medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Fake Publications in Biomedical Science: Red-flagging Method Indicates Mass Production

Bernhard A. Sabel[1*], Ph.D., Emely Knaack[1], Gerd Gigerenzer[2], Ph.D., Mirela Bilc[1], Ph.D.

1. Institute of Medical Psychology, Medical Faculty, Otto-von-Guericke University of Magdeburg, Magdeburg, Germany and Center for Brain and Behavioral Sciences (CBBS)

2. Max-Planck Institute for Human Development, Berlin, Germany

*corresponding author: Bernhard A. Sabel, Ph.D., Institute of Medical Psychology, Medical Faculty, Otto-von-Guericke University of Magdeburg, Leipziger Straße 44, 39120, Magdeburg, Germany; E-Mail: bernhard.sabel@med.ovgu.de

ABSTRACT

Background: Integrity of academic publishing is increasingly undermined by fake science publications massively produced by commercial "editing services" (so-called "paper mills"). They use AI-supported, automated production techniques at scale and sell fake publications to students, scientists, and physicians under pressure to advance their careers. Because the scale of fake publications in biomedicine is unknown, we developed a simple method to red-flag them and estimate their number.

Methods: To identify indicators able to red-flagged fake publications (RFPs), we sent questionnaires to authors. Based on author responses, three *indicators* were identified: "author´s private email", "international co-author" and "hospital affiliation". These were used to analyze 15,120 PubMed®-listed publications regarding date, journal, impact factor, and country of author and validated in a sample of 400 known fakes and 400 matched presumed non-fakes using classification (tallying) rules to red-flag potential fakes. For a subsample of 80 papers we used an additional indicator related to the percentage of RFP citations.

Results: The classification rules using two (three) indicators had sensitivities of 86% (90%) and false alarm rates of 44% (37%). From 2010 to 2020 the RFP rate increased from 16% to 28%. Given the 1.3 million biomedical Scimago-listed publications in 2020, we estimate the scope of >300,000 RFPs annually. Countries with the highest RFP proportion are Russia, Turkey, China, Egypt, and India (39%-48%), with China, in absolute terms, as the largest contributor of all RFPs (55%).

Conclusions: Potential fake publications can be red-flagged using simple-to-use, validated classification rules to earmark them for subsequent scrutiny. RFP rates are increasing, suggesting higher actual fake rates than previously reported. The scale and proliferation of fake publications in biomedicine can damage trust in science, endanger public health, and impact economic spending and security. Easy-to-apply fake detection methods, as proposed here, or more complex automated methods can help prevent further damage to the permanent scientific record and enable the retraction of fake publications at scale.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

INTRODUCTION

Trust in the integrity of academic publishing is a foundation of science, and lack of it damages its reputation (Behl, 2021; Byrne, 2019; Seifert, 2021; Else & Van Norden, 2021; Else, 2022; Byrne et al., 2022). Well-known cases of scientific misconduct by individual researchers include ghost and "honorary" authorships (Flanagin et al., 1998; Wislar et al., 2011; Frederickson and Herzog, 2021), cherry-picking, abstract spin, plagiarism of images (Bik et al., 2016), and outright data fabrication (Bik, 2020; Byrne and Christopher, 2020; Park et al., 2022). While individual fraud has been recognized for centuries, the recent emergence of commercial production of fake publications is a new and unprecedented development (Flanagin et al, 1998; Wislar et al., 2011; Mavrogenis et al., 2018; Byrne, 2019; Byrne and Christopher, 2020; Else and Van Norden, 2021; Sabel and Seifert, 2021; Sabel, 2022; Chawla Singh, 2022; Candal-Pedreira et al., 2022). The major source of fake publications are 1,000+ "academic support" agencies – so-called "paper mills" – located mainly in China, India, Russia, UK, and USA (Abalkina, 2021; Else, 2021; Pérez-Neri et al., 2022). Paper mills advertise writing and editing services via the internet and charge hefty fees to produce and publish fake articles in journals listed in the *Science Citation Index* (SCI) (Christopher, 2021; Else, 2022). Their services include manuscript production based on fabricated data, figures, tables, and text semi-automatically generated using artificial intelligence (AI). Manuscripts are subsequently edited by an army of scientifically trained professionals and ghostwriters. Although their quality is relatively low (Cabanac and Labbé, 2021), fake publications nevertheless often pass peer review in established journals with low to medium impact factors (IF 1-6) (Seifert, 2021). Some governments, funding bodies, and academic publishers are on the alert (Cyranoski, 2018; Mallapaty, 2020; Else, 2022; Candal-Pedreira et al., 2022), yet many scientists, journal editors, and learned societies appear to be surprisingly unaware that such publications exist at all.

Paper mill customers – students and scientists – are pressured to publish in SCI publications by their academic or government institutions or university-affiliated hospitals (Pérez-Neri et al., 2022). For example, the Beijing municipal health authorities require a fixed number of first-authored SCI articles for physicians to qualify for promotion (Else and Van Norden, 2021). Academic policies that count publications and value impact factors as surrogates for scientific excellence can force graduate students to fulfill SCI publication requirements and pressure scientists and physicians to meet publication quotas to attain salary increases, promotion, and/or scientific reputation. Paper mills are ready to help and offer their services to accomplish these goals.

We are aware of several instances where paper mills tried to promote their business by inviting journal editors to collaborate, as shown by this unsolicited email in 2022 from a paper mill to one of us who is editor of a biomedical journal (see Tab. 1 for interview excerpts):

> We are a well-known academic support institution from Guangzhou, China, which has been established for 8 years. … For reducing the publication time, we expect to cooperate with you in the future. Cooperation mode: we cite the content of your journal in our articles, thus increasing … your impact factor in 2022. You shall help us shorten the publication time. Payment: If an article is successfully published, we will pay for it at the price: IF*1,000 USD/article. For example, with

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

IF=2.36, total payment=2.36*1,000 USD=2,360 USD. And this price is negotiable.

