Reproducible Research in Statistics

Jessica Minnier – Knight BSR

June 16, 2016

What is Reproducible Research?

Reproducible = Replicable + Transparant

> *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) book "Reproducible Research with R and R Studio" and references therein, including Roger Bacon's "Opera quaedam hactenus inedita Vol. 1" from 1267

What are the different kinds of reproducibile research?

Enabling reproducibility can be complicated, but by separating out some of the levels and degrees of reproducibility the problem can become more manageable because we can focus our efforts on what best suits our specific scientific domain. Victoria Stodden (2014), a prominent scholar on this topic, has identified some useful distinctions in reproducible research:

***Computational reproducibility**: when detailed information is provided about code, software, hardware and implementation details.*

***Empirical reproducibility**: when detailed information is provided about non-computational empirical scientific experiments and observations. In practise this is enabled by making data freely available, as well as details of how the data was collected.*

***Statistical reproducibility**: when detailed information is provided about the choice of statistical tests, model parameters, threshold values, etc. This mostly relates to pre-registration of study design to prevent p-value hacking and other manipulations.*

ROpenSci Reproducibility Guide

Spectrum of Research

Stodden et al. (2013) place computational reproducibility on a spectrum with five categories that account for many typical research contexts:

- *Reviewable research: The descriptions of the research methods can be independently assessed and the results judged credible. (This includes both traditional peer review and community review, and does not necessarily imply reproducibility.)*
- *Replicable research: Tools are made available that would allow one to duplicate the results of the research. . .*
- *Confirmable research: . . . main conclusions of the research can be attained independently without the use of software provided by the author . . .*
- *Auditable research: Sufficient records (including data and software) have been archived . . .*
- *Open or Reproducible research: Auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of*

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 "The Real Reason Reproducible Research is Important" also see Peng (2011) "Reproducible research in computational science"

Early notions of 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.*
>
> *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**.*

- Claerbout and Karrenbach (1992) "Electronic documents give reproducible research a new meaning"
- Buckheit and Donoho (1995) "Wavelab and reproducible research"
- Schwab, Karrenbach, and Claerbout (2000) "Making scientific computations reproducible"

Current Issues and Discussion

How to Make More Published Research True

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."*

Availability of code in peer-reviewed journals

Stodden, Guo, and Ma (2013) "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals"

Table 1. Code Availability in the Journal of the American Statistical Association.

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

doi:10.1371/journal.pone.0067111.t001

Figure 1:

Reproducible research and Biostatistics (the journal)

Authors can choose to meet a subset of these criteria if they wish:

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.*

– Peng (2009) "Reproducible research and *Biostatistics*"

NIH requirements (beginning Jan 2016)

"Enhancing Reproducibility through Rigor and Transparency"

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 and Reproducibility" Policy
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 with NIH to encourage reproducibility

- Principles and Guidelines for 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
- A video/slide presentation about this topic and how it applies to grant applications and peer review can be found here: NIH Policy Rigor for Reviewers Presentation

NIH Principles and Guidelines for Reporting Preclinical Research

Journals should aim to facilitate the interpretation and repetition of experiments as they have been conducted in the published study.

- 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.
- journal assumes responsibility to consider publication of refutations of

