DATA ACQUISITION AND STATISTICAL ANALYSIS

Megan Branda, MS
Senior Research Instructor
Department of Biostatistics and Informatics
University of Colorado Denver
Megan.branda@ucdenver.edu
Reproducible / Replication

- **Reproducibility**: is the ability of an entire experiment or study to be duplicated, either by the same researcher or by someone else working independently.

- **Direct replication**: is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding, and is the means of establishing reproducibility of a finding with new data.

Soc. Psychol. 45, 137–141 (2014); Gen. Psychol. 13, 90–100 (2009)
Is all research reproducible? Should we expect all research to be reproducible?

We expect that a small percentage of research experiments/findings not to be reproducible. Why?
Is all research reproducible? Should we expect all research to be reproducible?

We expect that a small percentage of research experiments/findings not to be reproducible. Why?

How reproducible has research been?

A. >90%
B. 61-90%
C. 31-60%
D. <30%
A few examples from the literature

- One analysis estimates that 85% of biomedical research efforts are wasted
  

- In cell biology, industrial labs reported successful reproducibility of only 11% (@#$!) of the studies they attempted to replicate.
  

- In another drug development study, only 25% replication success was reported!
  
How to avoid!

- “asking questions at the design stage can save headaches at the analysis stage: careful data collection can greatly simplify analysis and make it more rigorous”
  

- Some of these studies (previous slide) did suggest a few practices that might be contributing to this lack of reproducibility:
  - Selective reporting
  - Selective analysis
  - And insufficient specification of the conditions necessary to obtain the results.

Threats to Reproducible Science

Nature Human Behaviour volume 1, Article number: 0021 (2017)
Objectives

■ Data Acquisition
  - Learn general guidelines for unbiased data collection
  ■ Best practices – how to find help!
  - Know the 3 top things to consider when storing data
  - Explain who owns research data and with who and how it should be shared

■ Statistical Analysis
  - Learn how to identify outliers and what to do with them
  - Recognize the trade-offs between suitable, better, and best methods for statistical analyses
  - Learn to identify common mistakes and deceptions when display and interpretation of results
DATA COLLECTION

Data Acquisition
Data – What to record!

- What was done
  - Data and results
- How it was done
  - Methods and materials
- When it was done
- Why it was done
- Who did it
- And the next steps

Records should be signed and dated!
Data Collection

1. Are there guidelines/regulations you need to follow?
   - **HIPPA, hazardous materials, copyrights, consent forms**

2. Use Appropriate methods!

   “Failure to find the effect could be due to either your experimental design or the lack of an effect, but you will not know which is true.”

   “Physical process of recording the data in some type of notebook (hard copy), computer file (electronic copy), or other permanent “record” of the work done.”

1. Define data items (clear set rules)

---

Taken from **ORI Introduction to RCR** (http://ori.hhs.gov/education/products/RCRintro/)
Do Not **Recreate the Wheel**

- When starting a new project use Google! Or a colleague!
  - *What resources are publically available*
  - *What has already been found to be a bad direction*
    - Search terms: Data collection, best practices, Data management, Guidelines
    - Experimental Design Assistant
      - [https://eda.nc3rs.org.uk/](https://eda.nc3rs.org.uk/)

- Organization
  - *Folder Organization*
    - ReadME
  - *Data Dictionary*
  - *Data Collection SOP*
  - *Define Roles and Responsibilities*
Welcome to the Experimental Design Assistant, a free resource from the NC3Rs to support researchers in the planning of animal experiments - ensuring robust study design and reliable and reproducible findings.

The EDA helps you build a machine-readable diagram representing your experimental plan. Tools include:

- Feedback and advice on your experimental plan
- Recommendation for the statistical analysis
- Support for randomisation and blinding
- Sample size calculation

Check the video demonstration and the user guide for general information on the EDA process. Find out more about the background for this project.

---

**Step 1**

**Login or Register**

Start using the EDA application

---

**Step 2**

**Plan your experiment as a diagram**

Check the examples and the user guide for more information

---

**Step 3**

**Critique your design**

The critique function enables you to get feedback and advice on your diagram; find more information here

---

**Step 4**

**Improve your design**

Modify your experimental plan based on feedback from the system
Dr. Z is mentoring a “promising” medical student over the summer in his research lab.

Student’s project:
- cancer cell line that requires 3 weeks to grow in order to test for a specific antibody
- the student has already written a short paper on his work

Dr. Z’s dilemma:
- after going over the raw data, some data were on pieces of yellow pads without clear identification from which experiment the data came
- some of the experiments were repeated several times without explanation as to way
- Dr. Z is not happy about the data, but doesn’t want to discourage the student from pursuing a career in research
Case Study
from Responsible Conduct of Research

- Dr. Z is mentoring a “promising” medical student over the summer in his research lab
- Student’s project
  - cancer cell line that requires 3 weeks to grow in order to test for a specific antibody
  - the student has already written a short paper on his work
- Dr. Z’s dilemma:
  - after going over the raw data, some data were on pieces of yellow pads without clear identification from which experiment the data came
  - some of the experiments were repeated several times without explanation as to way
  - she is not happy about the data, but doesn’t want to discourage him to pursue a career in research

- What is the primary responsibility of the mentor?
- Should the mentor write a short paper and send it for publication?
- Should the student write a short paper and send it for publication?
- If you were the mentor, what would you do?
DATA STORAGE

Data Acquisition
Data Storage

“Over time, data, as the currency of research, become an investment in research. If the data are not properly protected, the investment, whether public or private, could become worthless”

– ORI Introduction to RCR
What is considered safe?

