Reproducible Research in an Academic Environment

John Kloke
Center for Research on Health Care - Data Center
University of Pittsburgh

2 August 2011
Outline

▶ Literature Review
▶ Qualitative Data
▶ Discussion
▶ Summary
Baggerly & Berry (2011) “Reproducible Research”

- Journals must begin to demand RR.
- Investigator’s concerns: data ownership, intellectual property.


- In situations were supplying data is not possible authors should “justify the omission and assure independent reproducibility by alternative means.”
Associate Editor for reproducibility for *Biostatistics* will evaluate the criteria for evaluating the reproducibility of an article.

- **Data** made available on the journal’s website.
- **Code** are provided or linked to.
- **Reproducible** if the AER can execute the code and on data and get results presented in the paper.

At the time of the article only analysis in R were allowed.
Lit Review

- Launched a “reproducible research” initiative.
- Statement including whether the study protocol, data, or code are available.
- Sharing is not mandatory, but authors must state willingness to share.
- Interested in increasing the technical capacity of research teams, further development of data sharing, and support for RR.
Suggestions

1. Poor reproducibility is encouraged by the publication of significant results. Publish non-significant findings in respected e-journals.

2. A funding line for reproducibility studies. Should be cheaper, easier than the initial effort.

3. A continuous measure of reproducibility.
Methods
Informal survey of colleagues

- Examples of reproducible research and non-reproducible research (not nec. at Pitt).
- Examples of practices they implement.
- Examples of practices they would implement given unlimited time & resources.
Population
CRHC Data Center Statisticians

- Director: Doris Rubio
- Faculty: Kaleab Abebe, Jim Bost, Joyce Chang, John Kloke, Doug Landsittel, Charity Moore, Dana Tudorascu
- Staff (MS): Elan Cohen, Diane Comer, Irina Karpov, Jie Li
- Graduate Research Student (GSR): KL2 (1), CTSI (2), CER Supplement (2)
Examples of non-reproducible research

- Have a graduate student run all analyses (even if they are good).
- Investigator doing part of the analysis (with point and click software).
- Having many different datasets.
- Doing *anything* in Excel.
- (Stata) log file, script file
Examples of reproducible research

- Sweave, odfWeave, R2wd, SASweave, StatWeave, etc.
- One well documented (SAS, R, Stata) program with code for all values in the paper.
- SAS’s ods for tables output to excel, rtf, pdf
- matrix + xtable in R
- Stata: outreg, esttab
- Drop unneeded variables from the analysis dataset (decreases chances of using the wrong one).
- Shared file system: secure server, dropbox, Share Point
- Software repository: write packages, macros, etc.
- Version control (e.g. subversion)
Data Center is the Coordinating Center for (PI: Jim Bost; CO: Kaleab Abebe) HALT Polycystic Kidney Disease.

- NIDDK requires SAS Code & Frozen datasets (RR)
- Data monitoring is a place where RR can help. Monthly recreation of reports (on AEs).
- Do as much in the database as possible.
Other issues related to reproducibility

- Data quality
  - double data entry
  - DB errors (HT & WT switched).
- Investigators do preliminary selection of data. e.g. biomarkers that “looked like they showed something”
- Normalization/data clean by investigator
- How similar are study participants to the general population? e.g. Ad on bus: $1225 for 11 nights work
Discussion

Why not Sweave every project?

- The right tool for the right job. I don’t see a one-size fits all approach when projects vary so much in size and complexity.
- With several iterations of the manuscript, outputting to Sweave seems like overkill.
- Becoming proficient is Sweave takes time.
- Investigators don’t use latex.
  - Should we teach investigators to *weave?*
  - Do we take over editorial roll?
  - Hire *weavers?*
Data sharing issues

How to handle confidentiality and other data sharing road blocks?

▶ Independent audit? Similar to what JAMA requires for industry trials.
▶ Describe in the methods section. e.g. we used ODS to created Tables 1, 2, and 4.
▶ Simulated data set + Sweave. i.e. the the results could be reproduced.
Find a balance between accuracy and time commitment. Hard to expect GSRs to learn Sweave (i.e. latex & R) in a year (semester)

RR translates to convoluted code. compare 
\texttt{xtable(coef(lm(y\sim x)))} v. \texttt{lm(y\sim x)}

Conjecture: there is an inverse relationship between reproducibility and readability of code.

Teach reproducible research in MS programs.
Summary

- Need for good practices. Keep good records, code documentation.
- *weave is the gold standard, other tools can work too.
- Need to convince journals, funding agencies, investigators that RR is a good idea.
- Develop a checklist for reproducible research.
Cohort is well described. One could start with original dataset and follow steps to arrive at analysis dataset.

Values in tables produced by commercial software (e.g. ODS) or code is made available (R + xtable)

Values in text produced by software (*weave)

Data
- Original (w/ excluded patients, data errors)
- Analysis dataset
- Simulated
- Validation set.

Code
- Separate program file for each table, figure, section of the text
- Well documented.
Resources/Reading List (not comprehensive)

▶ ctspedia.org
▶ CRAN Task View: Reproducible Research