







Are Papers with Open Data More Credible? An Analysis of Open Data Availability in Retracted *PLoS* Articles

Michael Lesk¹ , Janice Bially Mattern² ,
and Heather Moulaison Sandy³  

¹ Rutgers University, New Brunswick, NJ 08901, USA
lesk@acm.org

² Villanova University, Villanova, PA 19085, USA

³ University of Missouri, Columbia, MO 65211, USA
moulaisonhe@missouri.edu

Abstract. Open data has been hailed as an important corrective for the credibility crisis in science. This paper makes an initial attempt to measure the relationship between open data and credible research by analyzing the number of retracted articles with attached or open data in an open access science journal. Using Retraction Watch, retracted papers published in *PLoS* between 2014 and 2018 are identified. Of the 152 total retracted papers, fewer than 15% attached their data. Since about half of the published articles have open data, and so few of the retracted ones do, we put forth the preliminary notion that open data, especially high quality and well-curated data, might imply scientific credibility.

Keywords: Retractions · Open data · Credibility

1 Science’s Credibility Crisis

By many accounts, the credibility of scientific research is in crisis (Saltelli and Fun-towicz 2017). Mounting evidence suggests that surprisingly few scientific studies are reproducible and/or replicable (Sayre and Riegelman 2018). In medical research, for instance, Prinz et al. (2011) found that Bayer could only replicate some 20–25% of 67 studies in the fields of cancer biology, women’s health, and cardiovascular diseases. Similar replication challenges are found in psychology (Bohannon 2015) and social science (Camerer et al. 2018).

In some cases, dubious results can be traced to sloppy analysis—for instance, computational errors, corner cutting, or statistical misinterpretation like ‘p-hacking’ (Benjamin et al. 2017). In others cases, the problem arises from data manipulation, plagiarism or outright research fraud. Marcus and Oransky (2015), from the Retraction Watch database (<http://retractiondatabase.org/>) and blog (<https://retractionwatch.com/>) report that “every day on average, a scientific paper is retracted because of misconduct” (p. A19). Some instances are high profile, including contrived data linking vaccinations to autism; tampered and falsified data on predictors of cancer; and fully fabricated data

on political opinions about same-sex marriage. When taken in combination with unwitting research errors, the credibility of scientific research is eroding.

The result is a crisis at multiple levels of society. First, at the level of research, it raises questions about how much undetected faulty research is out there. Research from a range of disciplines confirms that doubtful findings or even myths can and do become foundational knowledge, taken for granted through repeated citations (cf. Begley and Ellis 2012; Rekdal 2014). These mistakes waste effort and contribute to a skepticism of research in general.

Second, the eroding credibility of scientific research suggests a crisis for the role of science in policy. Although reverence for evidence-based decision and action has long meant that scientific research is an authoritative basis for social and political policy (Saltelli and Giampietro 2017), faulty science related to issues such as public safety, education, and economic development have begun to take a toll. Consider austerity policies that aim to reduce budget deficits through deep cuts to government spending; they ultimately impose great social costs, especially for society's most needy members. Yet, the science justifying these punishing policies is largely rooted in an analysis that has been exposed as sloppy and unjustified (Cassidy 2013). Anti-vaccine sentiment is also based on work which has been found faulty. This supports the idea that we have entered a post-factual world in which science is no longer trusted as the basis for policy.

The social legitimacy of scientific knowledge depends upon solving the problem of undetected faulty research. An ethos of open data can help by improving research transparency. Scientists can identify irreproducible results, catch sloppy computations, highlight and correct each other's errors, and ensure that they are building upon only the most reliable and verified findings (Molloy 2011; Gewin 2016; Leeming 2017).

2 Open Data in Science

Despite the potential to restore credibility to the scientific endeavor, open data sharing has not caught on the way its advocates might hope. Academics are notoriously reticent about openly sharing their data. As Fecher et al. (2015) point out, this reticence generally arises from a mismatch between taking an action (sharing) that benefits the greater good, and disincentives to sharing. For instance, where raw data discloses sensitive or confidential personal information about subjects, researchers' unwillingness to share data can be an ethical matter. Other resistance to sharing could be due to the commercial value of raw data, or because it is licensed.

Most of the reasons that researchers resist openly sharing data, however, are less simple. As the literature documents, many researchers are concerned about their results being called into question. After all, if their data are not available, nobody can attack their findings. To make data shareable is time consuming, requiring proper curation. Researchers are hesitant about the preparation required to make data intelligible to others (c.f. Tenopir et al. 2011). Some researchers are also concerned about data theft (Teixeira da Silva and Dobránszki 2015), and providing the proper metadata can be

challenging (Fecher et al. 2015; Sadiq and Indulska 2017). In short, researchers are more disposed toward activities that directly facilitate their research than to the data curation activities required for open data. Most universities are insufficiently incentivizing data sharing (Bolukbasi et al. 2013), and in the time it takes to curate data, another research team could publish the idea, “scooping” the work (Van Noorden 2014). In sum, there are a number of reasons data sharing is limited and remains limited.