- Is an encrypted flash drive safe?
- Dropbox
- OneDrive
- Sharepoint
- Your laptop?
Considerations When Storing Data/Research

■ Catastrophe
  - Lab notebooks are in a “safe” place
  - Electronic data are backed up and stored in a separate location
  - Samples are stored properly to avoid contamination

■ Confidentiality
  - Information on human subject – see HIPAA guidelines
  - Information on intellectual property

■ Period of retention
  - NIH generally requires 3 years after project end
  - Other agencies may require up to 7 years after project end
  - University of Colorado AMC requires 9 years after grant end
  - Other unforeseen uses...

Taken from ORI Introduction to RCR
(http://ori.hhs.gov/education/products/RCRintro/)
## University of Colorado Denver | Anschutz Medical Campus
### Record Retention Matrix

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Repository</th>
<th>Retention Period</th>
<th>Related Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loan Records</td>
<td>Bursars Office</td>
<td>3 years after pay-off</td>
<td>State Archives 34 CFR Sec. 74.53</td>
</tr>
<tr>
<td>Tax</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1098-T</td>
<td>Bursar Office</td>
<td>4 years</td>
<td>State Archives 34 CFR Sec. 74.53</td>
</tr>
<tr>
<td><strong>Grant and Research Records</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical Research Records</td>
<td>Department</td>
<td>2 years post marketing approval or</td>
<td></td>
</tr>
<tr>
<td>Protocols</td>
<td></td>
<td>IND withdrawal</td>
<td></td>
</tr>
<tr>
<td>Patient Records</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulatory Records</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Associated Contracts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accounting Records</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grant Project Research Records</td>
<td>Office of Grants</td>
<td>9 years after expiration of grant</td>
<td>State Archives Records Management Manual - Schedule 8</td>
</tr>
<tr>
<td>Activity Reports</td>
<td>and Contracts,</td>
<td>funding period or termination of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Academic Departments, Regulatory</td>
<td>contract and until no longer needed</td>
<td></td>
</tr>
<tr>
<td>Conflict of Interest Disclosures</td>
<td>Compliance or other</td>
<td>for reference.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>repository as designated.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary Reports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Working Papers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Related Documentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grants, Contracts, and Awarded</td>
<td>Department</td>
<td>6 years after the project becomes</td>
<td></td>
</tr>
<tr>
<td>Proposal Records</td>
<td></td>
<td>inactive and until no longer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>needed for reference or as otherwise</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>provided for by the award</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>documents.</td>
<td></td>
</tr>
</tbody>
</table>
DATA OWNERSHIP / SHARING

Data Acquisition
Ownership/Data Sharing

Who owns the data?

- Researchers
- Funders
  - Grants vs. Contracts
- Data Sources
  - Subjects
  - Countries

• Research Institutions

Illustration by David Zinn

Taken from ORI Introduction to RCR (http://ori.hhs.gov/education/products/RCRintro/)
Ownership/Data Sharing

Who owns the data?

- Researchers
- Funders
  - Grants vs. Contracts
- Data Sources
  - Subjects
  - Countries
- Research Institutions

“for the most part, NIH makes awards to institutions and not individuals”

- NIH Data Sharing Policy and Implementation Guidance

Illustration by David Zinn

Taken from ORI Introduction to RCR
(http://ori.hhs.gov/education/products/RCRintro/)
A few interesting quotes from the NIH Data Sharing Policy and Implementation Guidance on **Data Sharing**

“**Final research data** are recorded factual material commonly accepted in the scientific community as necessary to document, support, and validate research findings.”
A few interesting quotes from the NIH Data Sharing Policy and Implementation Guidance

“NIH expects timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset.”
Case Study
from *Responsible Conduct of Research*

Drs. K and W are conducting a NIH-funded long-term (25 years), observational study of the health of pesticide applicators.

- *Initial health assessment (health history, physical exam, blood and urine tests, DNA sample, and dust samples)*
- *Yearly health surveys and full health assessment every 4 years*

After the first 15 years:

- Published more than a dozen paper from the database
- Require a elaborate data-sharing agreement before releasing the data

Drs K and W’s dilemma is that they recently received requests for access to the database from:

- A pesticide company
- A competing research team
- A radical environment group with an anti-pesticide agenda
Drs. K and W are conducting a NIH-funded long-term (25 years), observational study of the health of pesticide applicators.

- Initial health assessment (health history, physical exam, blood and urine tests, DNA sample, and dust samples)
- Yearly health surveys and full health assessment every 4 years

After the first 15 years:

- Published more than a dozen paper from the database
- Require a elaborate data-sharing agreement before releasing the data

Drs. Kessenbaum and Wilcox's dilemma is that they recently received requests for access to the database from:

- A pesticide company
- A competing research team
- A radical environment group with an anti-pesticide agenda

QUESTIONS

■ How should Drs. K and W handle these requests to access their database?

