Interactive and Dynamic Administrative Reporting

October 5, 2018

David Schlueter, Ryan Jarrett, Richard Epstein, Michael Cull, Rameela Chandrasekhar, Kathy Gracey

Department of Biostatistics

Who are we?

- PhD Students in the Department of Biostatistics, Vanderbilt University Medical Center
- We work in collaboration with the Vanderbilt University Center of Excellence for Children in State Custody
- We regularly conduct analyses with MEGA, CANS, Medicaid data

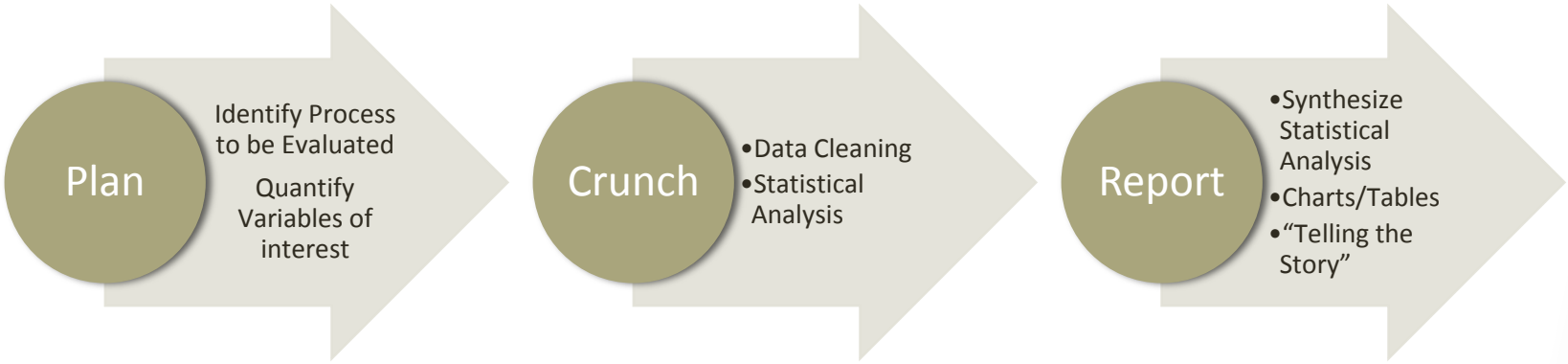
Challenges of our work

- We are primarily involved in statistical collaboration
- One challenge in our line of work is **effective communication** of complex statistical analyses
- Administrative data often have many variables of interest
- We would like our collaborators to have as complete of an understanding of the data as possible
- **Key points**
  - Information that can be conveyed **will always be limited by the reporting medium**
  - Big question of this talk: **How do we best present this vast amount of information?**

How do we construct reports, usually?

1. Identify process that is to be evaluated within the organization
2. Identify measurable outcomes by which process evaluation can be made
3. Leverage administrative data to make statistical inferences
    1. Data Cleaning
    2. Statistical Methods
4. Presentation of results
    - Word or Powerpoint

Standard Project Workflow



How does this process translate into real world practice?

1. Conduct analysis in statistical software such as SAS, R, Stata, etc.
   - Data is fed into software and processed separately from report
2. Generate figures and tables
   - Figures are typically generated in software package by itself
   - Tables can be made in most software, but formatting is often lacking (SAS, Stata)
     - Requires formatting in Word/Excel
3. Copy and paste into Microsoft Word and write report text around generated items

Problems with Static Reporting

- Workflow is not replicable
  - Analysis is not coupled with report
- Data changes from month/quarter year
- Loss in work time efficiency
- Each new data set requires re-calculation of quantities identified in the planning phase
- Addition of new analyses requires either generation of new report, or longer amounts of time invested in adding process to pre-existing report

Alternatives

- Scripted analysis using software package
  - Store a series of commands in a file that can be executed to exactly reproduce an analysis
  - Examples include SAS, R, Stata, SPSS
- Even though analysis can be reproduced for new data, the report cannot be reproduced
  - User must repeat steps to integrate statistical output into the text of report
- What are some alternatives to this workflow?

Reproducible Reporting Tools

- Recently, there has been much work in developing tools that combine typesetting and native code
- Main idea: Interweave scripted code with text
  - CODE + TYPESETTING
  - Make values that appear in text directly from the output of the statistical analysis
  - Statistical Analysis is performed in the same document as the text
  - Lightweight and portable file types
- Some examples?