This attempted corruption motivated us to quantitatively analyze the global scope of fake publishing. Because the problem is still perceived to be small (an estimated 1 of 10,000 publications, Tab. 2), publishers and learned societies are just beginning to adjust editorial, peer-review, and publishing procedures. Yet the actual scale of fake publishing remains unknown, despite the fact that the number of reports on paper mills are increasing.

To be able to estimate the scope of fake publishing, a method is needed to identify potential (red-flag) fake publications (RFPs). We therefore looked for potential indicators of fakes that are easy to use by reviewers, editors, and publishers and tested their feasibility for randomly selected neuroscience and medical journals. We then developed classification (tallying) rules for screening for potential fakes and determined their sensitivity and false alarm rates. Here, we report the proportion of RFPs as an upper estimate of the true number of fake publications.

METHODS

Exploration: To search for potential fake indicators, in Study 1, one of us (editor of a neurology journal) sent a questionnaire to the corresponding authors of a sample of suspicious published articles and, for control, to those of a sample of unsuspicious articles (see Table 3). Based on the differential willingness to respond, we identified potential indicators for fake publications. These indicators should satisfy the following patterns (see below):

(i) Authors of fake publications are reluctant to provide critical information as revealed by their response – or non-response – to the questionnaire by the editor,

(ii) the number of fake publications increases steadily over time, and

(iii) journals with a low to medium impact factor are most affected.

For Studies 1 to 6 we identified two easy-to-detect indicators, where a publication was labelled as RFP: if an author used a private email and had no international partner. We then tested these indicators in studies with 15,120 publications listed in PubMed® and Web of Science™ in the fields of neuroscience (including neurology) and (non-neuroscience) general medicine publications in an iterative way in a series of nine bibliographic studies (Fig. 1). Study 2 applied potential indicators to five neuroscience publications; Study 3 expanded this sample to estimate RFP growth from 2010-2020, including five journals in the field of general medicine. To estimate the 2020 incidence of RFPs, we increased the sample size in Studies 4 and 5. Finally, Study 6 checked the RFP rate in three open access journals.

Validation: Based on the potential indicators identified in Studies 1-6, in Study 7 we applied tallying rules to the same three indicators to a sample of publications (n=400) which have been proven fake because of fake gene sequences…., text or image plagiarism, or retractions (for retractions it was not possible to differentiate whether they were voluntary or forced):

(i) n=100 retractions (http://retractiondatabase.org/RetractionSearch.aspx?)
(ii) n=100 Tadpole paper mill items (https://docs.google.com/spreadsheets/d/1KXqTAyl4j-jVorFPMD2XRpr76LcIKJ0CVyIvRj0exYQ/edit#gid=0)
(iii) n=100 fake gene sequences; https://dbrech.irit.fr/pls/apex/f?p=9999:28
(iv) n=100 retractions from *Journal of Cellular Biochemistry* (Behl, 2021).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

These were matched with 400 papers presumed non-fake sampled from the same journals by selecting a fake publication's nearest-neighbor article (unless it was also a known fake). A limitation of this "matched sample" method is that we cannot be certain whether the publications presumed non-fake may not actually be fake as well. To test whether the efficacy of RFP detection can be increased, we explored in Study 8 citations of fake papers as a potential indicator by counting the number of RFPs in each reference list of 80 publications (40 proven fake and 40 presumed non-fake) (i.e., 2,594 additional publications were analyzed). For our subsequent sensitivity/ false-alarm rate analysis, we define a 10% RFP citation rate ("RFP citations in reference list >10%") and this as an additional (third) indicator.

Estimating RFP incidence: To estimate the rate of RFPs during the period of 2010-2020, data from Study 3 was analyzed bi-annually for medicine and neuroscience separately. Finally, in Study 9, RFPs were counted in neuroscience and medical journals to establish their within-country and across-countries (global) 2020 incidence.

The quantitative analysis was complemented by a qualitative one, where we analyzed >1,000 websites retrieved from Baidu and Google that advertise various editing services (search terms: "SCI-publication or –service," "essay writing service," "journal writing service," "SCI ghost writing"). We also interviewed a manager of a paper mill by email and in a subsequent (recorded) Zoom meeting (Tab. 1).

RESULTS

EXPLORATION OF FAKE INDICATORS

We searched for indicators that can be determined easily, quickly, and reliably by an editor on the basis of a submitted manuscript or publication alone. In an exploration phase, our search was guided by three hypotheses:

Hypothesis 1: Authors of fake publications will be unwilling to answer quality check surveys and provide original data. In Study 1, n=215 neurology articles were manually inspected by an experienced editor; 20.5% (n=44) were deemed suspicious. A questionnaire was sent to all authors and, for control, to 48 authors of non-suspicious papers. It contained questions that authors of fake papers might be reluctant to answer (e.g., "Are you willing to provide original data?" [only 1 author of 44 suspicious articles did] and "Did you engage a professional agency to help write your paper? [none did]; see Tab. 3). Despite repeated reminders with a warning that failure to reply – or replying inadequately – could trigger retraction, the response rate among suspected authors was only 45.4% (20/44) compared with 95.8% (46/48) for the control group. This survey provided the first indicators of red-flagged fake publications (RFP).

Hypothesis 2: Because paper mills are on the rise (Else and Van Norden, 2021), indicators uncovered in Study 1, if valid, should also increase each year. Study 2 analyzed the frequency of these indicators in five randomly chosen neuroscience journals, expanded in Study 3 to a larger sample of articles from those five neuroscience journals and an additional five medical journals bi-annually (2010-2020). The results show a rapid growth of RFPs over time in neuroscience (13.4% to 33.7%) and a somewhat smaller and more recent increase in medicine (19.4% to 24%) (Fig. 2). A cause of the greater rise of neuroscience RFPs may be that fake experiments (biochemistry, *in vitro* and *in vivo* animal studies) in basic science are easier to generate because they do not require clinical trial ethics approval by regulatory authorities.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Hypothesis 3: Because it is easier for paper mills to market their papers to journals with lower impact factors (IF; range 1-6) journals, our indicators, if valid, should also occur more frequently in such journals. Study 4 tested our indicators in an even larger sample of randomly selected journals included in the Neuroscience Peer Review Consortium. It red-flagged 366/3,500 (10.5%) potential fakes. In Study 5, RFPs were counted in 10 randomly chosen general (non-neuroscience) medical journals, where the 2020 rate was 23.8% in PubMed-listed journals with lower IF. Study 6 revealed that the RFP rate is even greater in open access (OA) journals (112/300 =40.3%) because it is often easier to publish there.

