted paper				
No	14,121	(98.0%)	8.167	(96.3%)
Yes	290	(2.0%)	312	(3.7%)
Reason for retraction				
Paper mill	167	(57.6%)	142	(45.5%)
Suspected paper mill	63	(21.7%)	101	(32.4%)
Concerns about data/images	13	(4.5%)	20	(6.4%)
Plagiarism	11	(3.8%)	2	(0.6%)
Error	8	(2.8%)	8	(2.6%)
Fabrication/falsification	3	(1.0%)	1	(0.3%)
Ethical issues	1	(0.3%)	0	(0.0%)
Authorship issues	0	(0.0%)	1	(0.3%)
Not specified	24	(8.3%)	37	(11.9%)

institution in China (54.0% of references and 78.3% of citations). Whereas 46.8% of the references had been published in first-quartile journals, 28.6% of the citations had been published in second-quartile journals. Journals in the oncology category were those in which the majority of references and citations were published. Oncotarget and PLOS One were the most frequently referenced journals in source-retracted papers, with 418 and 344 references, respectively. The International Journal of Molecular Science and Frontiers in Oncology cited the source-retracted papers the most, with 206 and 187 citations each.

Among the 416 source-retracted papers, 169 (41.9%) referenced at least 1 paper that had also been retracted and 178 (42.8%) were cited by at least 1 paper that was also later retracted. Of these, 145 of 169 (81.5%) referenced at least 1 paper retracted for originating from or being suspected of originating from paper mills. Meanwhile, 156 of 178 (87.6%) were cited by at least 1 paper retracted for originating from or being suspected of originating from paper mills. No differences were observed in the number of references or citations originating from or suspected of originating from paper mills, by publication modality or journal quartile of the source article (data not shown).

The reasons for retraction of retracted references and citations (290 and 312, respectively) are shown in Table 3. Among them, we found 473 papers retracted for originating from or being suspected of originating from paper mills. Eight source-retracted papers were identified with 4 or more references/citations to other paper mill papers or suspected paper mill papers.

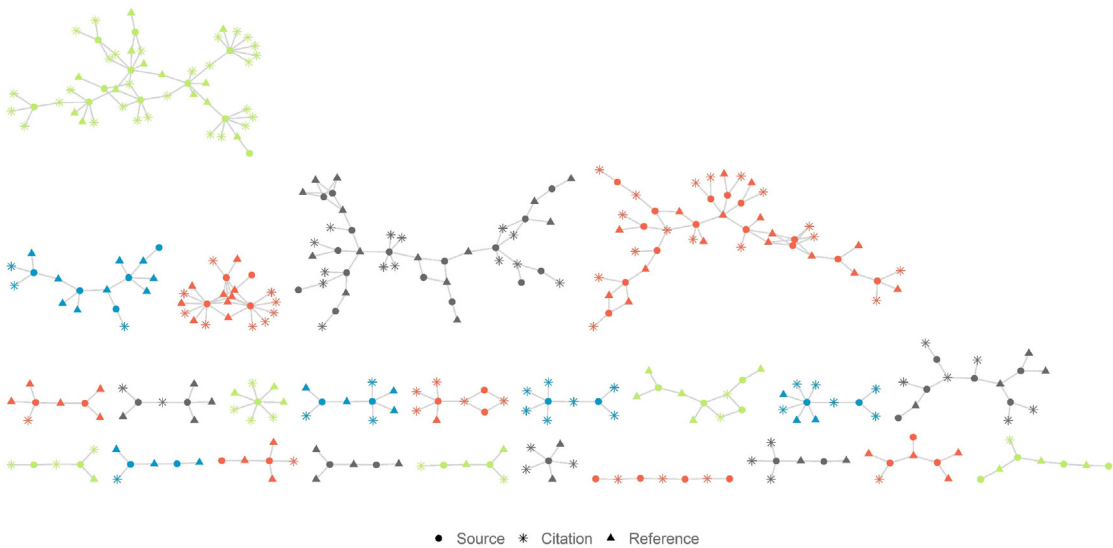

Figure. Network of source-retracted papers interconnected through their retracted references and retracted citations. This figure shows the clusters of related papers. Each cluster is shown in a different color to distinguish it (the color has no other meaning). Within each cluster, there are articles that have been retracted for originating from paper mills (source-retracted papers—all retracted) represented by a circle, articles that are included in the reference list of the source-retracted papers by a triangle (all retracted), and articles that have cited the source-retracted papers by an asterisk (all retracted).

