A practical guide for improving transparency and reproducibility in neuroimaging research

Krzysztof J. Gorgolewski and Russell A. Poldrack
Department of Psychology, Stanford University

Abstract

Recent years have seen an increase in alarming signals about the lack of replicability in neuroscience, psychology, and other related fields. To avoid a widespread crisis in our field and consequent loss of credibility in the public eye, we need to improve how we do science. This article aims to be a practical guide for researchers at any stage of their careers that will help them make their research more reproducible and transparent while minimizing the additional effort that this might require. The guide covers three major topics in open science (data, code, and publications) and offers practical advice as well as highlighting advantages of adopting more open research practices that go beyond improved transparency and reproducibility.

Introduction

The question of how the brain creates the mind has captivated humankind for thousands of years. With recent advances in human *in vivo* brain imaging, we how have effective tools to peek into biological underpinnings of mind and behavior. Even though we are no longer constrained just to philosophical thought experiments and behavioral observations (which undoubtedly are extremely useful), the question at hand has not gotten any easier. These powerful new tools have largely demonstrated just how complex the biological bases of behavior actually are. Neuroimaging allows us to give more biologically grounded answers to burning questions about human behavior. Questions that not only everyone can relate to (“why do we crave things?”, “how do we control learned responses?”, “how do we regulate emotions?” etc.), but that can also influence how we think about mental illnesses.

In addition to fantastic advances in terms of hardware we can use to study the human brain (function Magnetic Resonance Imaging, Magnetoencephalography, Electroencephalography etc.) we have also witness a many new developments in terms of data processing and modelling. Many bright minds have contributed to a growing library of methods that derive different features from brain signals. Those methods have widened our perspective on brain processes, but also resulted in methodological plurality [1]. Saying that there is no single way to analyze a single neuroimaging dataset is an understatement; we can confidently say that there many thousands ways to do that.

The passion driving brain researchers combined with methodological plurality can be a dangerous mix. As Richard Feynman said “The first principle is that you must not fool yourself — and you are the easiest person to fool.”. Recent years have seen an increase in alarming signals about the lack of replicability in neuroscience, psychology, and other related fields [2]. Neuroimaging studies are usually statistically underpowered due to the high cost of data collection, making the problem even harder. To avoid a widespread crisis in our field and consequently losing credibility in the public eye, we need to improve how we do science. This article aims to be a practical guide for researchers at any stage of their careers that will help them make their research more reproducible and transparent while minimizing the additional effort that this might require.

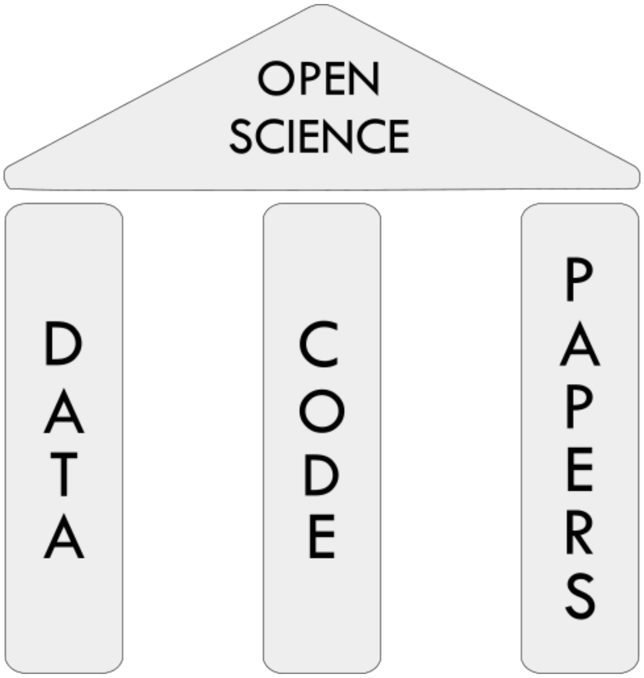

Figure 1. Three pillars of Open Science: data, code, and papers.