VALIDATION OF FAKE INDICATORS

To validate the fake indicators obtained in exploration studies 1-6, we computed sensitivity/false alarm rates by comparing a sample of known fakes (n=400) with a sample of presumed non-fakes (n=400). Then we combined the two best indicators ("author´s private email" and "hospital affiliation") to form a classification (tallying) rule: "If both indicators are present, classify as a potential fake, otherwise not" (the "AND" rule) (Katsikopoulos et al., 2020). Its detection sensitivity was 0.86 and the false-alarm rate 0.44. An "OR" classification rule ("If any of the indicators are present, classify as fake, otherwise non-fake") had a higher sensitivity (0.972) but also a high false-alarm rate (0.655). To explore possible improvements to our method for future studies, we added a third indicator ("RFP citations in reference list >10%") to the "AND" rule and tested it with n=80 publications in Study 8. This increased the sensitivity to 0.90 and reduced the false-alarm rate to 0.37.

Note that the tallying rule identifies likely fakes, but it cannot determine with certainty whether a given publication is actually (legally) a fake. Nevertheless, it is a reliable tool to red-flag scientific reports for further analysis and is a rational basis to estimate the upper value of fake publishing in biomedicine. Detecting fake papers in disciplines outside of biomedicine may of course require other indicators.

ESTIMATING THE INCIDENCE OF POTENTIAL FAKE PUBLICATIONS

The estimated incidence of RFPs in the Study 9 sample was 589/4,001, of which 328 (55.8%) were from China (Fig. 3; Tab. 4). The within-country percentages of RFPs vary considerably. Leading countries are Russia (48.3%), Turkey (47.5%), China (43.9%), Egypt (40.0%), and India (38.8%), with China – in absolute terms – as the largest contributor globally (55.8%). Note the large differences in fake rate between countries, even between neighboring countries with similar histories such as Russia (48.3%) and Ukraine (3.1%), probably due to different government policies (Fig. 3).

Given the 2020 global publication output of 1.33 million publications (Scimago) and an average of 28.8% RFPs in both fields of Study 3 (but 23.8% in Study 5), the 2020 RFP-incidence is approximately 383,000.

Assuming an average $10,000 price tag for a fake publication, the estimated annual revenue of paper mills is up to $3-4 billion. This revenue does not include non-Scimago journals (classified as "predatory journals"), which are probably more polluted by fakes, nor the open access and publication fees charged by academic publishers (approx. = $1 billion or an estimated 25% of paper mill revenue). Although the incidence of actual fake publications is expected to be smaller than that of RFPs if the number of false alarms exceeds that of

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

missed actual fakes, the overall estimate of the annual number of potential fake publications is considerable (>300,000), and it is on the rise.

QUALITATIVE ANALYSIS OF PAPER MILL STRATEGIES

A search on Baidu and Google uncovered that more than 1,000 paper mills openly advertise their services to "help prepare" academic term papers, dissertations, and articles intended for SCI publications. Most paper mills are located in China, India, UK, and USA, and some are multinational. These typically appear to use sophisticated, state-of-the-art AI-supported text generation, data and statistical manipulation and fabrication technologies, image and text pirating, and gift or purchased authorships. Paper mills fully prepare – and some guarantee – publication in an SCI journal and charge hefty fees ($1,000-$25,000; in Russia: $5,000) (Chawla, 2022) depending on the specific services ordered (topic, impact factor of target journal, with/without faking data by fake "experimentation"). An unsolicited meeting with a paper mill provided a rare and authentic inside view of their business practices (Tab. 1).

Paper mills employ science graduates, academicians, and (sometimes naïve) scientific consultants for editorial help who work in countries with high English aptitude (UK, USA, India). They also offer "rewards" (bribes) to editors for publishing their fabrications (Tab.1). We know of at least 12 such cases (two reported by editors, 10 acknowledged by an academic publisher who asked not to be identified). Here, editors were offered payment for each publication and were lured by a "citation booster" whereby paper mills offered to cite the "friendly" journal in their other fake articles. Although we do not know how many editors have received or accepted such bribes, it is an unprecedented and disturbing fraud-for-profit corruption of scholarly publishing.

DISCUSSION

The dramatic rise of fake science publishing is driven by an unscrupulously corrupt – and increasingly successful – paper mill industry responsible for an estimated 380,000 RFPs annually as of 2020. This indicates that the number of actual fakes is likely higher than currently known (Tab. 2). A 2020 estimate of 28.8% of RFPs in the present study indicates that the 2011 estimate of 0.1% fake publications in China (Hu and Wu, 2013) and the 1% listed in Table 1 is too low, and is closer to the 21-32% "honorary" and ghost authorship cases in biomedicine (Flanagin et al., 1998; Wislar et al., 2011) and the 5%-10% reported in a pharmacology (Seifert, 2021) and a cancer journal (Heck et al., 2021). It feeds a billion-dollar global industry, magnitudes higher than the $4.5 million monetary value estimated in 2011 (Hu and Wu, 2013).

Fake science publishing is known to originate mainly from China (Hu and Wu, 2013; Lei and Zhang, 2018; Mallapaty, 2020; Schneider, 2021), India (Elango, 2021), and Russia (Abalkina, 2021), and, as we showed, it has evolved into a rapidly growing industry of fake science publishing. Our analysis confirms the existence, continuous growth, and notable scope of fake publishing, with most red-flagged publications coming from China (55.8% in 2020).

