Reproducible Research in Computational Science

Roger D. Peng

Computational science has led to exciting new developments, but, contrary to the nature of the work, the potential for a study 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, high-dimensional data (1, 2). Large data sets lead 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, when computational science and the corresponding collection of large and complex data sets the notion of replication can be murky. 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 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 analogous noncomputational 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 its 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

---

**Fig. 1.** The spectrum of reproducibility.
Methodological Challenges in the Study of Primate Cognition

Michael Tomasello* and Josep Call*

Laboratory studies of primate cognition face the problem that captive populations of a species are not always comparable, and generalizations to natural populations are never certain. Studies of primate cognition in the field face the problem that replications are expensive and difficult, and again different populations are not always comparable. To help remedy these problems, we recommend the creation of data banks where primary data and videotapes may be deposited (perhaps as a requirement of publication) to facilitate cross-examination, replication, and, eventually, the pooling of data across investigators.

The basic methodological premise in the study of primate, including human, cognition is that the same overt behavior may be underlain by different mechanisms. Cognitive mechanisms are prototypically those in which organisms cognitively represent nonperceived situations and draw inferences that go beyond immediate perception, which enables organisms to make flexible behavioral decisions even in novel environmental situations. Research in primate cognition is aimed at discovering these cognitive mechanisms by observing their effects in adaptive action.

Much of the research in primate cognition is conducted with captive populations in the laboratory, where investigators can best control conditions experimentally. Field researchers sometimes argue that, in this setting, subjects are faced with cognitive problems that are not well matched with the cognitive abilities that have evolved in their species-typical environments. [For more on the debate between laboratory and field researchers, see the paper by Ryan in this issue (1).] Giving captive populations novel tasks, however, sometimes reveals cognitive skills not observed in nature. For example, gorillas in captivity are quite proficient at using tools in a variety of extractive foraging tasks, although they have not been observed to use them in the wild (2). Because all of the other great apes also use tools proficiently in the laboratory, this suggests that the common ancestor to all apes used tools (3)—an hypothesis that could never be formulated if gorillas were not tested for tool use in captivity. The main methodological challenge of experimental research in primate cognition, therefore, is to design problems that are novel for individuals—so
Reproducible Research in Computational Science
Roger D. Peng

Science 334 (6060), 1226-1227.
DOI: 10.1126/science.1213847