

Reproducible Research in Knight BSR

Jessica Minnier

April 20, 2016

Why reproducibility?

Claerbout's Principle

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.

(Buckheit and Donoho 1995; De Leeuw 2001)

(Jon F. Claerbout is the Cecil Green Professor Emeritus of Geophysics at Stanford University. He was one of the first scientists to emphasize that computational methods threaten the reproducibility of research unless open access is provided to both the data and the software underlying a publication. (Claerbout and Karrenbach 1992, Wikipedia)

Schwab, Karrenbach, Claerbout:

It takes some effort to organize your research to be reproducible.

We found that although the effort seems to be directed to helping other people stand up on your shoulders, the principal beneficiary is generally the author herself.

This is because time turns each one of us into another person, and by making effort to communicate with strangers, we help ourselves to communicate with our *future selves*.

(Schwab, Karrenbach, and Claerbout 2000)

also from M Shotwell and JM Álvarez' slides "Approaches and Barriers to Reproducible Practices in Biostatistics" and "Barriers to Reproducible Research and a Web-Based Solution" <http://biostatmatt.com/uploads/shotwell-interface-2011.pdf> and <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/MattShotwell/MSRetreat2013Slides.pdf>

How to Make More Published Research True

In J. P. Ioannidis (2014) "How to Make More Published Research True" in PLOS Medicine, the author writes a follow up to J. Ioannidis (2005) "Why most published research findings are false." He suggests reproducibility as one key component to the cause:

"To make more published research true, practices that have improved credibility and efficiency in specific fields may be transplanted to others which would benefit from them—possibilities include

- the adoption of large-scale collaborative research;
- *replication culture*;
- registration; sharing; *reproducibility practices*;
- better statistical methods;
- standardization of definitions and analyses;
- more appropriate (usually more stringent) statistical thresholds; and
- improvement in study design standards, peer review, *reporting and dissemination of research*, and training of the scientific workforce."

What is Reproducible Research?

Research results are replicable if there is sufficient information available for independent researchers to make the same findings using the same procedures.

In computational sciences this means: the data and code used to make a finding are available and they are sufficient for an independent researcher to recreate the finding.

In practice, research needs to be *easy* for independent researchers to reproduce.

– King (1995), Ball and Medeiros (2012), from Gandrud (2013)

Replicability has been a key part of scientific inquiry from perhaps the 1200s. It has even been called the “demarcation between science and non-science.”

– Gandrud (2013) and references therein, including Roger Bacon’s “Opera quaedam hactenus inedita Vol. 1” from 1267 <https://books.google.com/books?id=wMUKAAAAYAAJ>

Peng (2011) in Science: Reproducible Research in Computational Science

“Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.”

“Ultimately, developing a culture of reproducibility in which it currently does not exist will require **time and sustained effort** from the scientific community.”

– Peng (2011)

Peng (2009) in Biostatistics: Reproducible research and Biostatistics (the journal)

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 Associate Editor of Reproducibility 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.

– Peng (2009)

Reproducibility in Statistics

“Reproducibility is important because it is the only thing that an investigator can guarantee about a study.”
“a study can be reproducible and still be wrong”

“These days, with the complexity of data analysis and the subtlety of many claims (particularly about complex diseases), reproducibility is pretty much the only thing we can hope for. Time will tell whether we are ultimately right or wrong about any claims, but reproducibility is something we can know right now.”

“By using the word reproducible, I mean that the original data (and original computer code) can be analyzed (by an independent investigator) to obtain the same results of the original study. In essence, it is the notion that the data analysis can be successfully repeated. Reproducibility is particularly important in large computational studies where the data analysis can often play an outsized role in supporting the ultimate conclusions.”

