BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 3. 288–300
Advance Access published on 28 January 2011
doi:10.1093/bib/bbq084

Case studies in reproducibility

Torsten Hothorn and Friedrich Leisch

Submitted: 3rd September 2010; Received (in revised form): 29th November 2010

Abstract

Reproducible research is a concept of providing access to data and software along with published scientific findings. By means of some case studies from different disciplines, we will illustrate reasons why readers should be given the possibility to look at the data and software independently from the authors of the original publication. We report results of a survey comprising 100 papers recently published in *Bioinformatics*. The main finding is that authors of this journal share a culture of making data available. However, the number of papers where source code for simulation studies or analyzes is available is still rather limited.

Keywords: *Software; statistical analyzes; reproducible research; Sweave*

INTRODUCTION

'Wealthy men give women more orgasms'. This and similar headlines could be read in almost all newspapers around the world early in 2009. The finding was reported by Pollet and Nettle (2009) analyzing data gathered 1999–2000 in the Chinese Health and Family Life Survey. For scientists, it is natural to ask what evidence the authors can provide for such a finding. Pollet and Nettle [1] very carefully describe the data and the methods applied and their analysis meets the state-of-the-art for statistical analyzes of such a survey. Since the data are publically available, it should be easy to fit the model and derive the same conclusions on your own computer. It is, in fact, possible to do so using the same software that was used by the authors. So, in this sense, this article is fully reproducible. However, one fails performing the same analysis in R Core Development Team [27]. It turns out that Pollet and Nettle [1] were tricked by a rather unfortunate and subtle default option when computing AICs for their proportional odds model in SPSS. Herberich *et al.* [2] report on a reanalysis of this data set using a correct version of the AIC. It turns out that the male partner's income now receives less importance compared, for example, to the education level of the woman.

Global warming makes the headlines almost every day. One very influential finding, namely the 'hockey stick' pointing out dramatically increasing temperatures in the northern hemisphere in the second half of the 20th century, was published by Mann *et al.* [3]. The results presented in this article are also fully reproducible in the sense that, given the data, one can re-perform the statistical analysis to obtain the same results. However, the data processing and the statistical analysis itself were questioned by various authors, McIntyre and McKitrick [4] among them. This and other criticism lead to the 'hockey stick controversy' (http://en.wikipedia.org/wiki/Hockey_stick_controversy), which, at least partially, focuses on subtle details of statistical analysis techniques (such as centering data before performing principal component analyzes).

While a scientific debate on the relationship of men's wealth and women's orgasm frequency might be interesting only for a smaller group of specialists there is no doubt that the scientific evidence of global warming has enormous political, social and economic implications. In both cases, there would have been no hope for other, independent, researchers of detecting (potential) problems in the statistical analyzes and, therefore, conclusions, without access to the data. Furthermore, a reanalysis would be rendered difficult or even impossible without at least proper reference to the software used to derive the results. Acknowledging the many subtle

Corresponding author. Torsten Hothorn, Institut für Statistik, LMU München, DE-80802 München, Germany.
Tel: +49 89 2180 6407; Fax: +49 89 2180 5040; E-mail: torsten.hothorn@stat.uni-muenchen.de
Torsten Hothorn is Professor of Biostatistics at the Ludwig-Maximilians-Universität München in Germany. His research interests include regression modelling, variable and model selection, nonparametric statistics, and reproducible research.
Friedrich Leisch is Professor of Statistics at the Ludwig-Maximilians-Universität München in Germany. His research interests include statistical computing, finite mixture models, cluster analysis, and reproducible research.

choices that have to be made and that never appear in a 'Methods' section in papers, McIntyre and McKitrick [5] go as far as printing the main steps of their analysis in the paper (as R code). As a consequence of this and similar debates, Warren Washington, National Center for Atmospheric Research, in a Congressional Briefing (11 May 2010) (http://amstat.org/outreach/climatescience.cfm) demanded that 'All climate data should be freely available by others' and 'The scientific results must have reproducibility'.

Closer to our own field, namely Bioinformatics and Biostatistics, Ioannidis *et al.* [6] investigate the reproducibility of microarray gene expression analyzes published in a term of 4 years in *Nature Genetics*. They undertook the effort to reanalyze every paper under test by two independent teams of experts and state

> We reproduced two analyzes in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis.

A more detailed 'forensic' analysis of a published microarray study was published by Baggerly and Coombes [7]. It contains detailed case studies of problems when reproducing published results by others, which even led to the suspension of the authors of the original manuscript (http://cancerletter.com/articles/20100902). They make their own paper reproducible by providing extended supplementary electronic material. These documents were all created using Sweave, a utility for reproducible research with R, which is described in more detail in Sweave as a tool for reproducible research section. Note that such a 'forensic' analysis may be the starting point of a scientific discussion: Baggerly *et al.* [8] complain about lack of reproducibility of the results in Dressman *et al.* [9] backed up again by supplementary electronic material. Carey and Stodden [10] do forensic on the forensic and show that the truth is somewhere in the middle, arguments supported by package `dressCheck` are available from http://www.bioconductor.org.

So, in principle, the same issues as discussed above arise here: (i) Data need to be publically available for reinspection and (ii) the complete source code of the analysis is the only valid reference when it comes to replication of a specific analysis. One has to note, however, that in its original meaning, reproducibility refers to the whole experiment, i.e. all steps performed in the wet lab and in silico have to be reproducible by independent researchers. In the following, we will only focus on the in silico part of reproducibility, i.e. our own area of responsibility.

