

## Reply to: Recognizing and marshalling the pre-publication error correction potential of open data for more reproducible science

Ilias Berberi<sup>1\*</sup> and Dominique G. Roche<sup>1,2</sup>

<sup>1</sup> *Department of Biology, Carleton University, Ottawa, Canada*

<sup>2</sup> *Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland*

\*Author for correspondence: [iliasberberi@gmail.com](mailto:iliasberberi@gmail.com)

In a previous paper, we demonstrated a lack of association between open data requirements and post-publication error correction among journals in ecology and evolution<sup>1</sup>. To facilitate effective data review and error correction, we recommended improving the archiving quality of open datasets, sharing analytical code alongside datasets, and destigmatizing error correction among researchers and journal editors.

In response to our paper, Chen et al.<sup>2</sup> highlighted that mandatory open data policies also increase opportunities for detecting and correcting errors pre-publication. We welcome Chen et al.'s comment and acknowledge that we omitted discussing the important, positive impact that mandatory open data policies can have on various pre-publication processes. Our study design and the interpretation of our results were likely influenced by our prior experience of reporting data anomalies and research misconduct to journals, and witnessing first-hand the challenges of post-publication error correction<sup>3-5</sup>. As longstanding advocates of transparency and reproducibility in research, we would celebrate empirical evidence that data sharing mandates increase pre-publication error detection.

Chen et al. suggest that mandatory open data policies could lead researchers to better manage and share their data and code, thereby increasing opportunities for detecting errors pre-publication – for example, via feedback from co-authors and formal peer-review. Based on this logic, they argue that mandatory open data policies will foster studies with fewer errors at manuscript acceptance, resulting in fewer post-publication error corrections. We agree with Chen et al.'s reasoning and support the assertion that lowering the need for post-publication error correction is a key aim of good research data management and open data practices. Unfortunately, however, we would argue that – at present – there is limited evidence that open data meaningfully increase opportunities for pre-publication error detection and correction.

If mandatory open data policies facilitated identifying and correcting errors pre-publication, we would expect a decrease in retractions and corrections following policy adoption. Yet, contrary to this prediction, our data indicate that rates of error correction neither decreased nor increased after journals in ecology and evolution began requiring open data<sup>1</sup>. Positive impacts of open data on error correction both pre- *and* post-publication might obscure any change in these rates over time – for example, by simultaneously reducing and increasing the number of papers in need of

post-publication error correction. However, in our opinion, this effect is unlikely given that numerous barriers currently hinder the potential benefits open data mandates on pre-publication error correction.

First, journals in ecology and evolution overwhelmingly require open data at the time of manuscript publication, not submission<sup>6</sup>. This delayed requirement for open data impedes opportunities for data assessment and error detection pre-publication. A practical solution for improving journal open data policies is to require that data be archived and privately shared with editors and reviewers upon manuscript submission. Many data repositories support creating private links to archived datasets (e.g., Dryad, Zenodo, Figshare, OSF), making it possible for authors to confidentially share data and code with reviewers and editors prior to public release.

Second, journals that require data sharing often do not enforce their policy<sup>7</sup>, preventing reviewers from assessing datasets. For example, the journal *Behavioural Ecology* began requiring open data in 2016, but an estimated 32% of empirical studies published between 2016-2021 lack a publicly available dataset<sup>8</sup>. A similar trend was detected across more than 70,000 studies published in PLOS journals between 2019-2022<sup>9</sup>. Compliance issues with open data mandates are thought to result from unclear policy requirements as well as discrepancies between a policy's stated requirements and their interpretation by editors and authors<sup>7</sup>. Additionally, there is little consensus regarding how data policies should be enforced by ecology journals and whether editors or reviewers are responsible for holding authors to account<sup>10</sup>. The consequence of these systemic flaws is a struggle for reviewers to access and assess datasets associated with the manuscripts they are tasked to review.

Third, open data frequently do not meet minimum standards of quality allowing third parties to readily verify the analyses and results presented in a study<sup>7,11</sup>. Poor data archiving quality undermines error detection both pre- and post-publication. For example, in ecology and evolution, over 50 percent of open datasets are incomplete and difficult to understand and reuse<sup>12</sup>, with little to no evidence of improvement in data sharing practices between 2012-2019<sup>13</sup>. Importantly, while open data mandates have been shown to increase rates of data sharing<sup>14</sup>, they appear ineffective at achieving complete and reusable open datasets<sup>13</sup>. Improving the potential for pre- and post-publication error correction will require boosting both the accessibility and the archiving quality of open data, in line with the FAIR sharing principles<sup>15</sup>. These improvements will require training researchers in better data management and sharing practices<sup>13</sup> as well as more effective mechanisms for monitoring compliance with journal data policies<sup>10</sup>.

