Published in final edited form as:
Science. 2011 December 2; 334(6060): 1226–1227. doi:10.1126/science.1213847.

Reproducible Research in Computational Science

Roger D. Peng
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore MD 21205, USA

Abstract

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

The rise of computational science has led to exciting and fast-moving developments in many scientific areas. New technologies, increased computing power, and methodological advances have dramatically improved our ability to collect complex hig-dimensional data (1, 2). Large data sets have led to scientists doing more computation, as well as researchers in computationally oriented fields directly engaging in more science. The availability of large public databases has allowed for researchers to make meaningful scientific contributions without using the traditional tools of a given field. As an example of this overall trend, the Sloan Digital Sky Survey, a large publicly available astronomical survey of the Northern Hemisphere, was ranked the most cited observatory (3), allowing astronomers without telescopes to make discoveries using data collected by others. Similar developments can be found in fields such as biology and epidemiology.

Replication is the ultimate standard by which scientific claims are judged. With replication, independent investigators address a scientific hypothesis and build up evidence for or against it. The scientific community's "culture of replication" has served to quickly weed out spurious claims and enforce on the community a disciplined approach to scientific discovery. However, with computational science and the corresponding collection of large and complex data sets the notion of replication can be murkier. It would require tremendous resources to independently replicate the Sloan Digital Sky Survey. Many studies—for example, in climate—science require computing power that may not be available to all researchers. Even if computing and data size are not limiting factors, replication can be difficult for other reasons. In environmental epidemiology, large cohort studies designed to examine subtle health effects of environmental pollutants can be very expensive and require long follow-up times. Such studies are difficult to replicate because of time and expense, especially in the time frame of policy decisions that need to be made regarding regulation (2).

Researchers across a range of computational science disciplines have been calling for reproducibility, or reproducible research, as an attainable minimum standard for assessing the value of scientific claims, particularly when full independent replication of a study is not feasible (4–8). The standard of reproducibility calls for the data and the computer code used to analyze the data be made available to others. This standard falls short of full replication because the same data are analyzed again, rather than analyzing independently collected

To whom correspondence should be addressed. rpeng@jhsph.edu.

data. However, under this standard, limited exploration of the data and the analysis code is possible and may be sufficient to verify the quality of the scientific claims. One aim of the reproducibility standard is to fill the gap in the scientific evidence-generating process between full replication of a study and no replication. Between these two extreme end points, there is a spectrum of possibilities, and a study may be more or less reproducible than another depending on what data and code are made available (Fig. 1). A recent review of microarray gene expression analyses found that studies were either not reproducible, partially reproducible with some discrepancies, or reproducible. This range was largely explained by the availability of data and metadata (9).

The reproducibility standard is based on the fact that every computational experiment has, in theory, a detailed log of every action taken by the computer. Making these computer codes available to others provides a level of detail regarding the analysis that is greater than the analagous non-computational experimental descriptions printed in journals using a natural language.

A critical barrier to reproducibility in many cases is that the computer code is no longer available. Interactive software systems often used for exploratory data analysis typically do not keep track of users' actions in any concrete form. Even if researchers use software that is run by written code, often multiple packages are used, and the code that combines the different results together is not saved (10). Addressing this problem will require either changing the behavior of the software systems themselves or getting researchers to use other software systems that are more amenable to reproducibility. Neither is likely to happen quickly; old habits die hard, and many will be unwilling to discard the hours spent learning existing systems. Non–open source software can only be changed by their owners, who may not perceive reproducibility as a high priority.

In order to advance reproducibility in computational science, contributions will need to come from multiple directions. Journals can play a role here as part of a joint effort by the scientific community. The journal *Biostatistics*, for which I am an associate editor, has implemented a policy for encouraging authors of accepted papers to make their work reproducible by others (11). Authors can submit their code or data to the journal for posting as supporting online material and can additionally request a "reproducibility review," in which the associate editor for reproducibility runs the submitted code on the data and verifies that the code produces the results published in the article. Articles with data or code receive a "D" or "C" kite-mark, respectively, printed on the first page of the article itself. Articles that have passed the reproducibility review receive an "R." The policy was implemented in July 2009, and as of July 2011, 21 of 125 articles have been published with a kite-mark, including five articles with an "R." The articles have reflected a range of topics from biostatistical methods, epidemiology, and genomics. In this admittedly small sample, we have not yet encountered cases in which data and code were submitted for reproducibility review but results were not reproducible as claimed. It is encouraging that authors are taking advantage of the policy to make their work reproducible by others, but more work could be done to promote a broader adoption of the policy.

The fact that an analysis is reproducible does not guarantee the quality, correctness, or validity of the published results. The "R" kite-mark is meant to convey the idea that a knowledgeable individual has reviewed the code and data and was capable of producing the results claimed by the author. In cases in which questionable results are obtained, reproducibility is critical to tracking down the "bugs" of computational science. In cases with interesting findings, reproducibility can greatly facilitate building on those findings (12).