The rapid rise of the fake science industry is driven by SCI publication pressure on scientists, who are tempted to use paper mills that offer ghostwriting services at $1,000 to25,000 per publication, a profitable business model (Hu & Wu, 2013). Academic publishers acknowledge that the problem exists and are beginning to explore detection tools (Else, 2022; see also COPE & STM Committee on Publication Ethics, 2022). Chinese authorities, although aware of the situation (Cyranoski, 2018), have not yet resolved the problem. If quantity

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

of scientific output is the index for becoming the world leader in science, then paper mills contribute to reaching this goal. In fact, China has almost caught up with the US in publication output (see: www.nature.com/nature-index/country-territory-research-output).

Paper mills feed on the rising administrative practice to evaluate researchers mainly by the "publish-or-perish" criteria of counting papers and journal impact factors as a surrogate for evaluating actual research quality and content (Van Dalen and Henkens, 2012; Candal-Pedreira et al., 2022).

Fake academic publishing is a major driver of global science publishing growth and a growing problem for medical practice. For example, Byrne showed that 712 problematic papers were cited >17,000 times and estimated that about one quarter of them may misinform future development of human therapies (Park et al., 2022). Preclinical studies at biotech company Amgen, for example, could replicate the results of only 6/53 "landmark" articles, and at Bayer, only 14/67 were replicable in oncology, women's health, and cardiovascular medicine. The "replication crisis" slows down the development of life-saving therapies with an estimated financial loss of $28 billion annually by the pharmaceutical industry (Gigerenzer, 2018). Another example of how scientific fraud can affect medical practice is the report by Avenell et al. (2019). After assessing the citations of 12 retracted clinical trial reports in 68 systematic reviews, meta-analyses, guidelines, and clinical trials, they concluded that 13 out of the 68 reviews would likely have to change their conclusions if the retracted publications were removed.

It is important to keep in mind that our indicators provide a red flag, not legal proof, that a given manuscript or publication might be fake. However, it is the authors´ burden of proof to demonstrate that their science can be trusted. Whether this type of scientific misconduct is a conspiracy to commit injurious falsehood or a crime is for others to decide.

Fake science publishing is possibly the biggest science scam of all times, wasting financial resources, slowing down medical progress, and possibly endangering lives. The damage already done is unknown, and a realistic impact assessment of fake science is not yet available. The emergence of Chat-GPT and more sophisticated large language models might amplify the production of fake papers at less cost, although it will be difficult for paper mills to invalidate the indicators identified in this study (although deception via fake institutional email addresses also occurs).

Halting this development requires an immediate response. But what can be done? First, our simple detection tallying method can be used by reviewers and editors to red-flag potential fakes with or without additional indicators. Second, the academic community should consider revising its common practice to judge scientists´ productivity mostly (or solely) on surrogate quantitative criteria (publication numbers, H-factors, citation metrics, etc.) and instead evaluate the quality and relevance of their research (Van Dalen and Henkens, 2012). The European Research Council (ERC) has already taken a first step by asking researchers to refrain from listing impact factors in their applications, consistent with the San Francisco Declaration on Research Assessment (DORA). Thirdly, we need an advanced system to check scientific integrity, independent of academic publishers. Finally, learned societies, funding agencies, and governmental bodies should consider sanctioning fake polluted journals and their publishers.

Until science publishing fraud is largely eradicated, the collateral damage of fake science poses the risk that scientific analyses, experiments, and clinical trials will more likely fail, public health information will be less accurate or (intentionally) misleading, and presumably effective and safe therapies may not deliver what was promised. It also runs the risk that the public loses its trust in the honesty of science itself. Simple detection of fake

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

publications, as proposed here, or more complex automated methods can help prevent further damage to the permanent scientific record and enable the retraction of fake publications at scale. We propose a "call to action" to restore the integrity of our global knowledge base in biomedicine, science, and technology.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

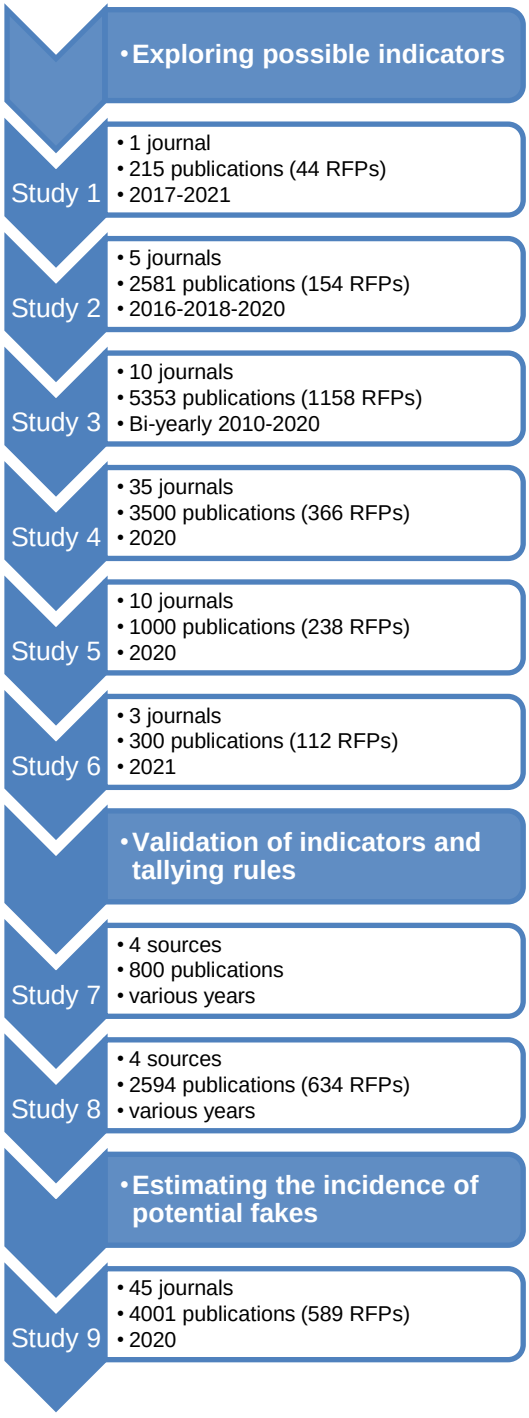

Figure 1. Screening plan for journals and publications. Studies 1-6 identified potential indicators of fake publications on the basis of hypotheses 1 to 3 (see text), Studies 7-8 validated the indicators and developed a classification (tallying) rule, and Study 9 estimated the incidence of potential fake publications on the basis of the tallying rule from Study 7. Of note, some publications were used in several screening studies as follows: S2/S1 n=91; S4/S2 n=461; S3/S2+S4 n=707; S9/S4+S5=4001. The total number of unique publications analyzed is n=15,120.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

