DATA ACQUISITION AND STATISTICAL ANALYSIS

Megan Branda, MS
Senior Research Instructor
Center for Innovative Design & Analysis (CIDA)
Department of Biostatistics and Informatics
University of Colorado Denver
megan.branda@ucanschutz.edu
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!
  
“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

Publish and/or conduct next experiment

Generate and specify hypothesis

Failure to control for bias

Design study

Low statistical power

Conduct study and collect data

Poor quality control

Analyse data and test hypothesis

P-hacking

Interpret results

P-hacking

Publication bias

HARKing

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!

Shamoo & Resnik. Responsible Conduct of Research 2015
DATA COLLECTION

1. Are there guidelines/regulations you need to follow?
   - HIPAA, 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.”
   - 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 publicly available
  - What has already been found to be a bad direction
    - Search terms: Data collection, best practices, Data management, Guidelines, lessons learned
    - Experimental Design Assistant

- Organization
  - Folder Organization
    - ReadME
  - Data Dictionary
  - Data Collection SOP
  - Define Roles and Responsibilities
Folders

1. Data
   a. Raw Data
   b. Data Dictionary
   c. Analytic files

2. Program
   a. Code used to generate raw data files (if applicable)
   b. Secondary files (if applicable)
   c. Code used to create analytic files
   d. Analysis Code/ Script
   e. Program Log from Statistical Program

3. Results
   a. Tables, charts, graphs, etc.
   b. Drafts by Date
   c. Preliminary Findings
   d. Final Report

4. Background
   a. Lit review
   b. Documentation from partner
   c. Relevant background
README

Project Name:
Executive Sponsor:
Champion:
Comirb#
Analyst:

**Background**

List all files and source of information, i.e. who sent it to you if applicable

**Data**

List each dataset in file, these should be analyzable datasets with a brief description of what they are for.

**Data/RawDate_DONOTDELETE**

All raw data provided on project is stored in here, these are not to be overwritten.

**NavLabDocumentation**

Protocol, intake form, SOW, COMIRB, All analytical plans are to be stored in this folder and each file listed.

**Program**

Each program to analyze the data should be listed, the purpose of the program and if multiple programs to get to a result are needed then document the order they need to be run in.
Randomization/Data Entry Instructions

REDCap will be used for data entry and randomization. Study coordinators at each site will randomize the patients prior to the clinical encounter with the enrolled clinician.

All study staff questions should be directed to xx and xx

All access for data entry or issues with software for data entry should be directed to the statistical team xx and xx

**Clinician enrollment:**

After a clinician has been consented, the study coordinator will assign the clinician a study id. You will have the clinician complete the baseline enrollment survey to obtain their demographics. This data is collected only at time of enrollment. If the clinician being consented does not have time to complete survey prior to participating in the trial with an enrolled patient, then the survey can be completed after the encounter. The data is to be entered into the redcap database called “SDM for AFib: Clinician Baseline Survey – AL & MS Sites” within 7 days of first enrolled patient otherwise considered overdue. A scanned copy of the consent along with a scanned copy of the completed survey will be uploaded to the database as well (for Park Nicollet, HCMC, U of Mississippi, and U of Alabama only). Once all clinicians have been enrolled for your site you will not need to use this again.

