support for Brazilian science

best practices guide in open and reproducible science

serrapilheira

best practices guide in open and reproducible science

Among the basic values of science is the idea that knowledge is part of the common heritage of humanity.[1] The advance of science depends on the validity and strength of its results as well as on the possibility of independently verifying and reproducing them.

Nonetheless, some aspects of the scientific environment set up barriers to data access and limit the reliability of what is published. Due to an incentive system based on publication in specialized journals that are usually associated with commercial publishers, less than half of all published articles are open access.[2] An even smaller fraction of these provide all of their data for independent verification.[3,4] This may be why recent surveys in fields like psychology and biomedical science suggest that most published results are not reproducible,[5-7] which implies wasted resources and delays in advancing science.

As an institution dedicated to supporting science in Brazil, Serrapilheira reaffirms its commitment to open science and its expectation that findings from research funded by the institute be published in a transparent manner so that they are accessible, verifiable and reproducible. The guidelines listed below are suggestions for best practices at different stages in the scientific process to achieve these goals.

Obviously, not all the recommendations apply directly to all fields of research. Theoretical physicists and mathematicians, for instance, need not follow orientations related to conducting experiments. On the other hand, guidelines on topics such as scientific publication are valid for all fields, albeit with their own specificities.

Upon developing a project:

- Perform a careful literature review before starting a project, be it experimental or theoretical. This step will help to avoid duplicating efforts and ensure that authorship of ideas and previous findings is duly attributed. The use of systematic review protocols with explicit research methodologies makes the process more impartial, reproducible and robust as a scientific contribution.[8]

- Registering your protocol before collecting data lends greater transparency to the process and makes it possible to verify whether data collection and analysis were carried out as planned in order to differentiate confirmatory from exploratory analyses.[9,10] Additionally, this practice allows identification of registered but unpublished studies.[11] This step is mandatory for clinical trials[12,13] and can be adopted in different ways in other fields of research. Repositories for protocol pre-registration include Open Science Framework and AsPredicted. Protocols may be made publicly available upon registration or embargoed until the data is published.

- One way of getting feedback before carrying out the study—when critiques and suggestions are more likely to have an impact on the project—is to have the protocols peer-reviewed, either as an independent article or as a preliminary submission in journals that accept the Registered Report format, where the methods are reviewed, thus providing preliminary approval of the study if the protocol is followed, regardless of the result.[14] A list of journals that accept this format is available at the Center for Open Science.

- Before collecting data, make a plan for statistical analysis that includes a sample size calculation. This avoids conducting experiments with low statistical power, which waste resources and generate less reliable results.[15] Many online calculators Power and Sample Size can be used to carry out these calculations.

- When designing your project, remember to include measures and controls for obtaining robust conclusions regardless of the result. It is important to establish criteria for differentiating a "negative" result (in which a certain experimental intervention does not generate the expected result) from a flawed experiment due to methodological problems, to ensure that the latter be sufficiently reliable for publication. Publishing negative results is a fundamental measure to avoid the distortion of scientific knowledge through publication bias.[16,17]

- Establish a data management plan that defines where the data will be stored during and after the project. Include a strategy for copying the data to have a backup and to ensure long-term accessibility, regardless of researcher mobility between laboratories/institutions. Many repositories, either tied to specific institutions or not, are available for this purpose —a list can be seen at FAIRsharing.org. The FAPESP website also offers guidelines on how to develop a data management plan.

Upon running experiments:

- Take measures to control bias, such as blinding experimenters to the identity of the groups while running experiments and evaluating outcomes. Even though the importance of these procedures has been recognized for decades, their adoption by various research areas, like preclinical animal research, is still unacceptably low for a procedure that should be routinely performed.[18]

- Adopt a system for recording protocols and results, such as a lab notebook or an electronic tool that does this job. Train those in charge of data collection and make it clear that recording protocols and results is mandatory and is the documentation of what was done and cannot be erased or adulterated. For more guidelines about laboratory notebooks, see the National Institutes of Health website.

- Try to get to know the intra- and inter-laboratory reproducibility of the methods you use. If they have not been established, consider the possibility of replicating key data from your project among different members of your laboratory or among different laboratories. Even though it may not be possible to do this for all data, taking this step can be a valuable practice for confirmatory experiments.[19]

- Establish protocols for quality control and checking of common methodological errors in order to periodically evaluate the quality of the data generated. If such procedures lead to excluding data or experiments, establish criteria so that this takes place before obtaining results and explicitly mention them in your description of methods to keep these exclusions from creating bias in the results.