These examples above make a very strong case that in quantitative research, the scientific method, i.e. the data-driven falsification of hypotheses, requires that qualified independent researchers gain access to data and are able to reproduce the analyzes. Acknowledging this fact, some journals started to implement policies and facilities to provide data and code along with published manuscripts. In the biostatistics community, *Biostatistics* is one of these journals. Keiding [11], in a letter to the editors of *Biostatistics*, raised concerns against an unreflected use of data and code published along with a specific manuscript. Especially a reanalysis that is more than a replication of the original analysis on a different computer requires subject-matter knowledge. We fully agree that there is danger that people start analyzing data in ways that are not consistent with the experiment this data comes from and, consequently, publishing misleading results. Of course, something like this might happen. However, the peer-review process of journals should prevent such papers from being published, and, if it fails to do so, readers can point out the shortcomings. It is even easier to spot questionable points in such analyzes since, by definition, the data are available. The 'boring' number crunching, i.e. a pure replication of an analysis presented in a paper, might turn out everything but boring as the stories sketched above indicate.

There is no reason why we should not try to implement high standards in reproducibility in our own disciplines, Bioinformatics and Biostatistics. The main focus of the present article is to shed some light on the status quo of reproducibility in Bioinformatics and Biostatistics. We base our investigation on a survey of research papers published in *Bioinformatics* and present a series of short case studies.

REPRODUCIBLE RESEARCH IN BIOINFORMATICS AND BIOSTATISTICS

Background

Most papers published in leading journals in the field, such as *Bioinformatics*, *Biostatistics* and others, present new methodology, and most authors provide descriptions of experiments and data that are analyzed using these new methods or mathematical proofs demonstrating correctness or optimality in some sense. These theoretical contributions are carefully checked by referees and associate editors (at least that is what we all expect). This thorough review process ensures the high quality of accepted papers. Theoretical results are often accomplished by empirical investigations. The majority of authors choose to analyze 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 summarizing results in figures and tables, and for the statistical analysis of real data. While referees try to do their best to detect inconsistencies in the theoretical parts of a paper, 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.

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, a paper 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 [12] 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.

This principle has been defined by Buckheit and Donoho [13] and is elaborated by Schwab *et al.* [14]. Translated into bioinformatical or 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 paper. In this sense, a piece of reproducible research is a paper that provides readers with all the material that is needed to produce the same results as described in the publication. It grants power to readers to look at the reported findings in the light of the 'raw' data and analysis and, potentially, to come to other interpretations or conclusions. These ideas have been widely adopted in statistics in the meantime, see for example Leisch and Rossini [15] or Gentleman and Temple Lang [16].

One hundred bioinformatics papers

Hothorn *et al.* [17] report on the proportions of papers published in *Biometrical Journal* that are, at least in principle, reproducible in the sense that the authors give readers access to data or computer code the published results have been derived from. Out of 56 regular papers published in Volume 50 of *BiometricalJournal*, 48 presented a data-driven example but only in 17 cases these data were accessible for readers. Only 8 papers offered computer code.

We expect things to be better in the field of Bioinformatics, because especially for proteomic and genomic data many journals have adopted a 'data must be published' policy. For example, http://www.mged.org/Workgroups/MIAME/journals.html gives a non-exhaustive list of journals requiring MIAME compliant data as a condition for publishing microarray based papers, including *Bioinformatics*, *Cell* and all *Nature* journals, to name a few.

In order to quantify the state of affairs in *Bioinformatics*, we drew a random sample of 100 papers from Numbers 1 to 7 of Volume 26 (corresponding to a total of 209 published papers). These 100 papers were evaluated according to the following criteria:

Manuscript type?: The type of the manuscript (original, discovery or application).

Data used?: Do the authors report results based on a quantitative analysis (yes, no)?

Data available?: Is this data accessible in some way (yes, no, upon request)?

Simulation shown?: Do the authors report results obtained from simulation experiments (yes, no)?

Simulation code available?: Is the computer source code of these simulations available (yes, no, upon request)?

Software described?: Was the software used to perform quantitative analyzes or simulation experiments described, i.e. its name given (yes, no)?

Versions described?: Was the version of the software used to perform quantitative analyzes or simulation experiments described (yes, no)?

Code available?: Was the source code of the software used to perform quantitative analyzes or simulation experiments available (yes, no)?

Five papers contained statements like data and/or code being 'available upon request'. We contacted the authors asking to share their data or code. All contacted researchers responded within a week and sent material.

Our sample consists of 50 original papers, 48 applications notes and 2 discovery notes. For the sake of simplicity, the 2 discovery notes are treated as original papers in the following.

The majority of original papers reported results based on some form of quantitative analysis involving data. A considerable amount of applications notes contained at least a data-based example. The distribution is shown in Figure 1A. Roughly half of the original papers dealing with data mentioned that these data are available to readers. Surprisingly, this proportion is even higher of applications notes; see Figure 1B. The culture of sharing data seems to be considerably larger in *Bioinformatics* compared with *Biometrical Journal*. One reason might be that many papers in *Biometrical Journal* discuss methodology for randomized clinical trials where data often is considered confidential by sponsors of such trials.

Roughly one third of the original papers present results of simulation studies. For application notes, this is rather unusual as expected, see Figure 1C, since application notes are restricted to two pages. However, the source code of these simulations is hardly ever accessible for inspection and evaluation by readers. Since simulation models are often quite complex and the verbal description of these models in published papers rather tense, it would be very hard for independent researchers to reproduce such results.

