Biometrical Journal and Reproducible Research

Torsten Hothorn[1,*], **Leonhard Held**[2] and **Tim Friede**[3]

[1] Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstraße 33, D-80539 München, Germany
[2] Institut für Sozial- und Präventivmedizin, Universität Zürich, Hirschengraben 84, CH-8001 Zürich, Switzerland
[3] Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK

Biometrical Journal serves the scientific community for over 50 years by publishing biostatistical innovations (Bergholt *et al.*, 2008; Victor *et al.*, 2008). These articles make a difference in many fields of our discipline, for example, in clinical trials methodology or survival analysis. Even more important, the journal helps to transfer new theoretical insights and statistical methods to medicine and life sciences. Although being already very successful in this direction, there remains much to be done to further increase the quality of articles published by *Biometrical Journal* and to increase the scientific impact of this research.

Almost all articles in *Biometrical Journal* present new methodology, and many authors provide mathematical proofs demonstrating correctness or optimality in some sense. These theoretical contributions are carefully checked by referees and Associate Editors. This thorough review process ensures the high quality of accepted articles. Theoretical results are often accomplished by empirical investigations. The majority of authors choose to analyse data previously published elsewhere to demonstrate the practical relevance of their contribution. The performance of new statistical methodology is often studied empirically through simulation experiments.

However, empirical investigations rely heavily on data and computer programs. Correct and efficient implementation in a computer program is crucial for simulation experiments, for summarising the results in figures and tables, and for the statistical analysis of real data. Although referees try to do their best to detect inconsistencies in the theoretical parts of an article, it is hard to comment on the validity of simulation experiments or the correctness of a specific data analysis. Indeed, due to the increasing complexity of statistical methodology with many control or tuning parameters, it is often impossible to reproduce published results. For example, in a recent letter to the Editor, Rubio and Pérez-Elizalde (2009) have identified numerical problems of a Markov Chain Monte Carlo (MCMC) analysis in Mendoza and Gutiérrez-Pena (1999). The article is perfectly valid from a theoretical point of view and the MCMC algorithm used is also correct in principle. However, convergence problems and perhaps too small Monte Carlo sample size have led to rather misleading results.

Moreover, for many practitioners, it may be hard or even impossible to apply a new method if no software is readily available. Under these circumstances, it is unlikely that a new method will make its way into applications. Thus, an article might not get the credit (usually measured by the number of citations) it deserves. Why? The reason is simple and best described by what de Leeuw (2001) coined *Claerbout's principle* after the geophysicist Jon Claerbout:

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

* Correspondence author: e-mail: Torsten.Hothorn@stat.uni-muenchen.de , Phone: +49-89-2180-6407, Fax: +49-89-2180-5040

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This principle has been defined by Buckheit and Donoho (1995) and is elaborated by Schwab, Karrenbach, and Claerbout, (2000). Translated into biostatistical terms this simply means: The scholarship does not only consist of theorems and proofs but also (and perhaps even more important) of data, computer code and a runtime environment which provides readers with the possibility to reproduce all tables and figures in a article. In this sense, a piece of reproducible research is an article that provides readers with all the materials that are needed to produce the same results as described in the publication. These ideas have been widely adopted in statistics in the meantime, see for example Leisch and Rossini (2003) or Gentleman and Temple Lang (2007).

An important question comes to our mind: Which proportion of articles in *Biometrical Journal* actually meets this high-quality standard? As statisticians, we have investigated this question by analysing a small questionnaire on all regular articles published last year in volume 50 of *Biometrical Journal* (omitting the special issue on *Multiple Comparison Procedures*). For each article, we recorded whether or not an article presents simulation results and one or more practical examples. In addition, we checked if access to data was given (either directly in the article or in the electronic form – we did not check references to other articles for the availability of the complete data). Last but not least, we looked for computer source code or binary programs implementing (parts of) the proposed methodology.

In total, 56 articles have been published in the five regular issues of volume 50. As summarised in Table 1, there are 53 articles with either simulations and/or illustrating examples. Only 17 of these articles provide access to data in some form and only eight articles give the readers the possibility to experiment themselves with computer code. Access to both code and data was given by six articles (Piepho, Richter, and Williams, 2008; Häducke, Pahlke, and Ziegler, 2008; Kuss, Blankenburg, and Härting, 2008; Bretz *et al*., 2008; Hothorn, Bretz, and Westfall, 2008; Li *et al*., 2008), corresponding to 11% of articles that at least potentially can be seen as *reproducible research*. That is not too bad, but there is much room for improvement.

