DATA ACQUISITION AND MANAGEMENT

Camille Hochheimer, PhD
Research Associate and Consulting Center Manager
Center for Innovative Design & Analysis (CIDA)
Department of Biostatistics and Informatics
University of Colorado Denver
camille.hochheimer@cuanschutz.edu

HOUSEKEEPING

Zoom Etiquette:

- Silence personal devices.
- Stay muted when not talking.
- Set up in a quiet location.
- Remain attentive. Avoid checking email/phone/web.
- Use the Chat function to ask questions or get technical help.
- Use your full name, not an alias.

Receiving credit for attendance:

To satisfy the NIH Requirement for Instruction in the Responsible Conduct of Research, the following are required in order to receive credit for attendance:

Attend the full 90 minutes of the training. Attending any 8 out of the 9 RCR seminars we offer will satisfy the NIH requirement.

Keep your video camera on throughout the session. NIH requirements for RCR training specify face-to-face discussion.

Participate interactively throughout the session. Participate in discussions, respond to polls, and sign the attendance sheet (link will be distributed in the Chat).

POLL: WHO IS THE AUDIENCE?

- 1. Select your primary university position or affiliation:
  - PI/Faculty
  - PRA/Researcher
  - Study Coordinator
  - Regulatory/Admin Support
  - Student/Trainee/Post-doc
  - Other

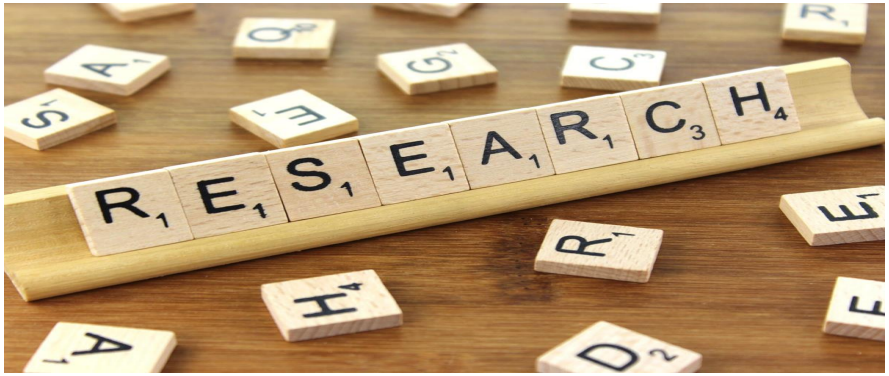
- 2. Are you primarily working on campus or remotely/from home?
  - On campus
  - Remotely

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. <u>Why?</u>

Poll: How reproducible has research been?

A FEW EXAMPLES FROM THE LITERATURE

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

  Lancet **383**, 101–104 (2014).

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

  Nature 483, 531–533 (2012)

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

  Nat. Rev. Drug Discov. 10, 712–713 (2011)

HOW TO AVOID!

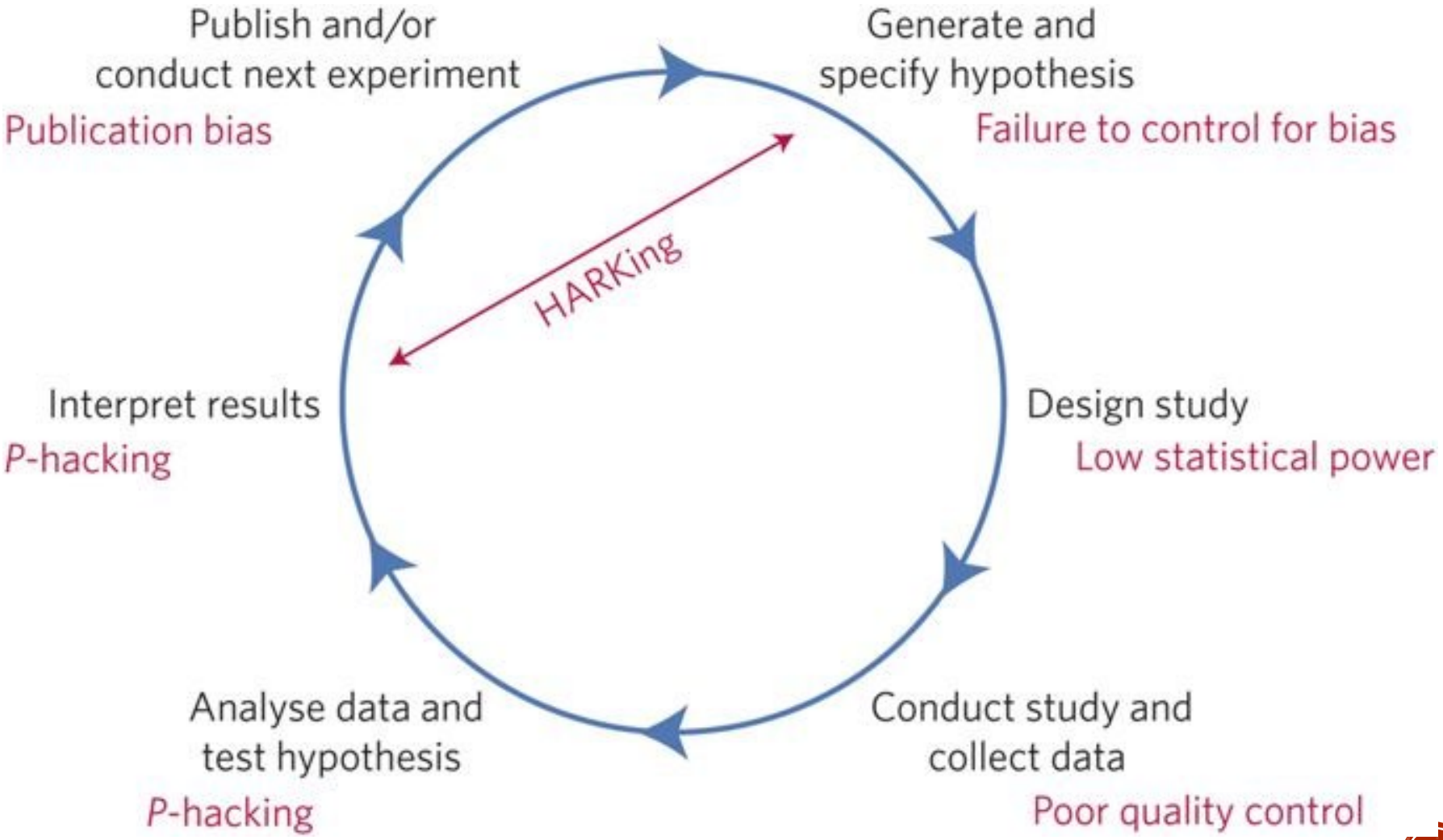
- "asking questions at the design stage can save headaches at the analysis stage: careful data collection can greatly simplify analysis and make it more rigorous"

  Kass, R. E. et al. Ten simple rules for effective statistical practice. PLoS Comput. Biol. 12, e1004961 (2016)

- 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.

Science 343, 229 (2014); Nat. Rev. Neurosci. 14, 365–376 (2013); PLOS ONE 5, e10068 (2010)

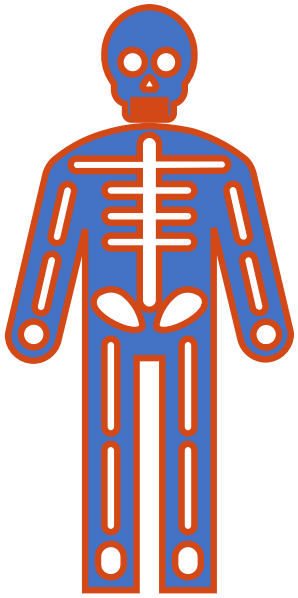
THREATS TO REPRODUCIBLE SCIENCE



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

OBJECTIVES

- Specimens
  - Common Errors
  - Guidance on best practices
- 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***
- A couple of guiding principles on statistical analysis!



SPECIMENS

Collection Considerations

Slides provided by Nassia Duncan, Lab Manager

- Procedure says to freeze sample immediately, but sample isn't frozen for three hours because of incident in OR. Incident is not recorded and testing proceeds as normal. Test results are not replicable because of degradation of the sample.
- Samples are pipetted from primary container into secondary storage container. PRA doesn't use a new pipette for each sample because it's not in the procedure. Contaminated samples are sent for testing. Results are not replicable.
- Saliva and urine samples are collected, placed in a bag, and shipped by air to a lab. The pressure changes force saliva and urine out of their containers, compromising the lids. All samples in the bag are contaminated and unusable. No data is collected.

SAMPLE INTEGRITY

Did you plan accordingly?

OOPS! 50K SAMPLES IN A FREEZER, WHICH ONE IS MINE?

- All samples are labeled with paper labels that are taped over with scotch tape for security before freezing. Adhesive fails in -80 freezer and tape peels off, taking all the ink with it. Samples are now unidentifiable. No data can be collected.
- Labeled samples are placed in an unlabeled box and put in the freezer. Two weeks later the lab attempts to locate samples among 300 other boxes. Samples are never found. No data is collected
- Each sample is labeled with permanent marker, but the handwriting is illegible. Lab makes best guess at reading the chicken scratch. Samples are mis-labeled and results are not replicable.
- Liquid samples are thrown in the freezer and land on their sides. Freezing expands the sample and pushes the lid off, exposing the sample to air and contamination. When the sample is thawed it leaks. Results are not replicable.

PROTECTING YOUR DATA STARTS AT SAMPLE COLLECTION

- Go through every step of the collection, shipping, and testing process. What is that sample going to go through?
  - What containers will be used? How are they sealed?
  - How will containers be labeled?
    - Label Everything Always!
      - Label primary and secondary containers. Keep a log of what you've collected and **where it is**. Who has it and when?
      - Keep labels consistent. The first collection of an experiment cannot change from "Baseline collection" to "Collection 0" six months into sampling.
    - Legibility. Can everyone at every stage of testing read that label?
    - Will labeling survive the freezer and multiple freeze/thaw cycles?
      - Many adhesives are non-functional if exposed to extreme cold or any moisture.
- The clearer and more detailed your collection procedure, the better off you are.

SAMPLES

Do you need to ship them?

- Is container watertight for biological samples? (And will **remain** so?)
- Will samples be shipped upright?
- What is the potential for leakage/breakage?
- Are different sample types packaged separately in case of leaks?
- Will samples be shipped by air? (pressure and temperature changes)
- Is shipping time sensitive? (dry ice, frozen samples, temperature sensitive samples)
- Have you followed federal guidelines for shipping biological or hazardous substances?

Data Acquisition

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 publically 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
  - Data Collection Software (e.g., REDCap)

FOLDER

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

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********

****

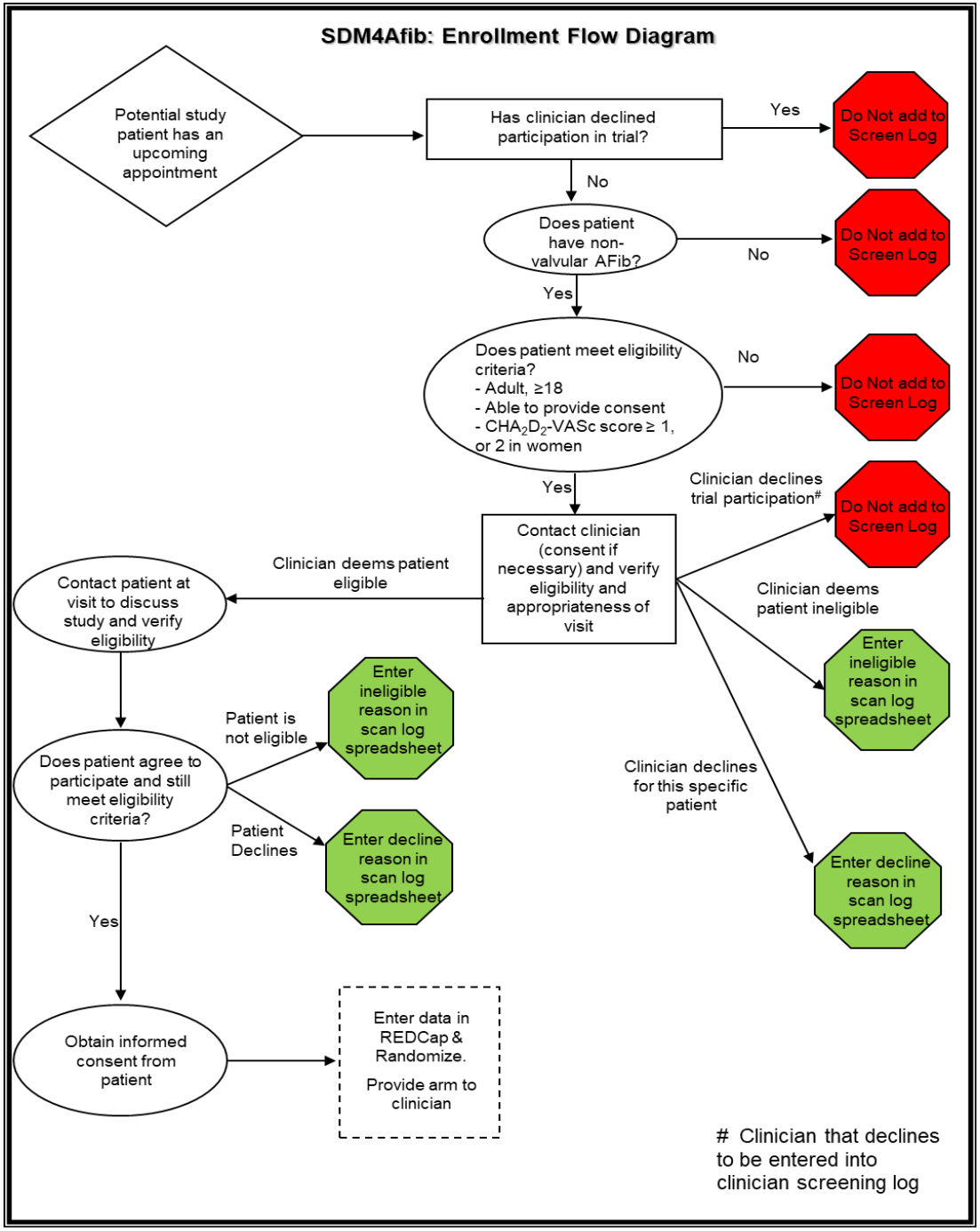
Protocol, intake form, SOW, COMIRB, All analytical plans are to be stored in this folder and each file lised.

****

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.


SDM4Afib: Enrollment Flow Diagram
Potential study patient has an upcoming appointment
Has clinician declined participation in trial?
Yes
Do Not add to Screen Log
No
Does patient have non-valvular AFib?
No
Do Not add to Screen Log
Yes
Does patient meet eligibility criteria?
- Adult, ≥18
- Able to provide consent
- CHA2D2-VASc score ≥ 1, or 2 in women
No
Do Not add to Screen Log
Yes
Contact clinician (consent if necessary) and verify eligibility and appropriateness of visit
Clinician declines trial participation#
Do Not add to Screen Log
Clinician deems patient eligible
Contact patient at visit to discuss study and verify eligibility
Clinician deems patient ineligible
Enter ineligible reason in scan log spreadsheet
Patient is not eligible
Enter ineligible reason in scan log spreadsheet
Does patient agree to participate and still meet eligibility criteria?
Clinician declines for this specific patient
Patient Declines
Enter decline reason in scan log spreadsheet
Enter decline reason in scan log spreadsheet
Yes
Obtain informed consent from patient
Enter data in REDCap & Randomize.
Provide arm to clinician
Clinician that declines to be entered into clinician screening log

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.

Field Name	Comment
Enrolled Patient ID	This is the ID generated by the study coordinator.
Staff Initials	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 (-).
Clinic Location	Choose enrolling site
Medication Cohort	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.
CHADS-VASc Score	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.
Arm Assignment^	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.
Signed consent form	Upload a .pdf or .doc copy of the patients signed consent form.
Patient eligibility CRF	Provided a .pdf or .doc of the completed eligibility CRF.
Patient excluded after randomization	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.

Welcome to REDCap!

REDCap is a secure web platform for building and managing online databases and surveys. REDCap's streamlined process for rapidly creating and designing projects offers a vast array of tools that can be tailored to virtually any data collection strategy.

REDCap provides automated export procedures for seamless data downloads to Excel and common statistical packages (SPSS, SAS, Stata, R), as well as a built-in project calendar, a scheduling module, ad hoc reporting tools, and advanced features, such as branching logic, file uploading, and calculated fields.

Learn more about REDCap by watching a 🎞 brief summary video (4 min). If you would like to view other quick video tutorials of REDCap in action and an overview of its features, please see the Training Resources page.

Please note that any publication that results from a project utilizing REDCap should cite grant support (**NIH/NCATS Colorado CTSA Grant Number UL1 TR002535**).

NOTICE: If you are collecting data for the purposes of human subjects research, review and approval of the project is required by your Institutional Review Board.

If you require assistance or have any questions about REDCap, please contact REDCap Admin.

For REDCAP TIPS, click here to visit the UC Denver REDCap Information website.

REDCap Features

Build online surveys and databases quickly and securely in your browser - Create and design your project using a secure login from any device. No extra software required. Access from anywhere, at any time.

Fast and flexible - Go from project creation to starting data collection in less than one day. Customizations and changes are possible any time, even after data collection has begun.

Advanced instrument design features - Auto-validation, calculated fields, file uploading, branching/skip logic, and survey stop actions.

e-Consent - Perform informed consent electronically for participants via survey.

Diverse and flexible survey distribution options - Use a list of email addresses or phone numbers for your survey respondents and automatically contact them with personalized messages, and track who has responded. Or create a simple link for an anonymous survey for mass email mailings, to post on a website, or print on a flyer.

REDCap Mobile App - Collect data offline using an app on a mobile device when there is no WiFi or cellular connection, and then later sync data back to the server.

Data quality - Use field validation, branching/skip logic, and Missing Data Codes to improve and protect data quality during data entry. Open data queries to automatically identify and resolve discrepancies and other issues real-time.

Custom reporting - Create custom searches for generating reports to view aggregate data. Identify trends with built-in basic statistics and charts.

Export data to common analysis packages - Export your data as a PDF or as CSV data for easy analysis in SAS, Stata, R, SPSS, or Microsoft Excel.

Secure file storage and sharing - Upload and share any type of file with anyone in the world through the File Repository feature or Send-It tool. Also works with exports and other built-in file uploading features.

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 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

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

Poll: How long do you have to store study material (Period of retention) per AMC guidelines?

CONSIDERATIONS WHEN STORING DATA/RESEARCH

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

- 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…

University of Colorado Denver | Anschutz Medical Campus
Record Retention Matrix
10/4/2017

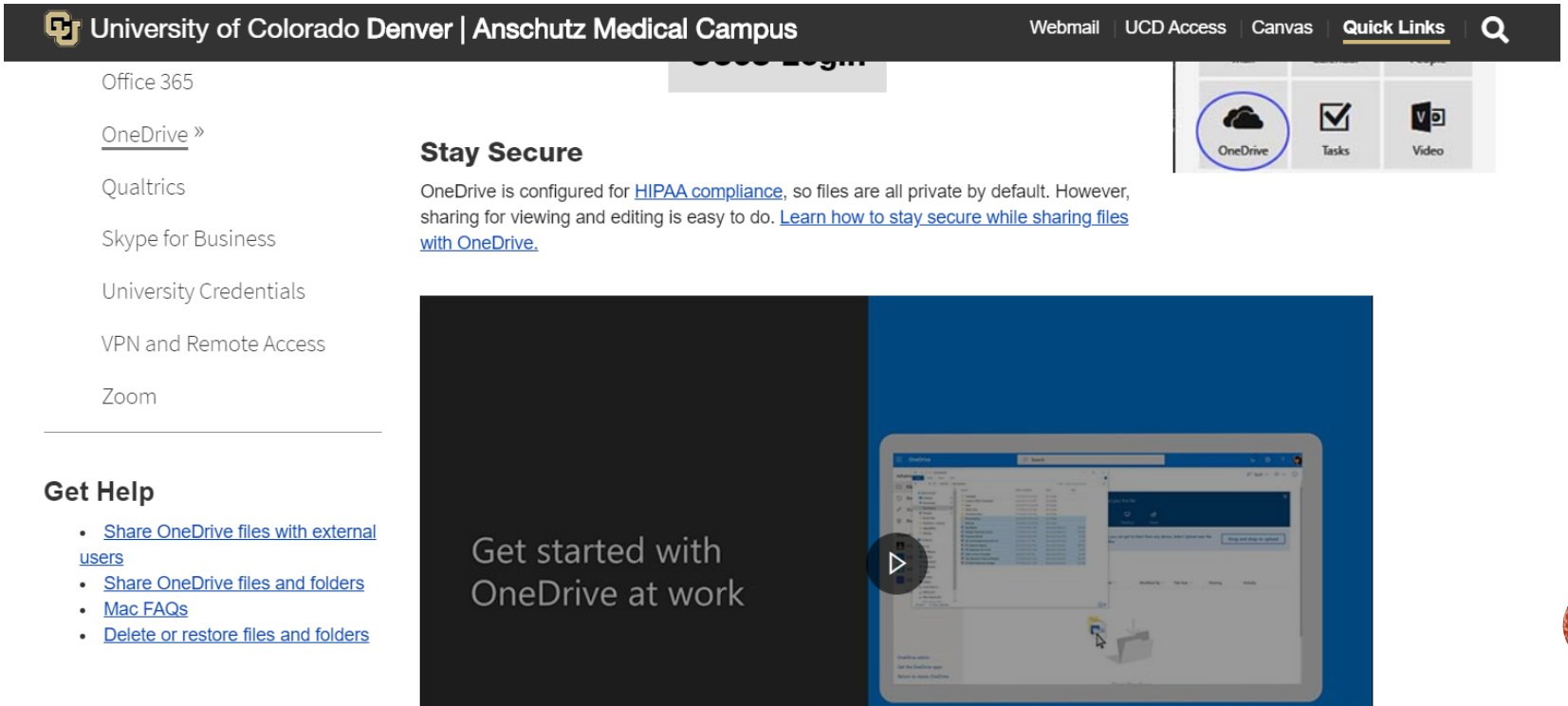
DocumentType		Repository	Retention Period	Related Authority
	Loan Records	Bursars Office	3 years after pay-off	State Archives 34 CFR Sec. 74.53
Tax				
	1098-T	Bursar Office	4 years	State Archives 34 CFR Sec. 74.53
Grant and Research Records				
Clinical Research Records Protocols Patient Records Regulatory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

Poll: Where can study data be safely stored?

https://www.cu.edu/sites/default/files/RecordRetentionUCD.pdf

MAKE SURE YOU VERIFY!

- Email safe guards: https://www.ucdenver.edu/offices/office-of-information-technology/secure-campus/encryption
- OneDrive: https://www.ucdenver.edu/offices/office-of-information-technology/software/how-do-i-use/onedrive



DATA OWNERSHIP / SHARING

Data Acquisition

OWNERSHIP/DATA SHARING

Who owns the data?

- Researche
- Funders
  - Grants v
- Data Sour
  - Subjects
  - Countrie
- Resea
  Institutions

"for the most part, NIH makes awards to institutions and not individuals"
– *NIH Data Sharing Policy and Implementation Guidance*

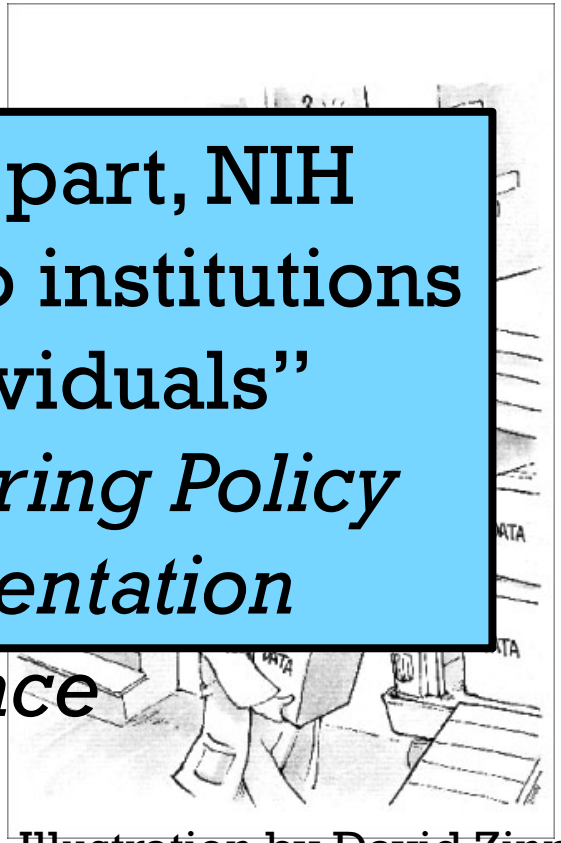


Illustration by David Zinn

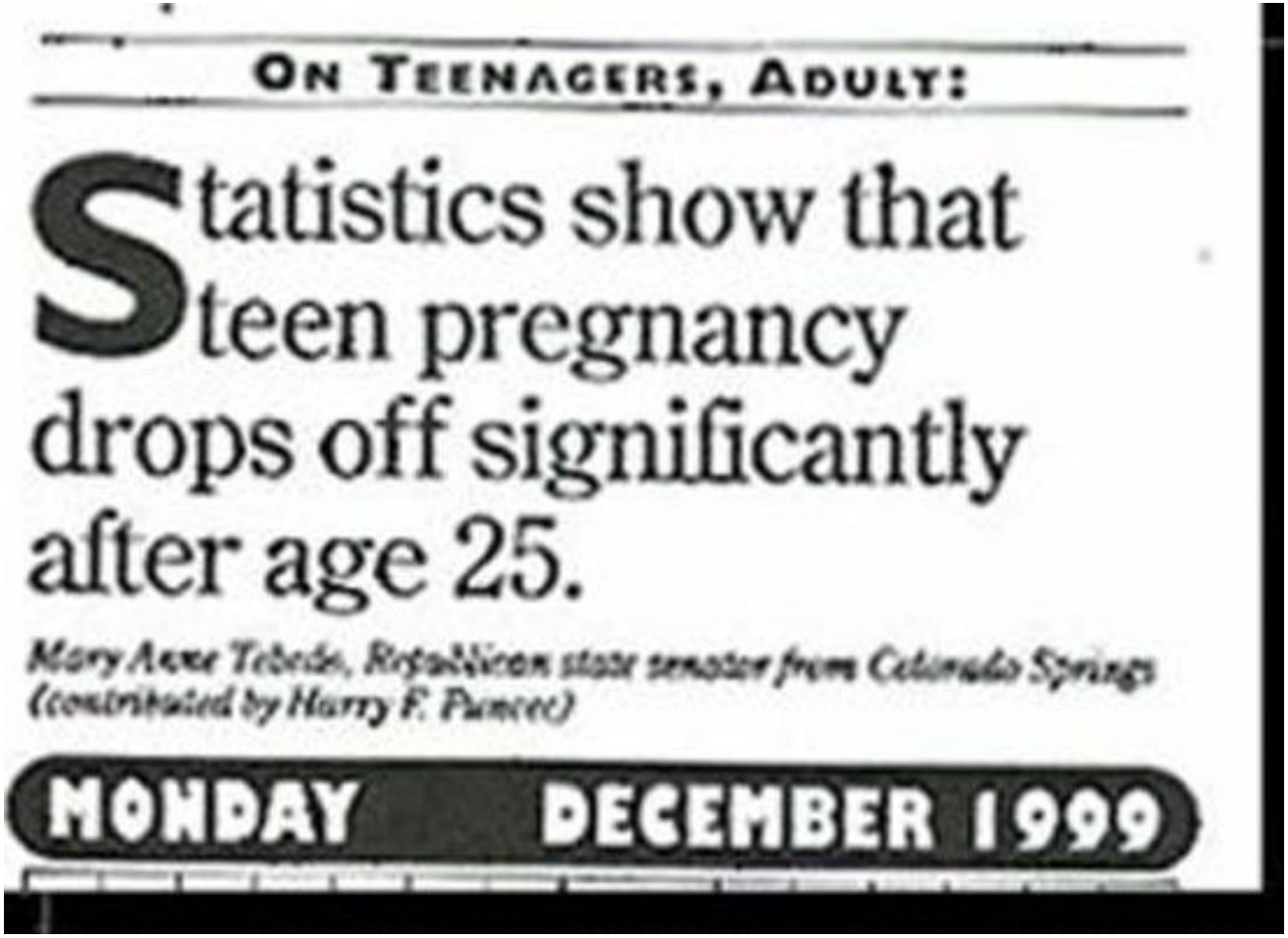
Taken from *ORI Introduction to RCR* (http://ori.hhs.gov/education/products/RCRintro/)

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."



Statistical Analysis

TIPS FOR REPRODUCIBLE STATISTICAL ANALYSES

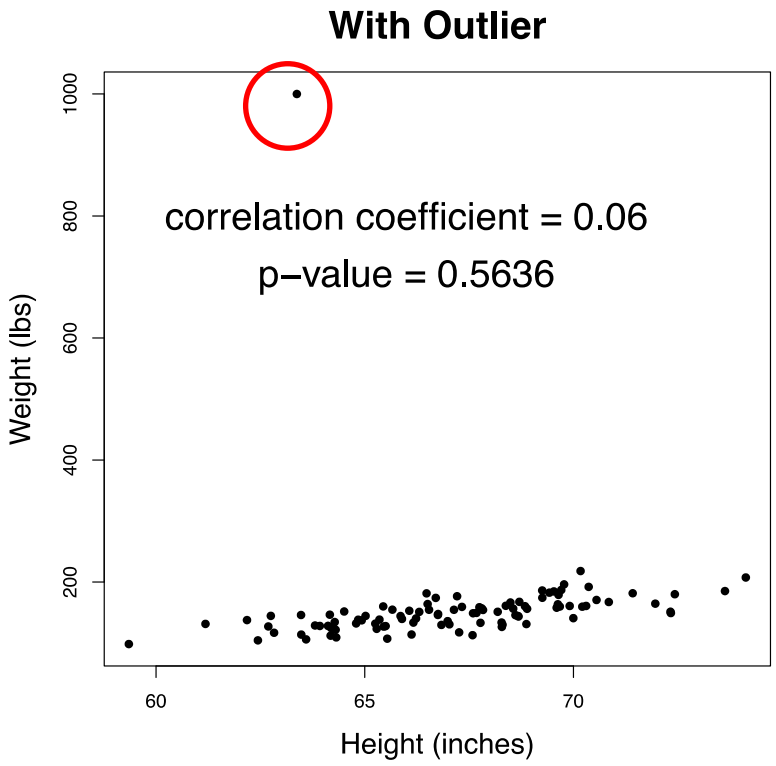
1. ALWAYS keep a version of the "most raw" data
   - Record when and where it was created, so you can easily tell if it has been changed since creation
2. Version Control
   - GitLab
3. Use a scripting language
   - Programs like R and SAS allow you to follow your steps exactly if you (or someone else) 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

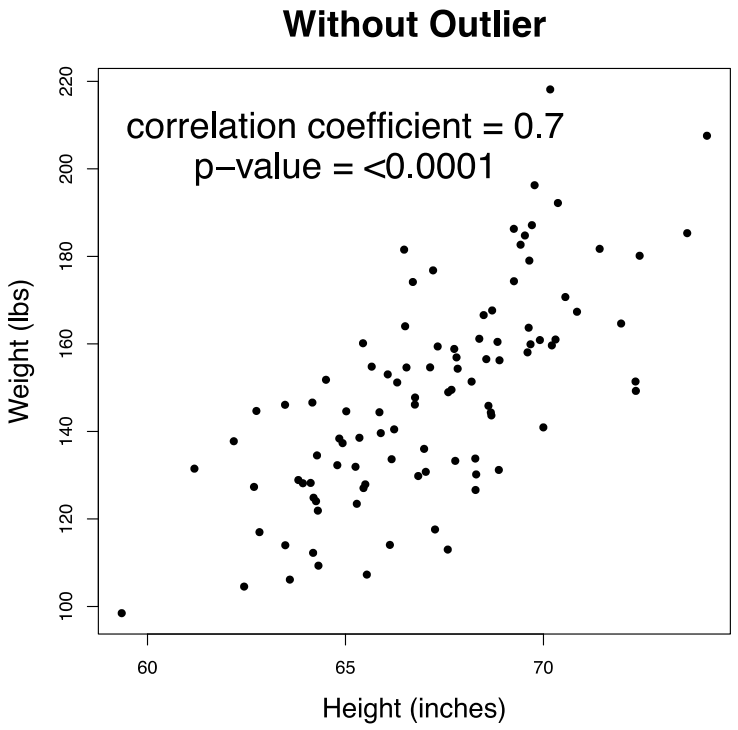
OUTLIERS

Statistical Analysis

Sign the Attendance Sheet now by clicking the link in the Chat.

OUTLIERS


With Outlier
correlation coefficient = 0.06
p–value = 0.5636
Weight (lbs)
1000
800
600
400
200
60
65
70
Height (inches)


Without Outlier
correlation coefficient = 0.7
p–value = <0.0001
Weight (lbs)
220
200
180
160
140
120
100
60
65
70
Height (inches)

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

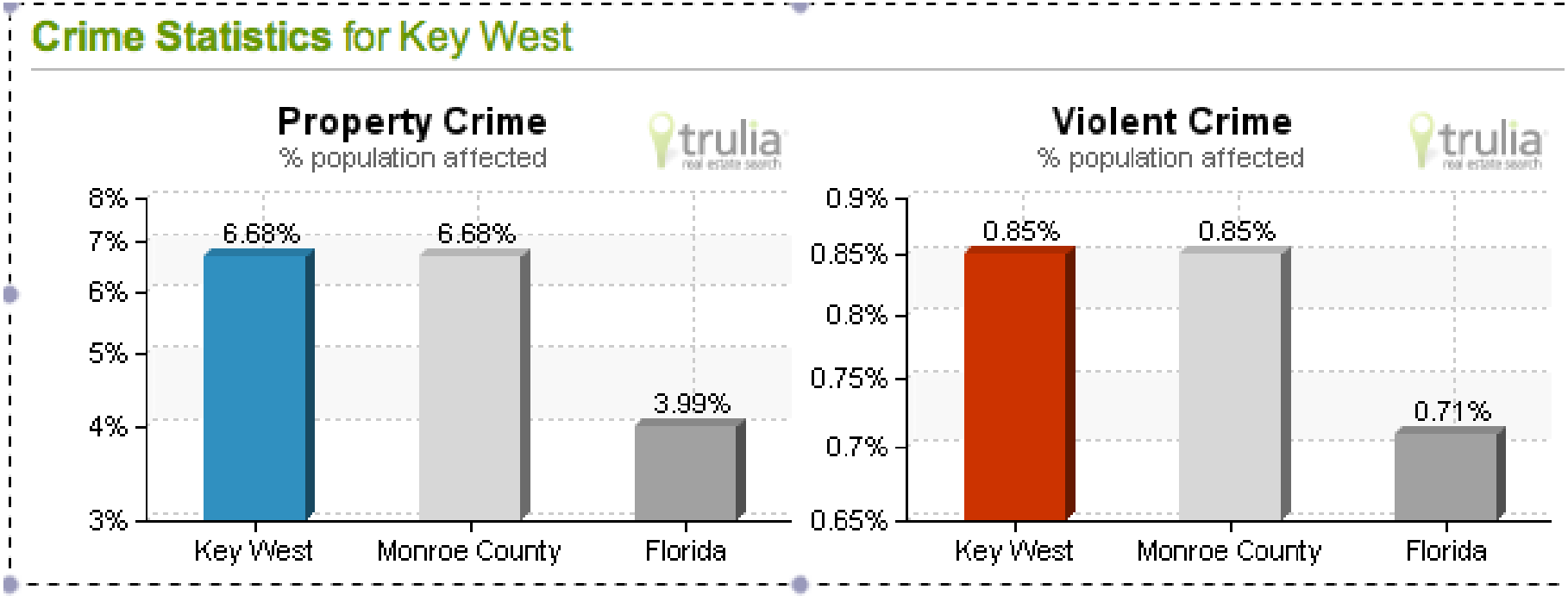
- 20 questions (Likert scale)
- 10 open-ended questions
- 480 surveys administered (320 responses)

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

For more info: "Caring about carelessness: Participant inattention and its effects on research"

DISPLAYING RESULTS



IS THERE VALUE IN P-VALUES?

Poll: What is the definition of a p-value?

- P-values tell us how likely you are to get the same result assuming that your null hypothesis is true
- They are commonly misinterpreted
- In most cases you cannot directly compare p-values
- You can compare effect sizes and these are usually more relevant when discussing your results

P-HACKING

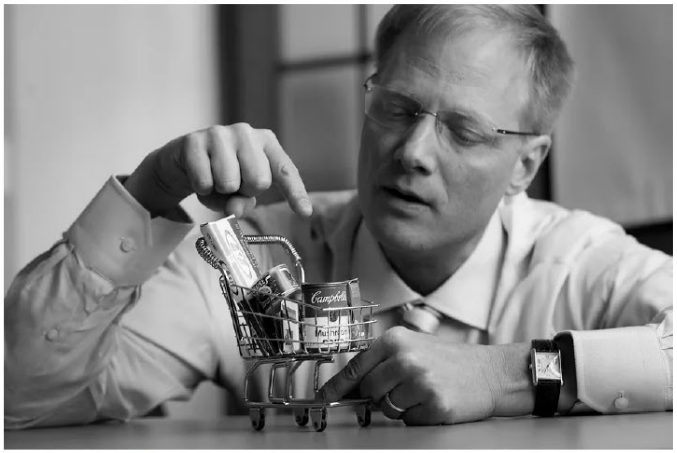
11/5/2018 A top Cornell food researcher has had 15 studies retracted. That's a lot. - Vox

Vox

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



Brian Wansink just had six papers retracted from top journals. | Jason Koski

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.'

HELPFUL RESOURCES

- UCLA IDRE - Statistical methods across multiple programming languages with real world problems with explanations of code and interpretation of output
  - https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

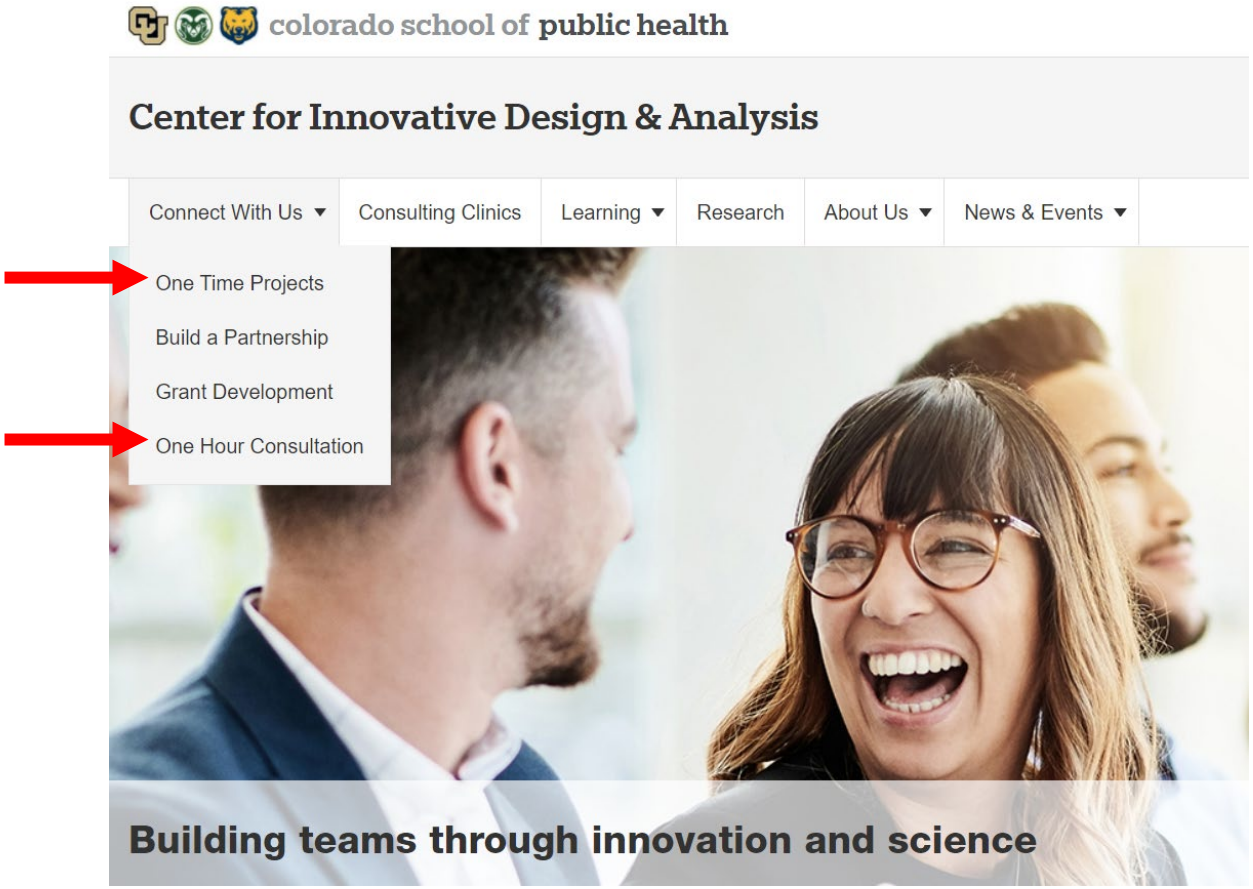
- Nature – Points of Significance Articles
  - https://www.nature.com/collections/qghhqm/pointsofsignificance

KEY ITEMS TO REMEMBER

- Do not do anything to your data that you are not willing to explain in a publication!
- Document everything
- Store your data securely
- Plan analyses before you collect data
- Know your assumptions with your data
- Do not work as an island – when in doubt, ask!

NEED MORE STATS HELP?

- https://coloradosph.cuanschutz.edu/research-and-practice/centers-programs/cida



ACKNOWLEDGEMENTS/REFERENCES

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

Nassia Dunn

Dr. Brandie Wagner

Megan Branda

https://research.cuanschutz.edu/regulatory-compliance

References

Responsible Conduct of Research by Adil E Shamoo and David B. Resnick. Second Ed. Oxford University Press, 2009.

NIH Data Sharing Policy and Implementation Guidance (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm), March 5, 2003.

Ethical Guidelines for Statistical Practice, American Statistical Association (http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx), April 2016.

Introduction to RCR – 6. Data Management Practices, Office of Research Integrity, US Department of Health and Human Services (http://ori.hhs.gov/education/products/RCRintro/c06/0c6.html), Revised Edition August 2007