Module 1 - Lesson 03

Reproducible Principles, Practices & Examples

Melinda K. Higgins, PhD.

2017-09-28

Outline

- Reproducible Principles
- Standard Practices
- Journalism - 538.com
- Telling Stories with Data
- Transparency - Journal of Biostatsatistics
- Speed - 2001 outbreak of *E.Coli 0104:H4*

Reproducible Principles - Process & Structure

- Organization
- Clear Documentation
- Standardized
- Centralized
- Efficiency

10 Simple Rules for Reproducible Computational Research[1]

1. For every result, keep track of how it was produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Version of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When POssible in Standardized Formats
6. For Analyses that include randomness, note underlying random seeds
7. Always Store Raw Data Behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

1. Sandve, G.K.; Nekrutenko, A.; Taylor, J.; Hovig, E. (2013) "Ten Simple Rules for Reproducible Computational Research" PLOS Computational Biology, 9(10).https://doi.org/10.1371/journal.pcbi.1003285

Standard Practices

Think about your own work…

- What do you want to automate?
- What could you re-use?
  - code, files, formatting, graphics, logos, header, footer, boilerplate
- What should you share with your team?
- What do you find yourself doing over and over?
  - correcting or reformatting
- If you won the lottery today (and left your job), what do you need to tell your replacement so they can pick up where you left off and complete your current tasks?

Journalism - 538.com

538.com http://fivethirtyeight.com/ hosts stories and opinion pieces covering poll analyses, politics, economics, health, popular culture, and sports. The founder, Nate Silver, and the 538 team are best known for their political polling and forecasting during the United States Presidential and related elections since 2008. ESPN now owns 538.com (as of 2013) retaining Nate Silver as the Editor-in-Chief.

Most of their articles provide references and links to the original data sources plus details on how their figures, analyses and statistical models were developed. They also host the data, code and details behind their analyses on Github https://github.com/fivethirtyeight/.

We will work with some of these datasets in our exercises later in this course and work with the `fivethirtyeight` R package https://cran.r-project.org/web/packages/fivethirtyeight/.

Telling Stories with Data

Andrew Flowers (economist, data scientist, journalist and former writer for fivethirtyeight.com) presented "Finding and Telling Stories with R" at the 2017 RStudio Conference (Orlando, FL).

The webinar recording of his presentation is available online https://www.rstudio.com/resources/videos/finding-and-telling-stories-with-r/.

In his presentation, he highlights the various aspects of "data journalism" and importance of workflow, data processing and transparency in analysis and communication - all key aspects of reproducibility. Andrew Flowers is also a contributor to the `fivethirtyeight` R package.

Transparency - Journal of Biostatsatistics

"Our reproducible research policy is for papers in the journal to be kite-marked **D** if the data on which they are based are freely available, **C** if the authors' code is freely available, and **R** if both data and code are available, and our Associate Editor for Reproducibility is able to use these to reproduce the results in the paper. Data and code are published electronically on the journal's website as Supplementary Materials."

https://academic.oup.com/biostatistics/pages/General_Instructions

Example of an article marked **R**:

- Air pollution and health in Scotland: a multicity study; by Duncan Lee; Claire Ferguson ; and Richard Mitchell; Biostatistics, Volume 10, Issue 3, 1 July 2009, Pages 409–423, https://doi.org/10.1093/biostatistics/kxp010

Speed - 2001 outbreak of

In 2001 there was an outbreak of *E.Coli 0104:H4* that killed 50 people in Europe http://dx.doi.org/10.5524/100001.

Researchers at BGI *(formally the Beijing Genomics Institute)* worked in collaboration with the Medical Center in Hamburg-Eppendorf to rapidly sequence the genome of the pathogen. Given the severity of the outbreak, the team announced and released the genome via Twitter to the world-wide community of microbial genomicists.

A Github repository was established https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki to "crowdsource" analysis and research to find a treatment.

People started contributing their work in under **24 HOURS** and within **5 DAYS!!** a bacterial agent was proposed to kill the pathogen.

Next in Lesson 04 …

Breakdown of Reproducible Components