Reproducible Research: A Range of Response

David Banks
Department of Statistical Science, Duke University
banks@stat.duke.edu

Recently, there has been widespread media attention to cases of problematic scientific research. Marc Hauser, an influential Harvard primatologist, was found to have committed academic misconduct (Miller, 2010). Hien Tran of the California Air Resources Board falsely claimed to have received a doctorate in statistics from the University of California at Davis, casting doubt on his estimates of the effect of pollution on mortality (*Union Tribune*, 2009) . Duke University initiated three clinical trials based on publications by Joseph Nevins and Anil Potti, whose analyses were deeply flawed—their papers have been withdrawn and the trials discontinued (Goldberg, 2010). And Kenneth Cuccinelli, the attorney general of Virginia, suspects that environmental scientist Michael Mann committed a criminal act in his analysis of global warming data; Cuccinelli has subpoened the University of Virginia for Mann's research notes (*Washington Post*, 2010).

Peng (2009) points up the problem clearly:

> The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is "reproducible research", which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses.

As a former editor of the *Journal of the American Statistical Association*, my own sense is that very few applied papers are perfectly reproducible. Most do not come with code or data, and even if they did, I expect a careful check would find discrepancies from the published paper. The reasons are innocent:

code written by graduate students is continually tweaked and has sketchy documentation. The exact data cleaning procedures are not perfectly remembered when the final version of the paper is written, or may be muddled by miscommunication among multiple authors. And even if a conscientious researcher provided a full description of every cleaning step, every model fitting choice, and all aspects of variable selection, the resulting paper would be so long and tedious that no doubt the foolish editor would demand that it be shortened.

The scientific community, and especially statisticians, have responded to the problem of scientific validity in significant ways. Ambroise and McLachlan (2002) pioneered modern forensic statistics; this subdiscipline reworks the analyses behind important publications, and (too often!) discovers fatal mistakes. Baggerly, Morris and Coombes (2004) discredited results from a proteomics study of ovarian cancer by Petricoin and Liotta (2002). Tibshirani (2005) re-examined previously published results on genetic prediction of follicular lymphoma, and found that the authors were mistaken. Baggerly and Coombes (2009) were instrumental in discovering the errors in the research by Nevins and Potti. And, of course, in some sense a small army of statisticians at the Food and Drug Administration have been doing forensic statistics for decades, checking the work of pharmaceutical companies that make New Drug Applications.

Statisticians are natural leaders in this kind of analytic review. But we are inadequate to the scale of the problem. The main drawback to use of forensic statistics to improve reproducibility is that it is so labor-intensive. Only a handful of studies can be deeply checked, and these are usually the ones that make the most spectacular claims. Additional barriers are that such review is slow, the work is not valued unless it discredits important publications, the process of obtaining the original data and code can be onerous, and there is not much joy in calling foul on the work of respected scientists and colleagues.

Traditional mechanisms for maintaining the quality of science rely upon the offices of editors, coauthors, and institutions. Editors are supposed to ensure that papers are closely scrutinized for correctness before publication. But this is a fiction—editors do the best that they can, but the process depends upon volunteer reviewers, all of whom are vastly overcommitted already. Standard refereeing would not discover the kinds of flaws that occurred in most of the high-profile cases that have roiled the media. And there is a subtle downside to editorial responsibility—it places an official stamp of approval on the paper. The public, and often even scientists, place too much trust in that stamp, which may contribute to the outcry and opprobrium that attends the discovery of honest, or even sloppy, mistakes.

Coauthors are probably the most effective enforcement mechanism for scientific quality. If the senior partner is meticulous and principled, it sets the tone for the entire collaboration. But beyond the tone, there is little that coauthors can do. Realistically, they cannot microcheck each other's code, and for multidisciplinary research, as in bioinformatics, the division of labor often implies that no member of the team is competent to check the work of the others. For very large research groups, perhaps working over

long time spans, accountability is especially difficult. And in cutting-edge research, the race for priority necessarily rewards short-cuts. [Peter Bickel told me of a story about Millikan and Harvey's oil drop experiment. Millikan and German scientists led by Ehrenhaft were competing to discover the unitary charge of the electron; it is a technically difficult experiment, and Ehrenhaft's group was not able to consistently reproduce their results. Neither could Millikan, but he cherry-picked his best experiments, reported those, and won the Nobel prize (Niaz, 2000).]