The link (https://epideque.shinyapps.io/networks/) shows the network of papers retracted for any reason, restricted to clusters made up of more than 5 elements inter-related via their citations and references. The network is also showed in Figure: source-retracted papers are represented by a circle, references by a triangle and citations by an asterisk (Fig). Note should be taken of the 3 major clusters formed by 53, 48, and 44 inter-related sourced-retracted papers through their retracted citations and references.

An additional analysis restricted to clusters made up of more than 45 elements, including all references and citations regardless of whether or not they had been retracted, yielded a large cluster of 2530 inter-related papers. This cluster links 174 sourced retracted papers (41.8% of those included), 203 references and citations retracted for originating from or being suspected of originating from paper mills (42.9% of those identified) (Table 3), and 38 references and citations retracted for any other reason (27.3% of those identified). This network can be seen in the link in the Supplementary Material and a reduced version of it in the link provided (https://epideque.shinyapps.io/networks/).

4. Discussion

The results of this study highlight the way in which original papers retracted in 2022 for originating from paper mills are interconnected through their references and citations. Furthermore, the most frequent cause of retraction among the references and citations of original paper mill papers retracted in 2022 is affiliation with such organizations, suggesting a possible manipulation of citations. To our knowledge, this is the first study to analyze the references and citations of retracted papers originated from paper mills.

The included retracted paper mill papers in this study display the typical characteristics of these types of papers, previously described in other studies [2,3,6,7,11]. Most of the authors are from China and are affiliated with hospitals. In addition, it has been observed that only 3.7% of e-mail correspondence provided by the authors is associated with an institution (university, hospital or research center, or other). This characteristic has been previously identified as a "red flag" by the International Committee of Medical Journal Editors and the STM Integrity Hub [1], as well as by previous studies [2,13]. Identification of the typical characteristics of paper mill papers may be crucial for addressing the problem, as they can serve as an early warning signal for editors and reviewers and may therefore prevent their publication.

The results of this study appear to confirm the links among paper mill papers via the papers they reference and the papers that subsequently cite them. According to our results, 41.9% of paper-mill papers referenced at least 1 retracted paper, while 42.8% received citations from retracted papers. This is likely a conservative estimate, as only about 2 out of every 10,000 published articles are retracted [14]. The first network constructed in this study contains only retracted articles. It shows the connections between the articles retracted in 2022 and their references and citations that were also retracted. We observed some cases where several articles retracted in 2022 are related because they have a retracted citation or reference in common. It is noticeable that most of these retracted citations and references are also retracted for originating from paper mills. This suggests that articles from paper mills cite and reference each other. As stated previously, these organizations do not only mass-produce manuscripts, but increase their visibility through citations and references, which also provides very positive feedback to their ‘‘clients’’ through increasing their curriculum vitae and their Hirsch index.

However, the first network has a limitation: not all articles that should be retracted are retracted. It is therefore important to note that references and citations that have not been retracted do not necessarily mean that they do not come from paper mills, but rather that they may not yet have been identified as such. Restricting the network to establishing relationships between those items that have been identified (retracted) as such can provide a limited point of view. With this in mind, we created the second network showing the source-retracted papers, with all their references and citations (bearing in mind that some of the papers that have not been retracted are also likely to be from mills but have not yet been identified). This network contains a large cluster with more than 2500 publications, that includes 174 source-retracted papers plus their references and citations, of which 203 have been retracted because they originated or were suspected to have originated from paper mills. This suggests the existence of a relationship between these source-retracted papers and the references and citations retracted for originating from paper mills, and it is plausible that they may in fact originate from the same paper mill organization. It is logical to think that papers from the same paper mill organization would tend to cite each other. It is important to be aware of the limitations of this second network. While it gives us a broader perspective, it may also introduce noise into the data by including references and citations that are not necessarily related to unethical practices.

Our findings are in line with previous studies. Sabel et al [2] fitted 2 models for detecting fraudulent publications generated by paper mills. In the first model, they included the variables, ‘‘noninstitutional email’’ and ‘‘affiliation with a hospital’’, though the percentage of false positives was too high. In the second model, they added the variable, ‘‘$>10\%$ of references to other paper-mill papers’’, and the percentage of false positives fell from 44% to 37%, while sensitivity increased. Although the difference is small, this could suggest that the presence of references to other paper mill papers may improve algorithms

designed to detect such papers. In addition, an article concluded that articles citing retracted articles are more likely to be retracted and states that coming from a paper mill organization was one of the main reasons for retraction of these papers [15].