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

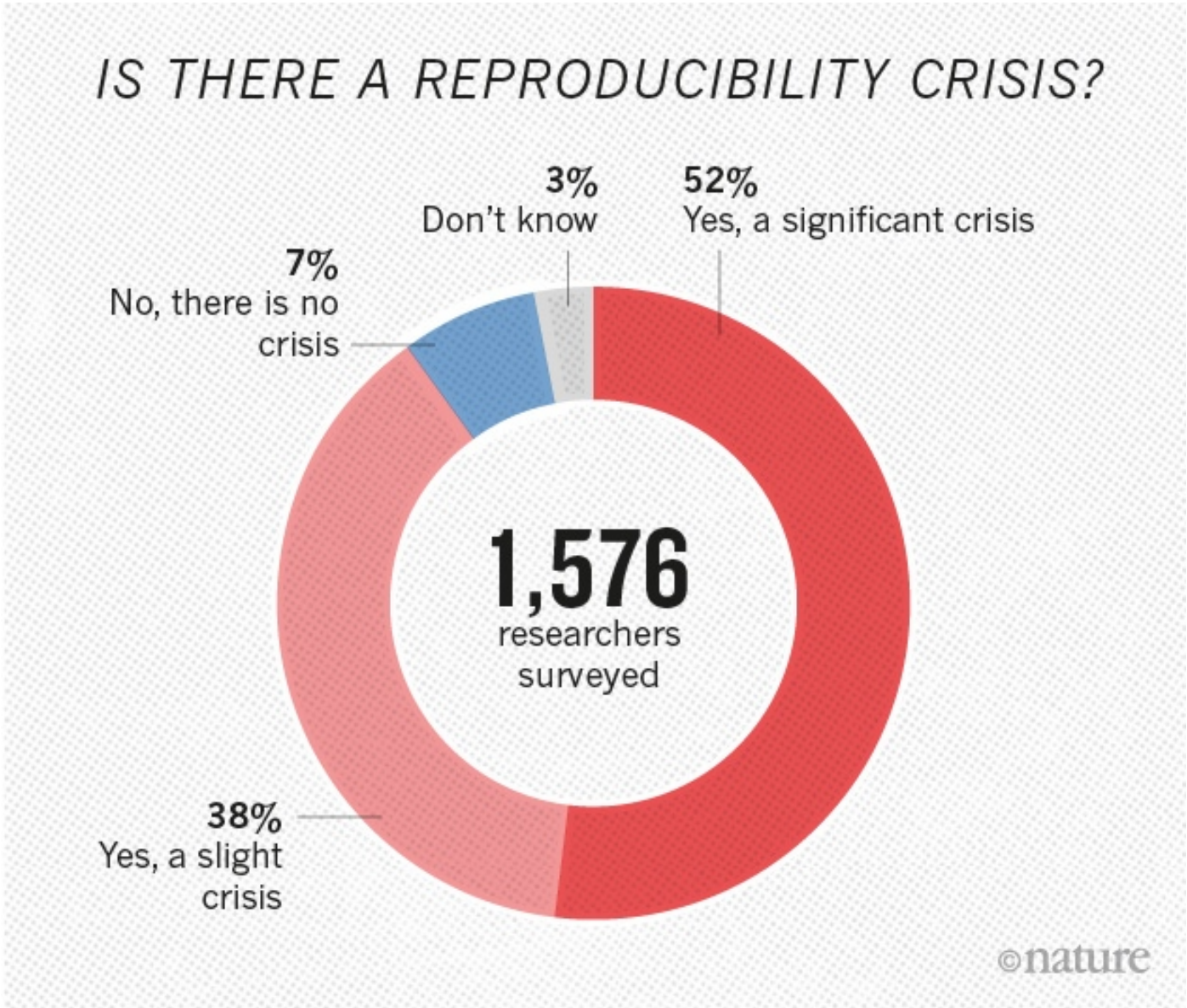
Nature series on "Challenges in Irreproducible Research"

Nature has a website containing editorials, features, news, and articles on various topics related to reproducibile research: Nature special: Challenges in Irreproducible Research
Including

- a checklist for authors of Nature papers described in the 2013 announcement "Announcement: Reducing our irreproducibility"
- R Nuzzo (2014) Nature News Feature "Scientific method: Statistical errors" on "P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume."

Nature series on "Challenges in Irreproducible Research"

- May 25 Editorial "Reality check on reproducibility"



