Resource Sharing to Improve Research Quality

Ghassan B. Hamra, PhD; Neal D. Goldstein, PhD; Sam Harper, PhD

Transparency and openness are vital for strengthening the scientific process. However, there is no clear agreement in the scientific community about the elements necessary to qualify scientific research as a transparent and open process. Historically, the description of study methods and results within individual academic publications has been treated as sufficient for establishing transparency; that is, based solely on the written description of study procedures and analytic techniques, a third party can be *assumed* to have all the information needed to reproduce the results of an individual study if the data were available. The core philosophy of *reproducible* research is slightly different and challenges this assumption. Rather than relying on the written report, reproducible research culture demands access to data and analytic code used to produce study results. In this scenario, anyone should be able to exactly reproduce the tables, figures, and evidence presented in a given article. The push for reproducible research and current publication practices do not conflict in their goal of disseminating high-quality research, but differ in the means of obtaining it.

An idea related to reproducible research, and one that is considerably more complex, is *replication*, or the process by which others might attempt to use entirely different data to corroborate, or refute, the findings of prior published research. Replication is related to reproducibility in the sense that independent investigators attempt to see whether reported findings are consistent. We treat replication as a self-correcting step in science: if prior study findings are expected to be replicable and cannot be replicated by independent parties, then at that point we might call into question those findings. While self-correction is natural in science, it is not the norm,[1] and reports have suggested that the extent to which study findings cannot be replicated is alarming, leading to the so-called replication crisis.[2] Many related reasons have been put forward to explain the replication crisis, including misaligned incentives in academia, the file drawer effect,[3] p-hacking,[4] overreliance on null hypothesis significance testing,[5] and even outright falsification of data. Some have suggested that our existing assumptions about what qualifies as transparent and open in science may be insufficient and that addressing this can safeguard against further replication crises.

In this commentary, we discuss the importance of transparency and openness, focusing on the 2 major elements necessary for reproducibility: the data and analytic code used to produce the results in a published research report. We highlight how greater openness can support more reliable findings (in the long run) by allowing checks for robustness of study findings. Next, we discuss the Transparency and Openness Promotion (TOP) guidelines. The *Journal of the American Heart Association* (*JAHA*), like many, signed on to the TOP guidelines, which were designed to encourage the sharing of research resources, an essential element for reproducibility in science. Finally, we highlight resources that are available to make data and analytic code sharing accessible to all researchers, addressing privacy concerns as well. By supporting reproducible research via transparency and openness, we can accelerate the identification and translation of scientific evidence to population and clinical interventions.

From the Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD (G.B.H.); Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Drexel University, Philadelphia, PA (N.D.G.); Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada (S.H.).

Correspondence to: Ghassan B. Hamra, PhD, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St, Baltimore, MD 21205. E-mail: ghassanhamra@jhu.edu

J Am Heart Assoc. 2019;8:e012292. DOI: 10.1161/JAHA.119.012292.

Assessing Reliability Through Resource Sharing

Reliability of study findings is often assumed by consumers and producers of academic research; that is, there is a perception that if a study has surmounted the hurdles of peer review, its findings must be of a sound scientific foundation. A number of large-scale replication efforts have challenged the assumption of reliability of peer-reviewed study findings. A project by the Open Science Collaboration attempted to replicate the findings of 100 studies from 3 of the top journals in psychology by reproducing the findings of the original studies.[6] The authors found that, on average, effect estimates were halved in magnitude, and only 47% of results with $P<0.05$ were repeated compared with original reported results.[7] Similarly, weaker than reported effect sizes were found in a large-scale replication project in economics.[8] The challenge of replicating results is further supported by an evaluation of published reanalyses of randomized clinical trials data[9]; here, the authors noted that 35% of reanalyses led to different interpretations of study findings compared with those of the original publications. An ongoing replication project of cancer biology research has shown mixed results, with only 5 of 14 important results being replicated, to date.[10] No such effort has been conducted specifically for cardiovascular research, but the message is clear: deeper scrutiny of study findings often leads to changes in the original study's conclusions.

There are many arguments put forward for why study results are difficult to replicate. When authors use $P<0.05$ as a criterion for reporting findings, 1 obvious argument is that authors might analyze data with the goal of meeting an arbitrary statistical cut point rather than conducting the most rigorous analysis (sometimes called p-hacking).[11] This can lead to decisions about statistical tools used and coding of variables that may not be appropriate. Related is the file drawer effect (ie, publication bias),[3] where analyses that do not meet an arbitrary statistical cut point are deemed unworthy of publication (or even submission),[12] regardless of study quality. In the worst case, researchers simply falsify data to support their hypotheses; well-known examples from cardiovascular research include the body of work by Dr John Darsee,[13] and the laboratory of Dr Piero Anversa that falsified data to suggest that heart muscle could be repaired with stem cells.[14] We argue that greater transparency and openness, discussed below, are necessary to address replication problems; because when study resources, namely, data and statistical code, are made available, author decisions can be more easily scrutinized and subjected to alternative assumptions and sensitivity analyses to assess robustness of findings. In the case of fraud, resource sharing provides an opportunity to scrutinize data to potentially identify improprieties though it provides no guarantee that false data will be identified.