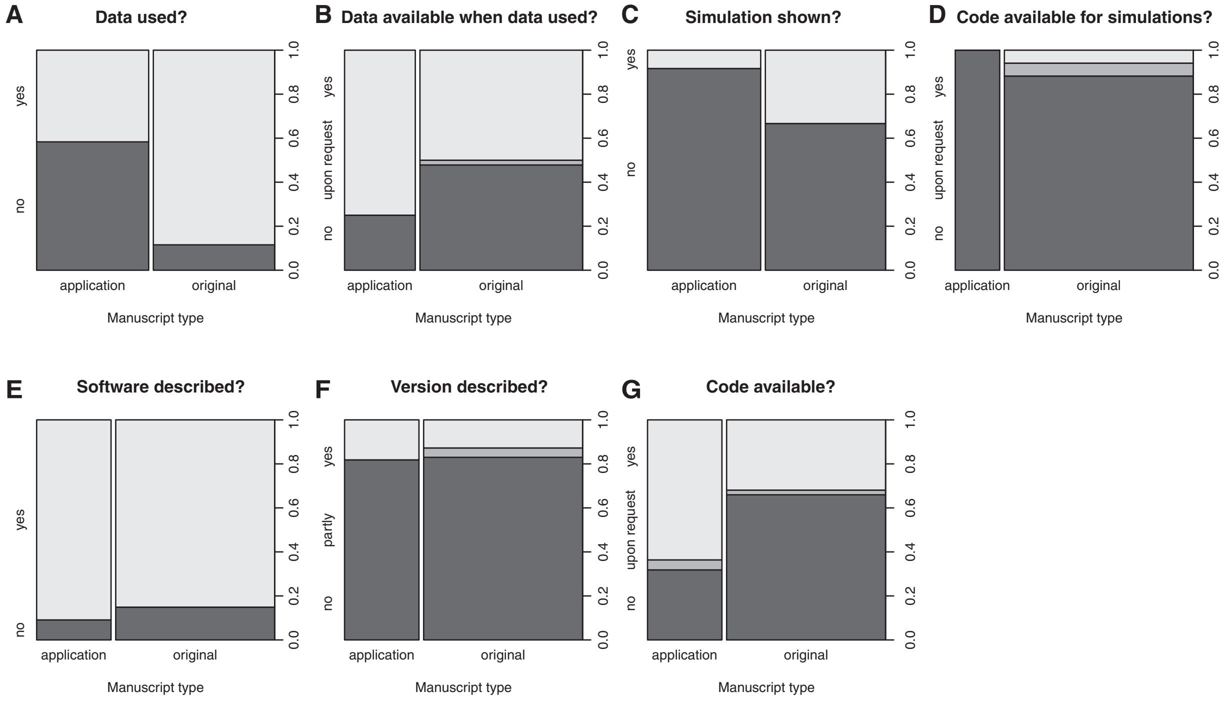

Figure 1: (**A**) Proportions of papers presenting results of quantitative analyzes. (**B**) Proportions of papers giving access to data when presenting results of quantitative analyzes. (**C**) Proportions of papers with simulation experiments. (**D**) Proportions of papers with simulation experiments that give access to the simulation code. (**E**) Proportions of papers describing software used to produce results. (**F**) Proportions of papers describing version of software used to produce results. (**G**) Proportions of papers giving access to the source code of software used to produce results.

We now turn our attention to the 69 papers that presented quantitative results, either based on data or simulation experiments. The vast majority of these papers describe the software, i.e. mention what software package or programming environment was used to produce the results (Figure 1E). However, most authors fail to also report version information (Figure 1F). Thus, it might become hard to reproduce certain results in a few years when, potentially, software will have undergone major changes.

Finally, and most important, the fraction of papers published by *Bioinformatics* that give readers access to the source code of software is roughly one third for original papers and two thirds for applications notes (Figure 1G). Since applications notes often advertise software, this result is not very surprising. Compared to the figures of *Biometrical Journal*, the number of papers that are, at least in principle, reproducible is rather high.

CASE STUDIES

We only evaluated the 100 *Bioinformatics* papers with respect to whether data and/or software are actually available for further inspection and analysis. However, the lesson learned from the two cases sketched in the introduction, numerical reproducibility is not the end but the beginning of, potentially, formulating a productive criticism. To make things more concrete, we take a closer look at six papers and comment on challenges that the reader has to face when trying to reproduce the reported findings.

Summer school biometric society

From 10 November to 12 November 2008 we organized a workshop on 'Reproducible Research & Software Validation' (http://www.statistik.lmu.de/RFSV08/, web page in German) within the annual series of summer schools of the German region of the Biometric Society. The lectures and practicals were given by Soren Hojsgaard (University of Aarhus), Anthony Rossini (Novartis) and the authors of this article. The 34 participants were mostly PhD students and young postdocs with a degree in statistics or a related field, working either in academia or industry.

In a hands-on session all participants were split into groups of 2–4 people, and each group was given two papers:

Paper set A: Heiden *et al.* [18] and Bailer *et al.* [19] contain empirical studies as they are typically published in medical journals. The statistical analyzes are shortly described in a few sentences in the papers, no code or data are available to the public.

Paper set B: Baggerly *et al.* [20] and Zuyderduyn [21], both from *BMC Bionformatics*, use (mixtures of) generalized linear models for SAGE data. Data sets and R code for the analyzes are available as supplementary electronic material, the R code is also included in the appendix of the printed version.