Upon analyzing results:

- Keep your data organized and document the analyses carried out, just as you would do with experimental findings. If you use your own code for analysis, make sure it is documented and understandable to third parties. Consider integrating your code with your results by using tools like Jupyter, R Markdown, knitr or Sweave, which allow one to easily repeat analyses and simulations.

- When using inferential statistics, remember that this process makes assumptions about data and requires hypotheses to be established a priori — and ideally registered beforehand. We are not opposed to the use of exploratory analyses; however, they must be clearly described as such and set apart from analyses designed to test a priori hypotheses.[20] Data-driven statistical models, such as those generated through machine learning, must be tested in sets of data that are different from those in which they were generated in order to avoid circularity in the analyses.[21,22]

- Avoid the common dichotomy between "significant" and "non-significant" results. If you use frequentist statistics, provide exact P-values and interpret them in terms of the plausibility of your hypothesis and the statistical power of the experiment before considering that a result is probably true.[23,24] If desired, this type of logic can also be formalized in Bayesian analyses that take the a priori probability attributed to a hypothesis into consideration.

- Keep in mind that statistical significance and effect magnitude are different concepts and consider alternatives for clearly expressing both of these attributes upon presenting your data, such as using effect size measures associated with confidence intervals.[25]

- Regardless of the results of your analyses, keep in mind that any inference based on a sample includes the possibility of error. Be candid when expressing uncertainty, both for positive and negative results, and address the limitations of your methods and analyses in discussing your results.

Upon publishing results:

- Use open access platforms for publication. These may include both open access journals (gold open access) and depositing of article versions in public repositories (green open access). For more information regarding open access, visit the Budapest Open Access Initiative and cOAlition S websites Regarding the green route, most publishers have policies that are compatible with the depositing of pre- and post-print versions— to check the polices of a particular publisher or journal, visit SHERPA/RoMEO. Upon publishing or depositing an article, use sharing options that guarantee that the article can be used and redistributed, such as Creative Commons licenses.

- Should you choose open access journals that charge publication fees (noting that this is not the only way of guaranteeing open access), aim to make sure that they have a real peer review system and do not fall under the category of "predatory journals." A list of open access journals that meet minimum quality control standards can be found at Directory of Open Access Journals.

- Deposit your articles as preprints before or upon submission in repositories such as arXiv, bioRxiv, chemRxiv, PsyArXiv, medRxiv and others (a wide-ranging list is periodically updated at Research Preprints). Using preprints accelerates the scientific process and is compatible with submitting articles to most journals (to search specific policies, see SHERPA/RoMEO ou a Wikipedia). After peer review, update your preprint so that it is as close to the final publication as possible— this will also act as a way of achieving green open access.

- Citing preprints in scientific articles and commenting on preprints (i.e. post-publishing peer review) is also encouraged. Furthermore, preprints should be considered to be valid scientific production for evaluating projects and researchers. For more information about preprints, especially in the life sciences, visit ASAPbio.

- Describe your methodology as broadly as possible so that other researchers can replicate your results. If the journal has space restrictions or word limits, provide supplementary material or deposit the protocol in a repository, ideally using platforms made for this purpose such as protocols.io. If there are reporting guidelines available for your field of research, use them as a checklist on what needs to be described. For a collection of guidelines in the field of the health and life sciences, visit the Equator Network.

- Fully share the data from your research upon submitting your article. The data can be included as supplementary material or deposited in repositories, be they institutional, general (e.g. Zenodo, Dryad, figshare, Dataverse) or specialized in nature. A list of repositories by area is available at FAIRsharing.org. If there is consensus about standardizing data in your field, you should adopt structured formats. Otherwise, include a manual describing the data, as well as metadata to make them accessible to automated searches. In any case, make sure that your data have a DOI for citation. Merely stating that data will be made available upon contacting the authors is insufficient, as such a practice does not guarantee access. [27,28] For more information about the principles of depositing scientific data, check the FAIR principles.

- When it comes to sensitive data that cannot be shared in full, such as personal information of volunteers or data that requires protection due to intellectual property issues, consider whether they can at least be made partially available (e.g., by removing personal information to guarantee volunteers' anonymity) or after an embargo (e.g., after intellectual property is established). For more information about data privacy, see the Brazilian General Law for Data Protection and the guidelines of the European Commission regarding the protection of data used in research.