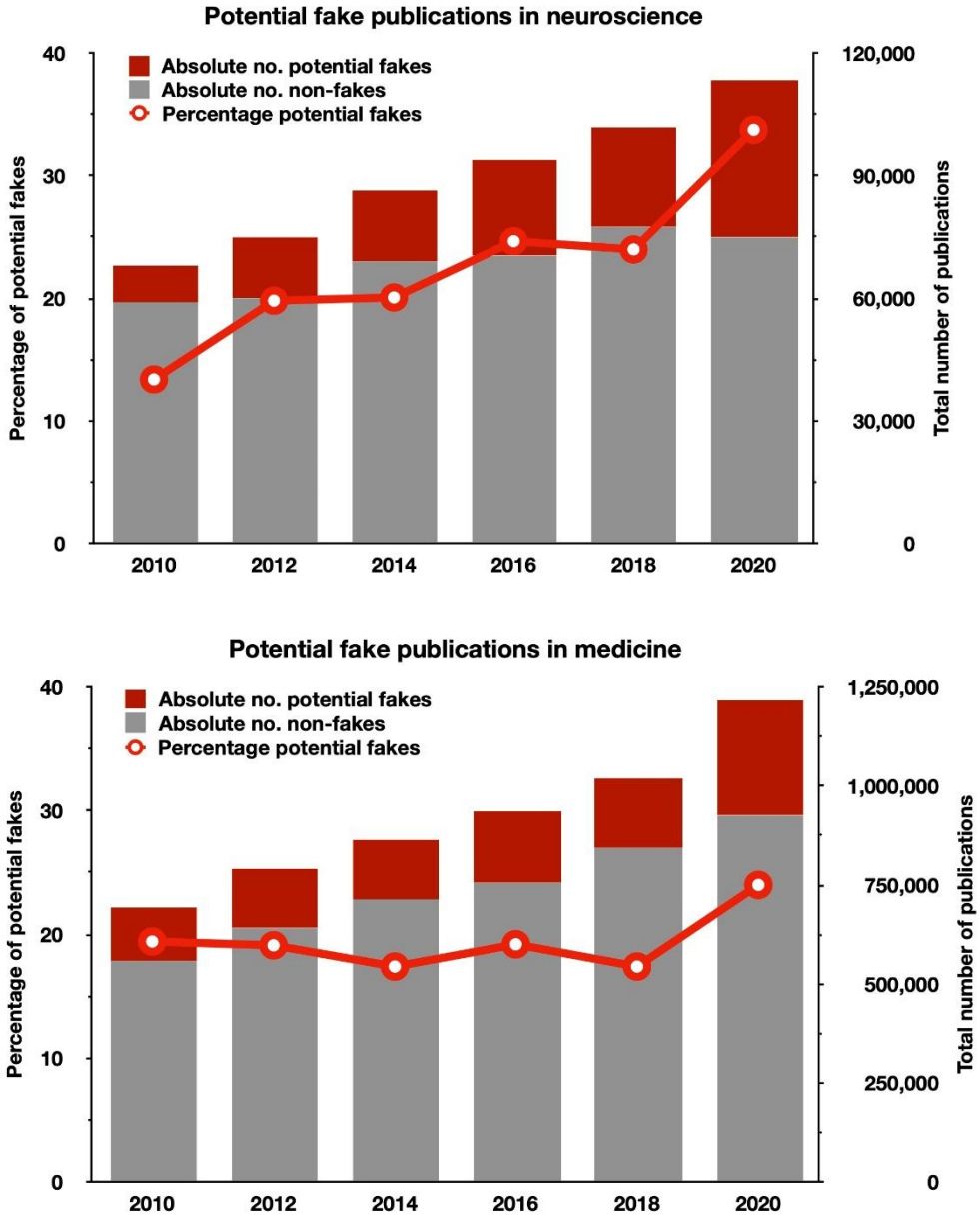

Figure 2. The rise of red-flagged fake publications (RFPs). Number of publications and percentage of red-flagged fake publications in medicine (top) and neuroscience (bottom) (Study 3). The red line (y1-axis) shows the percentage of red-flagged publications per year; the grey bars (y2-axis) show the estimates of non-fakes and the red bars those of potential fake publications, extrapolated to overall articles published in neuroscience (selected from journals of the Neuroscience Peer Review Consortium) and medicine (selected from scimagojr.com). Note the relatively slow increase in potential fakes in medicine and the rapid rise in neuroscience.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