The statistical analyzes of set A were done by Friedrich Leisch, hence we could easily provide data to the participants. The statistical methodology used is rather standard (Fisher's exact test, linear mixed effects models) and should be familiar to everybody with a degree in statistics. The second paper in set B use finite mixtures of Poisson regression models, which is computationally more 'exotic' and not as widely known as the models in set A. It cites and builds upon the first paper in set B and uses R package `flexmix` [22] for EM estimation of mixture models.

All groups were asked to reproduce the figures and tables they found in their papers on their own laptops. Not necessarily every single *P*-value, but at least up to the point where they felt confident that they could reproduce the rest of the paper. What the groups did not know is that the first paper in each set is rather trivial to reproduce (to get spirits up), and the second almost impossible within the given time of approximately 1 h for the group work. The main goal of the session was not to learn how to reproduce results, but to experience what makes reproduction easy or hard at first hand, and to discuss these issues afterwards in the common plenum of all groups.

Heiden et al. *[18]*

This article contains a typical case-control study, here with patients suffering from bipolar disorders and healthy controls. The research question was to investigate differences in genotypes and numbers of alleles

Table 1: Part of Table 2 from Heiden *et al.* [18]

Gene	Genotype	Cases (%)	Controls (%)	*p*
DRD2	B1/B2	4 (4)	6 (8)	
	B2/B2	98 (96)	73 (92)	0.336
DRD3	A1/A1	50 (49)	35 (44)	
	A1/A2	40 (39)	39 (49)	
	A2/A2	12 (12)	5 (7)	0.278

Table 2: Part of Table 2 from Bailer *et al.* [19]

	Effect	Value	SE	P
Binge eating	Group	1.29	5.52	.82
	Time	−8.90	1.97	‹.001
	Group × Time	3.08	2.84	.28
Vomiting	Group	8.21	6.04	.18
	Time	−7.20	1.97	‹.001
	Group × Time	−0.78	2.84	.78

for five pre-selected candidate genes between cases and controls. At first the workshop participants were puzzled that they were not given any data, but of course soon realized that the contingency tables like Table 1 printed in the paper are sufficient statistics for the rest.

Reproduction of results is hence very easy:

```
> DRD2 <- matrix(c(4, 98, 6, 73), ncol=2)
> fisher.test(DRD2)$p.value

[1] 0.3360815

> DRD3 <- matrix(c(50, 40, 12, 35, 39, 5),
  ncol=2)
> fisher.test(DRD3)$p.value

[1] 0.2779254
```

Bailer et al. *[19]*

This article is also a case-control study from the Medical University of Vienna. A guided self-help therapy was compared to cognitive-behavioral group therapy for the treatment of bulimia nervosa. Data about patients was collected at five different points in time: 1 is first contact, 2 start of treatment, 3 mid treatment (four weeks after 2), 4 end of treatment (four weeks after 3) and 5 a follow-up between half a year and 1 year later. Reproducing all results in the paper would have been too overwhelming, hence workshop participants were asked to only reproduce the last three columns of Table 2 from the article, see Table 2.

The corresponding analysis is described as follows in the paper:

> "... a mixed-effects linear regression analysis was performed (e.g., Pinheiro & Bates, 2000). We used random effects for intercept and time, fitting separate regression lines for each patient. We used group (i.e., guided self-help, CBT) as fixed effects and an interaction term for group and time. The fixed effects were evaluated for differences between the groups using standard t tests for the coefficients of a linear regression model. All computations were performed using the statistical computing environment R version 1.5.0 (http://www.R-project.org). ..."

The complete data set is rather large, workshop participants were given a subset with only the variables relevant for the task. As practitioners often do, the data were recorded by our cooperation partners from psychiatry in what is sometimes referred to as 'wide format', with one line per patient and several columns for the different points in time.

The README for the CSV file passed to the participants contained the following information:

GROUP: group therapy or self-help
BE4WBx: binge eating 4 weeks before time point x
VO4WBx: vomiting 4 weeks before time point x
LAX4WBx: use of laxatives 4 weeks before time point x
MEAL4WBx: meals 4 weeks before time point x
BMIx: body mass index at time point x

and the data file itself contains columns labelled BE4WB1, BE4WB2, BE4WB3, BE4WB4, BE4WB5, VO4WB1, etc.

Several things have to be figured out before a mixed effects model can be fitted to the data:

- Only time points 2, 3 and 4 (start, mid and end of treatment) are relevant for Table 2, this information is of course in the paper, but not in the statistical analysis section.
- By default R sorts factor levels alphabetically, and uses the first category as reference.
- The data have to be reshaped from wide to long format.

Most groups working on the task figured out the corresponding R code after some time:

```
> BULIMIA <- read.csv("bulimia.csv")
> BULIMIA$GROUP <- factor(BULIMIA$GROUP,
+                         levels=c("self-help", "group"))
> BE <- paste("BE4WV", 2:4, sep="")
> BE

[1] "BE4WV2" "BE4WV3" "BE4WV4"

> BEDATA <- reshape(BULIMIA[,c(BE,"GROUP")], varying=list(BE),
+                   direction="long", times=2:4,  v.names="y")
> summary(BEDATA)

       GROUP          time          y                id
 self-help:120   Min.   :2   Min.   :  0.00   Min.   : 1
 group    :123   1st Qu.:2   1st Qu.:  3.00   1st Qu.:21
                 Median :3   Median : 12.00   Median :41
                 Mean   :3   Mean   : 18.79   Mean   :41
                 3rd Qu.:4   3rd Qu.: 24.25   3rd Qu.:61
                 Max.   :4   Max.   :110.00   Max.   :81
                             NA's   : 59.00
```

The other groups were given this information in the middle of the session such that they could also try their luck with the mixed effects models.