How to deal with data

Data are probably the most important component of the scientific process. It not only allows the scientific community to validate the accuracy of scientific findings, but also empowers researchers to perform novel analyses or combine data from multiple sources. Papers accompanied by publicly available data are on average cited more often [3,4], while at the same time exposing fewer statistical errors [5]. Data sharing has been mandated by some grant funding agencies, as well as journals. Some also argue that sharing data is an ethical obligation toward study participants, in order to maximize the benefits of their participation [6]. Neuroimaging has a substantial advantage in terms of ease of data capture since data generation process is completely digital. In principle one could provide a digital record of the entire research process for the purpose of reproducibility. However, even though data sharing in neuroimaging has been extensively reviewed in [7] and [8] there is little practical advice on the topic.

Consent forms

Planning for data sharing should start at the ethical approval stage. Even though in the United States de-identified data can be freely shared without specific participant consent, the rules differ in other countries (and they may change in the upcoming revisions to the Common Rule, which governs research in the US). In addition it is more than fair to inform your participants about your intention to maximize their generous gift by sharing their data, and to allow them to withdraw from research if they don't wish to have their data shared. However, consent form language needs to be carefully crafted. To streamline the creation of consent forms with data sharing clauses, we have prepared a set of templates that can be easily inserted into existing consent forms after minor adjustments[1]. Those templates have been derived from existing consent forms of leading data sharing projects (Nathan Kline Institute Enhanced sample [9] and Human Connectome Project [10]) followed by consultations with bioethics experts. The templates come in two flavors: one for normal populations and generic data and one for sensitive populations and/or data. The latter splits the data into two sets: a publicly available portion and a portion that requires approval of a data sharing agreement in order to gain access to. We recommend using the restricted access version only for data and populations for which potential data re-identification would pose a substantial risk to the participants (for example in a study of HIV-positive subjects).

Data organization

To successfully share data one has to properly describe it. Even though some experimental details such as the MRI phase encoding direction may seem obvious for the researcher who obtained the data, they need to be clearly explained for external researchers. In addition, good data organization and description can reduce mistakes in analysis. While each experiment is different and may include unique measurements or procedures, most MRI datasets can be accurately described using one fairly simple scheme. Recently we have proposed such scheme - the Brain Imaging Data Structure (BIDS) [11]. It was inspired by the data organization used by OpenfMRI database, but has evolved through extensive consultations with the neuroimaging community. BIDS aims at being simple to adopt,and roughly follows existing practices common in the neuroimaging community. It is heavily based on a specific organization of files and folders and uses simple file formats such as NifTI, TSV and JSON. It does not require a database or any external piece of software for processing. A browser-based validator has been developed that allows one to easily check whether a dataset accurately follows the BIDS standard[2].

An additional benefit of using a standardized data organization scheme is that it greatly streamlines the data curation that is necessary when submitting data to a data sharing repository. For example datasets formatted according to BIDS undergo speedier and more streamlined curation process when submitted to OpenfMRI database [12].

Publishing data

Data should be submitted to a repository before submitting the relevant paper. This allows the author to point the readers to the location of the data in the manuscript. The manuscript can benefit from increased transparency due to shared data and the data itself can become a resource enabling future research.

[1] https://open-brain-consent.readthedocs.org/en/latest/ultimate.html
[2] http://incf.github.io/bids-validator

The most appropriate places for depositing data are field specific repositories. Currently in human neuroimaging there are two well recognized repositories accepting data from everyone: FCP/INDI [13] (for resting state fMRI only) and OpenfMRI [12] (for any datasets that includes MRI data). Field specific repositories have the advantage of more focused curation process that can greatly improve the value of your data. They also increase data discoverability since researchers search through them first when looking for datasets, and some are indexed by PubMed which allows the dataset to be directly linked to the paper via the LinkOut mechanism.

If for some reason field specific repositories are not an option we recommend using field agnostic repositories such as FigShare, Dryad, or DataVerse. When picking a repository one should think of long term data retention. No one can guarantee existence of a repository in the far future, but historical track record and the support of well established institutions can increase the chances that the data will be available in the decades to come.

Since data-agnostic repositories do not impose any restriction on the form in which you deposit your data nor do they check completeness you have to make sure that all of the necessary data and metadata are present. Using a data organization scheme designed for neuroimaging needs such as BIDS or XCEDE [14] can help you make sure data is represented accurately. In addition, it is a good idea to ask a colleague who is unfamiliar with the data to evaluate the quality and completeness of the description.

