Biostatistics (2010), **11**, 3, *pp.* 385–388
doi:10.1093/biostatistics/kxq028

An invitation to reproducible computational research

DAVID L. DONOHO

Department of Statistics, Stanford University, Stanford, CA 94305, USA
donoho@stanford.edu

1. Introduction

I am genuinely thrilled to see *Biostatistics* make a formal venture into computational reproducibility, and I congratulate the editors of *Biostatistics* on taking this much needed step. I find the policies being adopted by *Biostatistics* eminently practical, and I hope that many authors will begin using this option. In my comments, I will try to explain how I came to believe in the importance of reproducibility and why I think others may find it in their interest and in the community interest. I will then briefly mention some efforts in other disciplines.

2. My own reasons for working reproducibly

Computation-based science publication is currently a doubtful enterprise because there is not enough support for identifying and rooting out sources of error in computational work.

In my own experience, error is ubiquitous in scientific computing, and one needs to work very diligently and energetically to eliminate it. One needs a very clear idea of what has been done in order to know where to look for likely sources of error. I often cannot really be sure what a student or colleague has done from his/her own presentation, and in fact often his/her description does not agree with my own understanding of what has been done, once I look carefully at the scripts. Actually, I find that researchers quite generally forget what they have done and misrepresent their computations.

Computing results are now being presented in a very loose, "breezy" way—in journal articles, in conferences, and in books. All too often one simply takes computations at face value. This is spectacularly against the evidence of my own experience. I would much rather that at talks and in referee reports, the possibility of such error were seriously examined.

I was inspired more than 15 years ago by John Claerbout, an earth scientist at Stanford, to begin practicing reproducible computational science. See Claerbout and Karrenbach (1992). He pointed out to me, in a way paraphrased in Buckheit and Donoho (1995): "an article about computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." This struck me as getting to the heart of a central problem in modern scientific publication. Most of the work in a modern research project is hidden in computational scripts that go to produce the reported results. If these scripts are not brought out into the open, no one really knows what was done in a certain project; for example, tables and figures can be subtly miscomputed or mislabeled. I saw the good sense in making a commitment to transparency as the best way both to avoid and correct errors. Over the years, I have persuaded numerous graduate students and postdocs to work in a way that made clear precisely how each figure in a paper was being computed (Donoho *and others*, 2009). Over time, this effort has paid multiple dividends; I believe anyone who understands the process and the benefits will eventually be moved to practice it.

3. WHY SHOULD YOU WORK REPRODUCIBLY?

Computational reproducibility is not an afterthought—it is something that must be designed into a project from the beginning. One does need to develop a whole set of programming and research disciplines with the end result in mind and stick with them. A simple example is visible in the framework provided by *Biostatistics*: one has to plan to create a single script, in the R language, that generates all the figures and tables used in one's paper. Many individuals might otherwise work "interactively," and without planning for the end result. They might tend naturally to improvise commands at the keyboard to produce an analysis in bits and pieces, which they then transcribe and piece together into evidence for a paper. To be branded reproducible by *Biostatistics*, such an informal approach will never be accepted.

Does reproducibility sound like "extra work"? It can be, particularly when one is first trying to do it, that is, to break one's own previous nonreproducible habits. But here are some reasons we as "scientists" will want to work in a computationally reproducible way.

(a) *Improved work and work habits.* If we commit from the outset to make a given project computationally reproducible, we work differently, in a way that is much more transparent to ourselves and to others. We know where the project is headed—a published set of R scripts—and will have better focus while we are developing those scripts. Knowing that these scripts will be available to others, we will be motivated to polish and improve our scripts to raise them to a higher level of quality than we would have done if we thought "no one is looking."

(b) *Improved teamwork.* If we are part of a team that has committed to work this way, we will be able to communicate much more efficiently with other team members about our developing results; they can see what we are doing and propose improvements. Many questions team members may have about what we actually did in a computation will be settled much more efficiently and suggestions about what we might do differently will be much better focused. We will have much better confidence in the results we have produced with our colleagues.

(c) *Greater impact.* If we publish reproducibly, other researchers can easily use the methods we have developed. This has 2 benefits to us:

  (i) *Less inadvertent competition.* Others will be less likely to go to the expense and effort of developing competing methods for the same tasks, once our own solutions are transparently available. So our own work is less likely to be eclipsed rapidly.

  (ii) *More acknowledgement.* I recommend that all publishers of reproducible research adhere to the reproducibility licenses developed by Stodden (2009a, 2009b). Under that license, computationally reproducible work will be cited when it is used.

(d) *Greater continuity and cumulative impact.* If we publish reproducibly, we can more easily train students and postdocs when they join our teams, and we can much more easily benefit from their efforts after those students and postdocs leave our teams. In fact, a reasonable "get to know you" project for a new student assigns that student take a team's previously published project and adapt it to a new setting. If that project was developed under the reproducibility standard, it will be dramatically easier for the new student to understand and generalize the code.