Going forward, some improvements can be expected. First, new mandates require that U.S. federally funded research results must be made available with open access to the relevant data, although compliance has been limited (Nelson 2009). As well, some journal publishers have increasingly required that authors share the data that supports their research articles. For instance, in 2013, *PLoS* (Public Library of Science: <https://www.plos.org/>) established a policy that authors must provide open access to their research data unless they receive permission from the editor (“Editorial and Publishing Policies” 2018; Teixeira da Silva and Dobránszki 2015). Not all journals have such requirements. The social sciences, for example, have been slower. Few journals have policies that encourage open data (cf. *International Studies Quarterly*) with even fewer requiring it (cf. *The Journal of Politics*). Advocates are pressuring for change but scholars remain ambivalent. An ethos of open data remains nascent, at best.

Reticence to publish datasets does not mean that faulty research is going unattended. Retractions are nonetheless rare, and can be due to a number of possible problems (Marcus and Oransky 2015) and many are not due to data problems. Articles can be retracted because they have been previously published, because ethical principles of human subjects research were not followed, or because of complaints about the soundness of the methodology. Plagiarism, self-plagiarism, and “salami slicing” are other reasons that articles might be retracted. Sloppy computation, corner cutting, and more flagrant research misconduct can also lead to retractions or sanctions of some kind. Although misconduct is possible to detect without open data, it is easier if the data can be inspected.

Such verification, however, can be experienced as a threat, which creates an obstacle to an open data ethos in science. Researchers may worry that by opening their open data, their work is more likely to be retracted simply because it is easier to verify than that for which the data is not provided. But the reverse might also be true: articles with open data may be less likely to be retracted, precisely because in the onerous process of curating the data, authors are more likely to discover and address mistakes before publication. Peer’s (2018) process for validating work in her organization and avoiding common statistical mistakes exemplifies this suggestion.

2.1 Guiding Questions

The current research considers the relationship between open data and retraction by asking: Does open data put researchers at greater risk or does it help establish their contributions as real and meaningful, beyond reproach? Put differently, is open data a hindrance or help to the cultivation of more credible science?

3 Method

We focused on articles in *PLoS* because it (1) is entirely open access for the text of articles and (2) has a standard way of indicating when an article has attached data files. In addition, *PLoS* is a prolific publisher of scientific papers, with many articles that have been retracted¹ and subsequently tracked on Retraction Watch. “Retraction Watch is a Web site set up by two science journalists, Adam Marcus and Ivan Oransky, who have received international attention for tracking high-profile retractions of papers” (Bonnell et al. 2012, p. 2). Our study looks at 2014 to the present for a number of reasons. First, as mentioned, *PLoS* began requiring data in 2013. In addition, the U.S. National Science Foundation began requiring open data be made available on funded projects starting in 2011 (National Science Foundation 2016), which resulted in many projects of 3-year duration being published starting in 2014. Such projects required a “data management plan” explaining how the data collected would be made available to other researchers, often through deposit in an archive but sometimes by arrangement with the authors, and not necessarily publication in a journal. Another reason for focusing on 2014 onwards is to exclude the thousands of retractions resulting from the detection of published research papers discovered to have been generated by the Sci-Gen chatterbot (Labbé et al. 2015), a chatterbot that produces documents resembling scientific research papers, and that some journals and conferences published only to retract later.

PLoS articles were obtained from PubMed Central where they are available for download; Retraction Watch data was obtained through its website. Articles were matched via DOIs (digital object identifiers) and analyzed using Python.

4 Results²

Ultimately, very few papers are retracted from *PLoS* per year, as shown in Table 1.

Overall, about half the papers in *PLoS* have some kind of attached data. Table 2 shows the fraction of *PLoS* papers with data availability.

Of the retracted papers identified, fewer than average had accompanying data. Figure 1 show the counts of retracted articles by year, providing a red bar for all retracted papers and a green bar for those with open data. What is striking about this, is that in total about 27% of the retracted papers had open data. We did not examine the detailed reasons for retraction – as noted before, some retractions have nothing to do with data (e.g., plagiarism or publisher error) and many may have multiple causes.

¹ The *PLoS* retraction policy is at <https://journals.plos.org/plosone/s/corrections-and-retractions>, referencing and including the ICMJE (International Committee of Medical Journal Editors) rules given at <http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/scientific-misconduct-expressions-of-concern-and-retraction.html>. Note that a retraction does not necessarily imply either malfeasance or a significant error; articles can be retracted for a missing author, for example.