If the data accompanying the paper is very large or particularly complex you should consider writing a separate data paper about it [15]. A data paper is new type of publication dedicated purely to description of the data rather than its analysis. It will give you more space to describe the experimental procedures and data organization details, and also provides a mechanism for credit when the data are reused in the future. In addition you will get feedback about the dataset description through the peer review process. The list of journals that currently accept neuroimaging data papers includes but is not limited to: Scientific Data, Gigascience, Data in Brief, F1000Research, Neuroinformatics, and Frontiers in Neuroscience.

In addition to raw data we also encourage authors to share derivatives such as preprocessed volumes, statistical maps or tables of summary measures. Because other researchers are often interested in reusing the results rather than the raw data, this can further increase the reuse of the data. For example, statistical maps can be used to perform image-based meta analysis or derive regions of interest for new studies. For sharing statistical maps we encourage authors to use the NeuroVault.org platform [16].

How to deal with code

Neuroimaging data analysis has required computers since its inception. A combination of compiled or script code is involved in every PET, MRI, or EEG study, as in most other fields of science. The code we write to analyze data is a vital part of the scientific process, and similar to data, is not only necessary to interpret results, but can be also used to address new research questions. Therefore the sharing of code is as important as the sharing of data for scientific transparency and reproducibility.

Because most researchers are not trained in software engineering, the code that is written to analyze neuroimaging data (as in other areas of science) is often undocumented and lacks the formal tests that professional programmers use to ensure accuracy. In addition to the lack of training, there are few incentives to spend the time necessary to generate high-quality and well-documented code. Changes in the incentive structure of science will take years, but in the meantime, perceived poor quality of code and lack of thorough documentation should not prevent scientists from publishing it [17]. Sharing undocumented code is a much better than not sharing code at all.

An additional concern that stops researchers from sharing code is fear that they will have to provide user support and answer a flood of emails from other researchers who may have problems understanding the codebase. However, sharing code does not oblige a researcher to provide user support. One useful solution to this problem is to set up a mailing list (for example with Google) and point all users to ask questions through it; in this way, answers are searchable, so that future users with the same questions can find them via a web search. Alternatively one can point user to a community driven user support forum for neuroinformatics (such as NeuroStars.org) and ask them to tag their questions with a label uniquely identifying the software or script in question. Both solutions foster a community that can lead to users helping each other with problems, thus relieving some of the burden from the author of the software. In addition, since the user support happens through a dedicated platform there is less pressure on the author to immediately address issues than there would be with user requests send directly by email.

Many of the issues with code quality and ease of sharing can be addressed by careful planning. One tool that all research programmers should incorporate into their toolbox is the use of a Version Control System (VCS) such as git. VCS provides a mechanism for taking snapshots of evolving codebase that allow tracking of changes and reverting them if there is a need (e.g., after making a change that ends up breaking things). Adopting a VCS leads a to cleaner code base that is not cluttered by manual copies of different versions of a particular script (e.g, “script_version3_good_Jan31_try3.py”). VCS also allows you to quickly switch between branches - alternative and parallel versions of the codebase - to test a new approach or method without having to alter a tried and tested codebase. For a great introduction to git we refer the reader to [18]. We encourage scientists to use git rather than other VCS due to a passionate and rapidly growing community of scientists who use the GitHub.com platform, which is a freely available implementation of the git VCS system. In the simplest use case GitHub is a platform for sharing code (which is extremely simple for those who already use git as their VCS), but it also includes other features which make contributing to collaborative projects, reviewing, and testing code simple and efficient.

Striving for automation whenever possible is another strategy that will not only result in more reproducible research, but can also save a lot of time. Some analysis steps seem to be easy to perform manually, but that remains true only when they need to be performed just once. Quite often in the course of a project parameters are modified, list of subjects are changed, and processing steps need to be rerun. This is a situation in which having a set of scripts that can perform all of the processing steps automatically instead of relying on manual interventions can really pay off. There are many frameworks that help design and efficiently run neuroimaging analyses in automated fashion. Those include, but are not limited to: Nipype [19], PSOM [20], aa [21], and make [22]. As an example, for our recent work on the MyConnectome project we created a fully automated analysis pipeline, which we implemented using a virtual machine[3].