Knitr Package

- knitR
  - Package built on concept of integrating R code with typesetting language
  - LaTeX is a document preparation tool that uses plain text
    - Used heavily in mathematics due to ease of typesetting equations, section management, etc.
    - Learning curve is steep
  - R + LaTeX = Knitr

Knitr document basic structure

1. Define preamble (formatting of typeset document)
2. Intersperse R code in text by defining “chunks” of R syntax
   - Options include printing raw LaTeX code, defining figure captions, sizing, and much more
   - Code may be hidden or displayed within document
3. Environment variables may be printed in text via a variety of interface commands (e.g. \Sexpr{})
4. Let’s look at a very simple example!

```
1  \documentclass{article}
2
3  \begin{document}
4
5  This is a very simple demonstration of weaving R code with text.
6
7  <<chunk1,echo=FALSE>>=
8  x <- 113
9  y <- 13
10
11 z <- x - y
12 @
13
14 We know that the variable $x$ is \Sexpr{x}, the value of $y$ is \Sexpr{y} and the
   value of $z$ is \Sexpr{z}.
15
16
17 \end{document}
```

This is a very simple demonstration of weaving R code with text.

We know that the variable x is 113, the value of y is 13 and the value of z is 100.

Rmarkdown Package

- Rmarkdown
  - Similar to knitr, but with simpler syntax
  - Chunks are embedded within text
  - Analysis can be exported directly to HTML, pdf, or Word
  - Low barrier to entry (easy syntax and open source)
  - Portable
- Let's take a look at a very simple example in the R-Studio environment!

Many proprietary software packages have similar modules

- SAS +LaTeX = StatRep
- Stata has several including “estout” and “texdoc”
- Could also code native Javascript (very difficult and requires a high level of expertise in programming)

Benefits of Dynamic Workflow

1. Ease of report generation with new data
2. Transparency of statistical methods in result
   - Exact code can be provided in the output
3. Benefits of reproducibility of a script being tied directly to the report
   - New datasets can be uploaded and re-run using the same report template
   - Programming the dynamic document mitigates repeated work
     - Increase in efficiency

How has our workflow changed?

Drawbacks of static reporting

- While reproducible research is good, the aforementioned tools are **static**
- Administrative reports naturally grow
  - *Ad hoc* requests are commonplace and are eventually added to the report
  - Tables can start to become unwieldy
  - Information “overload”
- Report generation becomes a cyclical operation
- Graphics are static and non-interactive

Report Aesthetics

- Aesthetics are important for user understanding and interest
- Example: Graphical output
  - Static
  - “Exploration” impossible
  - Lacks user engagement
- Tables are restricted to what was planned *a priori*
  - If too many tables are added, report will become too long!
- Example: Let’s look at our CANS analysis in its PDF form

How do we manage information overload?

- One solution is to build dynamic and interactive applications
- HTML based applications allow the user to explore the data by interacting with it
  - Increases user engagement
  - Allows user to develop hypotheses
  - Reduces need for *ad hoc* requests
  - Organizes otherwise long documents
- How to organize?
  - Tabs within application
  - Dropdown menus to present only what is of interest
- E.g. Cross tabulation can be performed on variables selected from a dropdown menu

Software

- R-Shiny
  - Third package maintained by RStudio
  - Create reproducible and dynamic application that can be easily deployed to web
  - Applications are built in a block-wise fashion
  - Default “widgets” are very good, but user can program custom options if necessary
- Tableau
  - Proprietary, but a great alternative

Demonstration of R-Shiny applications created for the Vanderbilt COE

1. CANS Analysis
    1. Motivation
    2. Examples of Static Workflow
    3. Shiny Application
2. Problematic Prescribing Practices Application
    1. Motivation
    2. Shiny Application

Challenges of using these applications

- Requires programming knowledge/learning upfront
  - If user is familiar with R, the syntax is much easier to learn
  - Users are at the mercy of server accessibility
    - Internet is required
    - Pdf version of report can be made downloadable for offline access
- We have found that when many people are accessing the application at the same time, the applications can be slow

Data Security

- Data security is of utmost importance
  - Pricing structure available for hosting private application
  - Application is by default available for general viewing, however data is not required to be associated with application
  - Data can be uploaded manually by the user (via private server)
    - Requires access to data server
    - Upload/search widget is available in R-Shiny
  - Alternatively, data may be encrypted
    - Key is required to decrypt data and populate application
    - Hashing
    - Password generation for a select group of individuals
      - Sent through encrypted channel (i.e. Redcap)

Thank you!

- Are there any questions?

- Contact:
  - David: david.j.schlueter@vanderbilt.edu
  - Ryan: ryan.t.jarrett@vanderbilt.edu