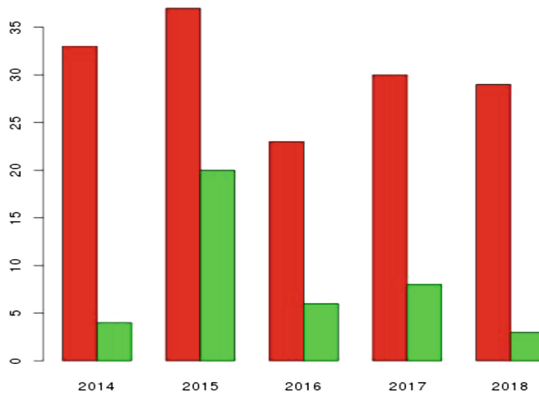
² Data for this project is available through the RUcore repository: <https://rucore.libraries.rutgers.edu/>.

Table 1. Number of articles and retracted articles in *PLoS*, by year (2014–2018).

Year	Total articles	Retracted articles	% Retracted
2014	35393	33	0.09%
2015	33449	37	0.1%
2016	26981	23	0.09%
2017	24670	30	0.1%
2018 (part year)	12693	29	0.2%
Total	133186	152	0.1%

Table 2. Percentage of articles published in *PLoS* that include accompanying data, by year (2014–2018).

Year	Total articles	Articles with data	% With data
2014	35393	6573	19%
2015	33449	17075	51%
2016	26981	13631	50%
2017	24670	11516	47%
2018 (part year)	12693	5696	45%

**Fig. 1.** Count of retracted articles in *PLoS*, by year (2014–2018). Red bars indicate retractions, green bars indicate retracted papers accompanied by open data. (Color figure online)

5 Discussion

From 2015 onward, roughly half the articles published in *PLoS* were accompanied by data, but the percentage did not increase. The steady level of open data—despite the expectations of data sharing—suggests that concerns over costs or misuse persist.

Yet, the risk to authors of re-examination appears to be overimagined: The articles with open data are not routinely being retracted because their data has been examined and found faulty. As demonstrated in Table 1, the risk of being accused of fraudulence

or sloppiness is very low (for example, for 2017 in all of *PLoS* there were more than 24,000 articles with data and only 30 were retracted).

The more onerous aspects of data curation may have inhibited some authors from preparing to distribute their data, but increasingly there is support for this activity. Many university libraries are able to assist with both organizing data and promising long-term storage (Heidorn 2011). Over time, making data available will become more expected, especially as funding organizations insist on it, and researchers should understand the benefits to them and to their field if they will be expected to take part.

As the findings of this paper suggest, (however preliminarily) it makes little sense to invoke fear of being criticized as a justification for withholding data.

6 Conclusion and Next Steps

In this short paper, we have examined the retraction rates for published articles in *PLoS* that include and do not include open data. More systematic research is needed to confirm our preliminary findings and we will continue to pursue it. We intend to keep using Retraction Watch as a source of information on rejected papers, and attempt to classify a larger variety of papers as with or without open data. Areas with a tradition of open data (astronomy, as an example) will be particularly interesting to observe, although these do not necessarily overlap with the areas that are readily available from archives such as PubMed. We believe that papers with high quality, well-curated open data will turn out to be more reliable and should deserve greater credibility, and be entitled to both more prestige in the university community and higher rankings in search engines, but acknowledge that much more study is needed to investigate these preliminary conclusions.

A number of good reasons for publishing data exist. Scientific results, whether they be about vaccine risks or climate change, are in doubt. If the credibility of our literature can be improved by including more of the data, everyone will be better off. Researchers might be incentivized to publish data since papers with open data get more citations (Piwowar and Vision 2013). Borgman (2012) discusses other advantages of open data, such as accelerating research if people can see each other's work.

This study's findings should encourage further research into the role of data sharing in promoting an ethos of transparency in science as a way of potentially ameliorating the credibility crisis. In the effort to establish with greater certainty the inverse relationship between open data and having one's findings impugned, clearer notations in articles about whether they make their data available would be helpful. For instance, the ACM (Association for Computing Machinery) does a good job of this e.g., with the badge. The preliminary results of this research therefore support further inquiry into the standardized recognition across publishers and publishing platforms for papers that attach or otherwise make their data open and available, and the quality and process of making that data available.