The main implication of this study is the contribution of the results to the detection of paper mills. Based on the evidence generated, it might be prudent to consider the presence of a retracted reference originating from a paper mill or being cited by a paper mill paper as a "red flag''. These can be important indicators or warning signals for scientific journals and must be borne in mind during the review process. Identification of such indicators may help prevent the publication of low-quality papers or papers that adopt ethically questionable practices. This indicator could be incorporated as a part of a checklist used to evaluate the trustworthiness of papers before they're published. In addition to the above indicator, the checklist should include other typical characteristics associated with paper mills, which have been previously identified in the scientific literature. In the same way as there are checklists to verify whether or not a journal is potentially predatory, such as ''Think.Check.Submit'' [16]; among others [17], this new checklist could help journals by improving the review procedures for scientific papers that are submitted to them for publication. Additionally, considering both references and citations can help to identify paper mill papers that have already been published.

A potential solution to identify paper mill papers more effectively could be the development of an algorithm, probably based on artificial intelligence, that assigns a probability percentage indicating the likelihood of an article being associated with a paper mill. While previous attempts had limitations [2], the STM Integrity Hub has been working on a tool to detect paper mill involvement in submitted manuscripts [18]. A minimum viable product of this tool, launched in April 2023 [19], is initially available only to STM Integrity Hub editors [20]. It remains unclear whether the tool has the capability to assign a percentage probability, which could significantly enhance its utility by allowing journals to establish "risk thresholds" and allocate resources more efficiently.

This study has some limitations. First, the networks analyzed were exclusively focused on references and citations of source-retracted papers, without including the references and citations of ''secondary papers'' (ie, the citations and references of source-retracted papers). This could underestimate the relationship of the different papers included. In future studies, consideration should be given to the inclusion of both directions in the network of references and citations, so as to form a more complete picture. Second, while metadata were successfully extracted from approximately 95% of the citations and references of source-retracted papers, there is a small percentage of missing data, though this is considered minimal. Third, the protocol and analysis plan were not previously published. An additional limitation is the absence of a control group. Future studies should consider the inclusion of a control group of unretracted papers, with the aim of evaluating whether the results differ. Furthermore, to allow for minimum citation window, this study included source-retracted papers having a minimum period of 12 months after retraction, but this period may possibly prove to be insufficient. Lastly, citations of the papers retracted during 2022 included in this study are likely to continue to increase with time.

This study has a number of advantages. The use of an exhaustive database such as Retraction Watch enabled all retractions that occurred in 2022 to be systematically included. To our knowledge, Retraction Watch is the most comprehensive retraction database, though the possibility of there being some missing data cannot be ruled out, which may limit the generalization of our findings. A further important advantage is the longitudinal nature of the study, in the sense that the source-retracted article (paper retracted for being generated by a paper mill) includes its references (previously published papers) as well as citations of the source article (subsequently published papers), which allows for a better temporal approach to the cluster of papers generated by a single paper mill. Lastly, we feel that the visual representation system used is very useful for understanding the structures of references and citations and may be of value for future studies on this topic.

In conclusion, the results of this study demonstrate that papers originating from paper mills cite and reference other papers originating from paper mills. The presence of these relationships in the citations and references may provide useful clues for identifying possible fraudulent publications. This novel methodology could be employed to identify articles potentially associated with paper mills. Early detection of paper mill papers is crucial for preserving the integrity of science. In the face of the proliferation of these types of papers, which will inevitably increase with time, editorial teams must put robust mechanisms in place for their detection, without in any way neglecting the supervision of papers which are suspected of originating from paper mills and have already been published.

CRediT authorship contribution statement

Cristina Candal-Pedreira: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Carla Guerra-Tort:** Writing – review & editing, Software, Methodology, Data curation. **Alberto Ruano-Ravina:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Fabián Freijedo-Farinas:** Writing – review & editing, Methodology, Data curation. **Julia Rey-Brandariz:** Writing – review & editing, Data curation. **Joseph S. Ross:** Writing – review

& editing, Methodology. **Mónica Pérez-Ríos:** Writing − review & editing, Methodology.

Data availability

Part of the data that support the findings of this study are available from Retraction Watch. Restrictions apply to the availability of these data, which were used under contract license for this study. If required, the data will be made available on request.