The last point can also be taken as a reason we as "educators" should work reproducibly: it makes our Ph.D. and postdoc advising more effective,

There are also reasons we as "taxpayers" would want our sponsored research expenditures to result in computational reproducibility.

(a) *Stewardship of public goods.* An ever larger part of the work effort being funded by public research expenditures is computational effort. Good stewardship of the massive volume of work product we are now "buying" with public funds requires that the work product is not "lost" as soon as the

project is over. However, this is exactly what happens in many cases today. Funded investigators often cannot reproduce their own work soon after the project is over. Under the standard of reproducibility, we plan so that this would not happen, and by and large it does not happen. I still hear frequently from users who are able to reproduce today results my co-authors and I published 18 years ago.

(b) *Public access to public goods.* In principle, the goal of publicly funded sponsored research is to make "knowledge" available. Increasingly, such "knowledge" now concerns the behavior of certain algorithms and software under certain conditions. It should in principle be possible for other scientists to access this "knowledge"; however, journal publication does not begin to provide the detail and depth of transmission that would be required.

Recall Claerbout's claim "an article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." Those who are sticklers for linguistic accuracy will find that "scholarship" is not really appropriate in this quotation. A better choice would have been "knowledge." Computational reproducibility is the means by which the published "knowledge" becomes really available to a widespread audience.

4. To learn more about reproducibility

Reproducible research is now developing rapidly—it is almost a movement at this point—and there is a lot of interesting material you may wish to learn more about. The work by Gentleman, Temple Lang, and others has really put reproducibility on the map in biostatistics. The reader may wish to know about efforts in other disciplines.

The IEEE magazine *Computing in Science and Engineering* had a special issue on this topic in early 2009, covering fields ranging from earth sciences to signal processing. My own group's contribution to that issue (Donoho *and others*, 2009) assessed our more than 15-year-long effort in this direction (Buckheit and Donoho, 1995; Donoho and Huo, 2005) and summarized the (very positive) citation impact that we believe results from reproducibility, as well as the many arguments pro and con that we have heard. The technical report version of that paper (Donoho *and others*, 2008) even included a section describing what the National Science Foundation and similar science funding agencies ought to be doing to encourage reproducibility. (Money talks!)

Recently, Vandewalle, Kovacevic, and Vetterli surveyed reproducible research in signal processing in IEEE *Signal Processing* magazine. Most notably, Vetterli's group introduced a reproducible research repository (http://rr.epfl.ch/) where one can publish one's code; the repository surveys users for their opinion on the reproducibility of a given paper—and summarizes the results for other users to see!

Jill Mesirov has developed a system for reproducible research in Computational Biology (Mesirov, 2009). Her system addresses researchers doing gene association studies, and who are used to microsoft office and similar tools. Her system's user interface is less technical and nerdy than R scripts and perhaps also more likely to be adopted in Computational Biology.

Clearly, much progress on this issue is being made; I expect that over time, more and more researchers will want to participate. I hope eventually some sort of critical mass will be reached so that computational reproducibility of the sort being introduced by *Biostatistics* is viewed as an essential component of scientific publication. This will provide a substantial improvement in the reliability of the published literature and also a substantial increase in the impact of the published literature. My congratulations to the editors for having the vision and energy to start this initiative.

Acknowledgment

Conflict of Interest: None declared.

REFERENCES

BUCKHEIT, J. AND DONOHO, D. L. (1995). Wavelab and reproducible research. In: Antoniadis, A. (editor), *Wavelets and Statistics*. New York, NY: Springer, pp. 55–81.

CLAERBOUT, J. AND KARRENBACH, M. (1992). Electronic documents give reproducible research a new meaning. In: *Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics*, pp. 601–604.

DONOHO, D. L. AND HUO, X. (2005). Beamlab and reproducible research. *International Journal of Wavelets, Multiresolution and Information Processing* **2**, 391.

DONOHO, D. L., MALEKI, A., UR-RAHMAN, I. SHAHRAM, M. AND STODDEN, V. (2008). Fifteen years of reproducible research in computational harmonic analysis. *Technical Report*. Department of Statistics, Stanford University, Stanford, CA.

DONOHO, D. L., MALEKI, A., UR-RAHMAN, I., SHAHRAM, M. AND STODDEN, V. (2009). Reproducible research in computational harmonic analysis. *Computing in Science and Engineering* **11**, 8.

MESIROV, J. (2009). *Research Reproducibility through GenePattern*. http://themindwobbles.wordpress.com/2009/06/27/research-reproducibility-through-genepattern-ismb-dam-sig-2009/.

STODDEN, V. (2009a). Enabling reproducible research: open licensing for scientific innovation. *International Journal of Communications Law and Policy* **13**, 1–25.

STODDEN, V. (2009b). The legal framework for reproducible scientific research: licensing and copyright. *Computing in Science and Engineering* **11**, 35–40.

VANDEWALLE, P., KOVACEVIC, J. AND VETTERLI, M. (2009). Reproducible research in signal processing—what, why, and how. *IEEE Signal Processing Magazine* **26**, 37–47. [detailed record] [bibtex] [reproducible research].