If one now fits a mixed effects model to the data

```
> library("nlme")
> lme1 <- lme(y~time*GROUP, random=~time|id, data=BEDATA,
+             na.action="na.omit")
> round(summary(lme1)$tTable, 2)

                Value Std.Error  DF t-value p-value
(Intercept)     43.27      7.29 102    5.94    0.00
time            -8.90      1.97 102   -4.51    0.00
GROUPgroup      -4.87     10.37  78   -0.47    0.64
time:GROUPgroup  3.08      2.84 102    1.08    0.28
```

we see that the time and interaction effects are identical to Table 2, but the effect, standard error and *P*-value for group are completely different.

None of the participants figured out what was going on: Naming the three relevant time points '2', '3' and '4' is completely arbitrary, the only important thing is that they are equidistant (4 weeks difference in this case). So start of treatment was used as baseline for the model and set to zero:

```
> BEDATA$time <- BEDATA$time - 2
> lme2 <- lme(y~time*GROUP, random=~time|id, data=BEDATA,
+             na.action="na.omit")
> round(summary(lme2)$tTable, 2)

                Value Std.Error  DF t-value p-value
(Intercept)     25.47      3.90 102    6.53    0.00
time            -8.90      1.97 102   -4.51    0.00
GROUPgroup       1.29      5.52  78    0.23    0.82
time:GROUPgroup  3.08      2.84 102    1.08    0.28
```

Now we have reproduced the upper part of Table 2. Chances would probably have been better if intercepts would have been listed in Table 2, but in the original paper the table is already very large and they had to be omitted for space reasons. It is not intuitive that time has to be changed to 'repair' the group effect, although it is obvious afterwards that shifting time does change the intercept, not the slope of the regression line. Such small details are virtually impossible to infer from textual descriptions of statistical analyzes, only the complete code makes the analyzes truly reproducible with sensible working effort.

Note that this paper is not a particularly bad example or that the publication standards of the respective journal are lower than average for a typical medical journal. Thousands of articles are published each year which suffer from exactly the same problems. In fact, we have even experienced that reviewers recommended to reduce the amount of space given to 'technical details' in order to save space.

Baggerly et al. *[20]*

The R code for this paper is in the appendix of the paper and self-contained, data sets are recreated using R statements like

```
y <- c(0, 1, 1, 15)
```

All workshop participants agreed that reproducing results is a trivial copy and paste exercise. The only criticism towards the publisher is that they inserted a blank line after each statement, which makes the code very hard to read.

Zuyderduyn [21]

The online supplementary material for this paper available on the *BMC Bioinformatics* web page includes an Excel spreadsheet and an R code text file which defines function `pmm.confidence()`. The appendix of the paper contains three blocks of R code which can be extracted easily from the HTML version of the paper. Again each line of code or comment is followed by a blank line, making the code very hard to read (and convince us that the empty lines are inserted by the publisher, not the original author).

The code does not even parse as it is in the paper because it contains syntax errors like

```
fits [k]] <- fit
```

where opening and closing brackets do not match. This may well be caused when the publisher reformatted the code, we have experienced situations when publishers changed R statements like `x<-3` (assign three to variable `x`) to `x < - 3` (compare variable `x` with the value minus three) because 'it looks better' and were hard to convince to undo the change.

Fixing the syntax errors is easy enough, all workshop participants managed to do that in a small amount of time. Puzzlement was large when actually executing the code, because nobody could figure out how the code in the appendix exactly relates to the main text of the paper. It does things related to the paper, but we could not figure out exactly what or how to use the code examples to reproduce anything. That may well be our fault, but if more than 15 people with a Master or PhD in statistics (including the author of the R package used by the code) in several independent groups cannot easily find the link, then there are at least some gaps in documentation.

Lessons learned from this experiment is that mere availability of data and code for a paper means next to nothing. The code need not even parse, because obviously nobody tried to run the version that got published. It would also be great if the code produced actually the results in the paper, and not just something similar.

Benchmarking papers

For the 100 *Bioinformatics* papers we evaluated in 'One hundred bioinformatics papers' section the spectrum of reproducibility can be characterized by two publications: one being fully reproducible and one leaving the reader without any clue. Both compare different methods with respect to some quality criteria in a benchmarking study and can, therefore, be compared with each other.

Hanczar et al. *(23)*

Hanczar *et al.* [23] investigate the small-sample performance of estimates in receiver operator characteristics. The authors describe the theoretical constructs of interest and introduce a simulation model. The simulation setup is also described. Only very briefly are the classifiers introduced (linear discriminant analysis, support vector machines and radial basis function support vector machine) before the attention shifts to the results. The accompanying web page lists additional results. There is no hope to reproduce the findings reported by the authors because (i) the description of the simulation model is insufficient (it is unclear how 'irrelevant features' are distributed), (ii) a lack of information how the classifiers were tuned, and, most important, (iii) which software was used for fitting the classifiers. One might expect various implementations of linear discriminant analysis to perform similar but this is not the case for more complex methods, especially when the model is sensitive to the choice of hyperparameters. Allowing users to access the source code of this simulation experiment would be an appropriate way to solve these issues.