— Roger Peng’s 2014 blog post on Simply Statistics <http://simplystatistics.org/2014/06/06/the-real-reason-reproducible-research/>

NIH Requirements

NIH requirements (beginning Jan 2016)

“Enhancing Reproducibility through Rigor and Transparency” <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html>

1. *Scientific Premise*

- “describe the general strengths and weaknesses of the prior research being cited by the investigator as crucial to support the application.”
- experimental design/power of prior studies used for hypothesis generation, weaknesses include different populations/species, unblinded, not adjusting for confounders

2. *Rigorous Experimental Design*

3. *Consideration of Sex and Other Relevant Biological Variables*

- “sex is a biological variable that is frequently ignored in animal study designs and analyses”

4. *Authentication of Key Biological and/or Chemical Resources*

5. *Implementation*

NIH Rigor & Reproducibility Resources

Website: <http://grants.nih.gov/reproducibility/index.htm>

FAQs: <http://grants.nih.gov/reproducibility/faqs.htm>

NIH Training Module: https://grants.nih.gov/reproducibility/module_1/presentation.html

Note: Most of this is in regards to the science, design of experiment, chemical and biological methods. Essentially no language describing reproducibility of analyses or data management for data or results generated by the grant.

Journals unite to encourage reproducibility

NIH Principles and Guidelines for Reporting Preclinical Research

<http://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research>

NIH held a joint workshop in June 2014 with the Nature Publishing Group and Science on the issue of reproducibility and rigor of research findings

NIH Principles and Guidelines for Reporting Preclinical Research

- aim of facilitating the interpretation and repetition of experiments as they have been conducted in the published study
- journal should include policies for statistical reporting in information to authors
- no limits or generous limits for methods sections
- should use a *checklist* during editorial processing to ensure the reporting of key methodological and analytical information to reviewers and readers
- Data and material sharing
 - at the minimum, data sets on which the conclusions of the paper rely must be made available upon request (where ethically appropriate) during consideration of the manuscript (by editors and reviewers) and upon reasonable request immediately upon publication.
 - Recommend deposition of data sets in public repositories, where available
 - Encourage presentation of all other data values in machine readable format
 - Encourage sharing of software and require at the minimum a statement in the manuscript describing if software is available and how it can be obtained.

NIH Principles and Guidelines for Reporting Preclinical Research (ctd.)

- if the journal publishes a paper, it assumes responsibility to consider publication of refutations of that paper, according to its usual standards of quality.
- best practice guidelines for image based data and a description of biological material with enough information to uniquely identify the reagents
- These measures and principles do not obviate the need for replication and reproduction in subsequent investigations to establish the robustness of published results across multiple biological systems.

<http://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research>

Checklist: authors required to report

from NIH Guidelines & Landis et al. (2012) “A call for transparent reporting to optimize the predictive value of preclinical research”. Nature 490, 187–191.

- *Standards*: community-based standards (nomenclature etc) where applicable
- *Replicates*: report how often each experiment was performed, whether results were substantiated by repetition under a range of conditions. Sufficient information about sample collection must be provided to distinguish between independent biological data points & technical replicates.
- *Statistics*: Require statistics to be fully reported in the paper, including statistical test used, exact value of N, definition of center, dispersion & precision measures
- *Randomization*: (yes/no) & method, at a minimum for all animal experiments
- *Blinding*: were experimenters blind to group assignment & outcome assessment, at a minimum for all animal experiments.
- *Sample-size (SS) estimation*: was an appropriate SS computed during study design & include method; if no power analysis, how was SS determined?
- *Inclusion and exclusion criteria*: criteria used for exclusion of any data or subjects. Include any similar experimental results that were omitted from reporting for any reason, esp. if results don't support main findings of study; describe any outcomes/conditions that are measured/used & not reported in results section.

Nature series on “Challenges in Irreproducible Research”

Nature has a website containing editorials, features, news, and articles on various topics related to reproducible research: <http://www.nature.com/news/reproducibility-1.17552>

Including

- a checklist for authors of Nature papers: <http://www.nature.com/authors/policies/checklist.pdf> described in the 2013 announcement “Announcement: Reducing our irreproducibility” <http://www.nature.com/news/announcement-reducing-our-irreproducibility-1.12852>
- R Nuzzo (2014) “Scientific method: Statistical errors” on “P values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume.” <http://www.nature.com/news/scientific-method-statistical-errors-1.14700>

Literate Programming

Literate Programming

Literate programming is an approach to programming introduced by Donald Knuth in which a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated. (Knuth 1984)

Examples: Sweave, knitr (for R); SASweave, Statrep (for SAS); StatWeave (for STATA)