References

- Begley, C.G., Ellis, L.M.: Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012)
- Benjamin, D.J., et al.: Redefine statistical significance, 22 July 2017. <https://doi.org/10.31234/osf.io/mky9j>
- Bohannon, J.: Many psychology papers fail replication test. *Science* **349**(6251), 910–911 (2015)
- Bolukbasi, B., et al.: Open data: crediting a culture of cooperation. *Science* **342**(6162), 1041–1042 (2013)
- Bonnell, D.A., et al.: Recycling is not always good: the dangers of self-plagiarism. *ACS Nano* **6**(1), 1–4 (2012). <https://doi.org/10.1021/nl3000912>
- Borgman, C.L.: The conundrum of sharing research data. *J. Am. Soc. Inf. Sci. Technol.* **63**(6), 1059–1078 (2012)
- Camerer, C.F., et al.: Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018)
- Cassidy, J.: The Reinhart and Rogoff Controversy: A Summing Up. The New Yorker, New York (2013)
- Editorial and Publishing Policies. PLoS (2018). <https://www.plos.org/editorial-publishing-policies>
- Fecher, B., Friesike, S., Hebing, M.: What drives academic data sharing? PLoS ONE **10**(2), e0118053 (2015). <https://doi.org/10.1371/journal.pone.0118053>
- Gewin, V.: Data sharing: an open mind on open data. *Nature* **529**(7584), 117–119 (2016). <https://doi.org/10.1038/nj7584-117a>
- Heidorn, P.B.: The emerging role of libraries in data curation and e-science. *J. Libr. Adm.* **51**, 662–672 (2011)
- König, N., Børsen, T., Emmeche, C.: The ethos of post-normal science. *Futures* **91**, 12–24 (2017). <https://doi.org/10.1016/j.futures.2016.12.004>
- Labbé, C., Labbé, D., Portet, F.: Detection of computer generated papers in scientific literature <hal-01134598> (2015). <https://hal.archives-ouvertes.fr/hal-01134598>
- Leeming, J.: How will open data advance scientific discovery? Naturejobs Blog (2017). <http://blogs.nature.com/naturejobs/2017/10/25/how-will-open-data-advance-scientific-discovery/>. Accessed 9 May 2018
- Marcus, A., Oransky, I.: What’s Behind Big Science Frauds?. The New York Times, New York (2015). <https://www.nytimes.com/2015/05/23/opinion/whats-behind-big-science-frauds.html>
- Molloy, J.C.: The Open Knowledge Foundation: open data means better science. PLoS Biol. **9**(12), e1001195 (2011). <https://doi.org/10.1371/journal.pbio.1001195>
- The National Science Foundation Open government plan: 4.0, September 2016. <https://www.nsf.gov/pubs/2016/nsf16131/nsf16131.pdf>
- Nelson, B.: Data sharing: empty archives. *Nature* **461**, 160–163 (2009). <https://doi.org/10.1038/461160a>
- Peer, L.: Reproducible research practices at ISPS, 30 April 2018. <https://isps.yale.edu/news/blog/2018/05/reproducible-research-practices-at-isps>
- Piwowar, H.A., Vision, T.J.: Data reuse and the open data citation advantage. PeerJ: Bioinformatics and Genomics section, 1 October 2013
- Prinz, F., Schlange, T., Asadullah, K.: Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712 (2011)
- Rekdal, O.B.: Academic urban legends. *Soc. Stud. Sci.* **44**(4), 638–654 (2014). <https://doi.org/10.1177/0306312714535679>

- Sadiq, S., Marta Indulska, M.: Open data: quality over quantity. *Int. J. Inf. Manage.* **37**, 150–154 (2017). <https://doi.org/10.1016/j.ijinfomgt.2017.01.003>
- Saltelli, A., Funtowicz, S.: What is science's crisis really about? *Futures* **91**, 5–11 (2017). <https://doi.org/10.1016/j.futures.2017.05.010>
- Saltelli, A., Giampietro, M.: What is wrong with evidence based policy, and how can it be improved? *Futures* **91**, 62–71 (2017). <https://doi.org/10.1016/j.futures.2016.11.012>
- Sayre, F., Riegelman, A.: The reproducibility crisis and academic libraries. *Coll. Res. Libr.* **79**(1), 2 (2018). <https://crl.acrl.org/index.php/crl/article/view/16846/18452>
- Tenopir, C., et al.: Data sharing by scientists: practices and perceptions. *PLoS One*, 29 June 2011. <https://doi.org/10.1371/journal.pone.0021101>
- Teixeira da Silva, J.A., Dobránszki, J.: Potential dangers with open access data files in the expanding open data movement. *Publ. Res. Q.* **31**, 298–305 (2015). <https://doi.org/10.1007/s12109-015-9420-9>
- Van Noorden, R.: Publishers withdraw more than 120 gibberish papers: conference proceedings removed from subscription databases after scientist reveals that they were computer-generated. *Nature News* (February 24, 2014; Updated February 25, 2014)