Kirchner et al. *(24)*

Kirchner *et al.* [24] introduce a random forest and discrete mapping approach to the analysis of mass spectrometry data. The methods are evaluated and compared based on results obtained from analyzes of two proteomics data sets. The interested reader is referred to a web page offering access to the data and the R source code along with the necessary information needed to reperform the analyzes. This electronic material makes this paper fully reproducible. The future will show how long this will be the case. The computing infrastructure changes constantly (two minor releases of R per year and, potentially, a larger number of updates in the `randomForest` package). However, the source code of the corresponding versions is archived at the Comprehensive R Archive Network (http://CRAN.R-project.org) and allows interested readers to go back to the versions that were used for the original analysis for the time being.

SWEAVE AS A TOOL FOR REPRODUCIBLE RESEARCH

As we have seen in the examples in the introduction and the case studies above, availability of valid code for analyzes is crucial for reproduction of results. Textual descriptions are simply not sufficient. To assist authors to to be able automatically provide code for each version of a manuscript, it helps a lot if manuscripts and analysis code are tightly bundled. In this section we describe as an example one possible solution that connects our favorite data analysis and text processing systems, R and LATEX. It is also the tool used by Baggerly and Coombes [7] to make their 'forensic' research reproducible.

The traditional way of writing a report as part of a statistical data analysis project uses two separate steps: First, the data are analyzed, and afterwards the results of the analysis (numbers, graphs, ...) are used as the basis for a written report. In larger projects the two steps may be repeated alternately, but the basic procedure remains the same. Statistical software supports this in a number of ways: graphs can be saved as PDF or WMF which in turn can be included in LATEX or Word documents, similarly for tables. The basic paradigm is to write the report around the results of the analysis, and often makes it hard to exactly reproduce results even for the original author as time goes by.

The purpose of Sweave ([25]; part of every R installation) is to create reports which can be updated automatically if data or analysis change. Instead of inserting a prefabricated graph or table into the report, the master document contains the *R code* necessary to obtain it. When run through R, all data analysis output (tables, graphs, ...) is created on the fly and inserted into a final LATEX document. The report can be automatically updated if data or analysis change, which allows for truly reproducible research.

Sweave source files are regular noweb files [26] with some additional syntax that allows control over the final output. Noweb is a simple literate programming tool which allows to combine program source code and the corresponding documentation into a single file. These consist of a sequence of code and documentation segments, called *chunks*. Different command line programs are used to extract the code ('*tangle*') or typeset documentation together with the code ('*weave*').

A small Sweave example is shown in Figure 2, which contains one code chunks embedded in simple LATEX markup. '<<...>>=' at the beginning of a line marks the start of a code chunk, while a '@' at the beginning of a line marks the start of a documentation chunk. Sweave translates this into a regular LATEX document, which in turn can be compiled by `latex`, see Figure 3.

All code chunks found in an Sweave document are evaluated by R in the order they appear in the document. Within the double angle brackets one can specify options that control how the code and the corresponding output are rendered in the final document. The example shown uses option `keep.source=TRUE` such that line breaks and comments of the code are kept. Other options allow to suppress showing input and/or output, or declare

```
Most groups working on the task figured out the
corresponding R code after some time:
<<keep.source=TRUE>>=
BULIMIA <- read.csv("bulimia.csv")
BULIMIA$GROUP <- factor(BULIMIA$GROUP,
                        levels=c("self-help", "group"))
BE <- paste("BE4WV", 2:4, sep="")
BE

BEDATA <- reshape(BULIMIA[,c(BE,"GROUP")], varying=list(BE),
                  direction="long", times=2:4,  v.names="y")
summary(BEDATA)
@
The other groups were given this information in
the middle of the session such that they could also
try their luck with the mixed effects models.
```

Figure 2: Parts of the Sweave file generating 'Bailer *et al.* [19]' section.

```
Most groups working on the task figured out the
corresponding R code after some time:
\begin{Schunk}
\begin{Sinput}
> BULIMIA <- read.csv("bulimia.csv")
> BULIMIA$GROUP <- factor(BULIMIA$GROUP,
+                         levels=c("self-help", "group"))
> BE <- paste("BE4WV", 2:4, sep="")
> BE
\end{Sinput}
\begin{Soutput}
[1] "BE4WV2" "BE4WV3" "BE4WV4"
\end{Soutput}
\begin{Sinput}
> BEDATA <- reshape(BULIMIA[,c(BE,"GROUP")], varying=list(BE),
+                   direction="long", times=2:4,  v.names="y")
> summary(BEDATA)
\end{Sinput}
\begin{Soutput}
       GROUP          time         y                id
 self-help:120   Min.   :2   Min.   :  0.00   Min.   : 1
 group    :123   1st Qu.:2   1st Qu.:  3.00   1st Qu.:21
                 Median :3   Median : 12.00   Median :41
                 Mean   :3   Mean   : 18.79   Mean   :41
                 3rd Qu.:4   3rd Qu.: 24.25   3rd Qu.:61
                 Max.   :4   Max.   :110.00   Max.   :81
                             NA's   : 59.00
\end{Soutput}
\end{Schunk}
The other groups were given this information in
the middle of the session such that they could also
try their luck with the mixed effects models.
```

Figure 3: Output when running Sweave on the code from Figure 2.

that the code generates a figure and we want to include the figure instead of the textual output.