The TOP Guidelines

In 2015, the TOP guidelines were established by the Center for Open Science as a means to promote and support more rigorous and defensible science by ensuring reproducibility of scientific findings and, hopefully, replication in subsequent studies.[6] Indeed, the National Science Foundation has described reproducibility as "... a minimum necessary condition for a finding to be believable and informative."[15] Put another way, study findings should be *reliable.* The failure of large-scale efforts to replicate study findings in different data across disciplines suggests this condition is not being met. Given its importance for informing future replication work, steps to enhance reproducibility seem necessary.

The TOP guidelines outline 3 levels of increasing stringency regarding sharing of research materials from study authors. *JAHA* is 1 of over 5000 signatories to the TOP guidelines, and has selected, as most journals have, the level 1 standard. This standard encourages, but does not require, the sharing of study materials by study authors with the greater scientific community. Rather, it requires that authors state the extent to which they are willing or able to share study resources for the purposes of reproducibility. Examples of language reflective of this abound in recent issues of *JAHA* and are presented below. Level 2 requires that authors deposit data and analytic code to a trusted repository and state explicitly whether resources will be provided elsewhere, and why. Level 3, the most stringent, requires that data and analytic code be posted to a trusted repository and that reported analyses be independently reproduced before publication.

Levels 2 and 3 guarantee that data and analytic code are available, and ensure that research results can be, or were, independently verified (ie, reproduced); they also allow others to check the impact of changes to model selection or variable coding decisions of the published work (ie, if study results are sensitive to alternative assumptions). But what of the level 1 guidelines? While they provide a structure to encourage the sharing of data and analytic code and require authors to address the issue of transparency, they provide no guarantee that study resources are available or easily accessible. To illustrate, we present a brief summary of data sharing statements by authors in the *JAHA*. We reviewed the text from Epidemiology articles published in the *Original Research* section of the journal from the beginning of 2019 through the end of April 2019.

Of 39 articles reviewed, the majority (n=30, 77%) expressed willingness to share study data upon request or noted that data are publicly available. However, only 5 of these explicitly stated that analytic code or methods were available for purposes of reproducibility, upon request. The remainder either made no statement about the availability of analytic code or methods (n=20), stated that materials could be requested but did not indicate whether all requests would be honored (n=2), or stated that the analytic methods were available within the article (n=3); of this latter group, we note that the authors likely intended this to indicate that the Methods section as described constitutes the analytic methods rather than the analytic code used to create the study results. Finally, 9 of the 39 articles (23%) explicitly stated that study resources would not be made available for purposes of reproducing study results. In many cases, investigators articulated concerns of study participant privacy when stating that data must be requested for purposes of reproducibility. In the case of publicly available data, statements were fairly obvious: that any investigator could access the data via its original source.

While most investigators stated a willingness to provide study data, only a single study expressed a willingness to share analytic code. Many authors stated that all the necessary information to reproduce study results, given the study data, were available within the article via the Methods section, which speaks to the prevailing norms we discussed above. However, the availability of analytic code is as important for reproducibility as the data used. This is because of the inherent limitations of published methods descriptions.

First, Methods sections are rarely an exhaustive description of the many steps that lead from raw data to analytic data sets. Raw data are only the starting point, and nearly always go through cleaning and processing steps that lead to the analytic data set used for a publication. In some situations, multiple raw data sources are combined to come up with a final analytic data set. This is especially notable for large-scale studies that utilize complex designs such as the National Health and Nutrition Examination Survey. This study spans many years and includes many measures of individual health including dietary recall, physical activity measures, and biomarkers. Linking across data sets with unique identifiers often requires as many or more steps as taken for a study's statistical analysis.

Second, the steps that take an investigator from analytic data set to study results are often not described in sufficient detail. Even after cleaning the data, analysts go through many steps including selection of variables and their structures and decisions about which model results to report. Not reporting this information makes it nearly impossible to easily reproduce study results, making the availability of analytic code necessary for transparency.

