Be the Change You Seek in Science

Michael P. Milham, MD, PhD[1,2], Arno Klein, PhD[3]

[1]Center for the Developing Brain, Child Mind Institute, New York, New York.
[2]Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, New York, New York.
[3]MATTER Lab, Child Mind Institute, New York, New York.

Corresponding Author:
Michael Peter Milham, MD, PhD
Phyllis Green and Randolph Cowen Scholar
Child Mind Institute
101 East 56th Street
New York, NY 10022

ABSTRACT

Ongoing debates regarding the virtues and challenges of implementing open science for brain imaging research mirror those of the larger scientific community. The present commentary acknowledges the merits of arguments on both sides, as well as the underlying realities that have forced so many to feel the need to resist the implementation of an ideal. Potential sources of top-down reform are discussed, along with the factors that threaten to slow their progress. The potential roles of generational change and the individual are discussed, and a starter list of actionable steps that any researcher can take, big or small, is provided.

Keywords: Open Science, Data Sharing, Neuroimaging, Reproducibility, Transparency, Reform

After more than a decade of discussions and debates, both public and private, brain imaging research remains embroiled in controversies regarding open science. Myriad commentaries and calls for action have been published in journals of all impact levels, each helping to reinforce talking points for both sides of the debate (e.g., [1,2]). Open science advocates commonly draw attention to crises in reproducibility, transparency and rigor that can only be addressed through open sharing of materials, methods, data, results and software. They also point to the overwhelming demands of "big data" research to motivate greater sharing of data. Open science detractors express concerns about the various logistical demands of sharing software, data, results and knowledge (e.g., documentation, organization, curation, privacy protection, user support), as well as the potential loss of competitive advantage for their labs and trainees. They also note that sharing efforts commonly go uncited and unrewarded by institutions and funding agencies, and they continue to raise questions about the value of shared data. Data generators and analysts tend to fall on opposing sides of this debate based on their self interests. With the undeniable merits of each side of the debate, it is easy for scientists to sense an impasse reminiscent of geopolitical conflicts.

The debate is particularly vigorous in the brain imaging community for reasons that, at a high level, may represent polarizing factors across all of science. For example, one possible explanation is that the exorbitant cost of brain imaging research divides researchers into camps of haves and have-nots. Those with access to data and computational resources want to maximize their return on investment; those without want to add value and contribute their expertise, as well to advance their own analytic goals. Data acquisition costs are substantial, and data analysis costs continue to rise as spatial and temporal resolutions of data increase, in parallel with increasingly computationally intensive approaches. As such, the divide between haves and have-nots is only going to grow.

On more careful inspection, however, there are other factors that may be equally important in polarizing this debate. Like many fields, brain imaging is highly interdisciplinary, requiring expertise spanning engineering, physics, computer science, statistics, psychology, physiology, neuroscience, medicine, and so on. The reality is that the individual who designs and leads a brain imaging study (e.g., psychologist, psychiatrist) is probably not the most qualified to analyze the data, and vice versa. As highlighted in recent articles (e.g., [3]), the increasing scale and sophistication of the questions of modern neuroscience are creating the need for ecosystems in which data generators, tool makers, and data users all co-exist. Arguably, the rules for interaction, reward, and survival are yet to be worked out, thus helping to fuel the debate.

Yet another potential polarizing factor relates to the reproducibility crisis in science [4], which has certainly hit home in the brain imaging community. Challenged by the limited sample sizes available, researchers attempt to replicate one underpowered study with another, often using inappropriate statistical techniques to compensate for small sample sizes. Compounding the problem are the nonstandard data storage and analysis systems used in many labs, which are not well suited for the analysis of larger datasets; they increase the potential for errors as well as make sharing difficult.

What will it take to break the impasse? In order to begin addressing this question, we must first acknowledge that the current state of affairs does not adequately support science, let alone open science. As such, an effective resolution would need to support the advance of science in more efficient and effective ways, while also satisfying the different concerns and priorities of individual stakeholders in an open ecosystem. This challenge seems so great that many have turned to key organizing bodies in the community, such as publishers, academic institutions, funding agencies and professional associations, to provide guidance or to set or enforce standards.

However, turning to organizing bodies for change presupposes that they not only have everyone’s interests in mind, but the organizational ability and authority to either guide and influence, or enact and enforce, a resolution. This is not always the case. As noted in a recent editorial[5], when mandates such as post-publication availability of data are used as a means of effecting top-down change, some investigators are prone to disregard them and entities may be hesitant to enforce them, whether due to their own interests or fears of alienating the community. For many, incentives represent an appealing alternative to effecting top-down change. In this regard, funding agencies have created a number of funding mechanisms to facilitate the sharing of previously collected data, though there has been limited success in expanding open data. A few institutions such as the Allen Institute for Brain Science, the Child Mind Institute, INRIA, Janelia Research Campus, and the Montreal Neurological Institute have made open science a core component of their values. This is an undeniably important precedent, though it will take time for traditional academic institutions to grapple with the challenges of recognizing open science contributions in their tenure, promotion and degree-granting processes. Journals are helping through the creation of publication formats that explicitly recognize data generators for sharing (e.g., Data Descriptor and Resource formats), though their weight in academic evaluations has yet to be documented.

