Reproducible Research and Clinical Trials

Paul A. Thompson, Ph.D.

Methodology and Data Analysis Core
Sanford Research/USD
Sioux Falls, SD

May 29, 2012

SANFORD HEALTH

Sanford Research/USD

"Mr. Irrelevant"

SANFORD HEALTH

Sanford Research/USD

"Dr. Irrelevant"

SANFORD HEALTH

"Dr. Irrelevant"

SANFORD HEALTH

Sanford Research/USD

A large healthcare system in
South Dakota-North Dakota-Minnesota
"The largest rural-based healthcare system in the US"
Foci:

- Type I Diabetes
- Breast Cancer
- Rare diseases and diseases of childen
- Basic research for these efforts

Since 2006, research staff - 25 $\rightarrow$ 250

SANFORD HEALTH

The current state of play

in biostatistics, bioinformatics, and epidemiology involves . . .

- "big data"
- multivariate data reduction
- Function and structure defined simultaneously
- Lack of face validity in statistical results

Methodology for high-dimensional data in personalized medicine is a huge challenge to get right.
Intuition about results is less help

Recent cases

Several recent cases are illustrative

- Potti case at Duke
  - Details are pretty well-known
  - Data management problems
  - Data labeling issues
  - Management control issues
- Retraction Watch: email digest of issues in research
  - Image manipulation and modification
  - Subject consent and proper use of data

Reproducible Research

A response to increasing complexity of analysis

- Somewhat new idea
- Structured analysis approach
- All data available
- All code available for analysis

A Little History

- Begun with computational science
- Issue: Ensuring that processes could be redone
  - Without huge effort
  - By same person, different person, etc
  - After your post-doc goes to Carnavaca and does not return

Types of reproducible research

Many different things are discussed under this term

- Workflow managers - structuring and managing process
- Shared computational environments - repeating analyses
- Analysis management systems - scripting analysis
- Computational documents - text and code mixed

We are mostly interested in **Analysis management systems**
Danger at SCT: Reproducible research ideas are considered . . .

- entirely obvious and trivial
- insulting since everyone already does RR in their group
- not a scientific topic, just more data management

RR and pecking order of research

Pecking order or who gets to make rules about data analysis

- Science often begins with bench scientists
- who often do their own analysis since they LOOOOOOOOVVVVESSS them some statistical analysis
- Biostatistics inherits this at the time of the R01
- Who reviews the earlier science for data management issues Anil Potti on Line 1
- Data quality BEFORE trial is obviously crucial

Sanford Research

- 6 groups of 5-6 Ph.D. level scientists
- Analysis often done with interactive tools
- Biostatistics inherits this at the time of the R01
- We do analysis sometimes, not in other cases
- We have responsibility for components of analysis

SANFORD HEALTH

Reproducible Research Principles

There are several principles of Reproducible Research.
Analysis methods

- . . . should be transparent AND auditable
- . . . should show a clear path
  - from data to dataset
  - from dataset to figure
  - from dataset to table
- . . . should not rely on memory of analyst

When your post-doc or SDA takes a new position, how will you recover your vital analyses, papers, projects, or DSMB reports?

What is and is not RR?

- Consistent approaches
  - Use of make-like tools
  - Scripted programming
  - Full table generation from scripts
  - Full document constructed by script
  - LaTeX-R-sweave
  - SAS-StatWeave/SASWeave
- Inconsistent approaches
  - Interactive programming
  - "Drag-n-drop" table construction
  - Figures and values inserted by hand in documents

Critical Steps to the trial

- Before the trial
  - Data analysis by bench scientists
  - They sometimes have odd opinions about data acceptability
- During the trial
  - DSMB reports
  - Interim analyses by non-DC investigators
- After the trial
  - Datasets for specific analyses
  - NIH requirements for data archiving
  - Data archiving needs

Reproducible Research as defined now

- R + LaTeX= Reproducible Research
- This excludes basic scientists

LaTeX is a super tool, and is the quickest way to go from being a statistician to full-time document management IMHO, LaTeX is being somewhat supplanted by MS Word

SAS + Word

Reproducible research or a False Approach?

- SAS can be used as a Reproducible Research engine
- Many sites are not at this level
  - SAS treated as individual preference tool
  - Hard-coded locations, commands
  - Macros, formats, other tools not formally controlled

Sanford Research Methodology and Data Analysis Center

We are implementing a reproducible research approach

- All projects use a parallel structure
- All locations are defined relative to project root
- Tables, figures linked into Word

Structure of the SAS analysis subdirectory


```
root
   name-of-project
      Data
         Excel/
         SAS/
         Raw/
      Documents/
         Figures/
         Tables/
         References/
      Programs/
         master.sas
         macros.sas
         formats.sas
```

```
1  dm 'clear output;clear log;';options nosource mprint;ods html close;ods html;
2  %macro _overrun(_setup=1,_readin=0,_clean=1,_processa=0,_reporta=0,);
3  %if (&_setup) %then %do;
4              /* Customize by setting the basedir value */
5  %let _basev=x:\MADA_Center\Archive\name-of-base-subdir;
6              /* These are data locations     */
7  %let _exceldata=&_basev\Data\Excel;  %let _sasdata=&_basev\Data\SAS;
8  %let _rawdata=&_basev\Data\Raw;      libname sasdir "&_sasdata";
9              /* These are output locations  */
10 %let _miscdocs=&_basev\Data\Statistical results;
11 %let _figdocs=&_basev\Data\Figures;
12 %let _tabdocs=&_basev\Data\Statistical results;
13             /* These are SAS code inclusions */
14 %let _SASprog=&_basev\Programs;
15 %include "&_SASprog\formats.sas";%include "&_SASprog\macros.sas";  %end;
16 %if (&_readin) %then %do;
17 PROC IMPORT OUT=sasdir.dset DATAFILE= "&_exceldata\name-of-excel-file.xlsx"
18             DBMS=EXCEL REPLACE;
19      RANGE="SHEET1$"; GETNAMES=YES; MIXED=NO; SCANTEXT=YES; USEDATE=YES;
20      SCANTIME=YES; RUN;
21 %end;
22 %if (&_clean) %then %do;  * clean up data  %end;
23 %if (&_processa) %then %do;  * run processes here %end;
24 %if (&_reporta) %then %do;  * write reports here %end;
25 %mend _overrun;
26 %_overrun(_setup=1,_readin=0,_clean=0,_processa=0,_reporta=0);
```

Reproducible Research

- A new standard
- Certain projects require more use than others
- Discipline is important in using scripting tools
- Enforcement of standards requires management "buy-in"

Questions?

????

paul.thompson@sanfordhealth.org

Much taken from

Thompson, P. A. and Burnett, A. E. (2012). Ethics and reproducible research. Encyclopedia of Research Ethics.

Presentation prepared with beamer class, PaloAlto theme, SeaGreen color scheme

SANFORD HEALTH