Instead of executing the code and generating a manuscript one can also extract only the code, e.g. to use it as online supplementary material for a manuscript. Appendix 1 is automatically generated using the code in Figure 4. We never need to check wether the code is actually in sync with the analysis, because *it is the original code that does the analysis*, not a copy. Including this appendix is of course redundant because we have shown the code as part of the main text anyway. Setting option `echo=FALSE`

```
\section{Code for Section~\ref{sec:hoehenried}}
\label{sec:appcode}

<<echo=FALSE,results=hide>>=
Stangle("exp-hoehenried.Rnw")
@

\lstinputlisting{exp-hoehenried.R}
```

Figure 4: Automatic generation of code appendix.

on all code chunks would have suppressed the R commands in the main text for readers not familiar with R, like medical doctors. Similarly, Appendix 2 is also automatically generated using R function `sessionInfo()`, which is also recommended by Baggerly and Coombes [7].

The implementation of Sweave is very flexible, and several packages on CRAN extend it to other typesetting systems like OpenOffice or HTML. Giving a full survey of all possible combinations or solutions for other statistical software than R is beyond the scope of this article.

CONCLUSIONS

The discussion of results and interpretations drawn from data or simulation experiments is an integral part of scientific progress. Without data and source code for simulations being available, this process is rendered difficult or even impossible. The hockey stick controversy, for example, led to a better insight into climate data and the development of more appropriate statistical methods for analyzing such data, i.e. to a scientific progress which would not have been possible without access to the data. Access to data is therefore, except for very rare instances where data confidentially must be given more weight, mandatory. In the same spirit, a simulation study based on closed source code that can't be assessed and reperformed by other researchers gives only little evidence to a finding.

However, the fact that source code for analyzes or simulation experiments is available does not automatically lead to reproducibility. Someone has to actually run the code. On of us (T.H.) serves as 'Reproducible Research Editor' for *Biometrical Journal* since 2008 and checks source code submitted along with manuscripts on a regular basis. It rarely happens that code for analyzes or simulation experiments runs out-of-the-box. In the majority of cases, authors are asked to make corrections. Checking code, if available at submission time, is an additional burden on reviewers, but actually less work than checking a mathematical proof: If the code runs, it is a copy and paste exercise, if not, one has to go back to the authors. Checking that the code makes sense is of course a different question that one should be able to address also after the manuscript was published.

Submission of Sweave files or other reproducible document formats is not really necessary, in most cases code, data and detailed description of software environment are sufficient. However, in that case somebody has to manually check that the code actually produces the results in the paper. With reproducible documents this can be partly automated, reducing the burden on the reviewers. An open question is the usage of proprietary software: It should be easy enough to find reviewers with access to, e.g. MATLAB, SAS or SPSS. But what about more 'exotic' closed source software? Is closed source software even eligible for real reproducibility? When the source code of a specific software environment is not available for inspection by readers parts of the analysis (the choices made during implementation by a software vendor) cannot be inspected for correctness or their impact on the scientific results.

Suppose we live in the best of all worlds and every journal required statistical analyzes to be reproducible. We still lack a real definition what 'reproducible research' really means: Is it sufficient that reviewers and/or editors could reproduce at the time of publication? The whole community for at least one year? Software environments are moving targets and change all the time, but on the other hand many numerical FORTRAN routines written in the 1960s and 1970s are still in use today. So analysis code published now may be useful for much longer than we anticipate.

But some code will fail as time goes on. Cloud computing and virtual machines use images with complete operating systems and software applications. It would be a valuable service to the scientific community to take snapshot of these at regular time points and offer to run, e.g. 'R 1.0.0 as of February 29, 2000' on the web. Of course there is a chicken and egg problem involved, because the virtual machine running the images may change or cease to exist.

Future generations of scientists may be interested in how exactly we analyzed and drew conclusions from our data. Therefore, we should aim at documenting our scientific work as good as we can. An 'open access' archive for data and software is only the first step.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Reproducibiliy of numerical results requires publication of data and computer programs along with the classical scientific paper.
- Reproducibility issues recently receive a lot of attention in quantitative disciplines.
- The status quo of reproducibility in Bioinformatics and Biostatistics is investigated by means of a survey of papers published in *Bioinformatics* and case-studies.