Whether individuals are influenced by mandates or incentives, their commitment to open science must benefit themselves to succeed. In the long term, this will likely require a profound overhaul of the accreditation and financial models that currently control career advancement, scientific publishing and the review process. To even engage this prospect, it is important to evaluate three primary arguments in defense of open science, which would need to be reconciled against social and economic factors that oppose it. First, science benefits from being open. Second, open standards support sharing and reuse. Third, scientific collaborations benefit from the clear specification of open sharing expectations at their inception. Science is certainly replete with specifications, standardization committees and shared software packages that have influenced or become an integral part of many scientists' everyday work. However, many continue to question the value of sharing data, which lies at the heart of the three polarizing factors in the debate over open science listed above. Without shared data, there will continue to be divisions between haves and have-nots, between labs with data generators and labs without, and between those who can conduct reproducible results with large sample sizes and those who cannot.

Following a decade of questions about the true value of shared data, a recent publication in *Nature Communications* by the International Neuroimaging Data-sharing Initiative (INDI) team has demonstrated the impact of shared data on the brain imaging field [6]. A particular emphasis was placed on the outputs of grassroots, open data-sharing consortia, where contributors provided their own independently collected data for sharing, knowing that they would get back more than they gave. In many cases, the data contributions were incentivized by the data needs of clinical scientists in child and adolescent psychiatry - a field characterized by a scarcity of data resources and researchers despite an overwhelming disease burden and an undeniable need for objective measures of illness to guide clinical decision-making. For these contributors, the cost of small sample sizes and research silos is the missed opportunity to change the status quo of clinical practice by increasing the speed of discoveries. As shown in the INDI analysis, openly shared data are being used by individuals from a range of disciplines, for peer-reviewed publications, teaching, method and tool development, theses and more.

Complementing the analysis of the impact on the scientific literature of consortia-based sharing of multiple, small datasets was a similar analysis of the impact of two larger projects that were explicitly designed for data sharing - the Nathan Kline Institute-Rockland Sample and the NIH Human Connectome Project. These analyses are particularly relevant, as a growing number of other large-scale open data resources are emerging (e.g., Child Mind Institute's Healthy Brain Network, NIH ABCD Study, Alzheimer's Disease Neuroimaging Initiative III, NIH Human Connectome Lifespan Studies, UK Biobank). There are also a growing number of individual investigator datasets shared through INDI and other data-sharing initiatives, such as OpenfMRI/OpenNeuro, that are allowing users to address a broader range of questions in clinical and cognitive neuroscience. These will likely expand as funding mechanisms require investigators to agree to sharing up front as part of the grant mechanism. The European Commission and 11 European research funding organizations just created such an initiative to ensure that by 2020 all articles they support financially are immediately open access upon publication, and for participants in the Horizon 2020 Open Research Data Pilot, the same will be true of research data.

The reality is that the vast majority of us participate in some form of open science every day - it is just a matter of whether we are a contributor, a user, or both, and how actively. While the open science ideal would be that we would all fully operate as contributors and users, it is important to acknowledge that participation in any category makes a difference. Contributors create opportunities for others to inspect, evaluate and build upon their contributions, yielding new outputs that may not have ever been imagined, and at times facilitating the identification and correction of errors. Users help to improve the reproducibility and standardization of science in the field simply by using common resources.

Looking forward, it makes more sense to help individuals find opportunities to increase their participation in open science than to make them feel a need to defend against the implementation of an ideal. In this regard, a continued focus on growing the breadth, sophistication and ease of use of open tools will help to increase their use; similarly increasing the range, scale, quality and ease of access of open datasets will increase their representation in research. The greater challenge is clearly that of increasing motivations for contributing to open science, and remembering that no single game changer will tip the balance for researchers overnight. Each of the key stakeholders (e.g., funding agencies, institutions, journals) must meaningfully increase incentives and rewards for open science; and when not sufficient, consider the implementation of enforceable mandates. Funding agencies can do their part by increasing the number of funding mechanisms dedicated to the development and expansion of infrastructural support for open science; the archive, analysis software and standards mechanisms from the BRAIN initiative are excellent examples, though more are needed. Equally important is the need for funding mechanisms explicitly focused on the creation of open data resources for immediate sharing; such mechanisms would value the potential significance, innovation and quality of proposed data, rather than the specific analyses the data contributor chooses to perform. These mechanisms can work to motivate sharing by prioritizing the funding of proposals by investigators with a history of compliance with sharing requirements, possibly through the introduction of sharing compliance as an explicit scoring criterion.

Academic institutions have an opportunity to push the balance in favor of sharing by revising the degree-granting, promotion and tenure processes to explicitly encourage and reward successful sharing. Journals can do the same by prioritizing the acceptance of manuscripts from investigators with a history of compliance with their sharing policies, and if necessary, creating temporary blocks on submissions for those who fail to comply.

