Biostatistics (2009), **10**, 3, *pp.* 405–408
doi:10.1093/biostatistics/kxp014
Advance Access publication on xxx xx, xxxx

Editorial

As coeditors of *Biostatistics*, we wish to encourage the practice of making research published in the journal reproducible by others. The following invited piece by Roger Peng sets out our policy on this; Roger will be assuming the role of Associate Editor for reproducibility as set out in his piece.

While we consider reproducibility to be a desirable goal, we wish to emphasise that our policy is to encourage our authors to consider this as an opportunity that they may wish to take, rather than as a requirement that we impose upon them. All submissions to the journal will continue to be reviewed using our established system; the issue of reproducibility will be considered only when a paper had been accepted for publication on the basis of its scientific merit as judged by our peer-review process.

PETER J. DIGGLE, SCOTT L. ZEGER

Reproducible research and *Biostatistics*

ROGER D. PENG

1. INTRODUCTION AND MOTIVATION

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.

The need for publishing reproducible research is increasing for a number of reasons. Investigators are more frequently examining weak associations and complex interactions for which the data contain a low signal-to-noise ratio. New technologies allow scientists in all areas to compile complex high-dimensional databases. The ubiquity of powerful statistical and computing capabilities allows investigators to explore those databases and identify associations of potential interest. However, with the increase in data and computing power comes a greater potential for identifying spurious associations. In addition to these developments, recent reports of fraudulent research being published in the biomedical literature have highlighted the need for reproducibility in biomedical studies and have invited the attention of the major medical journals (Laine *and others*, 2007). Even without the presence of deliberate fraud, it should be noted that as analyses become more complicated, the possibility of inadvertant errors resulting in misleading findings looms large. In the examples of Baggerly *and others* (2005) and Coombes *and others* (2007), the errors discovered were not necessarily simple or obvious and the examination of the problem itself required

a sophisticated analysis. Misunderstandings about commonly used software can also lead to problems, particularly when such software is applied to situations not originally imagined (Dominici *and others*, 2002).

While many might agree with the benefits of disseminating reproducible research, there is unfortunately a general lack of infrastructure for supporting such endeavors. Investigators who are willing to make their research reproducible are confronted with a number of barriers, one of which is the need to distribute, and make available for an indefinite amount of time, the supplementary materials required for reproducing the results. Another is the lack of an "instruction manual" that indicates which materials are needed and what might be the most suitable formats for making data and computer code available. In this editorial, we describe the efforts that *Biostatistics* is making to promote reproducibility in biostatistical research.

2. REPRODUCIBILITY POLICY FOR *Biostatistics*

From this issue forward, *Biostatistics* is willing to work with authors to publish articles that meet a standard of reproducibility. The standard involves three different dimensions that we describe in greater detail below. The purpose of defining different dimensions of reproducibility is to provide some level of continuity between "not reproducible" and "reproducible." The journal has for some time now allowed and encouraged authors to place supplementary materials online via the journal's Web site and the reproducible research policy builds upon that framework. It should be noted that this policy is still in the early stages and it is likely that the details will evolve as we gain experience working with authors.

2.1 *Dimensions of reproducibility*

The Associate Editor for reproducibility (AER) will handle submissions of reproducible articles. Currently, the AER's involvement with a submission begins only when an article has been accepted for publication. The AER will consider three different criteria when evaluating the reproducibility of an article.

1. Data: The analytic data from which the principal results were derived are made available on the journal's Web site. The authors are responsible for ensuring that necessary permissions are obtained before the data are distributed.
2. Code: Any computer code, software, or other computer instructions that were used to compute published results are provided. For software that is widely available from central repositories (e.g. CRAN, Statlib), a reference to where they can be obtained will suffice.
3. Reproducible: An article is designated as reproducible if the AER succeeds in executing the code on the data provided and produces results matching those that the authors claim are reproducible. In reproducing these results, reasonable bounds for numerical tolerance will be considered.

