Reanalysis: the forgotten sibling of reproducibility and replicability

Matthew Faria[1], Steve Spoljaric[2] and Frank Caruso[2]✉

Ensuring reproducibility and replicability has been an issue in many scientific disciplines in the past decade. Here, we discuss another 'R' that has not gotten enough airtime — reanalysis. We cover how open science and a focus on enabling reanalysis also make the goals of reproducibility and replicability easier to achieve.

Scientists strive for reproducibility (running the same experiments and obtaining similar results) and replicability (running the same analysis on the same data and obtaining the same result) — although precise definitions for these terms vary[1]. More than 80% of surveyed chemists have reported failure to reproduce an experiment[2]; an initiative to reproduce cancer biology results found that somewhere between the majority and all experiments were not reproducible as described[3], and nanomedicine has seen an increased focus on reproducibility[4]. We believe that, especially in the applied biological and chemical sciences, another R is worth aiming for: reanalysis (providing sufficient experimental description and data so that researchers can apply new analysis methods to provided data). Here, we make a case for facilitating reanalysis; central to accomplishing this are open methodology and data (FIG. 1).

The case for facilitating reanalysis

Advancements in analysis are as vital to scientific progression as new methods to collect data. For instance, many types of DNA sequencing would be of little use without the algorithms that computationally assemble measured fragments. The complexity and scale of modern scientific data combined with the increased accessibility of artificial intelligence (AI) and computational resources mean that we should increasingly expect progress to be driven by improvements in analysis. Analysis development, whether new metrics, improved phenomenological models or extensively trained AI models, relies on access to good data.

The contribution of large data projects to science is well recognized — for instance, the ENCODE project is within the top 100 most-cited papers published in the past 10 years. But not every analytic development requires big data; well-reported libraries of smaller sets of experiments performed by singular researchers or small groups can be enormously valuable. Articles that share data into a repository and include a link to it have been found to have approximately 25% higher citation impact[5]. Clearly, work with well-reported data facilitates reanalysis and can serve as a resource long beyond the typical lifespan of a paper.

We wish to address two concerns that limit facilitating reanalysis: a fear of scrutiny and concern about the additional workload required.

In one study, more than 50% of scientists identified misinterpretation and scrutiny as a reason that data were not shared, including concern about different results or criticism[6]. We advocate for a change of perspective, where the discovery of different results from the same data is viewed as advancement. Different results can mean that new analysis methods have been proven, an interesting phenomenon has been found or an otherwise overlooked effect has been detected.

Second, organizing and opening up data are difficult and add additional burden to the scientific endeavour. A recent meta-analysis found that researchers lack the time, resources and skills to effectively share their data[7]. In fields with high publication rates, it can be easy to put aside the extra work required to enable subsequent reanalysis. Not every paper or project warrants a significant effort towards reanalysis; however, by adopting open data principles, laboratories and research organizations will be in a better position to provide the level of detail needed. Additionally, the steps required to adopt open principles facilitate reproducibility.

Practical tips to facilitate reanalysis

One of the hardest decisions to make is how raw should the data be? Typically, the rawer the data, the more data curation is required (but also the more potential for reanalysis). A way to address this is to provide at least one level deeper of raw data to that shown in a figure. For example, flow cytometry analysis often displays summary statistics, but raw flow cytometry standard files are seldom provided in published work; these files are highly valuable for reanalysis. Similarly, protein expression levels are often displayed in biomolecular corona work but the raw output from proteomic measurements is rarely provided.

[1] Department of Biomedical Engineering, The University of Melbourne, Parkville, Victoria, Australia.

[2] Department of Chemical Engineering, The University of Melbourne, Parkville, Victoria, Australia.