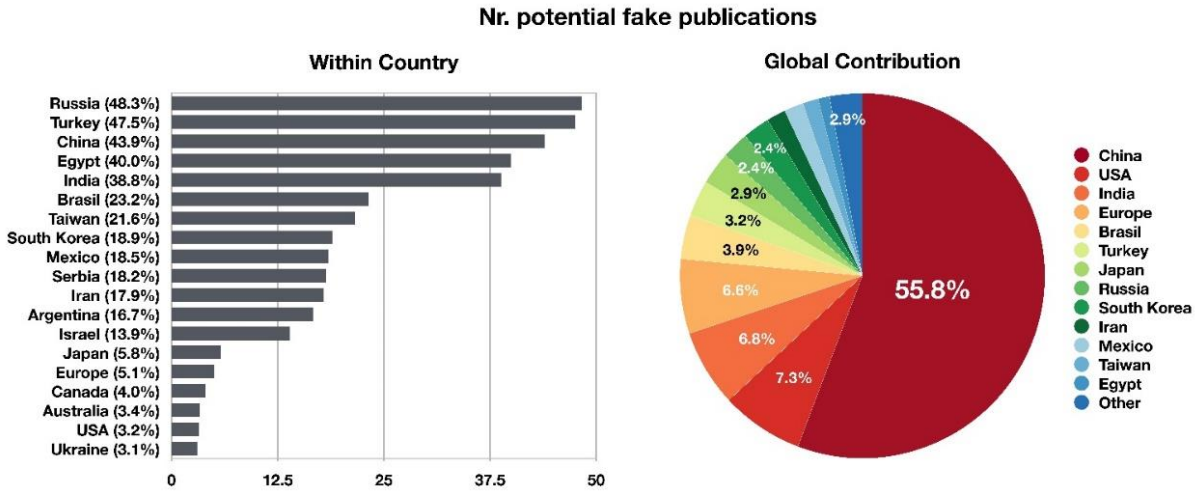

Figure 3. Estimating the global incidence of possible fakes in 2020
Left: Estimated percentage of red-flagged publications of each country´s publication output (based on Study 9). There are three clusters: countries with high (>30%), medium (10%-30%), or low (<10%) numbers of RFPs. Right: the contribution of different countries to the total number of potential fake publications (sample *N*=589). For absolute values, see Table 2. Interestingly, there are three clusters of the RFP rate which might reflect each nation´s average level of publication pressure: governmental (high), institutional (medium), or regular (low).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 1: Rendezvous with a paper mill – A true story

An unsolicited email from a paper mill to an SCI journal editor, and the following exchange (January–April 2022). The editor followed up the contact per email and a recorded Zoom meeting in order to detail the fraudulent business model, including the evidence of corruption.

Paper Mill

We are a well-known academic support institution from Guangzhou, China, which has been established for 8 years. Experts in our institution need to publish some research papers in SCI journals every month. For reducing the publication time, we expect to cooperate with you in the future. Cooperation mode: we cite the content of your journal in our articles, thus increasing … your impact factor in 2022. You shall help us shorten the publication time. Payment: If an article is successfully published, we will pay for it at the price: IF*1000 USD/article. For example, with IF=2.36, total payment=2.36*1000 USD=2360USD. And this price is negotiable.

What you can get:

1. An increased journal impact factor in 2022
2. Desirable payment we give

Editor

[Journals] are always happy to receive manuscripts for publication and always welcome suggestions of how to increase paper flow. …I have a few questions … so that I can better understand what to expect:

1. … How many scientists / publications have you supported and in which fields do you have experience? How about the field of medicine, especially neurology?
2. How can I be sure that the quality of the manuscripts is good enough for our journal? Do you have foreign native speakers who check the language thoroughly? …
3. How many manuscripts can I count on receiving from you per month? Is the number large enough to reach a noticeable growth of my journal in a relatively short period of time?
4. You mentioned a method to increase the impact factor. Can you explain how that works and what impact factor growth you can accomplish….? ….We are currently at around IF 2. How long would it take … to reach IF 5 or 6…?

Paper Mill

1. We have supported the publication of thousands of articles. We mainly have experience in biomedical fields, such as molecular biology, pharmacy, and tumor-related research.
2. We have people from native English-speaking countries to polish the manuscript and experts with biomedical background to review the content. We will arrange 3-4 colleagues to work with you.
3. We have more than 100 manuscripts every month. …
4. We are very familiar with a series of ethical norms of academic publishing, and we also know that the influencing factors of a magazine are related to the number of manuscripts published and the number of times articles are cited

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

by other scholars. Therefore, we will vigorously publicize and recommend the articles of the magazine to Chinese academics on our own publicity platform. Get maximum exposure. …

Editor

1. We are planning a special issue on "Neuromodulation… for neurorehabilitation." We are wondering if you are able to send us 8-12 manuscripts on the topic within the next 4-6 weeks ….
2. These days it is sometimes difficult to find qualified reviewers that are familiar with this field. Therefore, for each paper, would you be able to suggest at least 3 independent researchers knowledgeable in this field … willing to review the manuscript? ….
3. To estimate the growth potential of any collaboration: how many full-time employees does your agency employ? And do you have any access to freelancers, and if so, from which countries (language)?
4. Do you have any unique advantage over other agencies that have similar offers?
5. In order to evaluate the quality of your service, could you please send us … references of SCI-papers … you have helped to get published in the field of medicine?

….

Paper mill

1. We have our own molecular biology laboratory and have developed extensive cooperation with many universities in China. Therefore, we think we can send sufficient manuscripts on this topic.
2. First, we have colleagues who have worked at MDPI, AME Publishing Company, Hindawi, and so on, and they are responsible for suggesting qualified reviewers. Second, our company has many alumni from Sun Yat-sen University and the University of Hong Kong, and they will review our manuscripts more friendly [sic].
3. Our company currently has 100 full-time employees and more than 300 freelancers from the United States, Canada, Italy, Israel, Pakistan, Turkey, India, China and other countries.
4. Our advantage lies in the stable customer source, as well as the mature and efficient management system.
5. Here are the DOIs: …

[Note: the paper mill provided a list of 99 DOIs, 96 of which were detected by our indicators of fake publications]

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 2: Number of known fake publications in the biomedical literature