Declaration of competing interest

Dr Ross currently receives research support through Yale University from Johnson and Johnson to develop methods of clinical trial data sharing, from the Medical Device Innovation Consortium as part of the National Evaluation System for Health Technology, from the Food and Drug Administration for the Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation (CERSI) program (U01FD005938), from the Agency for Healthcare Research and Quality (R01HS022882), from the National Heart, Lung and Blood Institute of the National Institutes of Health (NIH) (R01HS025164, R01HL144644), and from Arnold Ventures; in addition, Dr Ross is an expert witness at the request of Relator's attorneys, the Greene Law Firm, in a qui tam suit alleging violations of the False Claims Act and Anti-Kickback Statute against Biogen Inc and is a Deputy Editor at the Journal of the American Medical Association. There are no competing interests for any other author.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111397.

References

[1] COPE & STM. Paper Mills - Research report from COPE & STM - English 2022.

[2] Sabel BA, Knaack E, Gigerenzer G, Bilc M. Fake publications in biomedical science: red-flagging method indicates mass production. medRxiv 2023. https://doi.org/10.1101/2023.05.06.23289563.

[3] Perez-Neri I, Pineda C, Sandoval H. Threats to scholarly research integrity arising from paper mills: a rapid scoping review. Clin Rheumatol 2022;41(7):2241−8.

[4] Sabel BA, Seifert R. How criminal science publishing gangs damage the genesis of knowledge and technology-a call to action to restore trust. Naunyn-Schmiedeberg's Arch Pharmacol 2021;394:2147−51.

[5] Byrne JA, Christopher J. Digital magic, or the dark arts of the 21(st) century-how can journals and peer reviewers detect manuscripts and publications from paper mills? FEBS Lett 2020;594(4):583−9.

[6] Candal-Pedreira C, Ross JS, Ruano-Ravina A, Egilman DS, Fernandez E, Perez-Rios M. Retracted papers originating from paper mills: cross sectional study. BMJ 2022;379:e071517.

[7] Else H, Van Noorden R. The fight against fake-paper factories that churn out sham science. Nature 2021;591:516−9.

[8] Van Noorden R. How big is science's fake-paper problem?. Nature; 2023. Available at: https://www.nature.com/articles/d41586-023-03464-x. Accessed June 11, 2024.

[9] Holst F. Increasing confidence and trust in research: cracking down on misconduct IOP Publishing. 2022. Available at: https://ioppublishing.org/news/increasing-confidence-and-trust-in-research/. Accessed August 14, 2023.

[10] González-Márquez R, Schmidt L, Schmidt BM, Berens P, Kobak D. The landscape of biomedical research. bioRxiv 2023. https://doi.org/10.1101/2023.04.10.536208.

[11] Abalkina A. Publication and collaboration anomalies in academic papers originating from a paper mill: evidence from Russia. arXiv 2021. https://doi.org/10.48550/arXiv.2112.13322.

[12] Kincaid E. Editorial board member dropped from journal site after Retraction Watch-Undark report links him to paper mill.: Retraction Watch. 2023. Available at: https://retractionwatch.com/2023/07/11/editorial-board-member-dropped-from-journal-site-after-retraction-watch-undark-report-links-him-to-paper-mill/. Accessed August 14, 2023.

[13] Seifert R. How Naunyn-Schmiedeberg's Archives of Pharmacology deals with fraudulent papers from paper mills. Naunyn-Schmiedeberg's Arch Pharmacol 2021;394:431−6.

[14] Brainad J, You J. What a massive database of retracted papers reveals about science publishing's 'death penalty'. Science; 2018. Available at: https://www.science.org/content/article/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty. Accessed June 11, 2024.

[15] Zhu H, Jia Y, Leung SW. Citations of microRNA biomarker articles that were retracted: a systematic review. JAMA Netw Open 2024;7(3):e243173.

[16] Think.Check.Submit. Available at: https://thinkchecksubmit.org. Accessed September 10, 2023.

[17] Cukier S, Helal L, Rice DB, Pupkaite J, Ahmadzai N, Wilson M, et al. Checklists to detect potential predatory biomedical journals: a systematic review. BMC Med 2020;18(1):104.

[18] Else H. Paper-mill detector put to the test in push to stamp out fake science. Nature; 2022. Available at: https://www.nature.com/articles/d41586-022-04245-8. Accessed September 10, 2023.

[19] STMPublishing News. STM Solutions releases MVP of new paper mill detection tool. 2023. Available at: https://www.stm-publishing.com/stm-solutions-releases-mvp-of-new-paper-mill-detection-tool/.

[20] Heck S, Bianchini F, Souren NY, Wilhelm C, Ohl Y, Plass C. Fake data, paper mills, and their authors: the International Journal of Cancer reacts to this threat to scientific integrity. Int J Cancer 2021;149:492−3.