Retractions from altmetric and bibliometric perspectives

Hadas Shema¹, Athanasios Mazarakis¹,², Oliver Hahn², and Isabella Peters¹,²

¹Web Science, ZBW - Leibniz Information Centre for Economics, Düsternbrooker Weg 120, Kiel,
²Kiel University, Christian-Albrechts-Platz 4, Kiel (Germany)

In this study we examine retractions in the context of their altmetric and bibliometric properties. We focus on differences between retractions with top altmetric scores and random retractions with altmetric presence, using a sample of top 100 retractions and another sample of 100 retractions with random altmetric scores. The findings show that retractions with top altmetric scores were published, on average, in journals with higher impact factor than retractions with random altmetric score and accumulated more citations. Misconduct was the main reason for retraction of articles in both samples, but more articles were retracted due to misconduct in the random sample (73) than in the top sample (50). Based on these findings, we suggest an association between retractions’ bibliometric and altmetric properties and the causes behind the retractions.

Literature Review

Altmetrics
A growing number of researchers use social media tools in their professional lives. Among those tools we can find social networks such as Facebook and Twitter, as well as academy-specific tools like Mendeley and the social network ResearchGate (Lemke et al., 2017; Van Noorden, 2014). Considering these developments, it is no wonder that alternative or complementary measurements of science, based on social media, have become part of the scientific landscape.

There is some evidence that social media attention relates to self-correcting procedures in scholarly literature. Brookes (2014) has compared papers whose data integrity has been questioned in public and those whose data integrity has been questioned, but the doubts have not been made public. He found that “public papers were retracted 6.5-fold more, and corrected 7.7-fold more, than those in the private set.”

Retractions
The Committee on Publication Ethics (COPE) guidelines regarding retractions define the retraction as “a mechanism for correcting the literature and alerting readers to publications that contain such seriously flawed or erroneous data that their findings and conclusions cannot be relied upon.” Additionally, COPE advises the consideration of retraction in cases of plagiarism,
redundant publication, or report of unethical research (Wager, Barbour, Yentis, & Kleinert, 2009). Also, a retraction is thought to be “an emerging institution that renders scientific misconduct visible” (Hesselmann, Graf, Schmidt, & Reinhart, 2017) and “a window into the scientific process” (Oransky & Marcus, 2010).

In the last two decades, the number of retractions has risen at a rate far exceeding the growth in the total number of published articles (see review in Hesselmann et al., 2017). Furthermore, many of the articles containing errors or misconducts have not been retracted so far (Bik, Casadevall, & Fang, 2016), which may lead to future retractions, though raising further the importance of retractions.

The journal’s impact, as presented by the journal impact factor (JIF), is in certain cases correlated with retractions (Fang & Casadevall, 2011). Fang et al. (2012) found fraud, suspected fraud or error as grounds for retraction were correlated with the JIF. However, duplicate publication or plagiarism were only slightly correlated with the JIF. Similar findings were reported by Madlock-Brown and Eichmann (2015), who found that the publications retracted for their results not being reproducible, for fraud and for error (in this order) were published in journals with higher average impact factor than publications retracted for issues of duplication and plagiarism.

**Error or misconduct?**

Earlier retraction studies, based solely on retraction notices, found that many, if not most, retractions were the results of errors, or of results by the authors which could not be reproduced (Nath, Marcus, and Druss, 2006; Wager and Williams, 2011). However, more recent studies classify most retractions as results of misconduct. Fang, Steen and Casadevall (2012) showed that in some retraction notices, the retraction seems to be the result of unintentional errors, while being, in fact, the result of misconduct. They found that “three-quarters were retracted because of misconduct or suspected misconduct, and only one-quarter was retracted for error.”

A possible reason for this difference between the classification of error and misconduct could be that the definition of misconduct changes between studies. While one study defined misconduct as fabrication, falsification, plagiarism, image/data manipulation and faked data/results/figures, but classified other problematic issues such as forged authorship and faked peer review as „other“ (Ribeiro & Vasconcelos, 2018), others choose to see all categories that are not honest error as misconduct (Moylan & Kowalczuk, 2016). It seems that the proportions of misconduct and error depend on the definition of each and on the sources of information the classification relies upon.