Published reports of known fake publications			
Reference	**Yrs. analyzed**	**No. of fakes**	**Type of analysis**
Carlisle (2021)	2007 - 2020	43	baseline data analysis
Fisher and Cox (2021)	2020	68	suspected fraudulent papers
Heck et al (2021)	2021	75	image forensics, raw data requests
Seifert (2021)	2019 - 2020	100	raw data requests
Nath et al (2006)	1982 - 2002	107	retraction analysis
Mallapaty (2020)	2017	107	retraction analysis
Behl (2021)	2017 - 2021	137	image forensics
Elango (2021)	1992 - 2020	161	retraction analysis
Bozzo et al (2017)	1946 - 2015	163	retraction analysis
Lei et al (2018)	1997 - 2016	260	retraction analysis
Abalkina (2021)	2019 - 2021	434	offers from paper mill website
Bik (2020)	2016 - 2020	633	image forensics
Bik et al (2016)	1995 - 2014	784	image forensics
Brown et al (2022)	2019	1,396	retraction analysis
Fang et al (2012)	2010	1,611	retraction analysis
Candal-Pedreira	2022	3,544	retraction analysis
Ataie-Ashtiani (2018)	2011 - 2017	4,931	retraction analysis
Total		**14,554**	

Note: Examples of prior studies that quantified the number of fake publications in the years 1982-2022 as identified by text or image plagiarism, data fabrication, paper mill offers, retractions, etc. (Note: duplicate counts could not be determined). Several of these references cite additional publications, and 96.8% of paper mill retractions come from Chinese institutions (Candal-Pedreira et al., 2022). While our list is not exhaustive, given that the total number of biomedical publications is >15 million in the last decade, the percentage of currently identified actual fake publications is on the order of 0.1% for this time period compared with our estimate of 28% in 2020 alone.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 3. Questionnaire sent to corresponding authors

Questionnaire on "Good Scientific Practice"
Name of first author with ORCID, if available
Name of corresponding author with ORCID, if available, and number of his/her SCI journal publications
Data collection and analysis of the work described in this publication complies with good scientific practice and data and their analysis are real. [Yes/No]
List institutional (!) (professional) email addresses of all co-authors (in the order of authorship).
Did you engage a professional agency to help write your paper? [Yes/No] If yes, please state the agency name, website and email.
If so requested by the editor, are you able and willing to provide the original data and images of your publication? [Yes/No]
Was the manuscript originally submitted simultaneously to other journals ? [Yes/No] If yes, specify all journals and the email address of their editorial offices.
Provide name and institutional email of your academic leadership: 1. Academic leader (e.g. President/ Dean) 2. Head of HR Department.
If your work received grant support, please specify the source(s) and provide contact information (name of institution/agency/ grant).
Please sign to confirm the following statement: *"I declare that I take the responsibility that the data and content of the above-mentioned publication are authentic and that they comply with the principle of good scientific practice."*

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 4. Country analysis of potential fake publications (RFP)

Country	No. publications screened	No. RFP	% RFP within country	Global RFP number by country (%)
Russia	**29**	**14**	**48.3**	**2.4**
Turkey	**40**	**19**	**47.5**	**3.2**
China	**747**	**328**	**43.9**	**55.8**
Egypt	**15**	**6**	**40.0**	**1.0**
India	**103**	**40**	**38.8**	**6.8**
Brazil	**99**	**23**	**23.2**	**3.9**
Taiwan	**37**	**8**	**21.6**	**1.4**
South Korea	**74**	**14**	**18.9**	**2.4**
Mexico	**54**	**10**	**18.5**	**1.7**
Serbia	**11**	**2**	**18.2**	**0.3**
Iran	**56**	**10**	**17.9**	**1.7**
Argentina	**12**	**2**	**16.7**	**0.3**
Israel	**36**	**5**	**13.9**	**0.8**
Japan	**293**	**17**	**5.8**	**2.9**
Europe	**772**	**39**	**5.1**	**6.6**
Germany	221	6		
Italy	144	11		
Spain	98	4		
France	88	2		
Netherlands	67	1		
Switzerland	50	3		
Sweden	45	1		
Poland	32	4		
Greece	15	4		
Portugal	12	3		
Canada	**25**	**1**	**4.0**	**0.2**
Australia	**89**	**3**	**3.4**	**0.5**
USA	**1347**	**43**	**3.2**	**7.3**
Ukraine	**162**	**5**	**3.1**	**0.8**
Sum	**4001**	**589**		**100%**

Note: Estimates of red-flagged (potential) fake publications (RFP), their rate by country (% RFP within country) and each country´s contribution to the RFP pool (n=589) as a percentage. The table shows the percentage of publications (red-flagged/number screened publications) in Studies 1-6 (2020, neuroscience and medicine). Countries with <10 publications are not included in the table; they are Ireland (N=9/1), Pakistan (N=8/2), Thailand (N=6/2), Belarus (N=1/6), South Africa (N=5/2), Saudi Arabia (N=5/2), Czech Republic (N=3/1), Malaysia (N=3/1), Nigeria (N=3/1), Tunisia (N=3/2), Romania (N=2/1), and Morocco (N=1/1).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

REFERENCES

Abalkina A (2021) **Publication and collaboration anomalies in academic papers originating from a paper mill: evidence from Russia** *arXiv preprint* **arXiv**:2112.13322

Alison Avenell A, Stewart F, Grey A, Gamble G, Bolland M (2019) **An investigation into the impact and implications of published papers from retracted research: systematic search of affected literature**. *BMJ Open* **9**(10):e031909.

Ataie-Ashtiani B (2018) **World map of scientific misconduct** *Science and engineering ethics* **24**:1657 / **Erratum to:** World Map of Scientific Misconduct *Science and engineering ethics* **24**:1653-1656

Behl C (2021) **Science integrity has been never more important: It's all about trust** *Journal of Cellular Biochemistry* **122**:694-695

Bik EM (2020) **The Tadpole Paper Mill, a "Science Integrity Digest" blog**. Source: https://scienceintegritydigest.com/2020/02/21/the-tadpole-paper-mill/

Bik EM, Casadevall A, Fang FC (2016) **The prevalence of inappropriate image duplication in biomedical research publications** *mBio* **7**:3:e00809-16

Bozzo A, Bali A, Evaniew N, Ghert M (2017) **Retractions in cancer research: a systematic survey** *Research integrity and peer review* **2**:1-7

Brown SJ, Bakker CJ, Theis-Mahon NR (2022) **Retracted publications in pharmacy systematic reviews** *Journal of the Medical Library Association* **110**:47-55