Perhaps the biggest barrier to reproducible research is the lack of a deeply ingrained culture that simply requires reproducibility for all scientific claims. Not unlike the culture of replication that persists across all scientific disciplines, the scientific community needs to develop a "culture of reproducibility" for computational science and require it of published claims. Another important barrier is the lack of an integrated infrastructure for distributing reproducible research to others. The current system is ad hoc with researchers in some fields having access to sophisticated central data repositories and researchers in other fields having few useful resources for sharing code and data. In many cases, a researcher does not have an obvious place to turn to make sure their work is reproducible and accessible by others. Journals' supporting online materials have some severe limitations, such as the inability to search and index available data.

Given the barriers to reproducible research, it is tempting to wait for a comprehensive solution to arrive. However, even incremental steps would be a vast improvement over the current situation. To this end, I propose the following steps (in order of increasing impact and cost) that individuals and the scientific community can take. First, anyone doing any computing in their research should publish their code. It does not have to be clean or beautiful (13), it just needs to be available. Even without the corresponding data, code can be very informative and can be used to check for problems as well as quickly translate ideas. Journal editors and reviewers should demand this so that it becomes routine. Publishing code is something we can do now for almost no additional cost. Free code repositories already exist [for example, GitHub (http://github.com) and SourceForge (http://sourceforge.net)], and at a minimum, code can be published in supporting online material. The next step would be to publish a cleaned-up version of the code along with the data sets in a durable non-proprietary format. This will involve some additional cost because not everyone will have the resources to publish data. Some fields such as genomics have already created data repositories, but there is not yet a general solution.

Last, the scientific community can pool its collective resources to create a DataMed Central and CodeMed Central, analogous to PubMed Central for all data, metadata, and code to be stored and linked with each other and with corresponding publications. Such an effort would probably need government coordination and support, but each would serve as a single gateway that would guide researchers to field-specific data and code repositories. Existing repositories could continue to be used and would interface with the gateway, whereas fields without existing infrastructure would be given access to these resources. The ultimate goal would be to provide a single place to which people in all fields could turn to make their work reproducible.

The field of science will not change overnight, but simply bringing the notion of reproducibility to the forefront and making it routine will make a difference. Ultimately, developing a culture of reproducibility in which it currently does not exist will require time and sustained effort from the scientific community.

References

1. Hanson B, Sugden A, Alberts B. Science. 2011; 331:649. [PubMed: 21310971]
2. Peng RD, Dominici F, Zeger SL. Am J Epidemiol. 2006; 163:783. [PubMed: 16510544]
3. Madrid JP, Macchetto D. Bull AAS. 2009 arXiv:0901.4552.
4. Schwab M, Karrenbach N, Claerbout J. Comput Sci Eng. 2000; 2:61.
5. Laine C, Goodman SN, Griswold ME, Sox HC. Ann Intern Med. 2007; 146:450. [PubMed: 17339612]
6. King G. PS: Polit Sci Polit. 1995; 28:444.
7. Gentleman R. Stat Appl Genet Mol Biol. 2005; 4:article 2.

8. Yale Law School Roundtable on Data and Code Sharing. Comput Sci Eng. 2010; 12:8.
9. Ioannidis JPA, et al. Nat Genet. 2009; 41:149. [PubMed: 19174838]
10. Mesirov JP. Science. 2010; 327:415. [PubMed: 20093459]
11. Peng RD. Biostatistics. 2009; 10:405. [PubMed: 19535325]
12. McCullough BD, McGeary KA, Harrison TD. Can J Econ. 2008; 41:1406.
13. Barnes N. Nature. 2010; 467:753. [PubMed: 20944687]

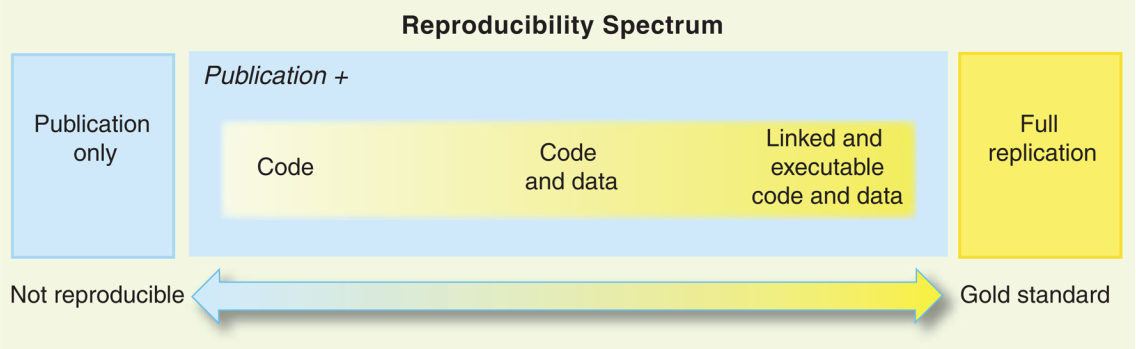

Fig. 1.
The spectrum of reproducibility.