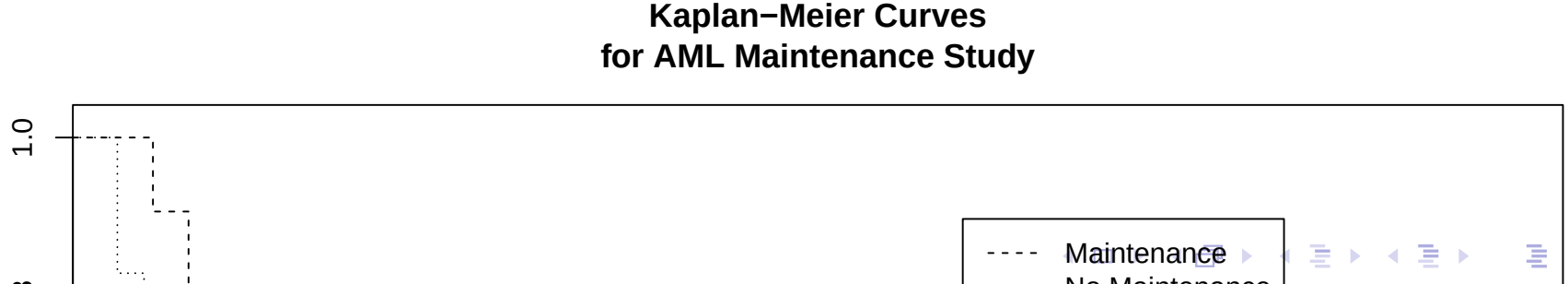
Reproducibility in Practice

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: knitr (for R), Sweave; SASweave, Statrep (for SAS); StatWeave (for STATA)
This is knitr (presentation made with knitr+RStudio):

```
library(survival)
leukemia.surv <- survfit(Surv(time, status) ~ x, data = aml)
plot(leukemia.surv, lty = 2:3)
legend(100, .9, c("Maintenance", "No Maintenance"), lty = 2:3)
title("Kaplan-Meier Curves\nfor AML Maintenance Study")
```



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) "Git can facilitate greater reproducibility and increased transparency in science."

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.***

Other tools

Tracking provenance. *Provenance refers to the tracking of chronology and origin of research objects, such as data, source code, figures, and results. (VisTrails, Kepler, Taverna, and several others)*
Automation *Several Unix tools are useful for streamlined automation and documentation of the research process, e.g. editing files, moving input and output between different parts of your workflow, and compiling documents for publication (shell programs, `make`)*
Capturing the computational environment *A substantial challenge in reproducing analyses is installing and configuring the web of dependencies of specific versions of various analytical tools. (VMware, VirtualBox, Docker, packrat for R)*

ROpenSci Reproducibility Guide

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) "Reproducible research and *Biostatistics*"

How to apply to ourselves?

Goals for BSR (for CCSG)

Aim to improve transparency and reproducibility in the research and analyses performed by BSR members.

Goals for BSR (for CCSG)

- Store all data and code in a central location
  - File/folder naming conventions for all new projects
  - All reports (+tables+figures) are saved in a final reports folder along with the source code to produce them.
- Generate reproducible reports for final results with knitr (for R code) and generate well-documented and easily reproducible results from SAS code.
- Version tracking system (git or another) whenever possible.
- Post commonly used code, R functions, and SAS macros to the online repository GitHub (`https://github.com/ohsu-knight-cancer-biostatistics`).
- Submit code for published results with manuscript or be willing share, also another BSR member should test it (20% of projects, the aim to reach 100% compliance by 2018).
- Develop a checklist for reproducible research based on the reporting standards presented by NIH for transparency and use this for every analysis report.
- Standard of protocols regarding reproducibility on our shared bridge site.
- Training in knitr and reproducible research guidelines

Develop a checklist

To do!

Adapt ROpenSci's Reproducibility Checklist?

Reproduciblity can occur at every step in the history of your project. How easy will it be for others or your future self to answer these questions?

Documentation

- *Is it clear where to begin? (e.g., can someone picking a project up see where to start running it)*
- *can you determine which file(s) was/were used as input in a process that produced a derived file?*
- *Who do I cite? (code, data, etc.)*
- *Is there documentation about every result?*
- *Have you noted the exact version of every external application used in the process?*
- *For analyses that include randomness, have you noted the underlying random seed(s)?*
- *Have you specified the license under which you're distributing your content, data, and code?*
- *Have you noted the license(s) for others peoples' content, data, and code used in your analysis?*

Necessary Goals

- 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. ROpenSci's Reproducibility & Writing Code Guide and "Best Practices for Scientific Computing" (Wilson et al. 2014)
- use html web-based output
  - see Matthew Shotwell's slides
  - 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)

"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)

Resources

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.

Online 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`
Learn git: `https://try.github.io/levels/1/challenges/1`

Websites/slides/blogs

ROpenSci's "Reproducibility in Science" guide: http://ropensci.github.io/reproducibility-guide/ including the reproducibility checklist http://ropensci.github.io/reproducibility-guide/sections/checklist/

Victoria Stodden's list of talks on various topics from "Reproducibility: Breakin' it Down" to "Legal Issues in Reproducible Research" http://web.stanford.edu/~vcs/Talks.html

Matthew Shotwell's slides (2011) "Approaches and Barriers to Reproducible Practices in Biostatistics". http://biostatmatt.com/uploads/shotwell-interface-2011.pdf

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

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/

Stodden (2014) "What scientific idea is ready for retirement?"

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

This presentation is made with 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.

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.