Byrne J (2019) **We need to talk about systematic fraud** *Nature* **566**:9

Byrne JA, Christopher J (2020) Digital magic, or the dark arts of the 21st century -how can journals and peer reviewers detect manuscripts and publications from paper mills? *FEBS letters* **594**:583-589

Byrne JA, Park Y, Richardson RAK, Pathmendra P, Sun M, Stoeger T (2022) **Critical Reviews and Perspectives Protection of the human gene research literature from contract cheating organizations known as research paper mills** *Nucleic Acids Research*: gkac1139

Cabanac G, Labbé C (2021) **Prevalence of nonsensical algorithmically generated papers in the scientific literature** *Journal of the Association for Information Science and Technology* **72**:1461-1476

Candal-Pedreira C (2022) **Retracted papers originating from paper mills: cross sectional study** *British Medical Journal* **379**:e071517

Carlisle JB (2021) **False individual patient data and zombie randomised controlled trials submitted to anaesthesia** *Anaesthesia* **76**:472-479

Chawla DS (2022) **How a site peddles author slots in reputable publishers' journals** *Science* **376**:231-232

Christopher J (2021) **The raw truth about paper mills** *FEBS letters* **595:**1751-1757

COPE & STM Committee on Publication Ethics (2022). **Paper Mills – research report** https://doi.org/10.24318/jtbG8IHL

Cyranoski D (2018) **China introduces sweeping reforms to crack down on academic misconduct** *Nature* **558**:171

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Elango B (2021) **Retracted articles in the biomedical literature from Indian authors** *Scientometrics* **126**:3965-3981

Else H (2021) **Scammers impersonate guest editors to get sham papers published** *Nature* **599**:361

Else H (2022) **Paper-mill detector put to the test in push to stamp out fake science** *Nature* **612**:386-387

Else H, Van Noorden R (2021) **The battle against paper mills** *Nature* **591**:516-519

Fang FC, Grant Steen R, 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**:17028-17033

Fisher L, Cox RJ (2021) **RSC Advances Editorial: retraction of falsified manuscripts** *RSC advances* **11**:4194-4195

Flanagin A, Carey, LA, Fontanarosa PB, Phillips SG, Pace BP, Lundberg GD, Rennie D (1998) **Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals** *Journal of the American Medical Association* **280**:222-224

Frederickson RM, Herzog RW (2021) **Keeping them honest: fighting fraud in academic publishing** *Molecular Therapy* **29**:889-890

Gigerenzer G (2018) **Statistical Rituals: The Replication delusion and how we got there** *Advances in methods and practices in psychological science* **1**:198–218

Heck S, Bianchini F, Souren N, Wilhelm C, Ohl Y, Plass C (2021) **Fake data, paper mills, and their authors: The International Journal of Cancer reacts to this threat to scientific integrity** *International Journal of Cancer* **149**:492-493

Hu Z-w, Wu Y-s (2013) **An empirical analysis on number and monetary value of ghostwritten papers in China** *Current Science* **105**:1230-1234

Katsikopoulos K, Şimşek Ö, Buckmann M, Gigerenzer G (2020) **Classification in the wild** *MIT Press* **208**, ISBN: 9780262045155

Lei L, Zhang Y (2018) **Lack of improvement in scientific integrity: An analysis of WoS retractions by Chinese researchers (1997–2016)** *Science and engineering ethics* **24**:1409-1420

Mallapaty S (2020) **China's research-misconduct rules target'paper mills' that churn out fake studies** *Nature* NEWS 21 August 2020

Mavrogenis AF, Panagopoulos GN, Megaloikonomos PD, Panagopoulos VN, Mauffrey C, Quaile A, Scarlat MM (2018) **Scientific misconduct (fraud) in medical writing** *Orthopedics* **41**:e176-e183

Nath SB, Marcus SC, Druss BG (2006) **Retractions in the research literature: misconduct or mistakes?** *The Medical journal of Australia* **185**:152-154

Park Y, West RA, Pathmendra P, Favier B, Stoeger T, Capes-Davis A, Cabanac G, Labbé C, Byrne JA (2022) **Identification of human gene research articles with wrongly identified nucleotide sequences**. *Life Science Alliance* **5**:e202101203

Pérez-Neri I, Pineda C, Sandoval H (2022) **Threats to scholarly research integrity arising from paper mills: a rapid scoping review** *Clinical Rheumatology* **41**:1-8

Sabel BA, Seifert R (2021) **How criminal science publishing gangs damage the genesis of knowledge and technology – a call to action to restore trust** *Naunyn-Schmiedeberg's Archives of Pharmacology* **394**:2147-2151

medRxiv preprint doi: https://doi.org/10.1101/2023.05.06.23289563; this version posted May 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Sabel BA (2022) **Paper Mills: global knowledge contamination by industrial-style fake science publishing**. Lecture abstract at the Conference of Academic Publishing in Europe (APE 2022); https://doi.org/10.26226/morressier.61d3a5d04a84e7b4701d9885

Schneider L (2021) **The Chinese paper mill industry: interview with Smut Clyde and Tiger BB8: forbetterscience.com** Available at: https://forbetterscience.com/2021/05/26/the-chinese-paper-mill-industry-interview-with-smut- clyde-and-tiger-bb8/

Seifert R (2021) **How Naunyn-Schmiedeberg's Archives of Pharmacology deals with fraudulent papers from paper mills** *Naunyn Schmiedeberg´s Archive of Pharmacol*ogy **394**:431-436

Tian M, Su Y, Ru X (2016) **Perish or publish in China: Pressures on young Chinese scholars to publish in internationally indexed journals** *Publications* **4**:9

Van Dalen HP, Henkens K (2012) **Intended and unintended consequences of a publish-or-perish culture: A worldwide survey** *Journal of the Association for Information Science and Technology* **63**:1282-1293

Wislar JS, Flanagin A, Fontanarosa PB, Deangelis CD (2011) **Honorary and ghost authorship in high impact biomedical journals: a cross sectional survey**. *British Medical Journal* **343**:d6128