- Should your research study use code developed by you or a third party, include it as supplementary information or deposit it in a repository like GitHub. Prioritize open source languages like R or Python, but make your code available even if you work with a proprietary language. Again, consider integrating the code into your results with tools like Jupyter, R Markdown, knitr or Sweave. Cite third-party codes and databases with their respective DOIs in your research in order to recognize them as valid scientific outputs.

- Attribute authorship in a way that is fair and compatible with the criteria in your area of research. Even though these may vary depending on the field, general guidelines can be found at the Comittee on Publication Ethics (COPE), and several areas and journals use the criteria of the International Committee of Medical Journal Editors. For greater clarity, we recommend that the contributions of each author be detailed in a separate section, ideally using standardized taxonomies like CRedIT.

- If you have any conflict of interest related to your results—be it financial or otherwise—be transparent by stating it when publishing and presenting results. Although conflicts of interest are inevitable at times, taking measures to minimize them contributes to the impartiality and integrity of your research findings. For more information and definitions regarding conflicts of interest, visit COPE.

Always:

- Engage in discussions about open and reproducible science, especially when it comes to training new scientists. Provide your students and collaborators with the right incentives, making it clear that you will evaluate them by the effort and rigor they invest in the research, and not for the result obtained. Raise issues related to scientific reliability in discussions with your research group and institution, and stay connected to literature on this topic.[29,30]

- When evaluating your peers as an editor, reviewer or committee member, take into account the evaluated researcher's scientific rigor, transparency and commitment to open science. Always be skeptical and consider whether the results you are evaluating are solid and reliable before asking about their impact.

- Consider science as a practice that needs constant review. Value, discuss or perform confirmatory replication studies in your field of research.[31] Register and publish your attempts at replication, regardless of whether they are successful. Engage in the evaluation of published science by participating in post-publication peer-review forums like PubPeer, and the comment sections of journals and preprint repositories. When considering working as a reviewer for journals, prioritize those that make revisions available,[32] either signed or anonymous. Be courteous and respectful at all times by giving constructive criticism that contributes to the advancement of science.

- Likewise, be open to criticism from colleagues and do not see critiques of your data as personal attacks. Be open to the possibility of correcting and reanalyzing your data by collaborating with your critics whenever possible. If mistakes or inconsistencies are found, be transparent when admitting them and use existing mechanisms for making corrections and retractions in the literature. Remember that admitting mistakes honestly is not a blemish on the researcher's reputation, and that the ethical consequences of not correcting them are far more serious.[33]

The recommendations made in this guide were developed by **Olavo Amaral** (Institute of Medical Biochemistry Leopoldo de Meis at the Federal University of Rio de Janeiro) and reviewed by **grantees of the Serrapilheira Institute** working in different fields of research. Coordination: **Cristina Caldas**.
Editing and review: **Maria Emilia Bender** and **John Ellis**.

Serrapilheira Institute
Director President: Hugo Aguilaniu
Director of Scientific Research: Cristina Caldas
Director of Science Outreach: Natasha Felizi

References

1. Merton RK. **The normative structure of Science**. In: Merton, RK. *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, University of Chicago Press, 1942. https://doi.org/10.2307/2576098

2. Piwowar H, Priem J, Larivière V et al. **The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles**. *PeerJ* 2018; 6: e4375. https://doi.org/10.7717/peerj.4375

3. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. **Public availability of published research data in high-impact journals**. *PLoS One* 2011; 6: e23457. https://doi.org/10.1371/journal.pone.0024357

4. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA. **Reproducible research practices and transparency across the biomedical literature**. *PLoS Biol* 2016; 14: e1002333 https://doi.org/10.1371/journal.pbio.1002333

5. Open Science Collaboration. **Estimating the reproducibility of psychological science**. *Science* 2015; 349: aac4716. https://doi.org/10.1126/science.aac4716

6. Begley CG, Ellis LM. **Drug development: Raise standards for preclinical cancer research**. *Nature* 2012; 483: 531–533. https://doi.org/10.1038/483531a

7. 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* 2012; 10: 712. https://doi.org/10.1038/nrd3439-c1

8. Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews of Intervention*. Version 5.1.0. The Cochrane Collaboration, 2011. Disponível em http://www.handbook.cochrane.org. https://doi.org/10.1002/9780470712184

9. Kimmelman J, Mogil JS, Dirnagl U. **Distinguishing between exploratory and confirmatory preclinical research will improve translation**. *PLoS Biol* 2014; 12: e1001863. https://doi.org/10.1371/journal.pbio.1001863

10. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. **The preregistration revolution**. *Proc Nat Acad Sci USA* 2018; 115: 2600-2606. https://doi.org/10.1073/pnas.1708274114