**Data**

We have downloaded the data of retracted publications from PubMed for publications with an official publication and retraction date between January 1, 2012 and August 2, 2017. We focused on data from 2012 onwards because Altmetric.com has started its data gathering on July 2011.
The PubMed data included 1700 retractions, out of which we have matched altmetric data for 919 publications. We then selected the 100 retractions with the highest altmetric scores for our top sample and 100 random retractions for our random sample. 16 observations appear in both samples. We collected the altmetric scores with help of the R package “rAltmetrics” (Karthik, 2017) on the 3rd and 5th of September 2018.

We collected from Web of Science the number of citations to the retractions in the sample which were published between the years 2012-2017 using the “All Databases” search. Those citations were collected between the 24th of July 2018 and the 1st of August 2018.

When observing the time between publication and retraction, we collected the publication date ahead of print, rather than on the official publication date, since altmetric indices begin to accumulate from the moment an article with an object identifier, such as DOI, is online.

We collected the publication and retraction dates using PubMed. In cases where PubMed did not offer an exact date, we searched the publisher’s website for the dates. In cases where there were no clear publication and/or retraction dates, but only mention of a month, we took the 15th of the month as the date. In case of version publication, as in F1000 papers, we took the online publication date of the first version. The time is measured in days.

**Results**

First, we compared the mean, median, skewness and standard deviation of the number of citations for the top and random retractions received, according to Web of Science (Table 1). Many papers received a low number of citations, but few papers were heavily cited. While papers within the top sample have on average a higher number of citations compared to the random sample, a higher standard deviation of the top sample suggests a disproportional increase of citations for higher rated (by altmetrics) papers.

**Table 1.** The mean, median, skewness and standard deviation of number of citations received by top and random retractions (Web of Knowledge)

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Median</th>
<th>sd</th>
<th>skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>0</td>
<td>265.00</td>
<td>29.08</td>
<td>13.50</td>
<td>38.78</td>
<td>2.89</td>
</tr>
<tr>
<td>Random</td>
<td>0</td>
<td>140.00</td>
<td>11.89</td>
<td>5.50</td>
<td>19.05</td>
<td>4.07</td>
</tr>
</tbody>
</table>
Second, we compared the same statistics of the journal impact factor for the publishing journals of the retracted articles in both samples (Table 2).

**Table 2.** The mean, median, skewness and standard deviation of the JIF of the publishing journal of retractions

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Median</th>
<th>sd</th>
<th>skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>1.27</td>
<td>44.41</td>
<td>14.61</td>
<td>6.48</td>
<td>14.43</td>
<td>0.97</td>
</tr>
<tr>
<td>Random</td>
<td>0.73</td>
<td>44.41</td>
<td>7.47</td>
<td>3.65</td>
<td>10.83</td>
<td>2.65</td>
</tr>
</tbody>
</table>

Performed Wilcoxon matched pairs signed rank tests for significant differences between means of both variables showed that the differences between variables were statistically significant on the 1%-level (p < .001) between both samples.

Our data suggests that top altmetric retractions tend to be published in higher impact journals, in comparison with random retractions, and tend to attract more citations.

**Misconduct or error?**

As we have noted in the literature review, the definitions of misconduct and error change from one study to another. In this study, we have chosen to classify as misconduct the articles where there has been an intention of one author or more to deceive the readers and as error where it was clear that the articles’ flaws were unintentional.

We classified the retractions according to their retraction notices, Retraction Watch entries and the Retraction Watch database, if those existed. Two coders each classified the samples separately, after which the classifications were compared and differences have been discussed until agreement has been reached. In cases where agreement between the coders could not be reached, the retractions have also been classified as „unclear“.

The „unclear“ category was used when the two coders both considered the reason behind the retraction as unknown, based on our sources, and the „other“ category was created for cases where the article has been retracted not due to error or misconduct of the authors, but different reasons (such as concerns of legal actions against the journal unless the article is retracted, though the article itself is considered valid). The „influenced by a third party“ category was used for
misconduct where the authors have not been at fault, such as a commentary piece which has been retracted because the article it commented about was fraudulent.