Some will question the likelihood of such changes from one or more of the stakeholders given the inertia that has dominated to date. Although understandable, it is important to note the generational shift that is underway. Many younger researchers, whose work has most directly benefited from open science, are slowly becoming the newest generation of leaders and reviewers. These individuals appear more willing to embrace the principles of open science than their predecessors, and will have the opportunity to bring about reform. Although slow, such generational shifts have many precedents for rehabilitating flawed systems throughout history.

Finally, it is worth addressing the question of what we each can do, individually, as we wait for change. I would suggest that at a minimum we can: (1) look for and seize opportunities to increase our role in the open ecosystem, big or small, and (2) increase the emphasis on maximizing the quality and reproducibility of our outputs through the adoption of common data collection, storage, and analysis standards, even if we do not intend to share at the present time. To help make these calls to action more concrete, we include a starter list in Box 1 for actionable steps that individuals can take in pursuit of these goals. We will continue to update these steps with the help of the community at https://matter.childmind.org/open-science.html.

As noted in the quote that inspired the title for this work - "We but mirror the world…If we could change ourselves, the tendencies in the world would also change.” Consistent with Mahatma Gandhi’s wisdom, if we each faithfully do our part, the collective will affect change, from the bottom up.

Box 1. Ways that researchers can promote the practice of open science today.

- Publications and presentations
  - Publish in open access venues and follow FAIR (findable, accessible, interoperable, reusable) principles.
  - When reviewing manuscripts or proposals, acknowledge where attempts are made in support of open science, and point out where greater efforts could be made toward more open science practices. Insist that they follow FAIR principles.
  - Publish data or software in open methods journals.
  - Boycott publishers/publications for review or submission that flout open standards.
  - Acknowledge, and actively promote, any and all uses of open science in one's presentations/publications/proposals/lectures and make it clear where people can access these resources.
  - When attending another's talk or lecture, ask how one can access any software/data/resources that were presented and if there are any usage restrictions.
- Social media
  - When commenting on others' scientific work or practices, stick to the science and do not engage in *ad hominem* attacks.
  - Do not take others' comments personally; respond where appropriate as it pertains to the science and request where appropriate for guidance toward better ways to practice open science.
  - Language is ambiguous and vague -- tactfully ask clarifying questions to help guide a discussion toward a useful resolution.
  - Publicly acknowledge contributions and thank contributors to open science projects whenever possible.
- Within one's home institution
  - Catalyze open science practices and projects through seminars, workshops, hackathons, contests (e.g., [7]), proposals, etc.
  - Join groups within one's institution to enact changes that promote evaluation/promotion criteria in support of open science practices.
  - Apply liberal licenses to software (e.g., Apache v2.0) and documentation (e.g., CC0) at the outset of a project.
  - When tasked with an assignment, big or small, opt for open methods where possible (for example, complete a homework assignment using Python, R, or Octave in a shareable Jupyter or R Notebook vs. using a proprietary, licensed product like Matlab).

  - Strive toward reproducibility (even for oneself in the future!) by providing self-contained software environments, example input/output data, and clear and updated documentation.
- Collaborations
  - Forge ties across labs even within an institution to make use of each other's data/software.
  - Collaborate with institutions that require open standards.
  - Clarify contributor roles at the outset of a publication or project to assign appropriate credit/accountability.
  - Make it very clear at the outset of a collaboration how open/shared software/data will be acknowledged/rewarded.
  - Publish a code of conduct for one's project to clarify roles and mechanisms for resolving disputes.
  - Clarify when data/software can be released at the outset of a project.
  - Use collaborative software engineering practices, with public discussions and issues (e.g., GitHub, GitLab, Apache Subversion).
  - Avail oneself of experts in alternative/complementary methods to reduce instrumentation bias (see [8], evaluate methods, and corroborate results.
  - Participate in interdisciplinary, open science and collaboration events that go beyond institutional boundaries (e.g., Brainhack; http://www.brainhack.org)

Acknowledgements. We would like to thank Satrajit Ghosh, Xavier Castellanos, Stan Colcombe, Alex Franco, Seok-Jun Hong, Aki Nikolaidis, and Joshua Vogelstein for their helpful comments and revisions in the preparation of this work, as well as Judith Milham, who acted as a consistent sounding board.

doi:10.20944/preprints201809.0543.v1

REFERENCES

1. Longo, D. L. & Drazen, J. M. Data Sharing. *N. Engl. J. Med.* **374,** 276–277 (2016).
2. FGED: Data Sharing and Research Parasites. Available at: http://fged.org/projects/data-sharing-and-research-parasites/
3. Vogelstein, J. T. *et al.* To the Cloud! A Grassroots Proposal to Accelerate Brain Science Discovery. *Neuron* **92,** 622–627 (2016).
4. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature News* **533,** 452 (2016).
5. Barron, D. How Freely Should Scientists Share Their Data? *Scientific American Blog Network* Available at: https://blogs.scientificamerican.com/observations/how-freely-should-scientists-share-their-data/
6. Milham, M. P. *et al.* Assessment of the impact of shared brain imaging data on the scientific literature. *Nat. Commun.* **9,** (2018).
7. Allen, G. I. *et al.* Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimers. Dement.* **12,** 645–653 (2016).
8. Tustison, N. J. *et al.* Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. *Front. Neurosci.* **7,** (2013).