This is knitr:

```
summary(cars)
```

```
##      speed      dist
## Min.   : 4.0    Min.   :  2.00
## 1st Qu.:12.0    1st Qu.: 26.00
## Median :15.0    Median : 36.00
## Mean   :15.4    Mean    : 42.98
## 3rd Qu.:19.0    3rd Qu.: 56.00
## Max.   :25.0    Max.    :120.00
```

How this presentation is made: Knitr + RStudio

This is an R Markdown presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

This is a document written in plain text (.Rmd file) with text and R code embedded with the special syntax. Within RStudio when you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Knitr

SASweave, Statrep

Version Control

Version Control

“Version control systems (VCS), which have long been used to maintain code repositories in the software industry, are now finding new applications in science. One such open source VCS, Git, provides a lightweight yet robust framework that is ideal for managing the full suite of research outputs such as datasets, statistical code, figures, lab notes, and manuscripts. For individual researchers, Git provides a powerful way to track and compare versions, retrace errors, explore new approaches in a structured manner, while maintaining a full audit trail. For larger collaborative efforts, Git and Git hosting services make it possible for everyone to work asynchronously and merge their contributions at any time, all the while maintaining a complete authorship trail.”

(Ram 2013)

Why Use Version Control?

Have you ever:

- Made a change to code, realised it was a mistake and wanted to revert back?
- Lost code or had a backup that was too old?
- Had to maintain multiple versions of a product?
- Wanted to see the difference between two (or more) versions of your code?
- Wanted to prove that a particular change broke or fixed a piece of code?
- Wanted to review the history of some code?
- Wanted to submit a change to someone else’s code?
- Wanted to share your code, or let other people work on your code?
- Wanted to see how much work is being done, and where, when and by whom?
- Wanted to experiment with a new feature without interfering with working code?

In these cases, and no doubt others, a version control system should make your life easier.

<http://stackoverflow.com/questions/1408450/why-should-i-use-version-control>

Git

Learn git: <https://try.github.io/levels/1/challenges/1>

Reproducibility in Practice

Peng (2009) in Biostatistics: Reproducible research and Biostatistics (the journal)

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 [...]

– Peng (2009)

Necessary

- File and folder naming conventions on server
 - Allows others to find files
 - Results folder needs to be the same
 - Could have other folders and files but at minimum the full reproducible results need to have consistent names
- All results should be able to be replicated from raw data based on files in a single folder
- Document versions of software used
- Do not type tables in by hand, must be generated from code
- Only alter data programmatically, chain of modification preserved
- Checklist for reproducibility to follow for all projects (including reports, manuscripts)

Ideals

- Literate programming
- Best practices for writing code (i.e. <http://ropensci.github.io/reproducibility-guide/sections/writingCode/>) and “Best Practices for Scientific Computing” (Wilson et al. 2014)
- use html web-based output
 - see Matthew Shotwell’s slides: <http://biostatmatt.com/uploads/shotwell-interface-2011.pdf>
 - nearly universal compatibility
 - persistent
 - images handled more naturally
- use **make** files to rerun analyses when certain files change
- version/revision control systems such as git for all files
- version control of data
- Software/package versions need to be maintained (i.e. packrat for R)

Develop a checklist

To do!

Adapt <http://ropensci.github.io/reproducibility-guide/sections/checklist/>

“Ten Simple Rules for Reproducible Computational Research”

- Rule 1: For Every Result, Keep Track of How It Was Produced
- Rule 2: Avoid Manual Data Manipulation Steps
- Rule 3: Archive the Exact Versions of All External Programs Used
- Rule 4: Version Control All Custom Scripts
- Rule 5: Record All Intermediate Results, When Possible in Standardized Formats
- Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- Rule 7: Always Store Raw Data behind Plots
- Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Rule 9: Connect Textual Statements to Underlying Results
- Rule 10: Provide Public Access to Scripts, Runs, and Results

(Sandve et al. 2013)

Recommended Books

Stodden, Victoria, Friedrich Leisch, and Roger D. Peng, eds. Implementing reproducible research. CRC Press, 2014.

Gandrud, Christopher. Reproducible Research with R and R Studio. CRC Press, 2013.

Xie, Yihui. Dynamic Documents with R and knitr. Vol. 29. CRC Press, 2013.