References

1. Pollet TV, Nettle D. Partner wealth predicts self-reported orgasm frequency in a sample of chinese women. *Evol Hum Behav* 2009;**30**:146–51.
2. Herberich E, Hothorn T, Nettle D, *et al*. A re-evaluation of the statistical model in Pollet and Nettle 2009. *Evol Hum Behav* 2010;**31**(2):150–51.
3. Mann ME, Bradley RS, Hughes MK. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* 1998;**392**:779–87.
4. McIntyre S, McKitrick R. Corrections to the Mann et. al. (1998) proxy data base and northern hemispheric average temperature series. *Energy & Environment* 2003;**14**(6): 751–71.
5. McIntyre S, McKitrick R. The M&M critique of the MBH98 Northern Hemisphere climate index: Update and implications. *Energy Environ* 2005;**16**(1):69–100.
6. Ioannidis JPA, Allison DB, Ball CA, *et al*. Repeatability of published microarray gene expression analyses. *Nat Genet* 2009;**41**(2):149–55.
7. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann Appl Stat* 2009; **3**(4):1309–34.
8. Baggerly KA, Coombes KR, Neeley ES. Run Batch Effects Potentially Compromise the Usefulness of Genomic Signatures for Ovarian Cancer. *J Clin Oncol* 2008;**26**(7): 1186–87.
9. Dressman HK, Berchuck A, Chan G, *et al*. An Integrated Genomic-Based Approach to Individualized Treatment of Patients With Advanced-Stage Ovarian Cancer. *J Clin Oncol* 2007;**25**(5):517–25.
10. Carey VJ, Stodden V. Reproducible research concepts and tools for cancer bioinformatics. In: Ochs MF, Casagrande JT, Davuluri RV, (eds). *Biomedical Informatics for Cancer Research*. New York: Springer, 2010;149–75.
11. Keiding N. Reproducible research and the substantive context. *Biostatistics* 2010;**11**(3):376–78. With discussion.
12. de Leeuw J. Reproducible research: The bottom line. Technical Report 2001031101, Department of Statistics Papers, University of California, Los Angeles, 2001. http://repositories.cdlib.org/uclastat/papers/2001031101 (29 August 2010, date last accessed).
13. Buckheit JB, Donoho DL. Wavelab and reproducible research. In: Antoniadis A, (ed). *Wavelets and Statistics*. New York: Springer, 1995.
14. Schwab M, Karrenbach M, Claerbout J. Making scientific computations reproducible. *Comput Sci Eng* 2000;**2**: 61–67.
15. Leisch F, Rossini AJ. Reproducible statistical research. *Chance* 2003;**16**(2):46–50.
16. Gentleman R, Temple Lang D. Statistical analyses and reproducible research. *J Comput Graph Stat* 2007;**16**(1):1–23.
17. Hothorn T, Held L, Friede T. Biometrical journal and reproducible research. *Biometrical Journal* 2009;**51**(4):553–5.
18. Heiden A, Schüssler P, Itzlinger U, *et al*. Association studies of candidate genes in bipolar disorders. *Neuropsychobiology* 2000;**42**(suppl 1):18–21.
19. Bailer U, de Zwaan M, Leisch F, *et al*. Guided self-help versus cognitive-behavioral group therapy in the treatment of bulimia nervosa. *J Eat Disorders* 2004;**35**(4):522–37.
20. Baggerly K, Deng L, Morris J, *et al*. Overdispersed logistic regression for SAGE: Modelling multiple groups and covariates. *BMC Bioinformatics* 2004;**5**(1):144.
21. Zuyderduyn S. Statistical analysis and significance testing of serial analysis of gene expression data using a poisson mixture model. *BMC Bioinformatics* 2007;**8**(1):282.
22. Leisch F. FlexMix: A general framework for finite mixture models and latent class regression in R. *J Stat Softw* 2004; **11**(8):1–18.
23. Hanczar B, Hua J, Sima C, *et al*. Small-sample precision of roc-related estimates. *Bioinformatics* 2010;**26**(6):822–30.
24. Kirchner M, Timm W, Fong P, *et al*. Non-linear classification for on-the-fly fractional mass filtering and targeted precursor fragmentation in mass spectrometry experiments. *Bioinformatics* 2010;**26**(6):791–97.
25. Leisch F. Sweave: Dynamic generation of statistical reports using literate data analysis. In: Härdle W, Rönz B, (eds). *Compstat 2002—Proceedings in Computational Statistics*. Heidelberg: Physica, 2002;575–80.
26. Ramsey N. *Noweb man page*. USA: University of Virginia, 1998.
27. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R foundation for statistical Computing, 2010 (15 October 2010, date last accessed).

APPENDIX 1

Code for ‘summer school biometric society’ section

```
###################################################
### chunk number 1:
###################################################
#line 83 "SweaveInput"
DRD2 <- matrix(c(4, 98, 6, 73), ncol=2)
fisher.test(DRD2)$p.value
DRD3 <- matrix(c(50, 40, 12, 35, 39, 5), ncol=2)
fisher.test(DRD3)$p.value


###################################################
### chunk number 2:
###################################################
#line 166 "SweaveInput"
BULIMIA <- read.csv("bulimia.csv")
BULIMIA$GROUP <- factor(BULIMIA$GROUP,
                        levels=c("self-help", "group"))
BE <- paste("BE4WV", 2:4, sep="")
BE

BEDATA <- reshape(BULIMIA[,c(BE,"GROUP")], varying=list(BE),
                  direction="long", times=2:4,  v.names="y")
summary(BEDATA)


###################################################
### chunk number 3:
###################################################
#line 182 "SweaveInput"
library("nlme")
lme1 <- lme(y~time*GROUP, random=~time|id, data=BEDATA,
            na.action="na.omit")

round(summary(lme1)$tTable, 2)


###################################################
### chunk number 4:
###################################################
#line 198 "SweaveInput"
BEDATA$time <- BEDATA$time - 2
lme2 <- lme(y~time*GROUP, random=~time|id, data=BEDATA,
            na.action="na.omit")

round(summary(lme2)$tTable, 2)
```

APPENDIX 2

Software version information

We used the following version of R (R Development Core Team, 2010) for data analysis:

- R version 2.12.0 (2010-10-15), i686-pc-linux-gnu
- Locale: LC_CTYPE=en_US.utf8, LC_NUMERIC=C, LC_TIME=en_US.utf8, LC_COLLATE=en_US.utf8, LC_MONETARY=C, LC_MESSAGES=en_US.utf8, LC_PAPER=en_US.utf8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.utf8, LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: nlme 3.1-97
- Loaded via a namespace (and not attached): grid 2.12.0, lattice 0.19-13