In the top 100 sample we have found that the main retraction reason for 50 of the articles was misconduct, for 40 of the articles the main reason was error, the „influenced by a third party“ category included 4 articles, the „other“ category 2 articles and the „unclear“ category 4 (Figure 1).

In the random 100 sample, the picture was different. We have found that misconduct was the main reason for retraction for 73 of the articles in the sample, while error accounted as the reason to only 19 of the retractions. The „unclear“ category included 6 articles, the „influenced by a third party“ category included 2 articles, and there were no articles in the “other” category (Figure 1).
Figure 1. The reasons behind retraction, top 100 and 100 random retractions

The results of a chi-squared test for the categories showed in Figure 1 reject the null hypothesis of independence ($\chi^2 = 19.99$, df = 4, $p < .001$).
Next, we compared the main causes of misconduct within the misconduct category (Figure 2 and Table 3).

**Figure 2.** Causes of retraction within the misconduct category for the top and random samples
Table 3. Causes of retraction within the misconduct category for the top and random altmetric retractions

<table>
<thead>
<tr>
<th>Causes of retraction</th>
<th>Top retractions (%)</th>
<th>Random retractions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data/image falsification</td>
<td>62</td>
<td>32.9</td>
</tr>
<tr>
<td>Plagiarism/duplication</td>
<td>26</td>
<td>41.1</td>
</tr>
<tr>
<td>Compromised peer review</td>
<td>0</td>
<td>13.7</td>
</tr>
<tr>
<td>Unethical conduct</td>
<td>6</td>
<td>8.2</td>
</tr>
<tr>
<td>Conflict of interests</td>
<td>4</td>
<td>1.4</td>
</tr>
<tr>
<td>Other misconduct</td>
<td>2</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Plagiarism/duplication was the most common reason for retraction in the random sample (Figure 2), present in 30 out of the 73 retractions that were retracted for misconduct (41.1%). In comparison, this reason occurred only in 13 out of the 50 retractions that were retracted for misconduct at the top sample (26%). When it comes to data/image falsification, the results are very different. Data falsification has been shown as at least one of the reasons for retraction only in 24 of the cases (32.9%) of the random sample but has been present in 31 of the retractions in the top sample (62%). We have not discovered any cases of compromised, or fake peer review at the top sample, but it occurred in 10 of the retractions in the random sample (13.7%). A chi-squared test showed that the difference between the misconduct categories is statistically significant ($\chi^2 = 22.80$, df = 6, $p < .001$).

Summary

Though the study was limited by the sample size, we have found differences between retractions with top altmetric score and retractions with random altmetric score. We show that while misconduct is the most prominent reason of retraction in both samples, it is considerably more prominent in the random sample than in the top sample. In many cases, top altmetric retractions that have been retracted due to misconduct have data falsification/fabrication as their cause of retraction, while the random retractions are often retracted due to plagiarism/duplication of text and images.

Our results regarding retractions due to error being more prominent in the top 100 retractions are in line with those of Steen and his colleagues regarding retractions published in high-impact
journals: “detection of error was significantly more likely in journals with a high IF” (Steen, Casadevall, & Fang, 2013), as well as those of Madlock-Brown and Eichmann (2015). The results suggest an association between altmetric and bibliometric properties of retractions – top altmetric retractions have been published, on average, in journals with higher impact factor and received more citations than the random sample.

We show that when dealing with retractions, the scholarly literature and social media have much in common. Prominent articles in prominent journals become, when retracted, highly visible failures, while articles published in lower impact journals, when retracted, pass “under the radar” of both the social media and the scholarly community.

**Acknowledgments:** Hadas Shema is an Alexander von Humboldt Foundation post-doctorate fellow. We thank Altmetric.com for the use of their data.
References
https://doi.org/10.1128/mBio.00809-16


https://doi.org/10.1016/j.maturitas.2009.09.018

https://doi.org/10.1136/jme.2010.040964