The most obvious ways to support transparency with study resources are (1) to make analytic data available, when it is reasonable to do so, with documentation of steps linking to it from raw data and (2) to provide software code that documents the steps that link analytic data to final study results, also documenting the steps taken in selecting and structuring variables. The clear benefits of this practice have been documented by others and focus on assessing the robustness of reported study results to alterations of the analysis. This serves to address potentially unfounded criticisms of a study and to identify problems with the reported results.

Many authors choose not to provide data because of privacy concerns. These concerns can relate to both institutional or investigator investment in the work, such as patents on products of research, or study participant privacy. Regarding the former, there is not much known about if and how resource sharing could be problematic; we recommend that investigators discuss concerns with their office of general counsel. In the case of study participant privacy, a helpful tutorial is available.[16] Briefly, there are likely few cases where institutional review board approvals could not be amended to reasonably allow for data sharing. Even in these cases, new procedures for creating synthetic data based on actual study data and statistical analyses show promise as a means of allowing for reproduction of study results without risking study participant anonymity.[17]

Resources to Support Transparency and Openness

Many efforts are under way to support and strengthen sharing study resources. There are repositories for both analytic code and data, and steps to ensure security of data that are made available for the purposes of reproducibility. We highlight 3 well-known examples, though there are many others: Dataverse, GitHub and the Open Science Framework. We note that all of these resources are ultimately under the control of the investigator, as opposed to a journal or institution, responsible for sharing data and analytic code.

Dataverse (https://github.com/) is an open source web application and repository for researchers to deposit their data and analytic code to provide easy access to interested parties. Researchers are guaranteed control of their data so that access is only provided to those with appropriate permissions. Equally, authors can specify allowable uses of the data, so others cannot simply download and publish findings as they see fit. To support researchers in their effort, acknowledgment via data citations are provided. The software is downloadable and data are kept at repositories at Harvard and, for some, at their research institute. Journals are also able to establish and maintain a Dataverse for interested authors.

GitHub (https://github.com/) is a resource targeted more towards software development. Many users of popular statistical software packages, such as R statistical software, utilize this site to provide source code for programs they are developing or wish to provide to the research community. It also serves as a repository for analytic code relevant for individual publications.[18]

Open Science Framework (OSF, https://osf.io/), developed and maintained by the Center for Open Science, is an open source platform for documenting and preserving all aspects of the research lifecycle. It supports collaborative projects, research logs and wikis, preregistration of study hypotheses, archiving of data and code, as well as providing a platform for preprints of research reports. To enable sharing and promote citation, each project, component, and file is given a unique, persistent uniform resource locator (URL). The platform also allows many third-party add-ons for connecting across a variety of platforms (eg, both Dropbox and GitHub can be linked to OSF). OSF is committed to open science, but also allows parts or even entire projects to be kept private.

Although all of these platforms are generally straightforward to use, for researchers not accustomed to making their materials available to others, it may seem daunting and burdensome. However, a number of introductory guides exist for how to incorporate open science practices into a lab or research group,[19–22] as well as "success" stories of researchers who feel their work has benefited from greater transparency.[23]

An important complement to these resources is support for resource sharing from both journals and research institutions. Many journals already provide a platform for making analytic code available via electronic supplements; a reasonable next step would be to provide authors with cloud space for making data available, when possible. It would be reasonable for journals to unite with resource-sharing groups to provide an integrated sharing platform for users. By formally offering resources to authors and researchers, journals can more easily support sharing practices that the TOP guidelines were designed to encourage.

Research institutions (university or otherwise) should integrate support for resource sharing. We previously noted that the incentive structure at many institutes is not supportive of resource sharing; indeed, it has no bearing on promotion and tenure at most academic institutes, and may be implicitly discouraged because of emphases on innovation and novelty. Recently, a meeting of stakeholders organized by the Meta-Research Innovation Center at Stanford led to promotion of 6 principles for evaluating scientists: 2 of these principles focus explicitly on openness and transparency practices, noting their connection to the TOP guidelines.[24]

Summary

Transparency and openness are a vital part of science; the minimum necessary components for achieving this are access to data and analytic code to allow others to assess the reliability of study findings via reproduction. There are resources available to support authors in this task that ensure protection of the data and to support acknowledgment of engagement with open science. Journals and research institutes can and should do more to align incentive structures with resource sharing to improve science.