Institutional mechanisms are probably the least-effective ways to promote reproducibility. Duke University's initial investigation of the work by Nevins and Potti was conspicuously poor, and may have better served the (myopically) perceived needs of the institution than cancer patients or science (Goldberg, 2010). In the David Baltimore case, two review boards (one at Tufts, chaired by Dr. Wortis, and one at MIT, chaired by Dr. Eisner) found against Margot O'Toole, the post-doctoral researcher who claimed that Imanishi-Kari's work was not reproducible (this essentially destroyed O'Toole's career). Years later, the Office of Research Integrity at the Department of Health and Human Services found Imanishi-Kari guilty of 19 counts of research misconduct; an appeals panel later reversed that decision, and the upshot is that the entire matter has become a Rorschach test in scientific ethics (cf. Kevles, 1996). Carnegie Mellon's prosecution of the Rimm Study, in which I was charged with misconduct, seems to me to have been improperly manipulated to protect the university from adverse publicity. Vergano (2010) reports that George Mason University and Rice University are investigating Ed Wegman and David Scott on charges of paraphrasing without attribution from the Wikipedia and various textbooks in their report to Congress on the statistics of global warming—I have to construe this as largely a reprisal for their support of academically unpopular climate skepticism. Banks (2003) provides other examples in which institutions have shown reckless fecklessness in addressing reproducibility concerns.

Since editors, coauthors, and institutions can do little to further reproducibility, statisticians have stepped forward with more innovative proposals. Roger Peng (2009) has urged that *Biostatistics* consider providing a "reproducibility review" for selected articles—this would entail a third party working to check the code, the data and the conclusions. Leisch (2002) has created a tool called Sweave that embeds R code into LaTeX documents so that readers/users can fully replicate an analysis, or perform alternative anlayses. Peng and Eckel (2009) extend this strategy, providing a set of tools for cached computation that promote sharing of data and code. More broadly, in bioinformatics some journals have adopted the MIAME (Minimum Information About a Microarray Experiment) standard for publication, which has pushed more data into the public sector and made studies more transparent. If statistics journals made submission of code a condition of publication, that might have similar salutary effects.

But I am skeptical. Ochsner et al. (2008) surveyed 20 journals that required data deposition for publication, and found that fewer than 50% of the authors had done so; they do not discuss the adequacy of metadata needed for reproducibility checks. More stringently, Ioannidis et al. (2009) examined all 18 quantitative microarray papers from *Nature Genetics* that appeared during the preceding two-year

interval when data deposition was required. Their goal was to reproduce the first quantitative result in each paper. In two cases, they succeeded; in ten cases, they found that the information provided was so incomplete that it was impossible to assess reproducibility.

I see a reproducibility standard as a noble aspiration. As Buckheit and Donoho (1995) noted:

> An article about computational science in a scientific publication is **not** the scholarship itself, it is merely the **advertising** of the scholarship. The actual scholarship is the complete software developmental environment and the complete set of instructions which generated the figures.

If this degree of documentation can be provided, it should be. But detailed documentation requires discipline and resources, both of which are barriers. There will be generational resistance; researchers would have to reconceptualize their approach, learn new software tools, and invest themselves more in process management at the expense of traditional research activities. Often statisticians do not own the data that they analyze. When there are legal requirements that data be placed in the public domain, there is still latitude for long delays, with little incentive to produce transparent, documented code. And even a careful and competent researcher may be reluctant to open up data to public potshots; honest mistakes can have severe reputational consequences, and one unkind colleague could create a chorus of criticism. (Partial rebuttals to my reservations are given in Donoho et al., 2009, but I am not persuaded.)

Instead, I have three small suggestions. None of these will solve the problem, but I think they may help improve the situation.

1. Poor reproducibility is tacitly promoted by a system that rewards publication of positive results. But in an era of electronic media, which eliminates the competition for shelf-space, it should be perfectly possible to publish null results. It is good science and honest work to e-publish a paper that says "None of the following 1,000 genes in my study had any statistical relationship to cancer." Our current research culture does not value this contribution as it should, but if such papers appear in respected e-journals, I think that may evolve.

2. Currently, there is no direct funding line for reproducibility studies. I can imagine that the NSF and the NIH might entertain proposals for replication of other researcher's studies. One argument is that if the original project was worth funding when the outcome was entirely speculative, it is certainly worthwhile to fund a follow-up study to replicate important results (especially since the follow-up funding might be much cheaper, since it could consist simply of an arm's length reanalysis of data that have already been expensively collected). A second argument is that such a program could have the socially desirable outcome of directing more funding to second-tier universities—it might take a top-drawer investigator to do the original study, but one need not be a genius to check the answer. One might even imagine that such money would go to support graduate students who would spend a semester or a year reproducing a study, under the supervision of a faculty member,

with obvious educational benefits. (I'm told that Victoria Stodden at Columbia University has developed a course in which statistics graduate students check published analyses.)