While automation can be very useful for reproducibility, the scientific process often involves interactive interrogation of data interleaved with notes and plots. Fortunately there is a growing set of tools that facilitate this interactive style of work while preserving a trace of all the computational steps, which increases reproducibility. This philosophy is also known as “literate programming” [23] and combines analysis code, plots, and text narrative. The list of tools supporting this style of work includes, but is not limited to: Jupyter (for R, Python and Julia)[4], R Markdown (for R)[5] and matlabweb (for MATLAB)[6]. Using one of those tools not only provides the ability to revisit an interactive analysis performed in the past, but also to share an analysis accompanied by plots and narrative text with collaborators. Files created by one of such systems (in case of Jupyter they are called Notebooks) can be shared together with the rest of the code on GitHub, which will automatically render included plots so they can be viewed directly from the browser without requiring installation of any additional software.

How to deal with publications

Finally the most important step in dissemination of results is publishing a paper. An essential key to increasing transparency and reproducibility of scientific outputs is accurate description of methods and data. This not only means that the manuscript should include links to data and code mentioned before (which entails that both data and code should be deposited before submitting the manuscript), but also thorough and detailed description of methods used to come to a given conclusion. As an author one often struggles with a fine balance between detailed description of different analyses performed during the project and and the need to explain the scientific finding in the most clear way. It is not unheard of that for the sake of a better narrative some results are omitted[7]. At the same time there is a clear need to present results in a coherent narrative with a clear interpretation that binds the new results with an existing pool of knowledge[8]. We submit that one does not have exclude the other. A clear narrative can be provided in the main body of the manuscript and the details of methods used together with null results and other analyses performed on the dataset can be included in the supplementary materials, as well as in the documentation of the shared code. In this way, the main narrative of

[3] https://github.com/poldrack/myconnectome-vm
[4] http://jupyter.org
[5] http://rmarkdown.rstudio.com
[6] https://www.ctan.org/pkg/matlabweb
[7] http://sometimesimwrong.typepad.com/wrong/2015/11/guest-post-a-tale-of-two-papers.html
[8] http://www.russpoldrack.org/2015/11/are-good-science-and-great-storytelling.html

the paper is not obfuscated too many details and auxiliary analyses, but all of the results (even null ones) are available for the interested parties. Often these extra analyses and null results may seem uninteresting from the author's point of view, but one cannot truly predict what other scientists can be interested in. In particular, the null results (which are difficult to publish independently) can contribute to growing body of evidence.

The last important topic to cover is accessibility of the manuscript. To maximize the impact of published research one should consider making the manuscript publicly available. In fact many funding bodies (NIH, Wellcome Trust) require require this for all manuscripts describing research they funded. Many journals provide an option to make papers open access, albeit sometimes at prohibitively high price (for example the leading neuroimaging journal - NeuroImage - requires a fee of $3000). Unfortunately most prestigious journals (Nature and Science) do not provide such option despite many requests from the scientific community. Papers published in those journals remain “paywalled” - available only through institutions which pay subscription fees, or through public repositories (such as PubMed Central) after a sometimes extended embargo period. The scientific publishing landscape is changing [24,25], and we hope it will evolve in a way that will give everyone access to published work as well as to the means of publication. In the meantime we recommend ensuring open access by publishing preprints at BioArxiv or Arxiv before submitting the paper to a designated journal. In addition to making the manuscript publicly available without any cost, this solution has other advantages. Firstly it allows the wider community to give feedback to the authors about the manuscript and potentially improve it. Secondly, in case of hot topics publishing a preprint establishes precedence on being the first one to describe a particular finding. Finally since preprints have assigned DOIs other researchers can reference them even before they will be published in a journal. Preprints are increasingly popular and vast majority of journals accept manuscripts that have been previously published as preprints. We are not aware of any neuroscience journals that do not allow authors to deposit preprints before submission, although some journals such as Neuron and Current Biology consider each submission independently and thus one should contact the editor prior to submission.

Summary

The scientific method is evolving towards a more transparent and collaborative endeavour. The age of digital communication allows us to go beyond printed summaries and dive deeper into underlying data and code. In this guide we hope to have shown that there are many improvements in scientific practice everyone can implement with relatively little added effort that will improve transparency, replicability and impact of their research.

Acknowledgements

This work was funded by Laura and John Arnold Foundation.