We have noted the importance of reproducibility to aid in avoiding future replication challenges. Others have noted the efficiency of taking advantage of resources for data and analytic code sharing. Specifically, providing data and analytic code to reproduce results at the time of publication ensures that investigators do not have to repeat steps they have already completed leading up to a publication; this includes preparing data for sharing with interested parties. This additionally opens up the potential for the peer-review process to vet analytic code in addition to the manuscript.[25] Data preparation is costly in terms of time and money, so providing the data and analytic code at a stage when they are already prepared can help avoid unnecessary use of resources.[26]

Resource sharing is a complement to other means of improving biomedical science, the most notable of which is preregistration of clinical trials. Indeed, a study of trials funded by the National Heart Lung and Blood Institute showed that the number of reported null findings increased after introduction of preregistration via clinicaltrials.org, which directly addresses the file drawer and p-hacking problems in publication.[27] A recent comment suggested that preregistration and eventual sharing of resources (including study data) were vital components of clinical trials to improve transparency and public trust.[28]

In closing, we believe that adopting a culture of transparency and openness will help advance science and, by reducing the number of nonreplicable findings, accelerate the identification of knowledge to inform clinical and population-level interventions to improve health.

Disclosures

None.

References

1. Ioannidis JP. Why science is not necessarily self-correcting. *Perspect Psychol Sci*. 2012;7:645–654.
2. Lash TL, Collin LJ, Van Dyke ME. The replication crisis in epidemiology: snowball, snow job, or winter solstice? *Curr Epidemiol Rep*. 2018;5:175–183.
3. Weinberg CR. Invited commentary: can issues with reproducibility in science be blamed on hypothesis testing? *Am J Epidemiol*. 2017;186:636–638.
4. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. *PLoS Biol*. 2015;13:e1002106.
5. Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. *Am J Epidemiol*. 2017;186:627–635.
6. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon TA, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T. Scientific standards. Promoting an open research culture. *Science*. 2015;348:1422–1425.
7. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349:aac4716.
8. Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, Kirchler M, Almenberg J, Altmejd A, Chan T, Heikensten E, Holzmeister F, Imai T, Isaksson S, Nave G, Pfeiffer T, Razen M, Wu H. Evaluating replicability of laboratory experiments in economics. *Science*. 2016;351:1433–1436.
9. Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ, Ioannidis JP. Reanalyses of randomized clinical trial data. *JAMA*. 2014;312:1024–1032.

10. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. *Elife*. 2014;3:e04333.
11. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*. 2011;22:1359–1366.
12. Franco A, Malhotra N, Simonovits G. Social science. Publication bias in the social sciences: unlocking the file drawer. *Science*. 2014;345:1502–1505.
13. Kochan CA, Budd JM. The persistence of fraud in the literature: the Darsee case. *J Am Soc Inf Sci*. 1992;43:488–493.
14. Maliken BD, Molkentin JD. Undeniable evidence that the adult mammalian heart lacks an endogenous regenerative stem cell. *Circulation*. 2018;138:806–808.
15. Bollen K, Cacioppo JT, Kaplan RM, Krosnick JA, Olds JL. Social, behavioral, and economic sciences perspectives on robust and reliable science: report of the subcommittee on replicability in science, advisory committee to the national science foundation directorate for social, behavioral, and economic sciences. 2015. Available at: https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.
16. Meyer MN. Practical tips for ethical data sharing. *Adv Methods Pract Psychol Sci*. 2018;1:131–144.
17. Shepherd BE, Blevins Peratikos M, Rebeiro PF, Duda SN, McGowan CC. A pragmatic approach for reproducible research with sensitive data. *Am J Epidemiol*. 2017;186:387–392.
18. Goldstein ND. Toward open-source epidemiology. *Epidemiology*. 2018;29:161–164.
19. Blischak JD, Davenport ER, Wilson G. A quick introduction to version control with Git and GitHub. *PLoS Comput Biol*. 2016;12:e1004668.
20. Gorgolewski KJ, Poldrack RA. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS Biol*. 2016;14:e1002506.
21. King G. An introduction to the dataverse network as an infrastructure for data sharing. *Sociol Methods Res*. 2007;36:173–199.
22. Ram K. Git can facilitate greater reproducibility and increased transparency in science. *Source Code Biol Med*. 2013;8:7.
23. Lowndes JSS, Best BD, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CC, Jiang N, Halpern BS. Our path to better science in less time using open data science tools. *Nat Ecol Evol*. 2017;1:160.
24. Moher D, Naudet F, Cristea IA, Miedema F, Ioannidis JPA, Goodman SN. Assessing scientists for hiring, promotion, and tenure. *PLoS Biol*. 2018;16:e2004089.
25. Ince DC, Hatton L, Graham-Cumming J. The case for open computer programs. *Nature*. 2012;482:485–488.
26. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JPA. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *BMJ*. 2018;360:k400.
27. Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One*. 2015;10:e0132382.
28. Hudson KL, Lauer MS, Collins FS. Toward a new era of trust and transparency in clinical trials. *JAMA*. 2016;316:1353–1354.

Key Words: replication • reproducibility • research ethics