3. it would be worthwhile to create a continuous measure of reproducibility, and score a random sample recent publications to determine the extent of the problem. We may be suffering from attention bias (which some people claim was Marc Hauser's failing in his primatology interpretations). Perhaps we are so dazzled by the occasional scandal that we overlook the fact that most research is quite sufficiently solid (although surely not perfect). On the other hand, the dazzling scandals may distract us from the possibility that nearly all applied statistical research has pretty serious flaws. Without some sort of baseline, it will be difficult to assess progress.

Clearly, the problem of reproducible research is fundamental to the integrity of science and public trust. Statisticians are among the leaders in this effort, but there are significant social barriers to change. Nonetheless, a wide range of approaches are being developed. We should be optimistic about the future, and proud of the role that statisticians have played in creating it.

Acknowledgement

I think Keith Baggerly of M. D. Anderson and Roger Peng of Johns Hopkins University for thoughtful comments and references.

References

Ambroise, C., and McLachlan, G. (2002). "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proceedings of the National Academy of Sciences* (USA), **99**, 6562–6566.

Baggerly, K., and Coombes, K. (2009). "Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology," *Annals of Applied Statistics*, **3**, 1309–1334.

Baggerly, K., Morris, J., and Coombes, K. (2004). "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics*, **20**, 777–785.

Banks, D. (2003). "Discussion of 'Emerging Ethical Issues in Statistical Publication," *Proceedings of the American Statistical Association: Section on Statistical Consulting*, 135–136.

Buckheit, J., and Donoho. D. (1995). "Wavelab and Reproducible Research," *Wavelets and Statistics*, A. Antoniadis, ed., Springer-Verlag, New York, N.Y., pp. 55–81.

Donoho, D., Maleki, A., Rahman, I., Shahram, M., and Stodden, V. (2009). "Reproducible Research in Computational Harmonic Analysis," *Computing in Science and Engineering*, **11**, 8–18.

Goldberg, P. (2010). "By Defending Potti, Duke Officials Become Target of Charges of Institutional Failure," *The Cancer Letter*, **36**, No. 28, July 23.

Ionnides, J., Allison, D., Ball, C., Coulbaly, I., Cui, X., Culhane, A., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G., Petretto, E., and van Noort, V. (2009). "Repeatability of Published Microarray Gene Expression Analyses," *Nature Genetics*, **41**, 149–155.

Kevles, D. (May 27, 1996). "Annals of Science: The Assault on David Baltimore." *The New Yorker*. http://www.newyorker.com/archive/1996/05/27/1996_05_27_094_TNY_CARDS_000374549.

Leisch, F. (2002). "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis," in Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 - Proceedings in Computational Statistics*. Physica Verlag, Heidelberg, pp. 575–580.

Miller, G. (2010). "Harvard Dean Confirms Misconduct in Hauser Investigation," *Science*, Aug. 20.

Niaz, M. (2000). "The Oil Drop Experiment: A Rational Reconstruction of the Millikan-Ehrenhaft Controversy and Its Implications for Chemistry Textbooks," *Journal of Research in Science Teaching*, **37**, 480–508.

Ochsner, S., Steffen, D., Stoeckert, Jr., C., McKenna, N. (2008). "Much Room for Improvement in Deposition Rates of Expression Microarray Datasets," *Nature Methods*, **5**, 991.

Peng, R. (2009). "Reproducible Research and *Biostatistics* (with discussion)," *Biostatistics*, **10**, 405–408. http://biostatistics.oxfordjournals.org/cgi/reprint/10/3/405

Peng, R., and Eckel, S. (2009). "Distributed Reproducible Research Using Cached Computations," *IEEE Computing in Science and Engineering*, **11**, 28–34.

Petricoin E., Ardekani, A., Hitt, B., Levine, P., Fusaro, V., Steinberg, S., Mills , G., Simone, C., Fishman, D., Kohn, E., and Liotta, L. (2002). "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer," *The Lancet*, **359**, 572–577.

Tibshirani, R. (2005). Letter, "Immune Signatures in Follicular Lymphoma," *New England Journal of Medicine*, **352**, 1496.

Union Tribune (2009). "The Air Board's Shame/Staff Never Revealed Internal Scandal Before Crucial Vote," *Union Tribune* editorial, Nov. 22.

Vergano, D. (November 21, 2010). "Experts Claim 2006 Climate Report Plagiarized." *USA TODAY*. http://www.usatoday.com/weather/climate/globalwarming/2010-11-21-climate-report-questioned_N.htm.

Washington Post (2010). "Ken Cuccinelli Seems Determined to Embarrass Virginia," *Washington Post* editorial, Oct. 6.