11. Goldacre B, DeVito N, Heneghan C, Lane S, Stephens R. **Introducing unreported clinical trial of the week**. *BMJ blog*, disponível em https://blogs.bmj.com/bmj/2018/03/29/it-is-time-to-fix-the-problem-of-unreported-clinical-trials/

12. Zarin D, Tse T, Williams RJ, Califf RM. **The ClinicalTrials.gov results database – update and key issues**. *N Engl J Med* 2011; 364: 852-860. https://doi.org/10.1056/nejmsa1012065

13. World Health Organization. **International Clinical Trials Registry Platform (ICTRP)**. Disponível em https://www.who.int/ictrp/en/

14. Chambers C, Munafò M et al. **Trust in science would be improved by study pre-registration**. *Guardian*, 5/6/2013. Disponível em: https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

15. Button KS, Ioannidis JP, Mokrysz C et al. **Power failure: why small sample size undermines the reliability of neuroscience**. *Nat Rev Neurosci* 2013; 14: 365-376. https://doi.org/10.1038/nrn3475

16. Dwan K, Altman DG, Arnaiz JA et al. **Systematic review of the empirical evidence of study publication bias and outcome reporting bias**. *PLoS One* 2008; 3: e3081. https://doi.org/10.1371/journal.pone.0003081

17. Moorthy VS, Karam G, Vannice KS, Kieny MP. **Rationale for WHO's new position calling for prompt reporting and public disclosure of interventional clinical trial results**. *PLoS Med* 2015; 12: e1001819. https://doi.org/10.1371/journal.pmed.1001819

18. Macleod MR, McLean AL, Kyriakopoulou A. et al. **Risk of bias in reports of in vivo research: a focus for improvement**. *PLoS Biol* 2015; 13: e1002273. https://doi.org/10.1371/journal.pbio.1002273

19. Mogil JS, Macleod MR. **No publication without confirmation**. *Nature* 2017; 542: 409-411. https://doi.org/10.1038/542409a

20. Kerr NL. **HARKing: hypothesizing after the results are known**. *Pers Soc Psychol Rev* 1998; 2: 196-217. https://doi.org/10.1207/s15327957pspr0203_4

21. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. **Circular analysis in systems neuroscience: the dangers of double dipping**. *Nat Neurosci* 2009; 12:535-540. https://doi.org/10.1038/nn.2303

22. Baştanlar Y, Özuysal M. **Introduction to machine learning**. *Methods Mol Biol* 2014; 1107:105-128. https://doi.org/10.1007/978-1-62703-748-8_7

23. Gigerenzer G. **Statistical rituals: the replication delusion and how we got there**. *Adv Meth Pract Psychol Sci* 2018; 1: 198-218. https://doi.org/10.1177/2515245918771329

24. Nuzzo R. **Statistical errors**. *Nature* 2014; 506: 150-152. https://doi.org/10.1038/506150a

25. Cumming G. T**he new statistics: why and how**. *Psychol Sci* 2014; 25: 7-29. https://doi.org/10.1177/0956797613504966

26. Clark J, Smith R. **Firm action needed on predatory journals**. *BMJ* 2015: 350. https://doi.org/10.1136/bmj.h210

27. Hardwicke TE, Ioannidis, JPA. **Populating the Data Ark: an attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles**. *PLoS One* 2018; 13: e0201856. https://doi.org/10.1371/journal.pone.0201856

28. Stodden V, Seller J, Ma Z. **An empirical analysis of journal policy effectiveness for computational reproducibility**. *Proc Natl Acad Sci USA* 2018; 115: 2584-2589. https://doi.org/10.1073/pnas.1708290115

29. Munafò MR, Nosek BA, Bishop DVM et al. **A manifesto for reproducible science**. *Nat Hum Behav* 2017 1: 0021. https://doi.org/10.1038/s41562-016-0021

30. Ioannidis JPA. **How to make more published research true**. *PLoS Med* 2014; 11: e1001747. https://doi.org/10.1371/journal.pmed.1001747

31. Nature. **Go forth and replicate!** *Nature* 2016; 536: 373. https://doi.org/10.1038/536373a

32. Polka JK, Kiley R, Konforti B, Stern B, Vale RD. **Publish peer reviews**. *Nature* 2018; 560: 545-547. https://doi.org/10.1038/d41586-018-06032-w

33. Rohrer JM, DeBruine LM, Heyman T, et cl. **Putting the self in self-correction**. *PsyArXiv* 2018. https://doi.org/10.31234/osf.io/exmb2