Fourth – as mentioned by Chen et al. – journals depend on the willingness of reviewers to assess submitted datasets during peer-review<sup>10</sup>. With increasing reports of academic fatigue and burnout among researchers<sup>16</sup>, it is unlikely that journals will compel reviewers to take on the extra responsibility of data review<sup>17</sup>. One recent and promising initiative undertaken by some journals in ecology and evolution is the appointment of *data editors* whose role focuses on assessing datasets associated with submitted manuscripts<sup>18</sup>. For example, *The American Naturalist* and the *Journal of Evolutionary Biology* combine machine-learning algorithms (DataSeer.ai) and manual review to identify and assess datasets before manuscripts are sent out for review<sup>19,20</sup>. Broadly appointing data editors across journals in ecology and evolution would bolster opportunities for data review and error detection pre-publication.

Our original study findings and the additional insights provided by Chen et al. outline promising future directions for improving how we assess the impact of open data policies on research practices. Importantly, the potential benefits of open data are not limited to error correction, and can extend to greater collaboration and equity in research, more efficient knowledge creation and mobilization, and greater public trust in science<sup>21–24</sup>. Empirical studies on these important topics are urgently needed to inform evidence-based open science policy and practice.

### Author Contributions

I.B. and D.G.R. wrote and edited the manuscript.

### Competing Interests

D.G.R. is a member of the Canadian Institute for Ecology and Evolution's Living Data Project and a co-founder and Past-President of the Society for Open, Reliable and Transparent Ecology and Evolutionary Biology ([www.sortee.org](http://www.sortee.org)). I.B. declares no competing interests. D.G.R. was supported by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement no. 838237-OPTIMISE.

### References

1. Berberi, I. & Roche, D. G. No evidence that mandatory open data policies increase error correction. *Nat Ecol Evol* **6**, 1630–1633 (2022).
2. Chen, R. S., Berthelsen, A. L., Brejon Lamartinière, E., Spangenberg, M. C. & Schmoll, T. Recognising and marshalling open data's pre-publication error correction potential for more reproducible science - A comment on Berberi & Roche. *Nat Ecol Evol* (2023).
3. Clark, T. D. *et al.* Scientific misconduct: The elephant in the lab. A response to Parker et al. *Trends Ecol Evol* **31**, 899–900 (2016).
4. Enserink, M. Sea of doubts. *Science (1979)* **372**, 560–565 (2021).
5. Enserink, M. Journal declines to retract fish research paper despite fraud finding. *Science* (2023) doi:10.1126/science.adh1493.
6. Fernández-Juricic, E. Why sharing data and code during peer review can enhance behavioral ecology research. *Behav Ecol Sociobiol* **75**, 103 (2021).
7. Christian, T. M., Gooch, A., Vision, T. & Hull, E. Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PLoS One* **15**, 1–15 (2020).
8. Roche, D. G. *et al.* Paths towards greater consensus building in experimental biology. *Journal of Experimental Biology* **225**, jeb243559 (2022).
9. Cadwallader, L., Morton, L. & Hryaszkiwicz, I. Trends in preprint, data, and code sharing, 2019-2022. *PLoS Blogs* <https://theplosblog.plos.org/2023/04/open-science-indicators> (2023).
10. Sholler, D., Ram, K., Boettiger, C. & Katz, D. S. Enforcing public data archiving policies in academic publishing: a study of ecology journals. *Big Data Soc* **6**, 2053951719836258 (2019).
11. Montgomerie, R. Data transparency 2020. *From the Editor's Desk of The American Naturalist* <http://comments.amnat.org/2021/01/note-since-fall-2020-robert-montgomerie.html> (2021).

12. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLoS Biol* **13**, e1002295 (2015).
13. Roche, D. G. *et al.* Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences* **289**, (2022).
14. Vines, T. H. *et al.* Mandated data archiving greatly improves access to research data. *FASEB Journal* **27**, 1304–1308 (2013).
15. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
16. Dance, A. Stop the peer-review treadmill. I want to get off. *Nature* **614**, 581–583 (2023).
17. Kratz, J. & Strasser, C. Data publication consensus and controversies. *F1000Res* (2014).
18. Thrall, P. H. *et al.* From raw data to publication: Introducing data editing at Ecology Letters. *Ecol Lett* **26**, 829–830 (2023).
19. Bolnick, D., Vines, T. H. & Montgomerie, R. Ensuring data and code archive quality: Why and how? <https://comments.amnat.org/2022/03/ensuring-data-and-code-archive-quality> (2022).
20. Lequime, S. Data editing at JEB. <https://jevbio.net/data-editing-at-jeb/> (2022).
21. Soeharjono, S. & Roche, D. G. Reported individual costs and benefits of sharing open data among Canadian academic faculty in ecology and evolution. *Bioscience* **71**, 750–756 (2021).
22. Haven, T., Gopalakrishna, G., Tjeldink, J., van der Schot, D. & Bouter, L. Promoting trust in research and researchers: How open science and research integrity are intertwined. *BMC Res Notes* **15**, 302 (2022).
23. Poisot, T., Mounce, R. & Gravel, D. Moving toward a sustainable ecological science: don't let data go to waste! *Ideas in Ecology and Evolution* **6**, 11–19 (2013).
24. Rosman, T., Bosnjak, M., Silber, H., Koßmann, J. & Heycke, T. Open science and public trust in science: Results from two studies. *Public Understanding of Science* **31**, 1046–1062 (2022).