■ Is it ethical to require people who request data to sign elaborate data sharing agreements?
Statistics show that teen pregnancy drops off significantly after age 25.

Mary Anne Tobe, Republican state senator from Colorado Springs (contributed by Harry P. Fument)

Monday, December 1999
Tips for Reproducible Statistical Analyses

1. ALWAYS keep a version of the “most raw” data
   - Record when and where it was created, so you can easily tell if it has been changed since creation

2. Version Control
   - GitLab

3. Use a scripting language
   - Programs like R and SAS allow you to follow your steps exactly if you (or someone else) had to redo your analysis
   - Easily execute and document QC steps
   - Avoid copy/paste errors

4. Add comments/notes directly to program
   - Why are you doing this step?
   - What is the goal of this step?

5. Export precise tables/figures from program
   - Avoid transposition errors
   - Save time/energy where changes are requested in initial steps
OUTLIERS

Statistical Analysis
Outliers

With Outlier

correlation coefficient = 0.06
p-value = 0.5636

Without Outlier

correlation coefficient = 0.7
p-value = <0.0001
Outlier Mitigation

1. Identify
   - 2 or 3 standard deviations
   - Unrealistic values
   - Inconsistent

2. Investigate
   - Was there a technical issue? typo? etc?
   - Is it even a possible true value?

3. Remediate with DOCUMENTATION
   - Make a rule and write it down

4. Sensitivity analysis
   - What would have happened if you hadn’t eliminated values? Is your result robust?
Anonymous survey of college students on opinion about academic integrity

- 20 questions (Likert scale)
- 10 open-ended questions
- 480 surveys administered (320 responses)

Issues:

1. 8 surveys appear as practical jokes (obscenities, additional numbers added to scale, etc.)
   - Some questions appear usable but some are not
2. 35 respondents appear to be confused about scale
   - They answer “5” when “1” is more logical given their other answers
3. 29 surveys have names on them when respondents were instructed not to do so
Case Study from *Responsible Conduct of Research*

Anonymous survey of college students on opinion about academic integrity

- 20 questions (Likert scale)
- 10 open-ended questions
- 480 surveys administered (320 responses)

**Issues:**

1. 8 surveys appear as practical jokes
   - Some questions appear usable but some do not

2. 35 respondents appear to be confused about scale
   - They answer “5” when “1” is more logical given their other answers

3. 29 surveys have names on them when respondents were instructed not to do so

**QUESTIONS:**

1. How should the researchers deal with these issues with their data?

2. Should they try to edit/fix surveys that have problems?

3. Should they throw away any surveys? Which ones?

4. How might their decisions concerning the disposition of these surveys affect their overall results?
SUITABLE, BETTER, AND BEST METHODS FOR ANALYSIS

Statistical Analysis
Methods for Statistical Analysis

- What is the norm in the field?
- A spectrum of alternative statistical methods

![Diagram showing increasing scope, monetary and time costs, and precision from bias to most statistically rigorous method.]

- Bias, Inappropriate method
- General method with stated assumptions
- Most statistically rigorous method that evaluates most/all assumptions

- Increasing scope
- Increasing monetary and time costs
- Increasing precision
A top Cornell food researcher has had 15 studies retracted. That’s a lot.

Brian Wansink is a cautionary tale in bad incentives in science.

By Brian Resnick and Julia Belluz  |  Updated Oct 24, 2018, 2:25pm EDT
DISPLAY AND INTERPRETATION OF RESULTS

Statistical Analysis
Displaying Results

Crime Statistics for Key West

**Property Crime**
- Key West: 6.68%
- Monroe County: 6.68%
- Florida: 3.99%

**Violent Crime**
- Key West: 0.85%
- Monroe County: 0.85%
- Florida: 0.71%
Displaying Results

Confident in obtaining first NIH R01 Grant?

Confident in Obama’s Economic Plan?

ABC News/Washington Post polls
Interpreting Results

- **Association vs. Causation**
  - *Causation can only be proven in a carefully designed and carefully controlled prospective study*
  - Eating more chocolate will not cause you to become a Nobel Laureate

- **Potential Confounding Issues**
  - *Confounding variable* – “extraneous variable in a statistical model that correlates with both the dependent variable and the independent variable” – Wikipedia
  - e.g., Coffee drinkers are more likely to get lung cancer
  - Smokers are more likely to be coffee drinkers and smokers are more likely to get cancer
Key Items to Remember

- Do not do anything to your data that you are not willing to explain in a publication!
- Document Everything!
- Secure Storage
- Plan before do!
- Do not work on an Island!
- Know your assumptions with your data!
- When in doubt, ask!
Acknowledgements/References

- Dr. Laura Saba
- Dr. Paula Hoffman
- Dr. Brandie Wagner
- ORC and CCTSI

http://www.ucdenver.edu/research/ORC/RI/Pages/dataacquisition.aspx

References