References

1. Carp J. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI

experiments. Front Neurosci. 2012;6: 149. doi:10.3389/fnins.2012.00149

2. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349. doi:10.1126/science.aac4716

3. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS One. 2007;2: e308. doi:10.1371/journal.pone.0000308

4. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ. 2013;1: e175. doi:10.7717/peerj.175

5. Wicherts JM, Bakker M, Molenaar D. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. Tractenberg RE, editor. PLoS One. 2011;6: e26828. doi:10.1371/journal.pone.0026828

6. Brakewood B, Poldrack RA. The ethics of secondary data analysis: considering the application of Belmont principles to the sharing of neuroimaging data. Neuroimage. 2013;82: 671–676. doi:10.1016/j.neuroimage.2013.02.040

7. Poline J-B, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, et al. Data sharing in neuroimaging research. Front Neuroinform. 2012;6: 9. doi:10.3389/fninf.2012.00009

8. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. Nat Neurosci. Nature Publishing Group; 2014;17: 1510–1517. doi:10.1038/nn.3818

9. Nooner KB, Colcombe SJ, Tobe RH, Mennes M, Benedict MM, Moreno AL, et al. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. Front Neurosci. 2012;6: 152. doi:10.3389/fnins.2012.00152

10. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, et al. The WU-Minn Human Connectome Project: an overview. Neuroimage. 2013;80: 62–79. doi:10.1016/j.neuroimage.2013.05.041

11. Gorgolewski KJ, Auer T, Calhoun VD, Cameron Craddock R, Das S, Duff EP, et al. The Brain Imaging Data Structure: a standard for organizing and describing outputs of neuroimaging experiments [Internet]. bioRxiv. 2015. p. 034561. doi:10.1101/034561

12. Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, et al. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform. 2013;7: 1–12. doi:10.3389/fninf.2013.00012

13. Mennes M, Biswal BB, Castellanos FX, Milham MP. Making data sharing work: The FCP/INDI experience. Neuroimage. Elsevier Inc.; 2012; doi:10.1016/j.neuroimage.2012.10.064

14. Gadde S, Aucoin N, Grethe JS, Keator DB, Marcus DS, Pieper S. XCEDE: An Extensible Schema for Biomedical Data. Neuroinformatics. 2012;10: 19–32. doi:10.1007/s12021-011-9119-9

15. Gorgolewski KJ, Margulies DS, Milham MP. Making data sharing count: a publication-based

solution. Front Neurosci. Frontiers; 2013;7: 9. doi:10.3389/fnins.2013.00009

16. Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, et al. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front Neuroinform. Frontiers; 2015;9. doi:10.3389/fninf.2015.00008

17. Barnes N. Publish your computer code: it is good enough. Nature. 2010;467: 753. doi:10.1038/467753a

18. Blischak JD, Davenport ER, Wilson G. A Quick Introduction to Version Control with Git and GitHub. PLoS Comput Biol. 2016;12: e1004668. doi:10.1371/journal.pcbi.1004668

19. Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform. 2011;5: 13. doi:10.3389/fninf.2011.00013

20. Bellec P, Courchesne SL, Dickinson P, Lerch J, Zijdenbos A, Evans AC. The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows. Front Neuroinform. 2012;6. doi:10.3389/fninf.2012.00007

21. Cusack R, Vicente-Grabovetsky A, Mitchell DJ, Wild CJ, Auer T, Linke AC, et al. Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. Front Neuroinform. 2014;8: 90. doi:10.3389/fninf.2014.00090

22. Askren MK, McAllister-Day TK, Koh N, Mestre Z, Dines JN, Korman BA, et al. Using Make for Reproducible and Parallel Neuroimaging Workflow and Quality-Assurance. Front Neuroinform. Frontiers; 2016;10. doi:10.3389/fninf.2016.00002

23. Knuth DE. Literate programming. CSLI Lecture Notes, Stanford, CA: Center for the Study of Language and Information (CSLI), 1992. 1992; Available: http://adsabs.harvard.edu/abs/1992lipr.book.....K

24. The future of publishing: A new page. Nature. 2013;495: 425. doi:10.1038/495425a

25. Popova K. Evolution of Open Access Policies and Business Models: Which Way Leads to The Future? In: Scicasts [Internet]. [cited 5 Feb 2016]. Available: https://scicasts.com/insights/2123-open-science/10333-evolution-of-open-access-policies-and-business-models-which-way-to-the-future/