

# AMSTATNEWS

The Membership Magazine of the American Statistical Association

- [Home](#)
- [About](#)
- [Editorial Calendar](#)
- [Submission Instructions](#)
- [PDF Archives](#)
  - [2008 Amstat News](#)
  - [2009 Amstat News](#)
  - [2010 Amstat News](#)
  - [2011 Amstat News](#)
- [Advertise](#)



[Home](#) » [Columns](#), [Featured](#), [Science Policy](#)

## Reproducible Research

1 January 2011 16,774 views 11 Comments

*This month's guest editors, Keith A. Baggerly and Donald A. Berry, make the case that journals have a key role to play in making research reproducible. Their call comes in the aftermath of attempts to reproduce the cancer research results of Duke's Anil Potti and Joseph Nevins, whose seemingly promising 2006 work led to three clinical trials. Baggerly and colleague Kevin R. Coombes were the lead figures to uncover not only an inability to reproduce the research, but many obstacles in attempting to do so. In November, Duke terminated all three trials and Anil Potti resigned.*  
~ Steve Pierson, ASA Director of Science Policy, [pierson@amstat.org](mailto:pierson@amstat.org)

### ***Contributing Editors***



Keith A. Baggerly is professor of bioinformatics and computational biology at The University of Texas MD Anderson Cancer Center in Houston, Texas.



Donald A. Berry is head of the division of quantitative sciences and chair and professor of the department of biostatistics at The University of Texas MD Anderson Cancer Center.

Research is reproducible if it can be reproduced by others. Of course, rerunning an experiment will give different results—an observation that gave rise to the development of statistics as a discipline. Our focus here is “reproducible research” (RR) in the sense of reproducing conclusions from a single experiment based on the measurements from that experiment.

In the 1990s, the geophysicist Jon Claerbout became frustrated with new students having great difficulty duplicating previous students’ research. Making further advances meant spending months, or even years, trying to reproduce previous advances. Reproducing computations can be agonizing, even with one’s own calculations. The problem is exacerbated by the passage of time. We suspect all applied statisticians have had such an experience.

Our definition of RR includes a complete description of the data and the analysis of that data—including computer programs—so the results can be exactly reproduced by others. RR is self-contained. This is obviously important in Claerbout’s motivating context. But it is as or more important in letting others judge the appropriateness of the analysis.

As data sets become more voluminous and complex, RR is increasingly important because our intuition fails in high dimensions. Investigators know how expression levels of some genes predict patient outcomes, but nobody understands the predictive efficacy of “signatures” involving hundreds of genes. Researchers may think they know, but we have (inadvertently) performed the null experiment of giving an investigator a list of random genes—purportedly prognostic—and having the investigator explain why they made sense. To use complex signatures, we must trust that the steps taken to produce them are accurate, or at least checkable.

Much scientific research cannot be reproduced, which can have severe consequences. We recently tried to reproduce reports that drug sensitivity “signatures” could predict patient response to cancer chemotherapy. Since the reported methods were incomplete, we employed “forensic bioinformatics” to reconstruct the methods used to derive the reported results. These reconstructions identified several basic errors, including drastic mistakes such as reversing “sensitive” and “resistant” labels. These errors initially went unnoticed because it was impossible to easily check the steps.

We published our reconstructions in September of 2009. Clinical trials in lung cancer at Duke University had begun two years before. Lack of clear reproducibility hampered corrective action. Trials were suspended in October of 2009 while Duke investigated our claims. Trials were restarted in January of 2010, although Duke didn’t release the investigation’s report or the data justifying the restart. A redacted report became available in May under the [Freedom of Information Act](#). It indicated that important problems of reproducibility were not adequately addressed. In the interim, we publicized new problems with the analysis.

In July, it was revealed that a principal investigator had allegedly made false claims (including a Rhodes scholarship) on his CV and in grant applications. Several statisticians wrote to Harold

Varmus, director of the National Cancer Institute, to request the trials be suspended again, and they were. Journals that had published the investigators' results began independent investigations. In October of 2010, the senior author requested retraction of a paper on which the trials were based, citing problems we reported to Duke in November of 2009.

Journals must begin to demand RR. Convincing them will not be easy. Authors have legitimate concerns, including data ownership, intellectual property, and identifying where data can be accessibly stored. Despite these concerns, there is growing recognition that RR is important. Spurred by the forensic episode cited above, the Institute of Medicine announced a review of omics-based tests in clinical trials in October of 2010. Sessions on RR are more common at scientific and editorial board meetings. Journal editors are aware that irreproducibility can damage credibility and are open to suggestions regarding improvements.

Simple enforcement of rules many journals already have (e.g., that data be posted) at even a cursory level would help. The journal *Biostatistics* has an associate editor for reproducibility who can assign grades of merit to conditionally accepted papers: D: data are available, C: code is available, and R: the AE could run the code and reproduce the results without much effort.

This last constraint requires authors to make their code user friendly, but we see no alternative. Few reviewers have the time or expertise to assess reproducibility, though they can check whether data and code are provided. In a September *Nature* [letter](#), we and others described information journals should elicit from authors. This includes data (raw, processed, and clinical, with indications of provenance), code, descriptions of nonscriptable steps, and prospectively defined analyses (in a protocol, for example), if any.

Motivated by the clear need, we have tried to make our work reproducible. We found the process less cumbersome than we feared. In bioinformatics and computational biology at MD Anderson Cancer Center, we now require that reports be prepared using [Sweave](#). This software uses R and Latex. On running Sweave files through R, code is evaluated and results formatted for proper inclusion. Given the report, if we have access to the same raw data, we can run the report and confirm that we obtain the same values. (See [example reports](#).) Our adoption of Sweave involved a few months of learning, but we now produce better reports faster because we have reusable templates and we are in the habit of planning for reproducibility from the outset. Alternatives to Sweave, including [GenePattern](#), are available.

Finally, RR is necessary, but not sufficient, for good science. It needn't contain the motivation for what was done, and the motivation may be data-dependent. For example, suppose we compare counts following two treatments, but the results are not statistically significant. So we take logarithms, after adding 0.01 to every count. Then, we tell the "complete" story, starting from taking logarithms. Or perhaps the data we used were "cleaned" before we got them. These potentially fatal biases will not be known by someone checking reproducibility, and they may not be known to the primary analyst. Difficulties with these and other "silent multiplicities" are described in a [2007 piece](#) in *Pharmaceutical Statistics* written by Donald A. Berry.

We exhort statisticians to join in making RR more widespread. Click [here](#) for further reading.



11 Comments »

★★★★★ (2 votes, average: 5.00 out of 5)