Authors can choose to meet a subset of these criteria if they wish. For example, an author may choose to release code showing how a particular method is implemented but may not have permission to publish the data. In such a case, the "code" criterion is satisfied, but the "data" and "reproducible" criteria are not. For authors interested in submitting materials satisfying the "reproducible" criterion, the journal is currently limiting submissions to those whose analyses are conducted using the R software environment. This limitation may change in the future and will generally be dependent on the resources of the journal and the AER. Papers that meet any or all of the above three numbered criteria will be kite marked D, C, and/or R on their title page in the journal.

3. Instructions for authors

In general, authors should indicate in their submission that they intend to submit supplementary materials specifically for the purposes of allowing others to partially or fully reproduce their work. Authors can submit analytic data sets, computer code, or both in support of their papers. Authors may additionally indicate which results of the paper can be reproduced using the submitted materials, although such an indication is not required. Unless an author indicates that the supplementary materials satisfy the "reproducible" criterion (see below), the submitted materials will not be checked for reproducibility but, at the discretion of the Editors, may still be posted to the journal's Web site with an indication that the author has contributed the materials for the purpose of reproducing the results.

When submitting data sets and code, authors should use open and documented formats rather than proprietary formats. Files containing computer code should be submitted in ASCII text format. While proprietary data formats may be standard in some subspecialties, we would prefer that an open alternative be submitted for the purposes of posting on the journal's Web site. Increasing the longevity and usefulness of the data and code is one important goal which is best supported by the use of open formats. The AER will work with authors to find appropriate formats for data and code submissions.

3.1 *Materials for satisfying the "reproducible" criterion*

To satisfy the "reproducible" criterion, authors should submit all the necessary materials so that the AER can execute the code on the analytic data sets and produce output similar to that obtained by the author. Currently, only submissions written using the R software environment will be accepted for satisfying this criterion. Authors should submit the following:

1. A "main" script which directs the overall analysis. This script may load data, other software, and call the necessary functions for conducting the analysis described in the article.
2. Other required code files, presumably called from the "main" script file.
3. External data or auxiliary files containing the analytic data sets or other required information.
4. A "target" file (or files) containing the results which are to be reproduced. Such a file could consist of an ASCII text file containing numerical results or a PDF file containing a figure. This will aid in the comparison of computed results with published results.

Although not required, authors are encouraged to use literate programming tools such as a combination of LaTeX and Sweave. Specifically for those using Sweave, submissions should include the following:

1. The original Noweb source for the article, typically a file with a .Rnw or .Snw extension.
2. The LaTeX file generated by the Sweave function, typically with file extension .tex.
3. Any data files or auxiliary code needed to execute the Sweave function successfully on the Noweb source file.
4. Any bibliographic database files (i.e. BibTeX files).

If the AER is able to reproduce the stated results, the submitted materials will be posted to the journal's Web site with an indication that the results in the corresponding paper are reproducible.

4. Example

In this issue, Duncan Lee and coauthors have published a paper ("Air pollution and health in Scotland: A multi-city study") along with all the necessary data and code for reproducing the principal findings presented in their paper. In the paper, the authors relate hospital admission counts to air pollution levels with spatial Poisson regression models. These models are fitted to the data within a Bayesian framework

using Markov chain Monte Carlo (MCMC) methods. The authors have provided the code implementing their MCMC sampler and have provided an "Overall script.R" file that directs the analysis.

REFERENCES

BAGGERLY, K., MORRIS, J., EDMONSON, S. AND COOMBES, K. (2005). Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *Journal of the National Cancer Institute* **97**, 307–309.

COOMBES, K., WANG, J. AND BAGGERLY, K. (2007). Microarrays: retracing steps. *Nature Medicine* **13**, 1276–1277.

DOMINICI, F., MCDERMOTT, A., ZEGER, S. L. AND SAMET, J. M. (2002). On the use of generalized additive models in time series studies of air pollution and health. *American Journal of Epidemiology* **156**, 1–11.

LAINE, C., GOODMAN, S. N., GRISWOLD, M. E. AND SOX, H. C. (2007). Reproducible research: moving toward research the public can really trust. *Annals of Internal Medicine* **146**, 450–453.