✉e-mail: fcaruso@unimelb.edu.au

https://doi.org/10.1038/s43586-022-00103-z

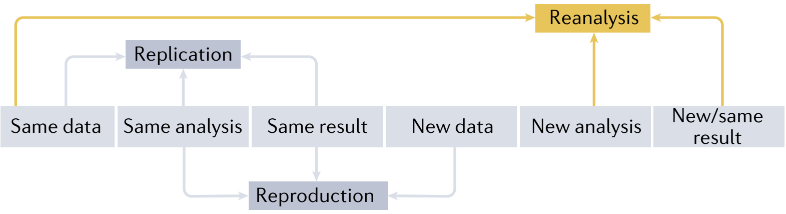

Fig. 1 | **Relationship between reproducibility, replicability and reanalysis.** Replication finds the same results with the same data and analysis. Reproduction finds the same results with the same analysis but with new data. Reanalysis applies new analysis to the same data and may (or may not) find the same results.

It helps to have an internal champion for open data and reanalysis within a group or as a close collaborator. Theoretical, mathematical and computational scientists are more aligned with developing new analysis techniques for existing data.

Machine-readable formats that describe an experiment, and the location of corresponding data, can be very valuable. Coming up with a reasonable format is within the reach of most scientists and makes data far easier to interpret and reuse. For instance, we have experimented with organizing libraries of flow cytometry time-course experiments with INI files[8]. Each INI file contains details of how the experiment was done and the location of the corresponding output data files. INI files are both human-readable and easy for a script to interpret; this has allowed interesting reanalysis of the same data library[9].

Planning should go beyond single papers or projects to longer timescales. Continuous accumulation of organized data is extremely valuable for research advancement. The use of presubmission group checklists[10] for data provision can help organize data over longer timescales and help to establish an internal culture of data organization and provision.

Needs and next steps

In materials science and chemistry, there is much work still to do to promote open data and improve the facilitation of reanalysis. How can the community meet this need? First, the type of data needed for reanalysis typically does not fit in the supplementary information section of a scientific publication. Instead, tools such as figshare and Zenodo and those from the Center for Open Science are useful to share larger, organized datasets. Better integration of these tools with scientific publications would make good data easier to find. Second, we should provide additional recognition for researchers, groups and organizations who take steps to facilitate reanalysis; efforts like the Research Symbiont Awards should be expanded. This is analogous to other initiatives to recognize core scientific endeavours outside of paper publication, such as cataloguing peer review by Publons or attempts by the Journal of Open Source Software to reward well-documented, usable open source scientific software. Third, additional attention to standards would be valuable, especially standardized formats and repositories for domain-specific data. These have proven valuable for protein structure, genetics and -omics data, just to name a few. Ultimately, the provision of good data to facilitate reanalysis is worth it, and we should celebrate those who do.

1. Plesser, H. E. Reproducibility vs. replicability: a brief history of a confused terminology. *Front. Neuroinf.* **11**, 76 (2018).
2. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
3. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**, e67995 (2021).
4. Leong, H. S. et al. On the issue of transparency and reproducibility in nanomedicine. *Nat. Nanotechnol.* **14**, 629–635 (2019).
5. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. & McGillivray, B. The citation advantage of linking publications to research data. *PLoS ONE* **15**, e0230416 (2020).
6. Kim, Y. & Stanton, J. M. Institutional and individual influences on scientists' data sharing practices. *J. Comput. Sci. Educ.* **3**, 47–56 (2012).
7. Perrier, L., Blondal, E. & MacDonald, H. The views, perspectives, and experiences of academic researchers with data sharing and reuse: a meta-synthesis. *PLoS ONE* **15**, e0229182 (2020).
8. Faria, M. et al. Revisiting cell–particle association in vitro: a quantitative method to compare particle performance. *J. Control. Release* **307**, 355–367 (2019).
9. Johnston, S. T., Faria, M. & Crampin, E. J. Isolating the sources of heterogeneity in nano-engineered particle–cell interactions. *J. R. Soc. Interface* **17**, 20200221 (2020).
10. Faria, M. et al. Minimum information reporting in bio–nano experimental literature. *Nat. Nanotechnol.* **13**, 777–785 (2018).

Competing interests
The authors declare no competing interests.