- Format for clinician ID’s:
  - example “AF*Site*Clinician*001”;
  - where site is either ‘M’=Mayo Clinic Rochester,
    - ‘E’=Mayo Clinic Emergency Department or
    - ‘P’=Park Nicollet or
    - ‘H’=HCMC or ‘
    - ‘J’=University of Mississippi or
    - ‘A’=University of Alabama
  - and the number starts at 001 and increases incrementally as you enroll clinicians.
  - Example: AFMC001, first enrolled clinician at Mayo
  - The numbers do not need to be incremental.
  - If you made an error in entry of the ID into REDCap, contact the statistical team for them to correct.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrolled Patient ID</td>
<td>This is the ID generated by the study coordinator.</td>
</tr>
<tr>
<td>Staff Initials</td>
<td>Enter the initials of the staff member at the site that is approaching the patient for enrollment: First, Middle, Last. If the middle initial is not known, enter a dash (-).</td>
</tr>
<tr>
<td>Clinic Location</td>
<td>Choose enrolling site</td>
</tr>
<tr>
<td>Medication Cohort</td>
<td>Start – Patients that have not taken an anticoagulant within 6 months of enrollment. Review – Has a prescription for an anticoagulant and taken medication within 6 months.</td>
</tr>
<tr>
<td>CHADS-VASc Score</td>
<td>Choose: For men with a score of 1 or women with a score of 2. Or For men with a score greater than 1 or greater than 2 for women.</td>
</tr>
<tr>
<td>Arm Assignment</td>
<td>This is the randomization tool, confirm that the medication cohort and CHADS-VASc score has been entered correctly prior to clicking the button for randomization. The tool will ask you to confirm the data prior to randomization and will provide an error if the required information is not obtained. After you have confirmed the data is accurate you will click the ‘randomize’ button. If the data is inaccurate you make your correction within that screen and then select ‘randomize’. The changes to the medication cohort and CHADS-VASc will be saved. The arm assignment that is provided is final.</td>
</tr>
<tr>
<td>Signed consent form</td>
<td>Upload a .pdf or .doc copy of the patients signed consent form.</td>
</tr>
<tr>
<td>Patient eligibility CRF</td>
<td>Provided a .pdf or .doc of the completed eligibility CRF.</td>
</tr>
<tr>
<td>Patient excluded after randomization</td>
<td>If a patient is found to be ineligible after consent then this field will be checked and the reason for ineligibility will need to be provided. If the patient consents and then withdraws* then check this field and provide the reason and note within the text field that they are a withdraw.</td>
</tr>
</tbody>
</table>
REASONS TO KEEP ACCURATE RECORDS

- Reproducibility
- Future analyses
- Investigations of misconduct
- Proving ownership of intellectual properties
- Others?
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 why
- Dr. Z is not happy about the data, but doesn’t want to discourage the student from pursuing a career in 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 why
- 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
“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
- Redcap
- Your laptop?
MAKE SURE YOU VERIFY!

- **Email safeguards:** [https://www1.ucdenver.edu/offices/office-of-information-technology/secure-campus/encryption](https://www1.ucdenver.edu/offices/office-of-information-technology/secure-campus/encryption)

- **OneDrive:** [https://www1.ucdenver.edu/offices/office-of-information-technology/software/how-do-i-use/onedrive](https://www1.ucdenver.edu/offices/office-of-information-technology/software/how-do-i-use/onedrive)

---

Stay Secure

OneDrive is configured for [HIPAA compliance](https://www1.ucdenver.edu/offices/office-of-information-technology/secure-campus/encryption), so files are all private by default. However, sharing for viewing and editing is easy to do. Learn how to stay secure while sharing files with OneDrive.
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

**10/4/2017**

<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 1098-T</td>
<td>Bursar Office</td>
<td>4 years</td>
<td>State Archives 34 CFR Sec. 74.53</td>
</tr>
</tbody>
</table>

### Grant and Research Records

<table>
<thead>
<tr>
<th>Record Type</th>
<th>Repository</th>
<th>Retention Period</th>
<th>Related Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Research Records</td>
<td>Department</td>
<td>2 years post marketing approval or IND withdrawal</td>
<td></td>
</tr>
<tr>
<td>Protocols</td>
<td></td>
<td></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 and Contracts, Academic Departments, Regulatory Compliance or other repository as designated.</td>
<td>9 years after expiration of grant funding period or termination of contract and until no longer needed for reference.</td>
<td>State Archives Records Management Manual - Schedule 8</td>
</tr>
<tr>
<td>Activity Reports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conflict of Interest Disclosures</td>
<td></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 Proposal Records</td>
<td>Department</td>
<td>6 years after the project becomes inactive and until no longer needed for reference or as otherwise provided for by the award documents.</td>
<td>State Archives Records Management Manual - Schedule 8</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

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

– NIH Data Sharing Policy and Implementation Guidance

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.”
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
**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 papers from the database
- Require an 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 Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Punce)
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?
**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 (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?
Work cited more than 20,000 times.
‘the misreporting of research data, problematic statistical techniques, failure to properly document and preserve research results and inappropriate authorship.’
DISPLAYING RESULTS

Crime Statistics for Key West

Property Crime
% population affected

- Key West: 6.68%
- Monroe County: 6.68%
- Florida: 3.99%

Violent Crime
% population affected

- Key West: 0.85%
- Monroe County: 0.85%
- Florida: 0.71%
HELPFUL RESOURCES

- UCLA IDRE - Statistical methods across multiple programming languages with real world problems with explanations of code and interpretation of output

- Nature – Points of Significance Articles
  - https://www.nature.com/collections/qghhqm/pointsofsignificance
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!
- Documentation!
- 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