Resources (classes)

Karl Broman’s class “Tools for Reproducible Research” at UWisconsin-Madison <http://kbroman.org/Tools4RR/>

“Reproducible Research” by Johns Hopkins on Coursera (Peng, Leek, Caffo) <https://www.coursera.org/learn/reproducible-research>

References and Resources (websites)

ROpenSci’s “Reproducibility in Science” guide: <http://ropensci.github.io/reproducibility-guide/> including the reproducibility checklist <http://ropensci.github.io/reproducibility-guide/sections/checklist/>

Matthew Shotwell’s slides (2011) “Approaches and Barriers to Reproducible Practices in Biostatistics”. <http://biostatmatt.com/uploads/shotwell-interface-2011.pdf>

ROpenSci’s blog post “Reproducible research is still a challenge” by R. FitzJohn, M. Pennell, A. Zanne, W. Cornwell, June 9, 2014, describes the experience of running an example analysis: <https://ropensci.org/blog/2014/06/09/reproducibility/>

StackOverflow question “Why should I use version control?” <http://stackoverflow.com/questions/1408450/why-should-i-use-version-control>

- example code and documentation: <https://github.com/richfitz/wood>
- example report using literate programming: <http://richfitz.github.io/wood/wood.html>

References and Resources (websites, more)

Karl Broman’s class “Tools for Reproducible Research” resource page <http://kbroman.org/Tools4RR/pages/resources.html> and “Why Reproducibility is Hard” <https://kbroman.wordpress.com/2015/09/09/reproducibility-is-hard/>

CRAN’s task view on Reproducible Research: <https://cran.r-project.org/web/views/ReproducibleResearch.html>

Frank Harrell’s wiki on statistical reporting: <http://biostat.mc.vanderbilt.edu/wiki/Main/StatReport>

University of Wisconsin-Madison’s Department of Statistics site on “Reproducible Research Tools”: <https://www.stat.wisc.edu/reproducible>

References

Ball, Richard, and Norm Medeiros. 2012. “Teaching Integrity in Empirical Research: A Protocol for Documenting Data Management and Analysis.” *The Journal of Economic Education* 43 (2). Taylor & Francis: 182–89.

Buckheit, Jonathan B, and David L Donoho. 1995. *Wavelab and Reproducible Research*. Springer.

Claerbout, Jon, and Martin Karrenbach. 1992. “Electronic Documents Give Reproducible Research a New Meaning.” In *Proc. 62nd Ann. Int. Meeting of the Soc. of Exploration Geophysics*, 601–4.

De Leeuw, Jan. 2001. “Reproducible Research. the Bottom Line.” *Department of Statistics, UCLA*.

Gandrud, Christopher. 2013. *Reproducible Research with R and R Studio*. CRC Press.

Ioannidis, John PA. 2014. “How to Make More Published Research True.”

Ioannidis, JPA. 2005. “Why Most Published Research Findings Are False.” *PLoS Med* 2 (8): e124.

King, Gary. 1995. “Replication, Replication.” *PS: Political Science & Politics* 28 (03). Cambridge Univ Press: 444–52.

Knuth, Donald Ervin. 1984. “Literate Programming.” *The Computer Journal* 27 (2). Br Computer Soc: 97–111.

Peng, Roger D. 2009. “Reproducible Research and Biostatistics.” *Biostatistics* 10 (3). Biometrika Trust: 405–8.

———. 2011. “Reproducible Research in Computational Science.” *Science (New York, Ny)* 334 (6060). NIH Public Access: 1226.

Ram, Karthik. 2013. “Git Can Facilitate Greater Reproducibility and Increased Transparency in Science.” *Source Code for Biology and Medicine* 8 (1): 7.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. “Ten Simple Rules for Reproducible Computational Research.”

Schwab, Matthias, Martin Karrenbach, and Jon Claerbout. 2000. “Making Scientific Computations Reproducible.” *Computing in Science & Engineering* 2 (6). AIP Publishing: 61–67.

Wilson, Greg, DA Aruliah, C Titus Brown, Neil P Chue Hong, Matt Davis, Richard T Guy, Steven HD Haddock, et al. 2014. “Best Practices for Scientific Computing.” *PLoS Biol* 12 (1). Public Library of Science: e1001745.