The problem of non-reproducible research has recently been tackled from two slightly different points of view. The first one is a computer science view, offering ideas such as Literate Programming (Knuth, 1992) and tools such as noweb (Ramsey, 1994) and Sweave (Leisch, 2002) to create documents that contain the whole scholarship, *i.e*. textual descriptions including theoretical derivations as well as data and computer code for the empirical part. Also very interesting is the current discussion regarding the quality of statistical analyses in medical journals, *i.e*. the more ''practical'' view on the problem. Three prominent journals, the *Annals of Internal Medicine*, the *American Journal of Epidemiology* and *Biostatistics*, are currently moving towards making research articles reproducible by asking authors to provide access to both data and computer code (Peng, Dominici and Zeger, 2006; Laine *et al*., 2007; Peng, 2009).

It is our aim to increase the quality, usefulness and scientific impact of *Biometrical Journal* articles through reproducibility. Of course, this is a long-term goal and insisting on full reproducibility is certainly too ambitious. Availability of data or software may even be impossible in some circumstances. However, we want to encourage authors to move with us in this direction and to submit data and software with the manuscript. In our view, there is no better advertising for new methodology than providing data and computer code for the reader. After acceptance of an article, supporting information will undergo an additional reproducibility check by the Associate Editor for

Table 1 Total numbers of articles presenting simulation studies or example analyses and giving access to data or code in issues 1–4 and 6 of volume 50.

	Simulation	Example	Data	Code
No	17 (30.4%)	8 (14.3%)	39 (69.6%)	48 (85.7%)
Yes	39 (69.6%)	48 (85.7%)	17 (30.4%)	8 (14.3%)

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Reproducible Research (currently Torsten Hothorn). He will help the authors to provide computer code which will be useful to readers and which will eventually increase the impact of their work.

Reference

Bergholt, A., Burger, H.-U., Hothorn, L. A. and Ziegler, A. (2008). Presidential address: 50 years Biometrical Journal. *Biometrical Journal* **50**, 5–7.

Bretz, F., Hsu, J., Pinheiro, J. and Liu, Y. (2008). Dose finding–a challenge in statistics. *Biometrical Journal* **50**, 480–504.

Buckheit, J. and Donoho, D. L. (1995). *Wavelets and Statistics*, Chapter Wavelab and Reproducible Research. Springer, New York.

de Leeuw, J. (2001). Reproducible research: the bottom line. *Technical Report 2001031101*. Department of Statistics Papers, University of California, Los Angeles. http://repositories.cdlib.org/uclastat/papers/2001031101.

Gentleman, R. and Temple Lang, D. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics* **16**, 1–23.

Häducke, O., Pahlke, F. and Ziegler, A. (2008). A general approach for sample size and power calculations based on the Haseman-Elston method. *Biometrical Journal* **50**, 257–269.

Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**, 346–363.

Knuth, D. E. (1992). *Literate Programming, volume 27 of CSLI Lecture Notes*. Center for the Study of Language and Information, Stanford, California.

Kuss, O., Blankenburg, T. and Härting, J. (2008). A relative survival model for clustered responses. *Biometrical Journal* **50**, 408–418.

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.

Leisch, F. (2002). Dynamic generation of statistical reports using literate data analysis, in: Härdle, W. and Rönz, B. (eds.), *COMPSTAT 2002 – Proceedings in Computational Statistics*. Physica, Heidelberg, 575–580.

Leisch, F. and Rossini, A. J. (2003). Reproducible statistical research. *Chance* **16**, 46–50.

Li, J., Nordheim, E. V., Zhang, C. and Lehner, C. E. (2008). Estimation and confidence regions for multi-dimensional effective dose. *Biometrical Journal* **50**, 110–122.

Mendoza, M. and Gutiérrez-Pena, E. (1999). Bayesian inference for the ratio of the means of two normal populations with unequal variances. *Biometrical Journal* **41**, 133–147.

Peng, R. D. (2009). Reproducible research and biostatistics. *Biostatistics* **10**, 405–408.

Peng, R. D., Dominici, F. and Zeger, S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology* **163**, 783–789.

Piepho, H.-P., Richter, C. and Williams, E. (2008). Nearest neighbour adjustment and linear variance models in plant breeding trials. *Biometrical Journal* **50**, 164–189.

Ramsey, N. (1994). Literate programming simplified. *IEEE Software* **11**, 97–105.

Rubio, F. J. and Pérez-Elizalde, S, (2009). Letter to the Editor. *Biometrical Journal* **51**, XXX–XXX.

Schwab, M., Karrenbach, M. and Claerbout, J. (2000). Making scientific computations reproducible. *Computing in Science and Engineering* **2**, 61–67.

Victor, N., Läuter, J., Ihm, P. and Dietz, K. (2008). Celebrating fifty years of the *Biometrical Journal*. *Biometrical Journal* **50**, 901–910.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim