DATA ACQUISITION AND MANAGEMENT

Matt Bolt, MS
Center for Innovative Design & Analysis (CIDA)
Department of Biostatistics and Informatics
Colorado School of Public Health
matthew.bolt@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 the statement that best describes you:
  - I am new to biomedical research
  - I have some previous experience with biomedical research
  - I have several years of biomedical research experience

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

SOURCE OF COURSE MATERIAL

- U.S. Department of Health & Human Services (executive umbrella over NIH, CDC, FDA, etc.)
  - Office of Research Integrity (ORI)
- ***ORI Introduction to the Responsible Conduct of Research***
- https://ori.hhs.gov/education/products/RCRintro/

THEMES OF PRESENTATION

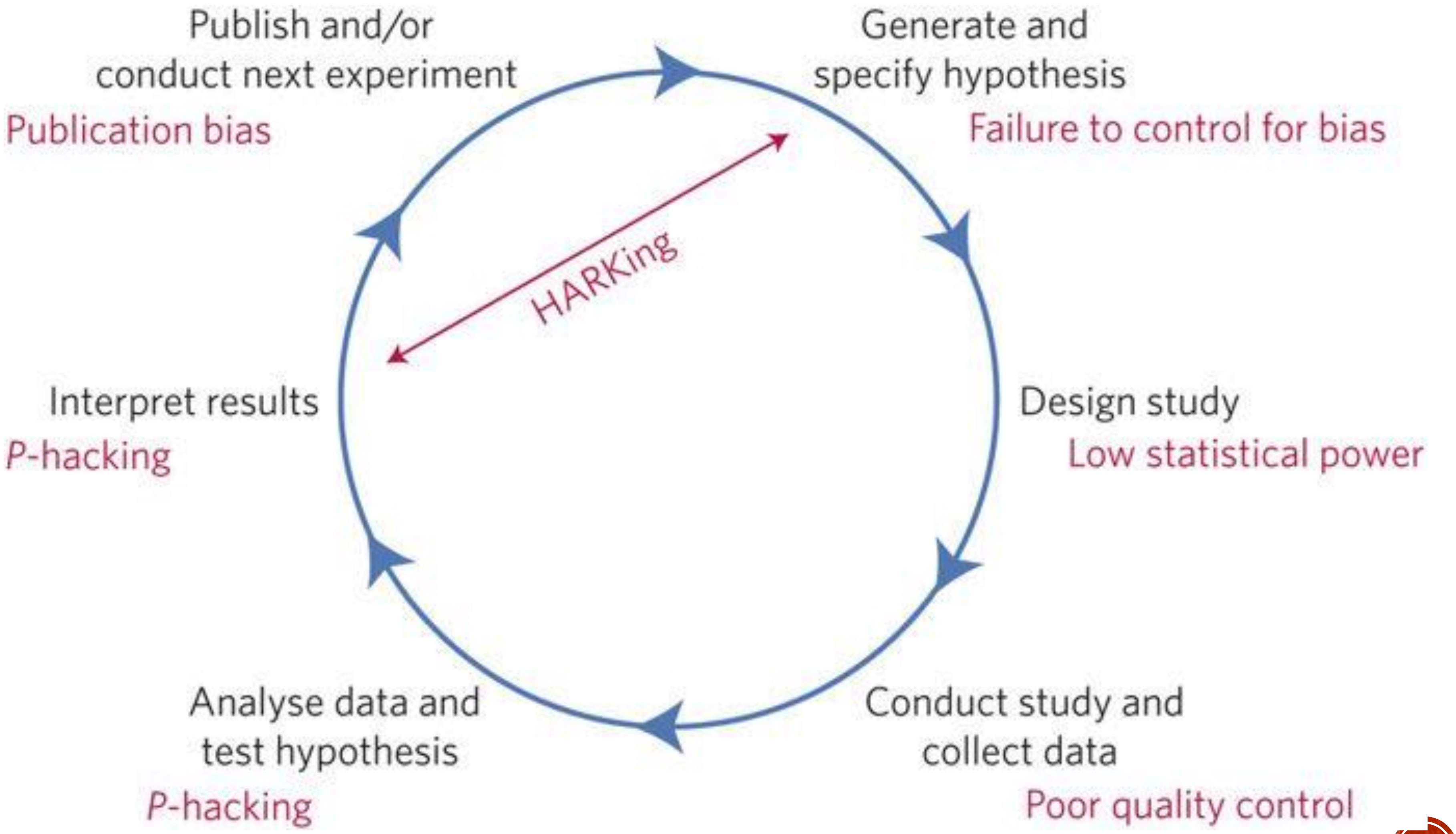
- **Replicability** is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

- **Reproducibility** is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis.

- **Generalizability** is the extent that results of a study apply in other contexts or populations that differ from the original one.

Reproducibility and Replicability in Science. National Academies, 2019.

THREATS TO REPLICABLE SCIENCE



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

Publish and/or conduct next experiment

Publication bias

HARKing

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

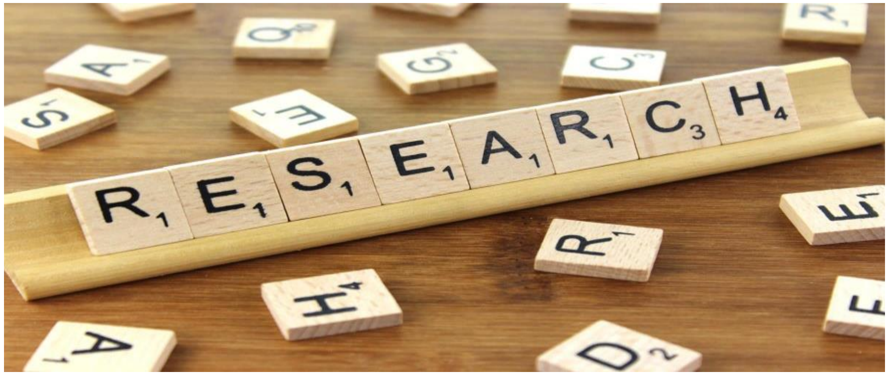
THREATS TO REPLICABLE/REPRODUCIBLE SCIENCE

- Sources of non-replicability/non-reproducibility:
  - Inadequate recordkeeping
  - Selective reporting and/or analysis
  - Cultural barriers
  - Obsolescence of digital artifacts
  - Lack of expertise to reproduce other's research

Reproducibility and Replicability in Science. National Academies, 2019.

- "…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)



Poll: How reproducible has research been?

A FEW EXAMPLES FROM THE LITERATURE

- Data access for reproducing results [is] available in 20%-25% of a sample of studies in biology.

  Prinz et al. (2011)

- From NIH-funded studies, only 12% [of studies] provided datasets in recognized repositories.

  Read et al. (2015)

- Out of 268 manuscripts on biomedical research, none provided access to all raw data used in the study.

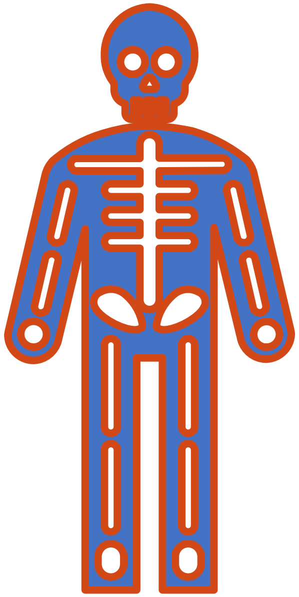
  Iqbal et al. (2016)

- Evaluation of the *PLOS ONE* open data policy revealed that 20% of published studies still had restrictions on data access.

  Byrne (2017)

OBJECTIVES OF PRESENTATION

- Specimens *(Replicability)*
  - Common Errors
  - Guidance on best practices
- Data Collection, Storage, and Ownership *(Replicability)*
  - 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 work and analysis *(reproducibility)*
  - Guiding principals for analysis
- Pitfalls



SPECIMENS

Collection Considerations for Replicable Results

Slides provided by Nassia Duncan, Lab Manager

CATASTROPHIC SCENARIOS: PART 1

- 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.

CATASTROPHIC SCENARIOS: PART 2

- 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. Why are these results not replicable?
- 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. Why are these results 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. Why are these results 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?
  - 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.

QUESTIONS TO CONSIDER

- 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 Collection

DATA COLLECTION

1. Are there guidelines/regulations you need to follow?
   - HIPAA, hazardous materials, copyrights, consent forms
2. Use appropriate methods!

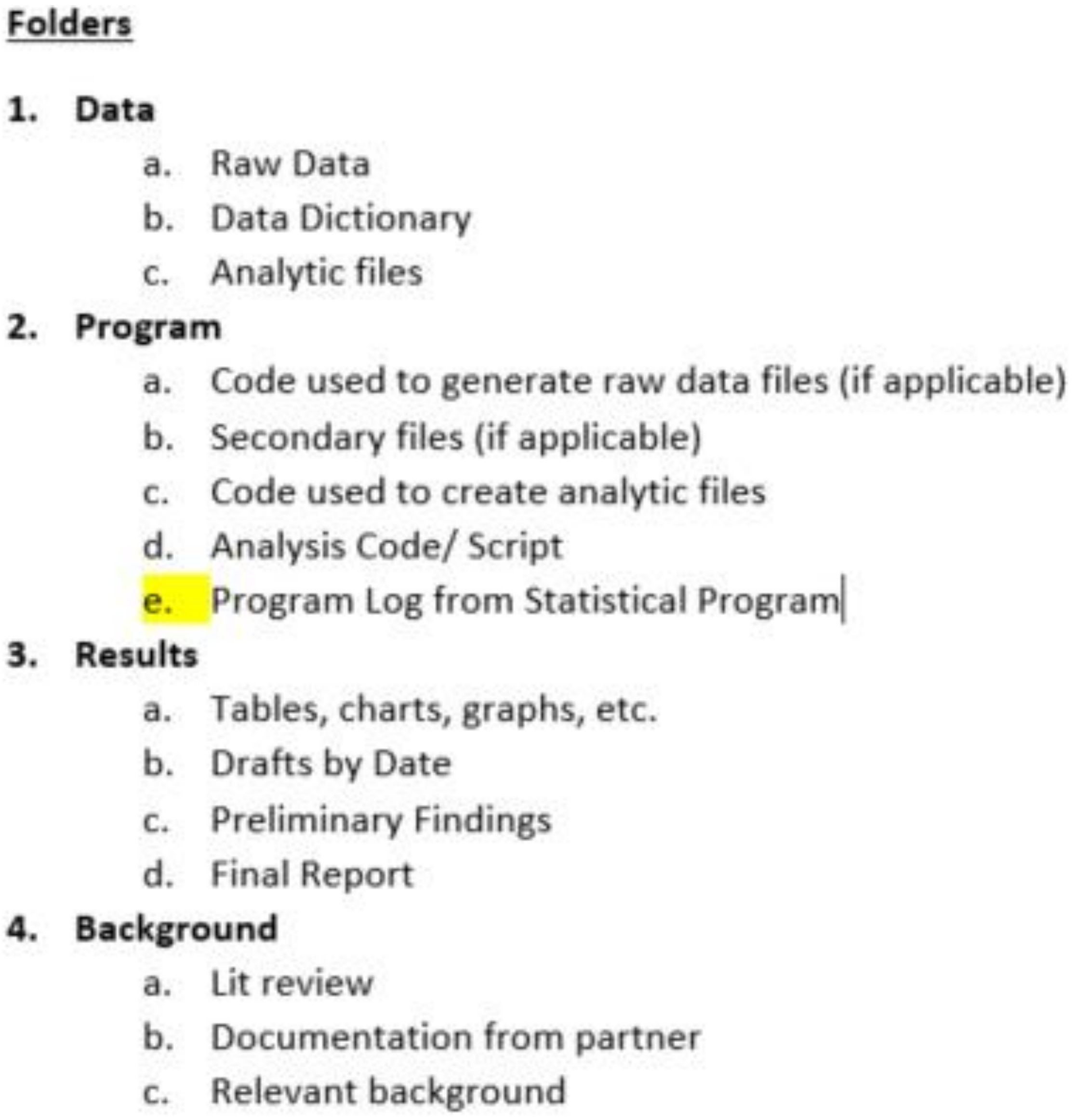
   "Failure to find the effect could be due to either your experimental design or the lack of an effect, but you will not know which is true."
3. Define data items and guidelines for recording them
4. Determine the physical process of recording the data
   - Notebook (hard copy)
   - Computer file or electronic notebook (electronic copy)

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

DO NOT RECREATE THE WHEEL

- When starting a new project use existing resources or a colleague!
  - What resources are publicly available or provided by your institution?
  - What processes do we know that *don't* work?
    - Search terms: Data collection, best practices, Data management, Guidelines, lessons learned
    - Experimental Design expertise
- Organization (next several slides)
  - Folder Organization
  - ReadME
  - Data Dictionary
  - Data Collection SOP
  - Defined Roles and Responsibilities
  - Electronic Laboratory Notebooks
  - Data Collection Software (e.g., REDCap)
- Make sure your data is usable for analysis
  - (European Spreadsheet Risks Interest Group)

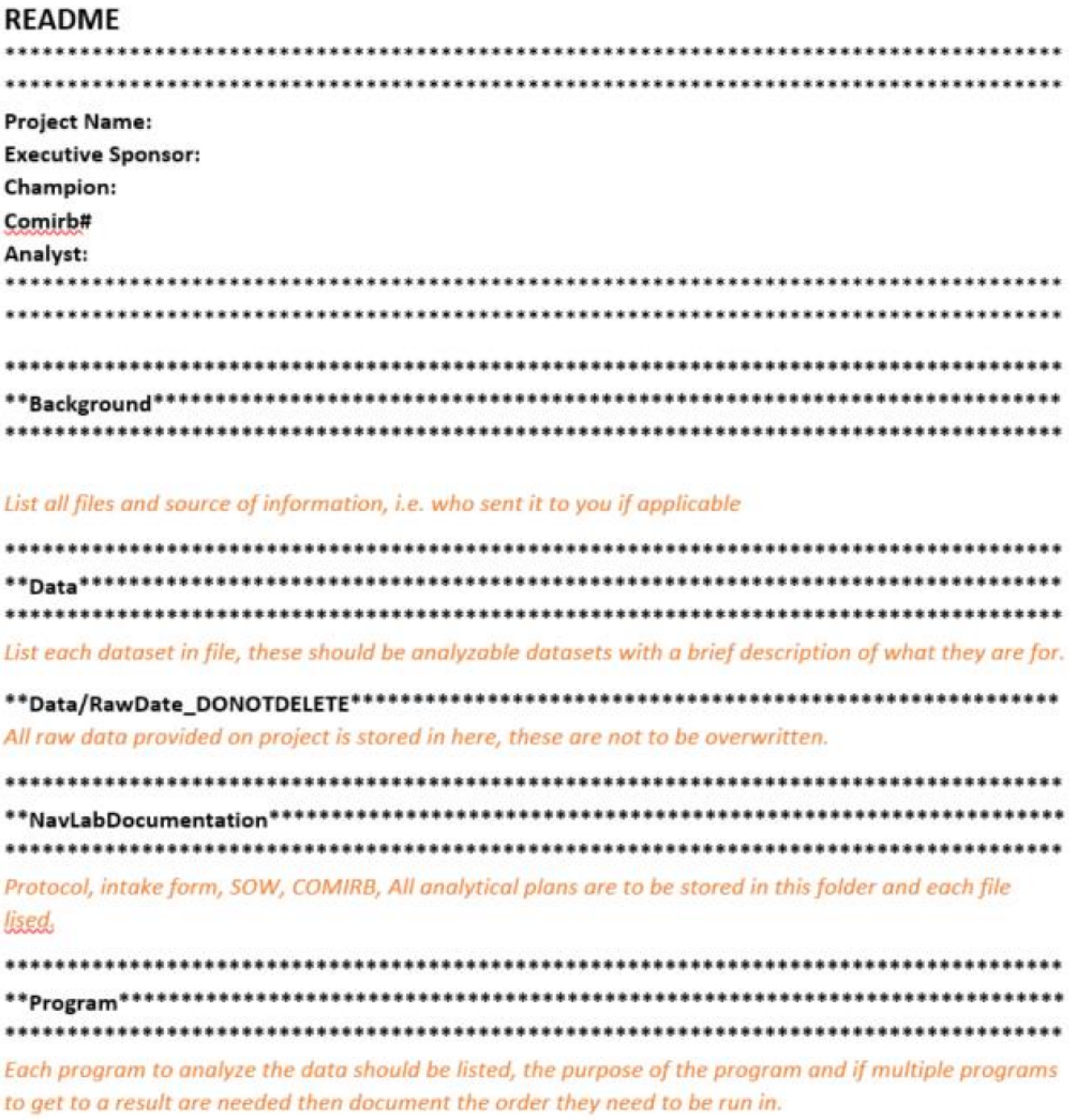
FOLDER ORGANIZATION



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.

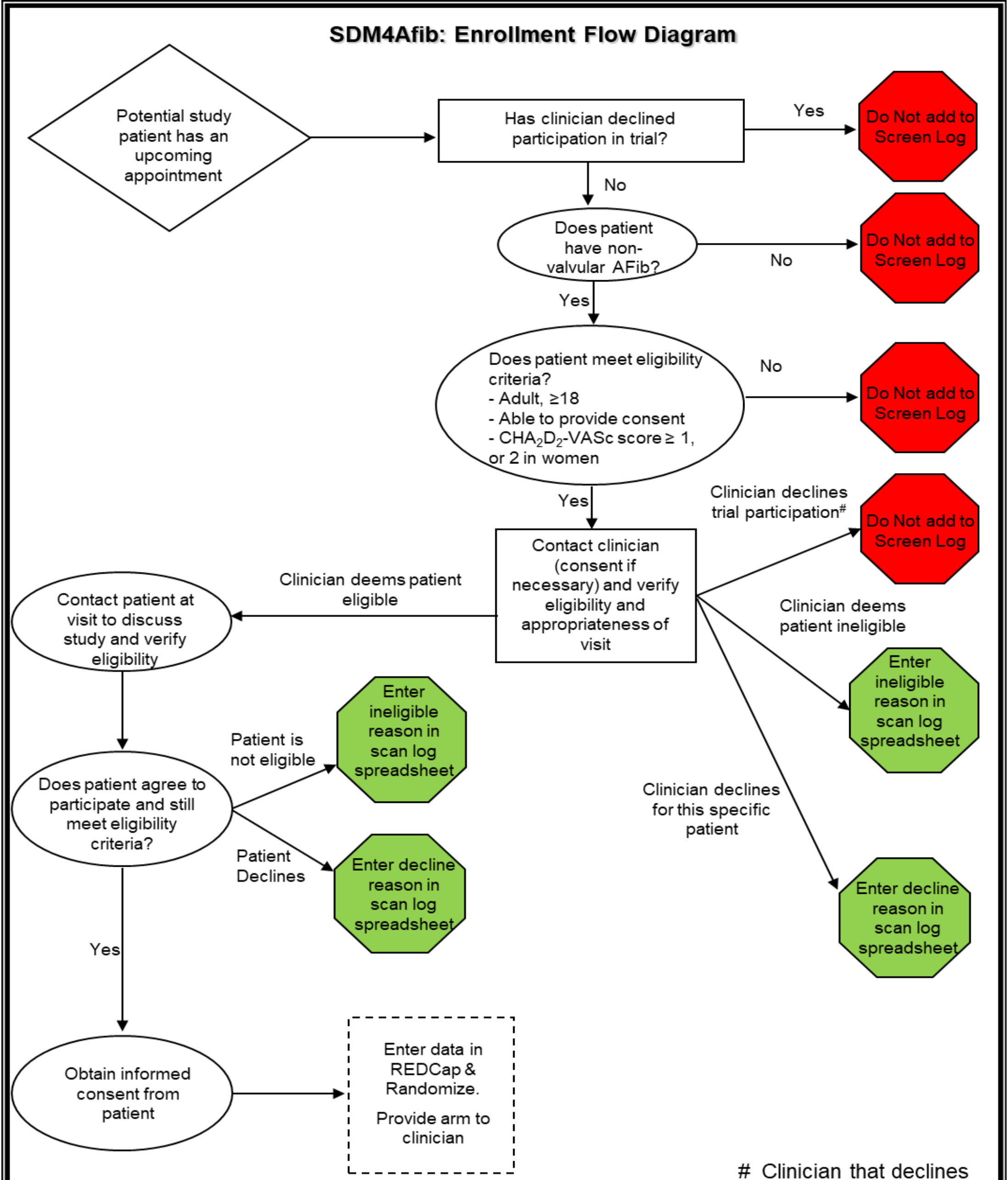
****

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
- CHA_2D_2-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 ineligible

Enter ineligible reason in scan log spreadsheet

Clinician declines for this specific patient

Enter decline reason in scan log spreadsheet

Clinician deems patient eligible

Contact patient at visit to discuss study and verify eligibility

Does patient agree to participate and still meet eligibility criteria?

Patient is not eligible

Enter ineligible reason in scan log spreadsheet

Patient Declines

Enter decline reason in scan log spreadsheet

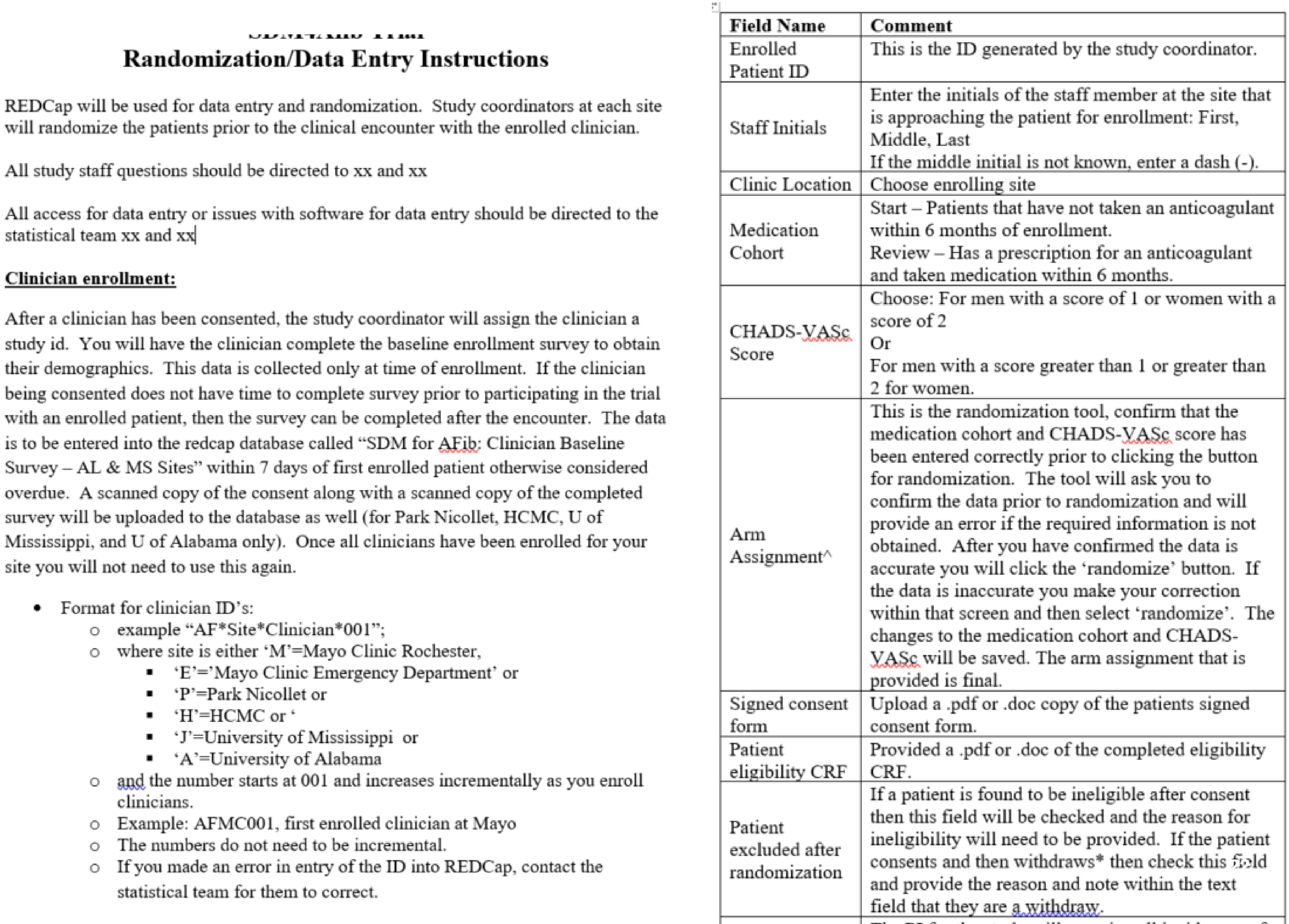
Yes

Obtain informed consent from patient

Enter data in REDCap & Randomize.

Provide arm to clinician

Clinician that declines to be entered into clinician screening log



SDM4Afib Trial

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.

ELECTRONIC LABORATORY NOTEBOOKS (ELN) & LABARCHIVES

- Electronic data management platforms that serve as repositories and organizational tools for research data
- Easier to follow data management plan
- Control access, organization, sharing, and archiving
- Secure
- LabArchives is provided to Anschutz faculty and staff for free, might require fee for students
  - Secure and cloud-based
  - Discipline agnostic
  - Shared notebooks
  - Enables compliance with data management requirements

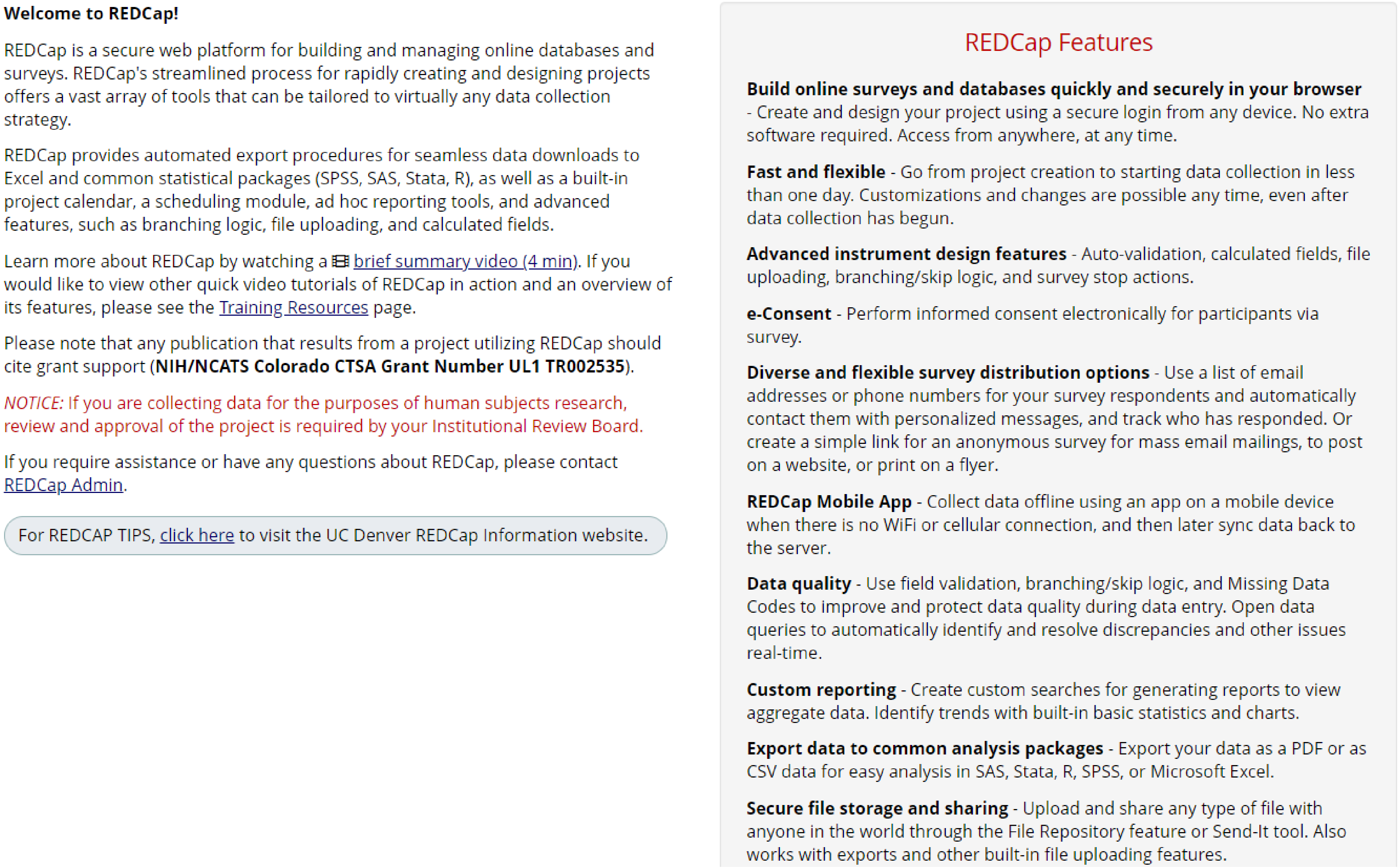
LABARCHIVES RESOURCES

- Website: https://research.cuanschutz.edu/university-research/labarchives
- Support: support@labarchives.com
- Tutorials, articles, etc: https://www.labarchives.com/labarchives-knowledge-base/
- Lunch and learn video: https://www.youtube.com/watch?v=eDYm-nQTAX4

DATA COLLECTION CASE STUDY
FROM *RESPONSIBLE CONDUCT OF RESEARCH*

- Dr. Z is mentoring a medical student over the summer in his research lab
- Student's project
  - Cancer cell line that requires 3 weeks to grow to test for a specific antibody
  - Student has already written a short paper on his work
- Dr. Z's dilemma:
  - Raw data on pieces of yellow pads without clear identification of the experiment from which the data came
  - Some experiments were repeated several times without explanation
  - Doesn't want to discourage student from pursuing 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?



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.

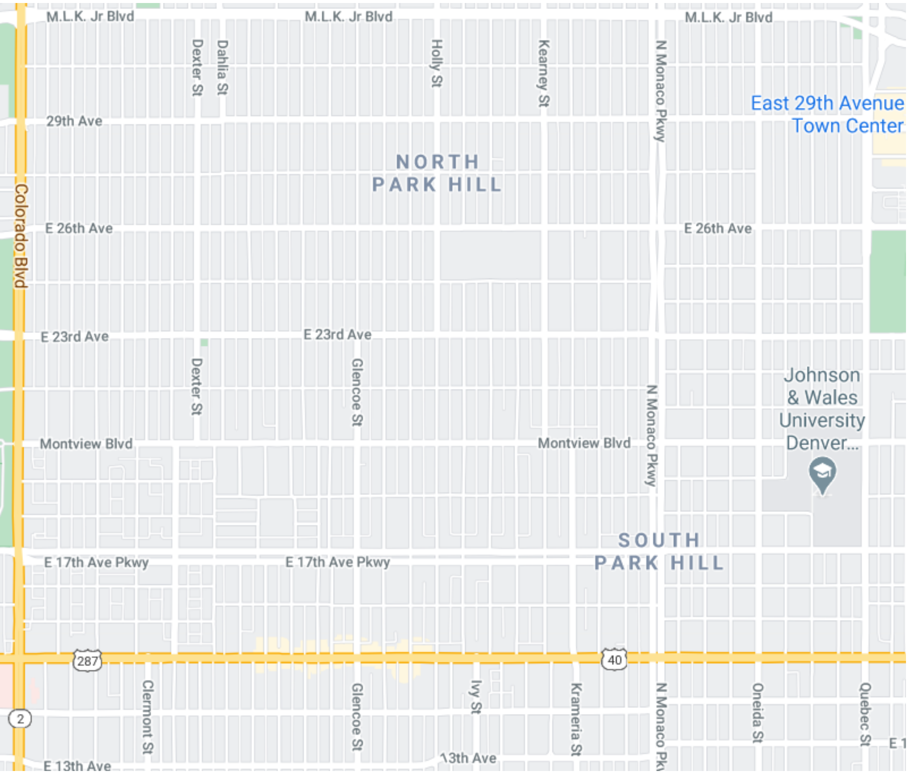
Data quality - Use field validation, branching/skip logic, and Missing Data Codes to improve and protect data quality during data entry. Open data queries to automatically identify and resolve discrepancies and other issues real-time.

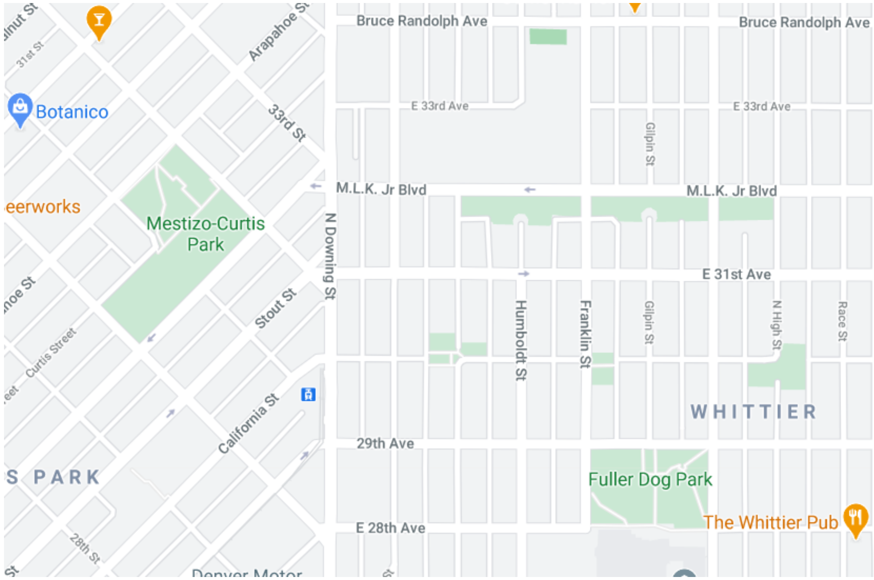
Custom reporting - Create custom searches for generating reports to view aggregate data. Identify trends with built-in basic statistics and charts.

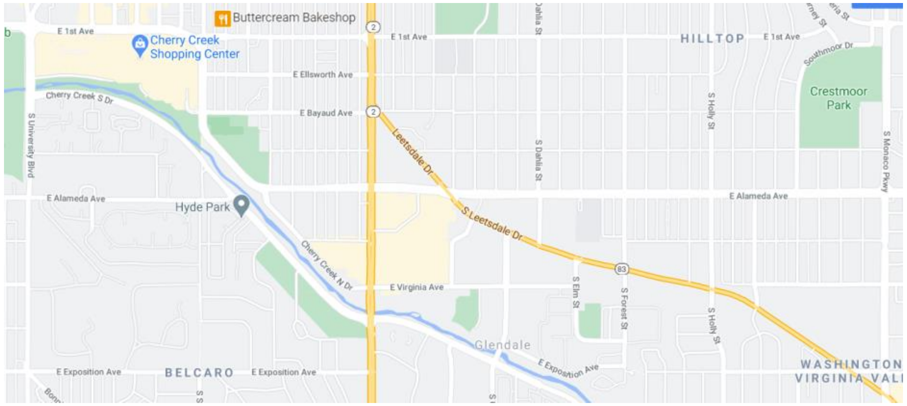
Export data to common analysis packages - Export your data as a PDF or as CSV data for easy analysis in SAS, Stata, R, SPSS, or Microsoft Excel.

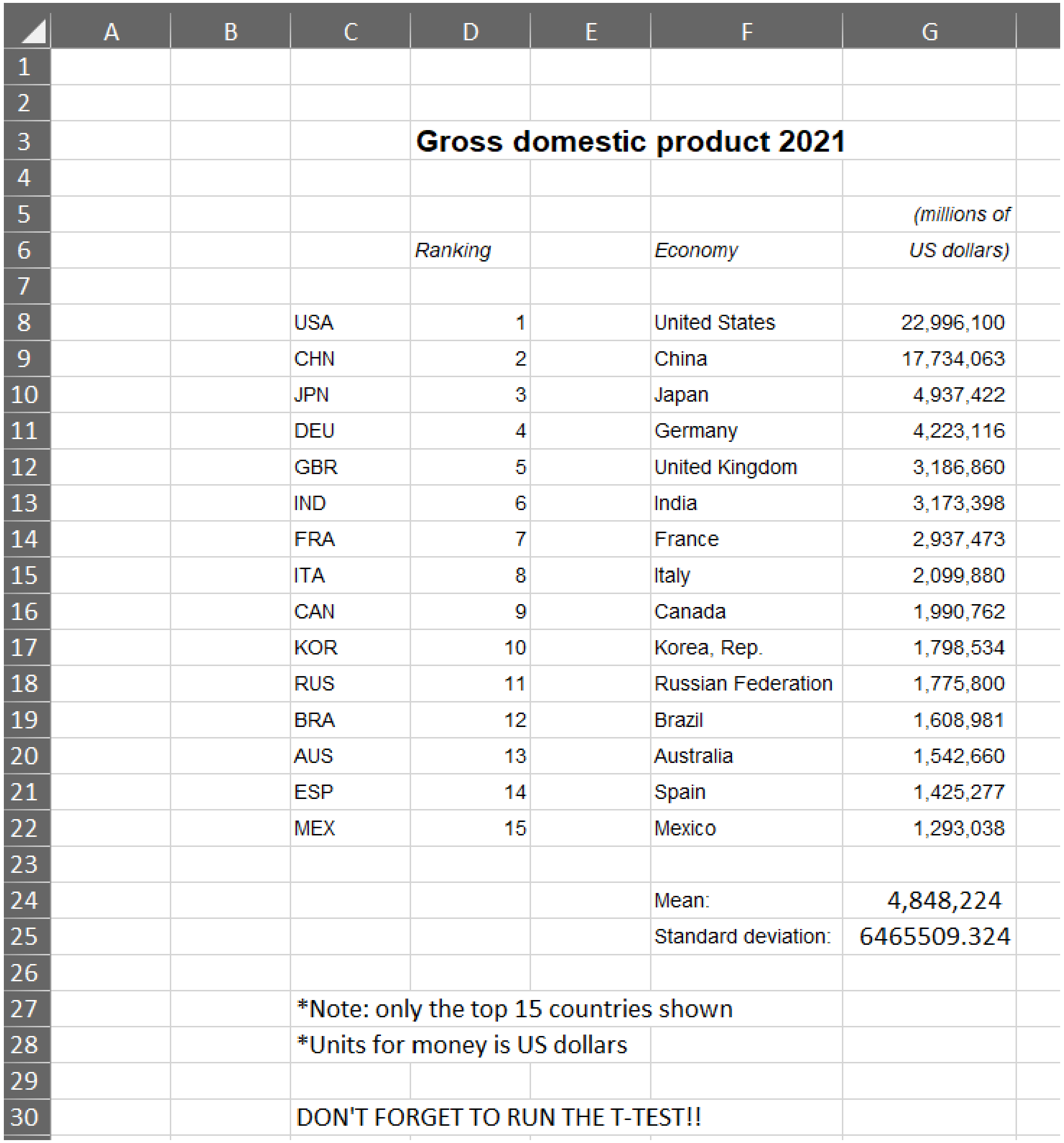
Secure file storage and sharing - Upload and share any type of file with anyone in the world through the File Repository feature or Send-It tool. Also works with exports and other built-in file uploading features.

REDCAP AND TIDY DATA


M.L.K. Jr Blvd
M.L.K. Jr Blvd
M.L.K. Jr Blvd
Dexter St
Dahlia St
Holly St
Kearney St
N Monaco Pkwy
East 29th Avenue Town Center
29th Ave
NORTH PARK HILL
Colorado Blvd
E 26th Ave
E 26th Ave
E 23rd Ave
E 23rd Ave
Dexter St
Glencoe St
N Monaco Pkwy
Johnson & Wales University Denver...
Montview Blvd
Montview Blvd
SOUTH PARK HILL
E 17th Ave Pkwy
E 17th Ave Pkwy
287
40
2
Clermont St
Glencoe St
Ivy St
Krameria St
N Monaco Pk
Oneida St
Quebec St
E 13th Ave
13th Ave


31st St
Arapahoe St
Botanico
33rd St
eerworks
Mestizo-Curtis Park
Stout St
Curtis Street
California St
S PARK
28th St
Denver Motor
N Downing St
Bruce Randolph Ave
Bruce Randolph Ave
E 33rd Ave
E 33rd Ave
Gilpin St
M.L.K. Jr Blvd
M.L.K. Jr Blvd
E 31st Ave
Humboldt St
Franklin St
Gilpin St
N High St
Race St
WHITTIER
29th Ave
Fuller Dog Park
The Whittier Pub
E 28th Ave


Buttercream Bakeshop
Cherry Creek Shopping Center
E 1st Ave
E 1st Ave
E 1st Ave
HILLTOP
Southmoor Dr
Crestmoor Park
E Ellsworth Ave
E Bayaud Ave
Cherry Creek S Dr
S University Blvd
Leetsdale Dr
S Dahlia St
S Holly St
S Monaco Pkwy
E Alameda Ave
E Alameda Ave
Hyde Park
S Leetsdale Dr
Cherry Creek N Dr
E Virginia Ave
S Elm St
S Forest St
S Holly St
83
2
Glendale
E Exposition Ave
E Exposition Ave
E Exposition Ave
BELCARO
WASHINGTON VIRGINIA VALE



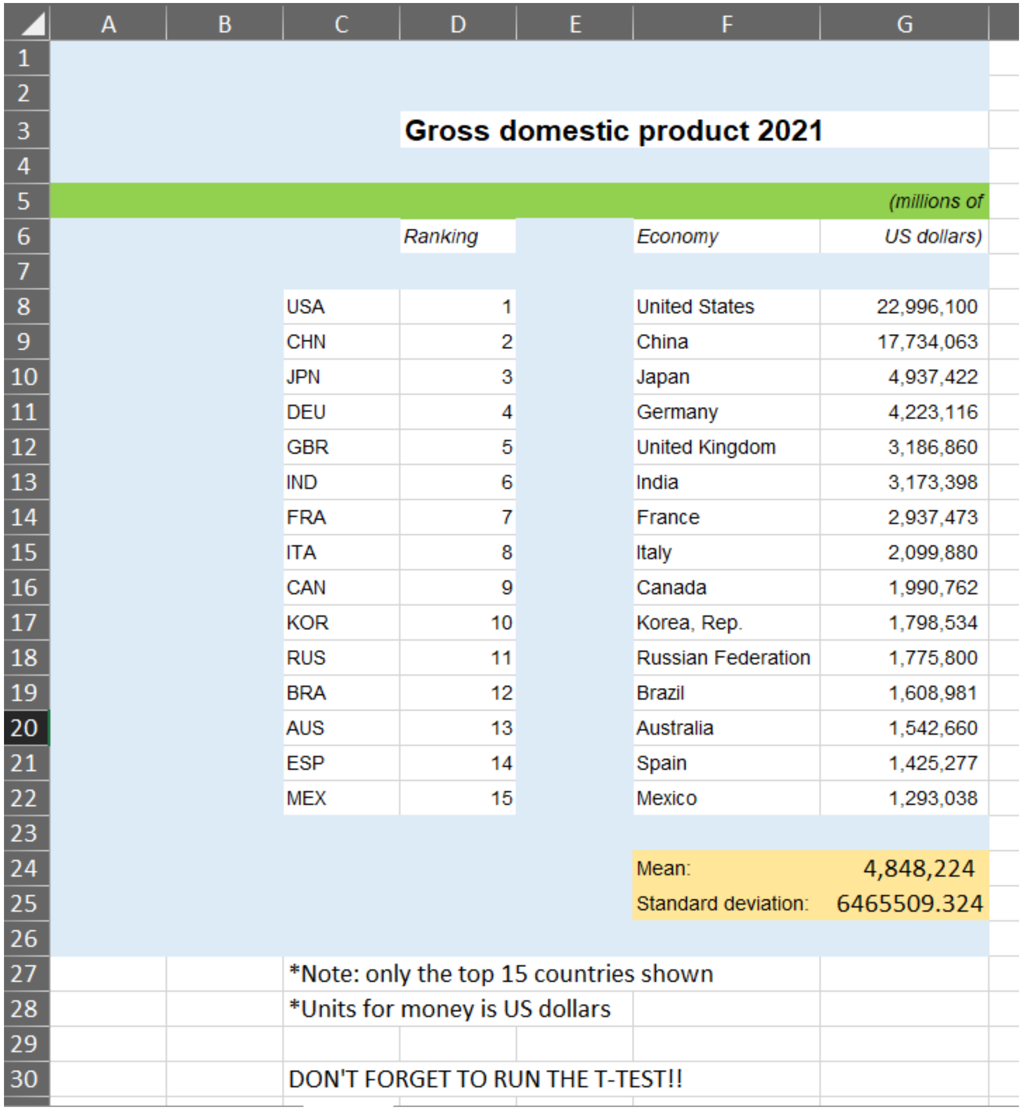
Gross domestic product 2021

	Ranking	Economy	(millions of US dollars)
USA	1	United States	22,996,100
CHN	2	China	17,734,063
JPN	3	Japan	4,937,422
DEU	4	Germany	4,223,116
GBR	5	United Kingdom	3,186,860
IND	6	India	3,173,398
FRA	7	France	2,937,473
ITA	8	Italy	2,099,880
CAN	9	Canada	1,990,762
KOR	10	Korea, Rep.	1,798,534
RUS	11	Russian Federation	1,775,800
BRA	12	Brazil	1,608,981
AUS	13	Australia	1,542,660
ESP	14	Spain	1,425,277
MEX	15	Mexico	1,293,038
		Mean:	4,848,224
		Standard deviation:	6465509.324

*Note: only the top 15 countries shown

*Units for money is US dollars

DON'T FORGET TO RUN THE T-TEST!!



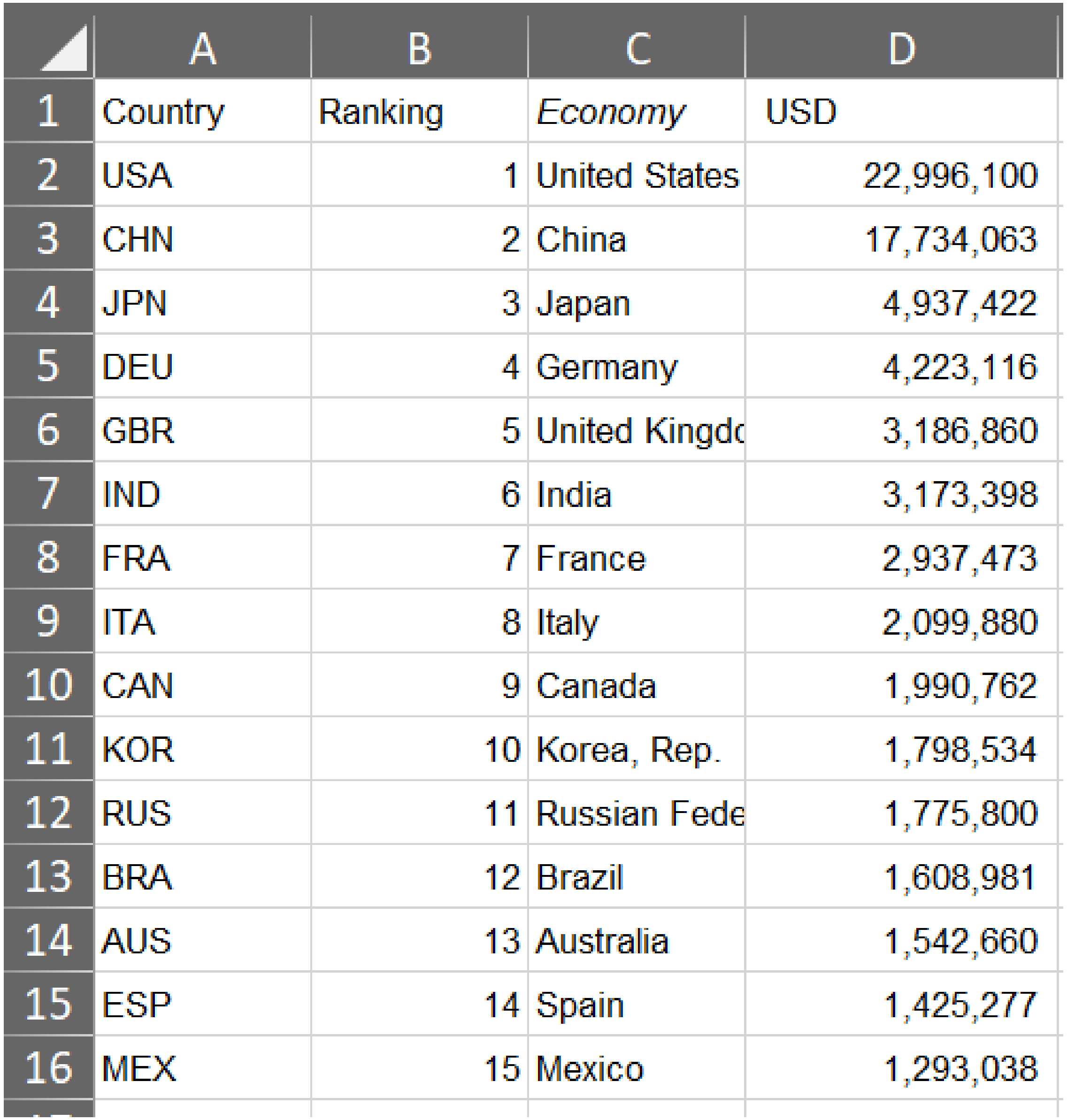
Gross domestic product 2021

	Ranking	Economy	(millions of US dollars)
USA	1	United States	22,996,100
CHN	2	China	17,734,063
JPN	3	Japan	4,937,422
DEU	4	Germany	4,223,116
GBR	5	United Kingdom	3,186,860
IND	6	India	3,173,398
FRA	7	France	2,937,473
ITA	8	Italy	2,099,880
CAN	9	Canada	1,990,762
KOR	10	Korea, Rep.	1,798,534
RUS	11	Russian Federation	1,775,800
BRA	12	Brazil	1,608,981
AUS	13	Australia	1,542,660
ESP	14	Spain	1,425,277
MEX	15	Mexico	1,293,038
		Mean:	4,848,224
		Standard deviation:	6465509.324

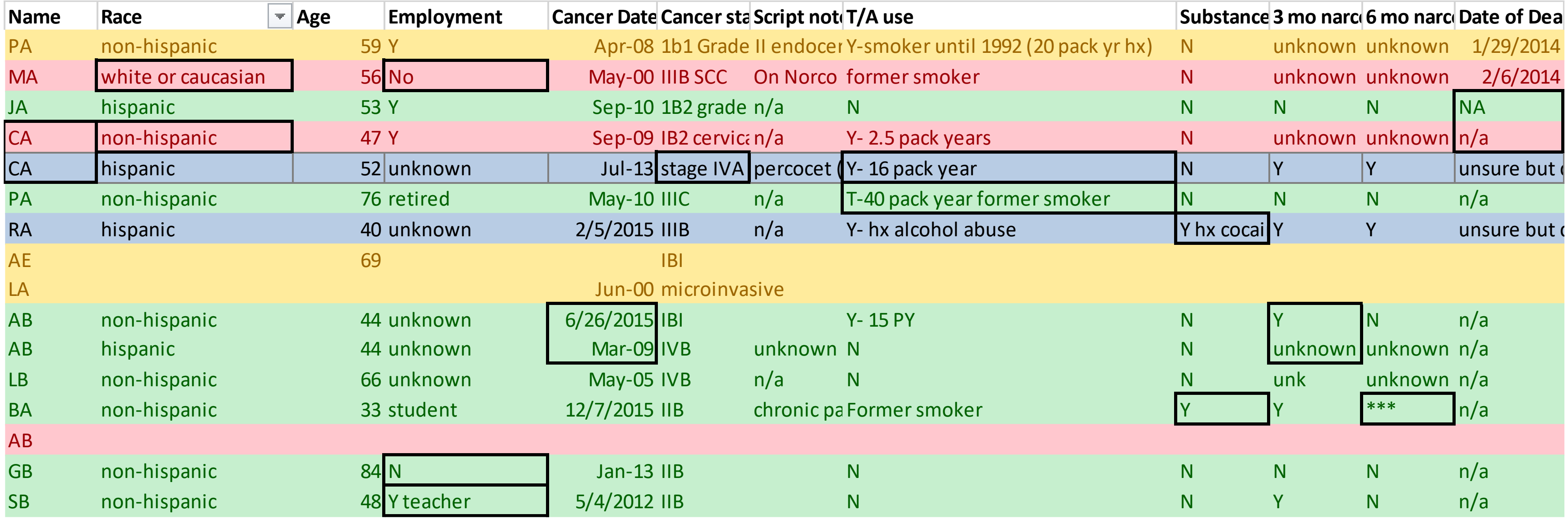
*Note: only the top 15 countries shown

*Units for money is US dollars

DON'T FORGET TO RUN THE T-TEST!!



	A	B	C	D
1	Country	Ranking	*Economy*	USD
2	USA	1	United States	22,996,100
3	CHN	2	China	17,734,063
4	JPN	3	Japan	4,937,422
5	DEU	4	Germany	4,223,116
6	GBR	5	United Kingdo	3,186,860
7	IND	6	India	3,173,398
8	FRA	7	France	2,937,473
9	ITA	8	Italy	2,099,880
10	CAN	9	Canada	1,990,762
11	KOR	10	Korea, Rep.	1,798,534
12	RUS	11	Russian Fede	1,775,800
13	BRA	12	Brazil	1,608,981
14	AUS	13	Australia	1,542,660
15	ESP	14	Spain	1,425,277
16	MEX	15	Mexico	1,293,038



Name	Race	Age	Employment	Cancer Date	Cancer sta	Script not	T/A use	Substance	3 mo narc	6 mo narc	Date of Dea
PA	non-hispanic	59	Y	Apr-08	1b1 Grade II endocer		Y-smoker until 1992 (20 pack yr hx)	N	unknown	unknown	1/29/2014
MA	white or caucasian	56	No	May-00	IIIB SCC	On Norco	former smoker	N	unknown	unknown	2/6/2014
JA	hispanic	53	Y	Sep-10	1B2 grade	n/a	N	N	N	N	NA
CA	non-hispanic	47	Y	Sep-09	IB2 cervica	n/a	Y- 2.5 pack years	N	unknown	unknown	n/a
CA	hispanic	52	unknown	Jul-13	stage IVA	percocet (	Y- 16 pack year	N	Y	Y	unsure but c
PA	non-hispanic	76	retired	May-10	IIIC	n/a	T-40 pack year former smoker	N	N	N	n/a
RA	hispanic	40	unknown	2/5/2015	IIIB	n/a	Y- hx alcohol abuse	Y hx cocai	Y	Y	unsure but c
AE		69			IBI						
LA				Jun-00	microinvasive						
AB	non-hispanic	44	unknown	6/26/2015	IBI		Y- 15 PY	N	Y	N	n/a
AB	hispanic	44	unknown	Mar-09	IVB	unknown	N	N	unknown	unknown	n/a
LB	non-hispanic	66	unknown	May-05	IVB	n/a	N	N	unk	unknown	n/a
BA	non-hispanic	33	student	12/7/2015	IIB	chronic pa	Former smoker	Y	Y	***	n/a
AB											
GB	non-hispanic	84	N	Jan-13	IIB		N	N	N	N	n/a
SB	non-hispanic	48	Y teacher	5/4/2012	IIB		N	N	Y	N	n/a

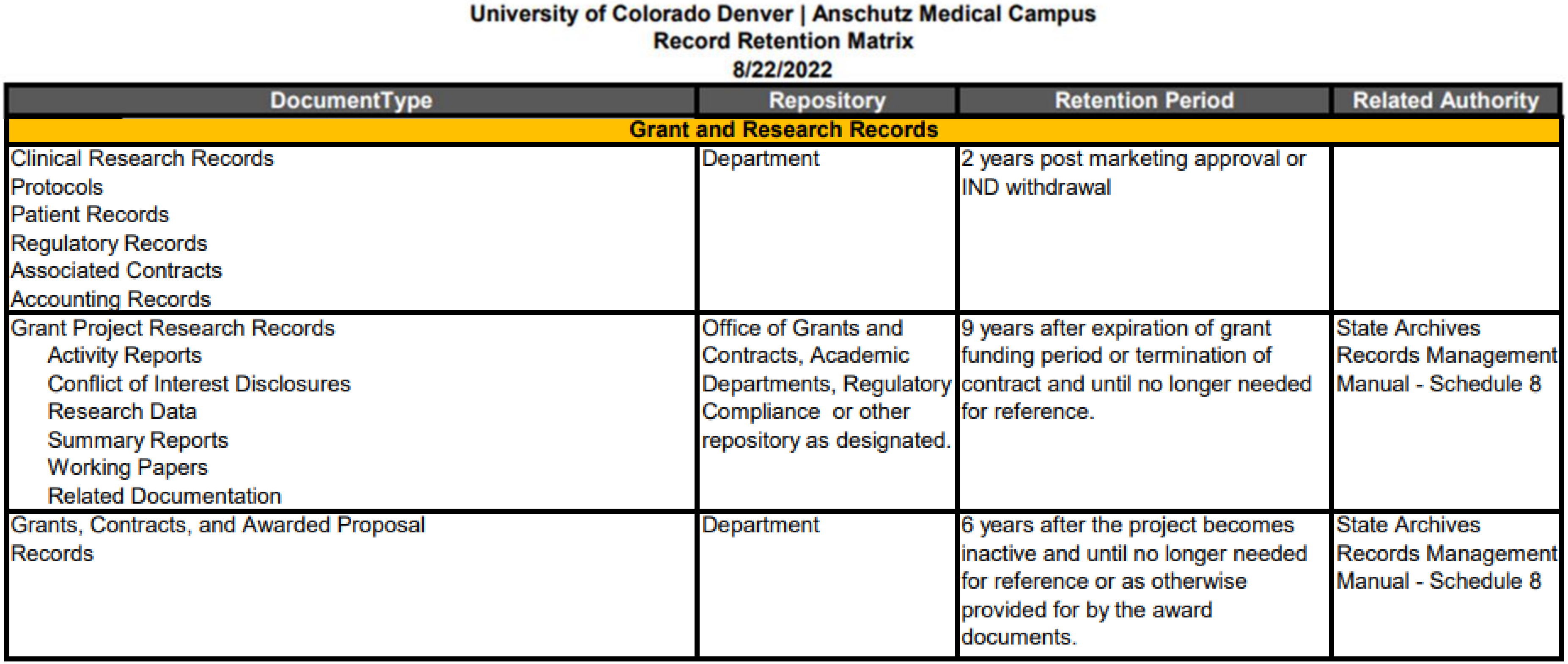
DATA STORAGE

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?



University of Colorado Denver | Anschutz Medical Campus
Record Retention Matrix
8/22/2022

DocumentType	Repository	Retention Period	Related Authority
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

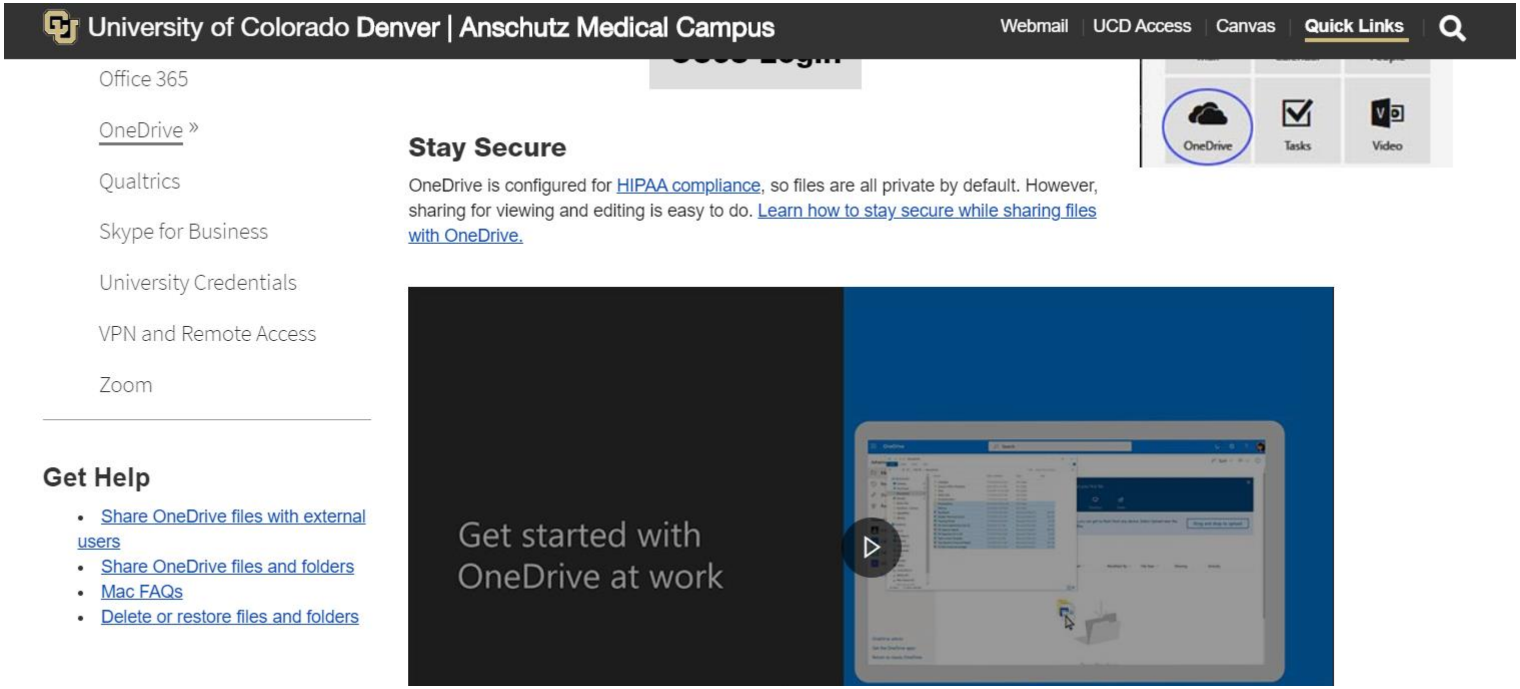
CONSIDERATIONS WHEN STORING DATA/RESEARCH RECORDS

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 *(mice)*
- 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…

POSSIBLE ELECTRONIC DATA STORAGE LOCATIONS

- Email safe guards: https://www.ucdenver.edu/offices/office-of-information-technology/secure-campus/encryption
- OneDrive: https://www.cuanschutz.edu/offices/office-of-information-technology/tools-services/for-faculty/detail-page/microsoft-onedrive-for-business



Stay Secure

OneDrive is configured for HIPAA compliance, 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.

Get Help

- Share OneDrive files with external users
- Share OneDrive files and folders
- Mac FAQs
- Delete or restore files and folders

CONSIDERATIONS FOR IMAGING DATA

- Human imaging data should be treated as identifiable data
- This should be stored on the Isilon server
- You may need to help pay for storage space depending on the file size
- Computing should be done on a server behind the university firewall
- Non-human subject data does not require HIPAA-compliant storage but will require sufficient storage space

DATA OWNERSHIP / SHARING

OWNERSHIP/DATA SHARING

Who o

- Rese
- Fund
  - Gr
- Data
  - Sul
  - Co

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

- Research Institutions

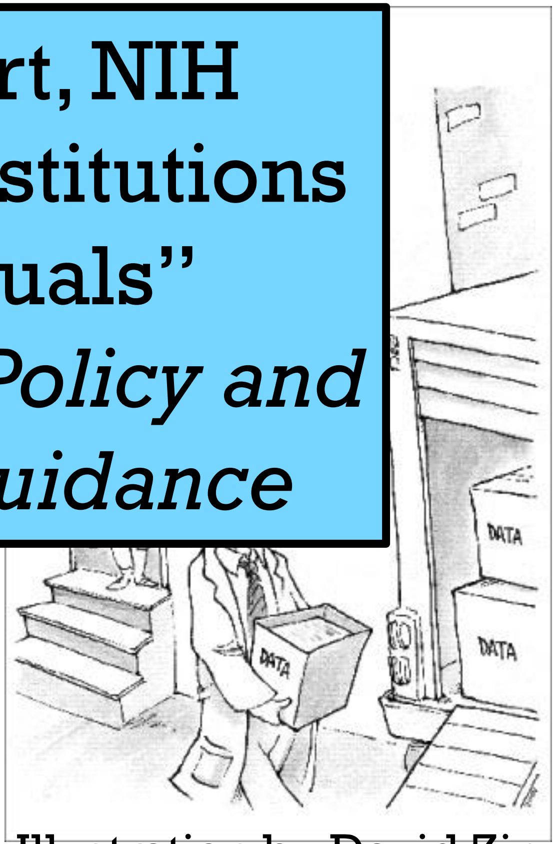


Illustration by David Zinn

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

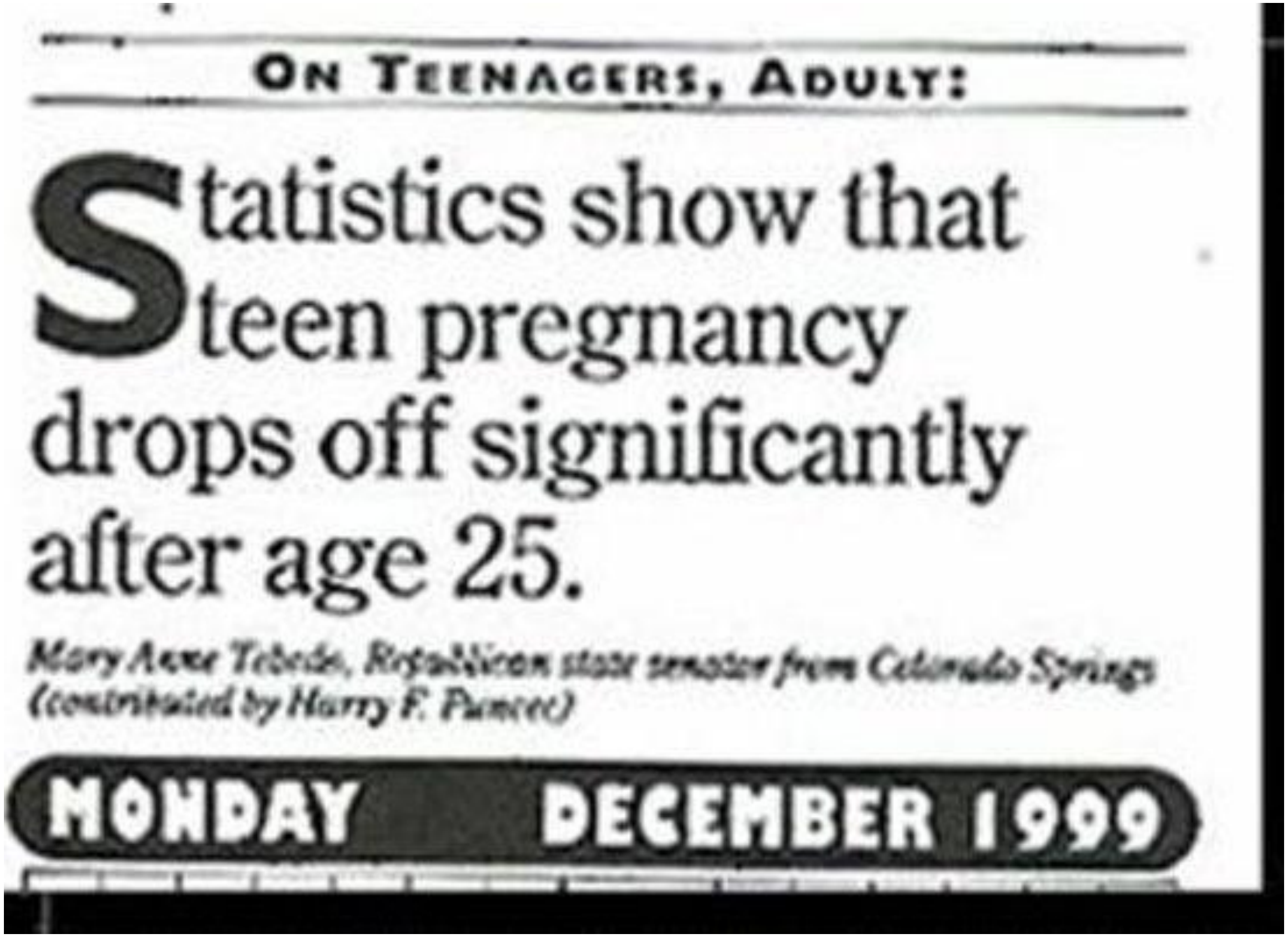
NIH DATA MANAGEMENT AND SHARING POLICY (DMSP) – STARTED JANUARY 25, 2023

- Plan and budget for managing and sharing data
  - Identify appropriate repository
  - Develop a Data Management and Sharing (DMS) plan
  - Estimate funding needed to execute this plan
- Submit a DMS plan when applying for NIH funding
  - Will be reviewed by NIH program staff, not peer review
  - Must be approved prior to award
- Comply with plan during study
  - Provide updates on DMS activities in annual progress reports
  - Work with program officer to review and approve modifications

https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview

DMSP RESOURCES AT AMC

- The Research Informatics Office (RIO)
  - Webinars and Town Halls
  - Definitions of required elements
  - Templates
  - Budgeting and compliance
- NIH webinars
- Budgeting – NIH FAQ
  - You MUST budget for this: can’t cover this with other FTE



ON TEENAGERS, ADULT:

Statistics show that teen pregnancy drops off significantly after age 25.

Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Puncec)

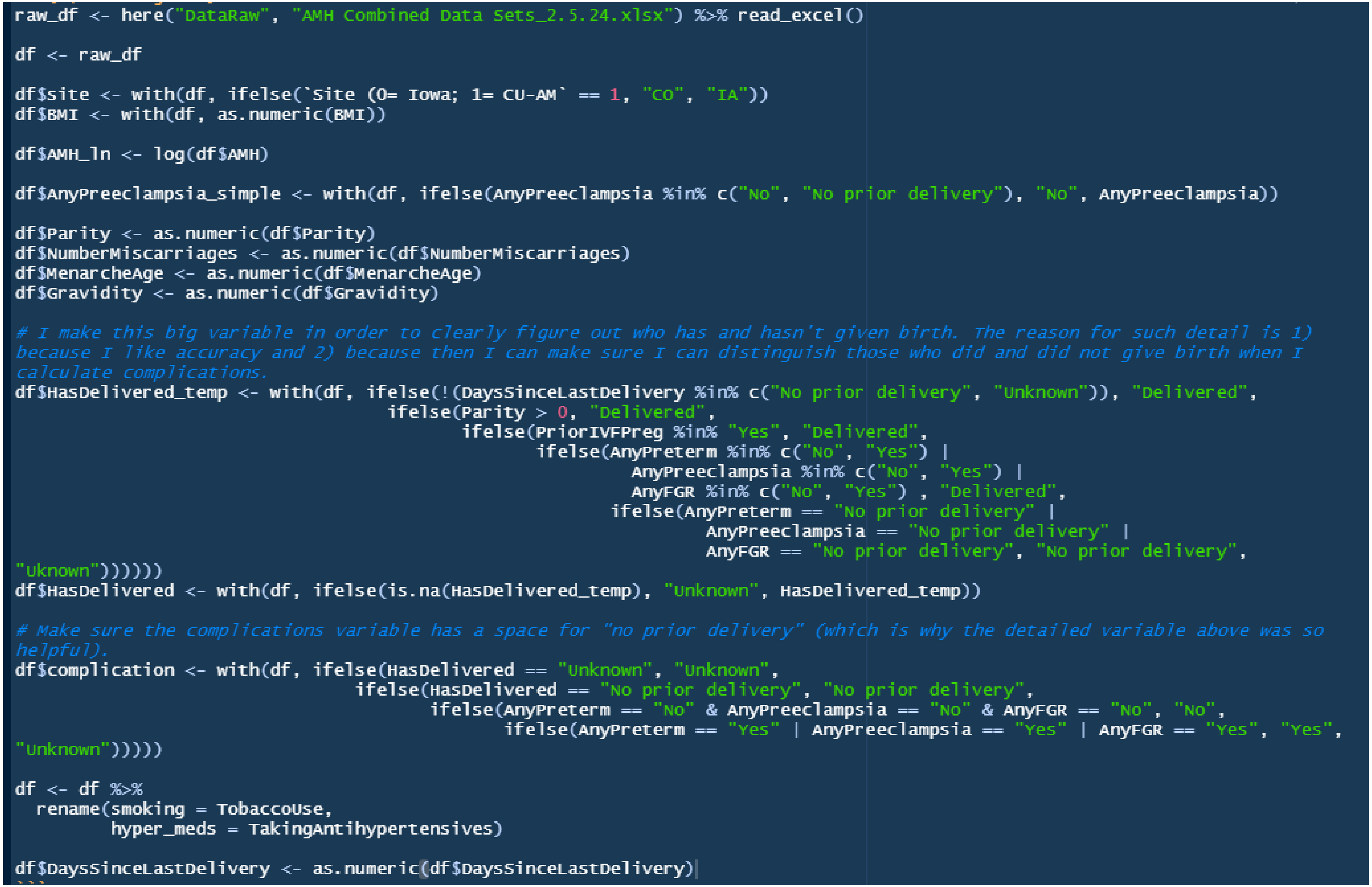
MONDAY DECEMBER 1999

Statistical Analysis

(replicability -> reproducibility)

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. Use a scripting language
   - Programs like R and SAS allow you to follow your steps <u>exactly</u> if you (or someone else) had to redo your analysis
   - Easily execute and document QC steps
   - Avoid copy/paste errors
   - NO EXCEL! POINT-AND-CLICK = NOT GREAT!
3. Add comments/notes directly to program
   - Why are you doing this step?
   - What is the goal of this step?
4. Use Version Control (i.e. GitHub or GitLab)
5. Export precise tables/figures from program
   - Avoid transposition errors
   - Save time/energy when changes are requested

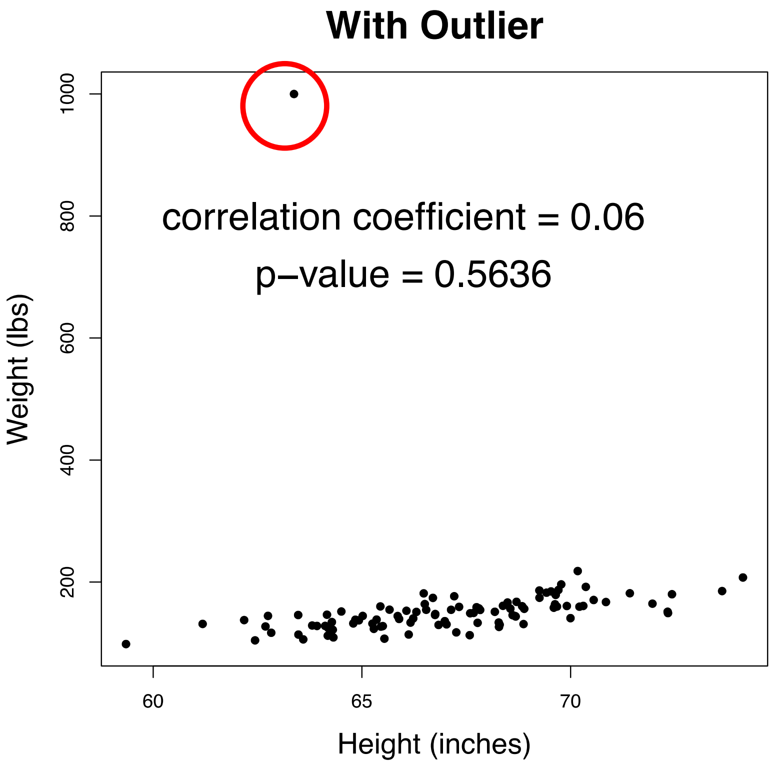


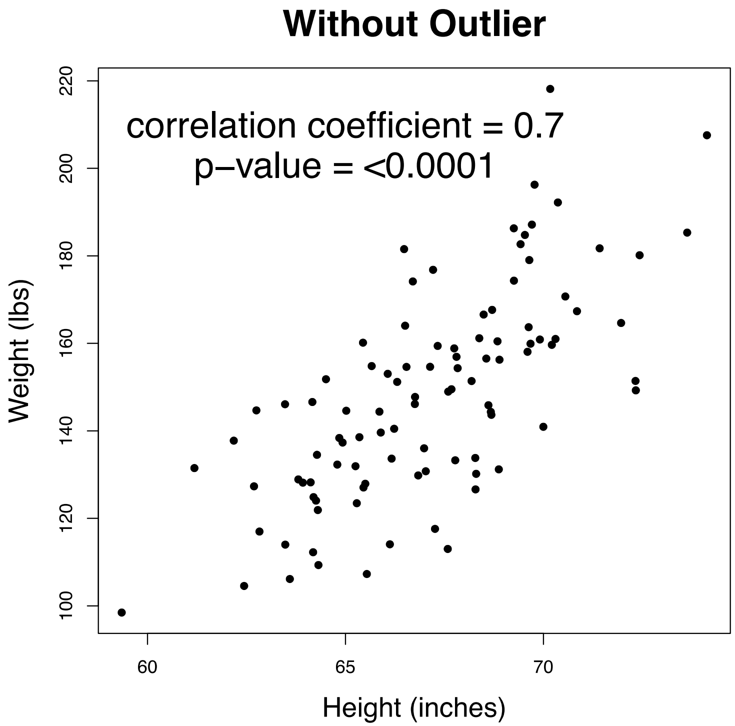
```
raw_df <- here("DataRaw", "AMH Combined Data Sets_2.5.24.xlsx") %>% read_excel()

df <- raw_df

df$site <- with(df, ifelse(`Site (0= Iowa; 1= CU-AM` == 1, "CO", "IA"))
df$BMI <- with(df, as.numeric(BMI))

df$AMH_ln <- log(df$AMH)

df$AnyPreeclampsia_simple <- with(df, ifelse(AnyPreeclampsia %in% c("No", "No prior delivery"), "No", AnyPreeclampsia))

df$Parity <- as.numeric(df$Parity)
df$NumberMiscarriages <- as.numeric(df$NumberMiscarriages)
df$MenarcheAge <- as.numeric(df$MenarcheAge)
df$Gravidity <- as.numeric(df$Gravidity)

# I make this big variable in order to clearly figure out who has and hasn't given birth. The reason for such detail is 1) 
because I like accuracy and 2) because then I can make sure I can distinguish those who did and did not give birth when I 
calculate complications.
df$HasDelivered_temp <- with(df, ifelse(!(DaysSinceLastDelivery %in% c("No prior delivery", "Unknown")), "Delivered",
                                   ifelse(Parity > 0, "Delivered",
                                          ifelse(PriorIVFPreg %in% "Yes", "Delivered",
                                                 ifelse(AnyPreterm %in% c("No", "Yes") |
                                                          AnyPreeclampsia %in% c("No", "Yes") |
                                                          AnyFGR %in% c("No", "Yes") , "Delivered",
                                                        ifelse(AnyPreterm == "No prior delivery" |
                                                                 AnyPreeclampsia == "No prior delivery" |
                                                                 AnyFGR == "No prior delivery", "No prior delivery",
"Uknown"))))))
df$HasDelivered <- with(df, ifelse(is.na(HasDelivered_temp), "Unknown", HasDelivered_temp))

# Make sure the complications variable has a space for "no prior delivery" (which is why the detailed variable above was so 
helpful).
df$complication <- with(df, ifelse(HasDelivered == "Unknown", "Unknown",
                                ifelse(HasDelivered == "No prior delivery", "No prior delivery",
                                       ifelse(AnyPreterm == "No" & AnyPreeclampsia == "No" & AnyFGR == "No", "No",
                                              ifelse(AnyPreterm == "Yes" | AnyPreeclampsia == "Yes" | AnyFGR == "Yes", "Yes", 
"Unknown")))))

df <- df %>%
  rename(smoking = TobaccoUse,
         hyper_meds = TakingAntihypertensives)

df$DaysSinceLastDelivery <- as.numeric(df$DaysSinceLastDelivery)
```

OUTLIERS

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)
200
400
600
800
1000
60
65
70
Height (inches)


Without Outlier
correlation coefficient = 0.7
p–value = <0.0001
Weight (lbs)
100
120
140
160
180
200
220
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?

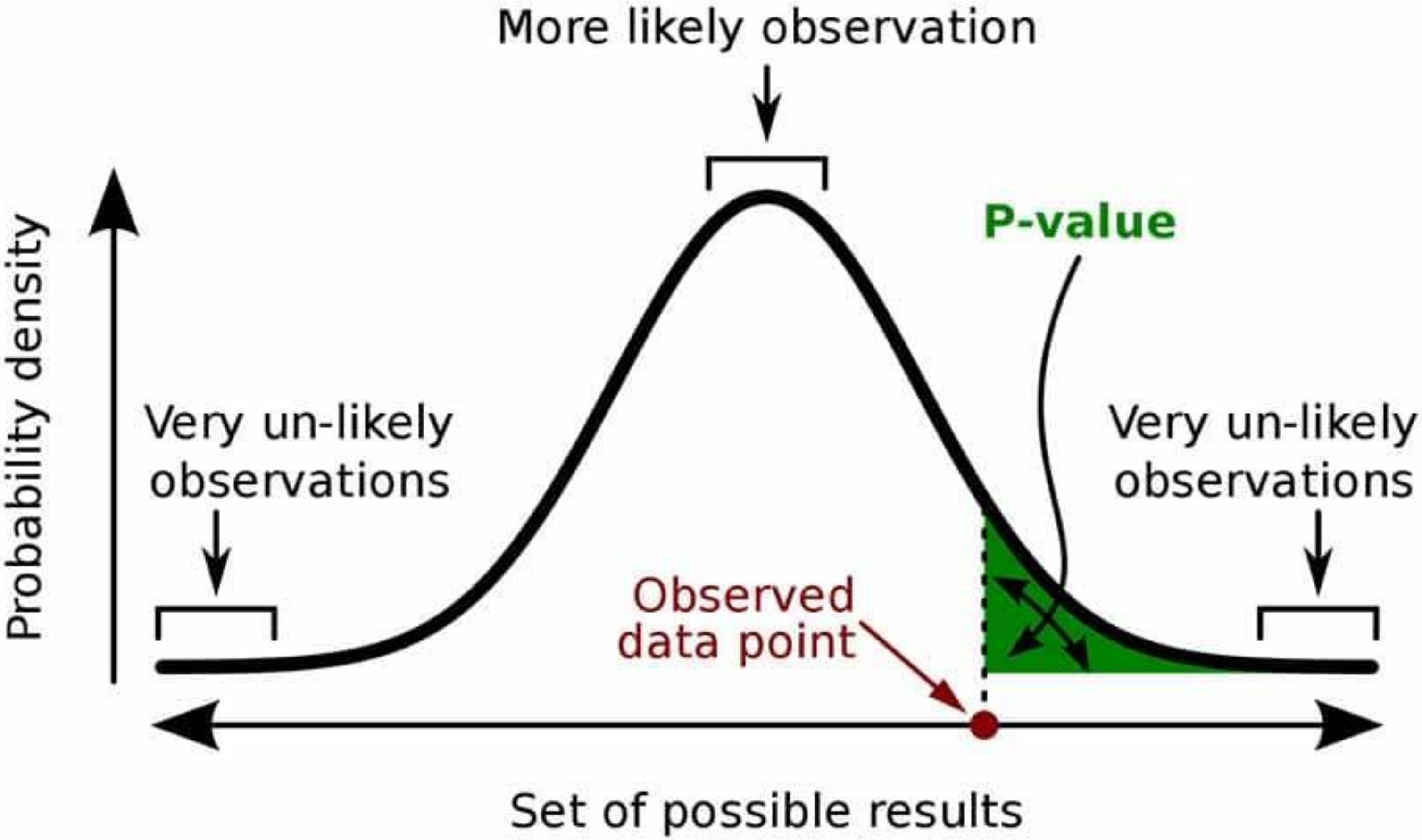
P-VALUES

DO YOU KNOW WHAT A P-VALUE IS?

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

WHAT IS A P-VALUE?

- P-values tell us how likely you are to get the same result (or one more extreme) assuming that your null hypothesis is true
- They are commonly misinterpreted
- Null hypothesis: the hypothesis you are trying to *disprove*
  - “Adult males and females have the same mean height”
  - “There is no correlation between smoking status and development of lung cancer.”



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

- https://www.simplypsychology.org/p-value.html

MORE INTUITION FOR P-VALUES

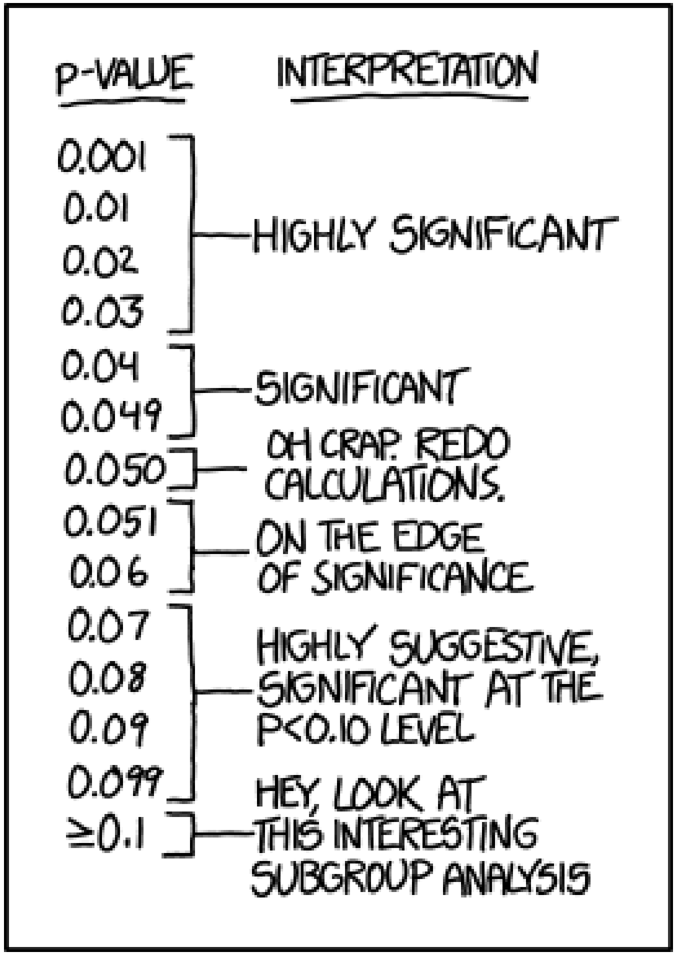
- https://www.rossmanchance.com/ISIapplets.html

IS THERE VALUE IN P-VALUES?

- In most cases you cannot directly compare p-values (must be from exact same sample and same outcome)
- Statistical significance does not indicate clinical significance

WHAT IS P-HACKING?

- Stop collecting data once $p<0.05$
- Report only measures or outcomes with $p<0.05$
- Use covariates to get $p<0.05$
- Exclude participants to get $p<0.05$
- Transform data to get $p<0.05$


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

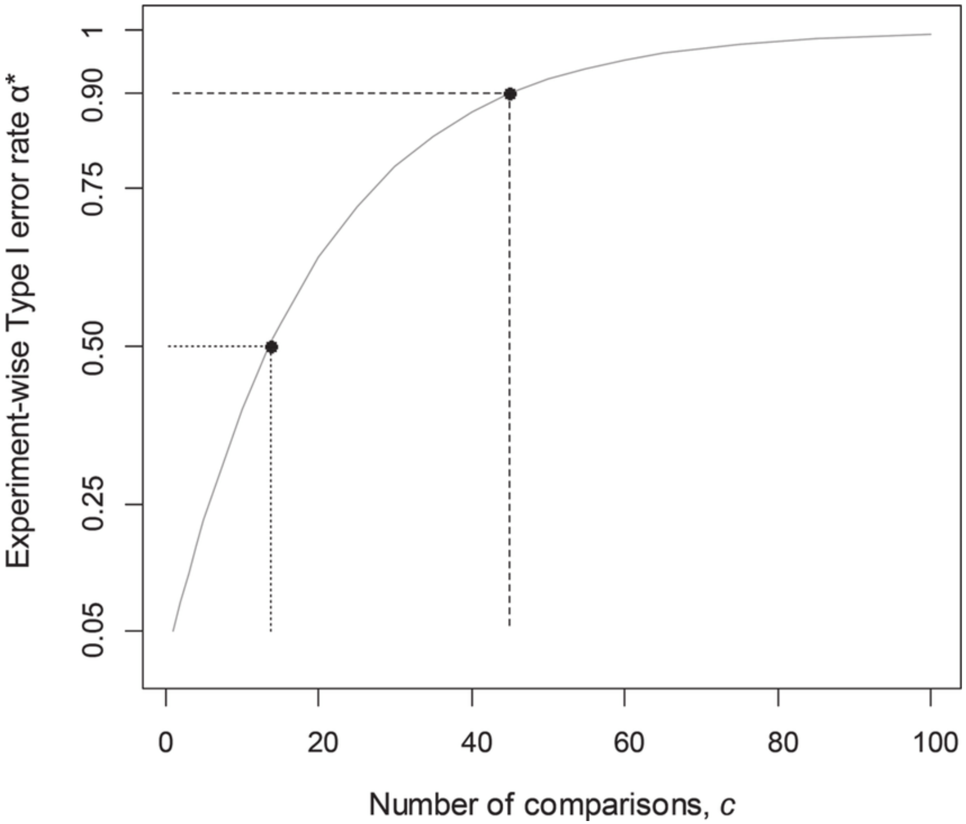
Please, do not do this ↑

WHAT HAPPENS WHEN YOU P-HACK

The $p<0.05$ threshold assumes a 5% type I error rate, but…

⬆ As the number of statistical comparisons increases

⬆ The type I error rate increases

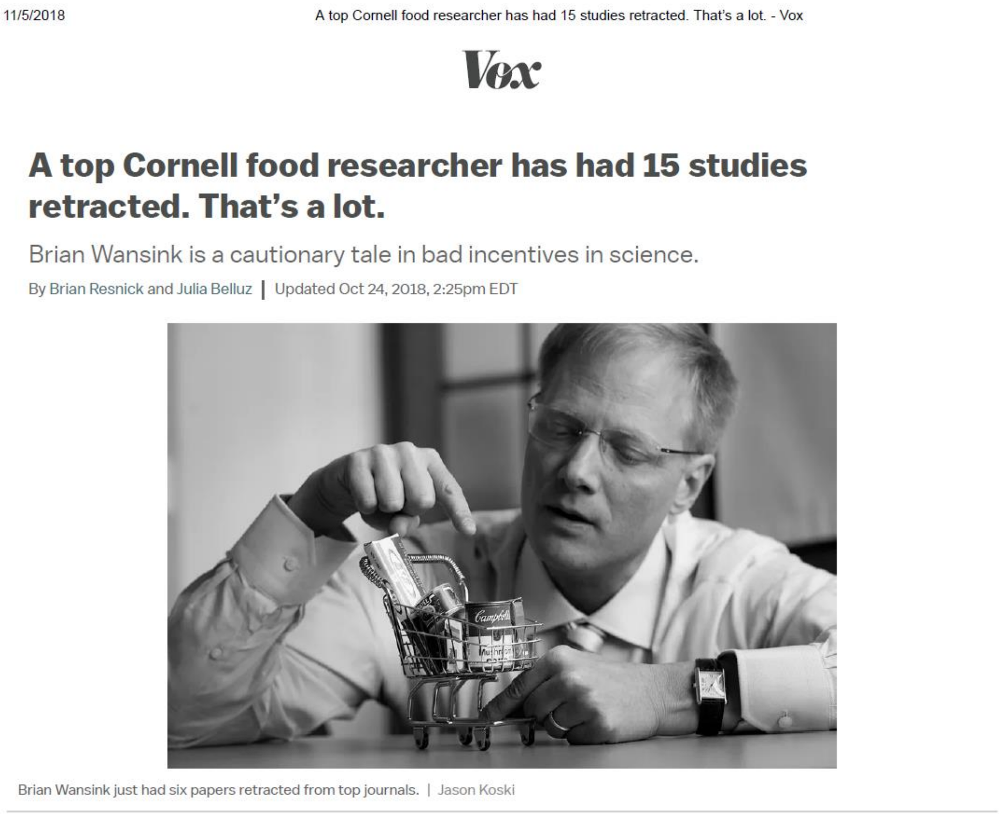


Bello, Nora & Renter, David. (2018). Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. Journal of Dairy Science. 101. 10.3168/jds.2017-13978.

HOW TO AVOID P-HACKING

- Determine a primary analysis a priori and stick to that analysis
- Adjust for multiple comparisons
- Interpret your findings in an exploratory context instead of a confirmatory context
- Compare effect sizes, which are usually more clinically relevant when discussing your results

THE DARK SIDE OF 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

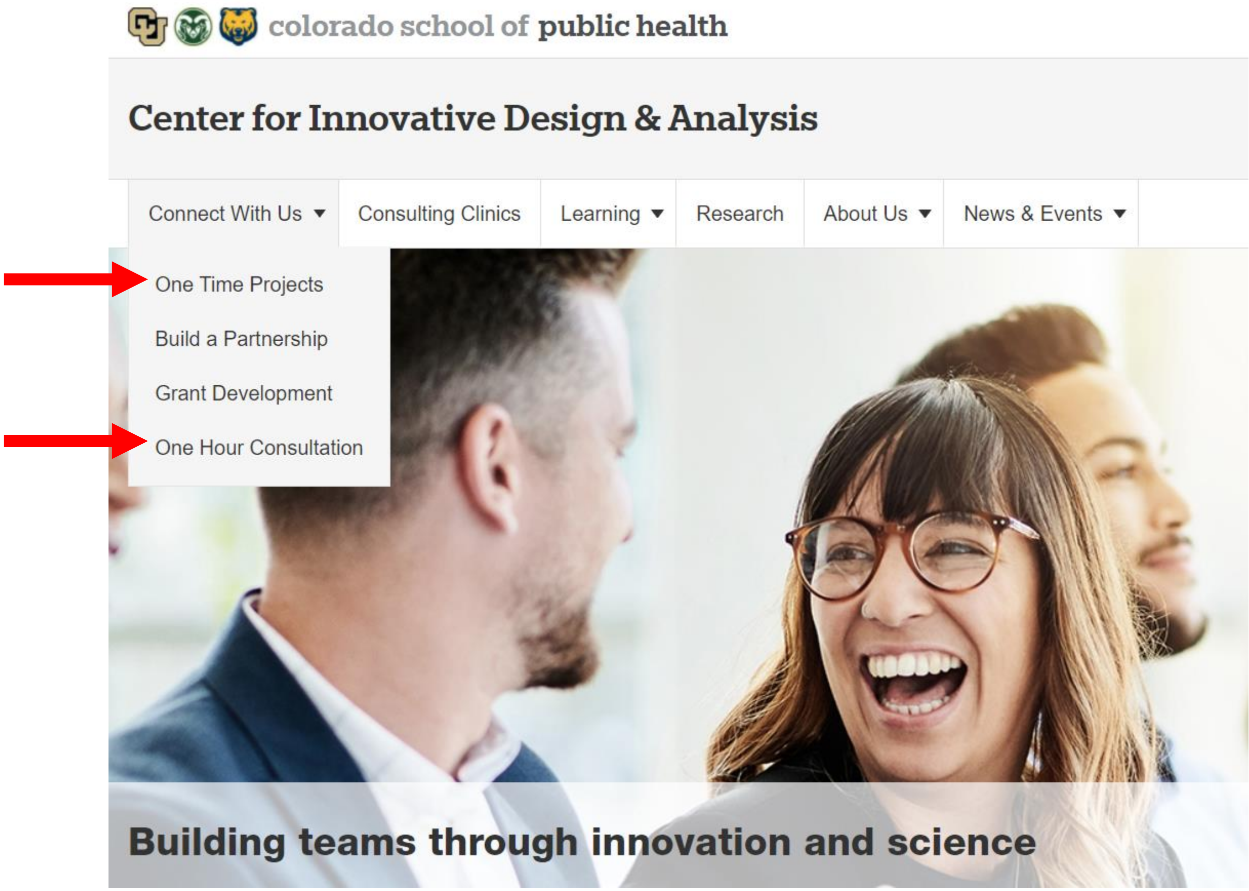
- Biostats4U – Statistical resources for non-statisticians
  - https://biostats4you.umn.edu/
- UCLA IDRE - Statistical methods across multiple programming languages with real world problems with explanations of code and interpretation of output
  - https://stats.idre.ucla.edu/other/mult-pkg/whatstat/
- 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
- Dr. Camille Hochheimer

Nassia Dunn

Dr. Brandie Wagner

Megan Branda

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

References

Reproducibility and Replicability in Science. (*https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science*) National Academies, 2019.

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

NIH Data Management and Sharing Policy (https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview), January 2023.

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