UVA Health
School of Medicine

Clinical Research Education Series

Accreditation

- The University of Virginia School of Medicine and School of Nursing designates this live activity for a maximum of **1.0** ***AMA PRA Category 1 Credits.***^{TM} Physicians should claim only the credit commensurate with the extent of their participation in the activity.
- The University of Virginia School of Medicine and School of Nursing awards **1.0 contact hour** for nurses who participate in this educational activity and complete the post activity evaluation.
- The University of Virginia School of Medicine and School of Nursing awards **1.0 hour of participation** (consistent with the designated number of *AMA PRA Category 1 Credit(s)*^{TM} or ANCC contact hours) to a participant who successfully completes this educational activity. The University of Virginia School of Medicine and School of Nursing maintains a record of participation for six (6) years.

UVA Health
School of Medicine

Disclosures

- The speaker, Sarah Ratcliffe has no personal or professional financial relationships with a commercial entity producing healthcare goods or services.
- The planning committee members, Katie Rea, MSCR, BSN, RN, CCRC; Lori Elder, RN, BSN, CCRA; Kimberly Underwood, BS, CCRC; and Maria Davenport, MPH, CCRP; have no personal or professional financial relationships with a commercial entity producing healthcare goods or services.

Activity Code: 146405

Instructions for claiming credits/contact hours can be found here.

UVA Health
School of Medicine

UNIVERSITY of VIRGINIA
SCHOOL of MEDICINE

A Statistical Perspective of Study Data Management

Sarah J. Ratcliffe, PhD

Professor & Director, Division of Biostatistics

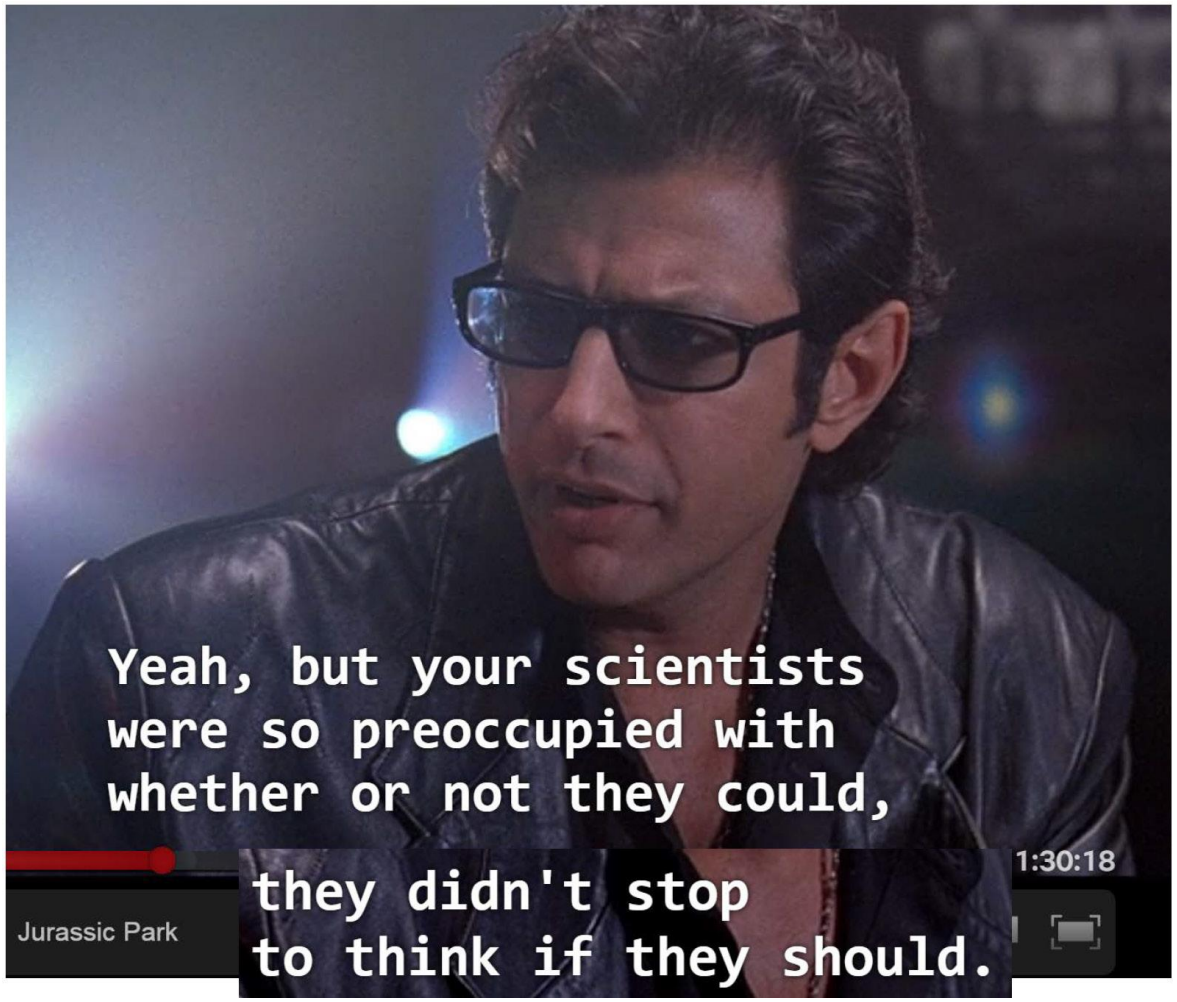
Senior Vice Chair for Research, Department of Public Health Sciences

Email: sratcliffe@virginia.edu

7 August 2023

Introduction

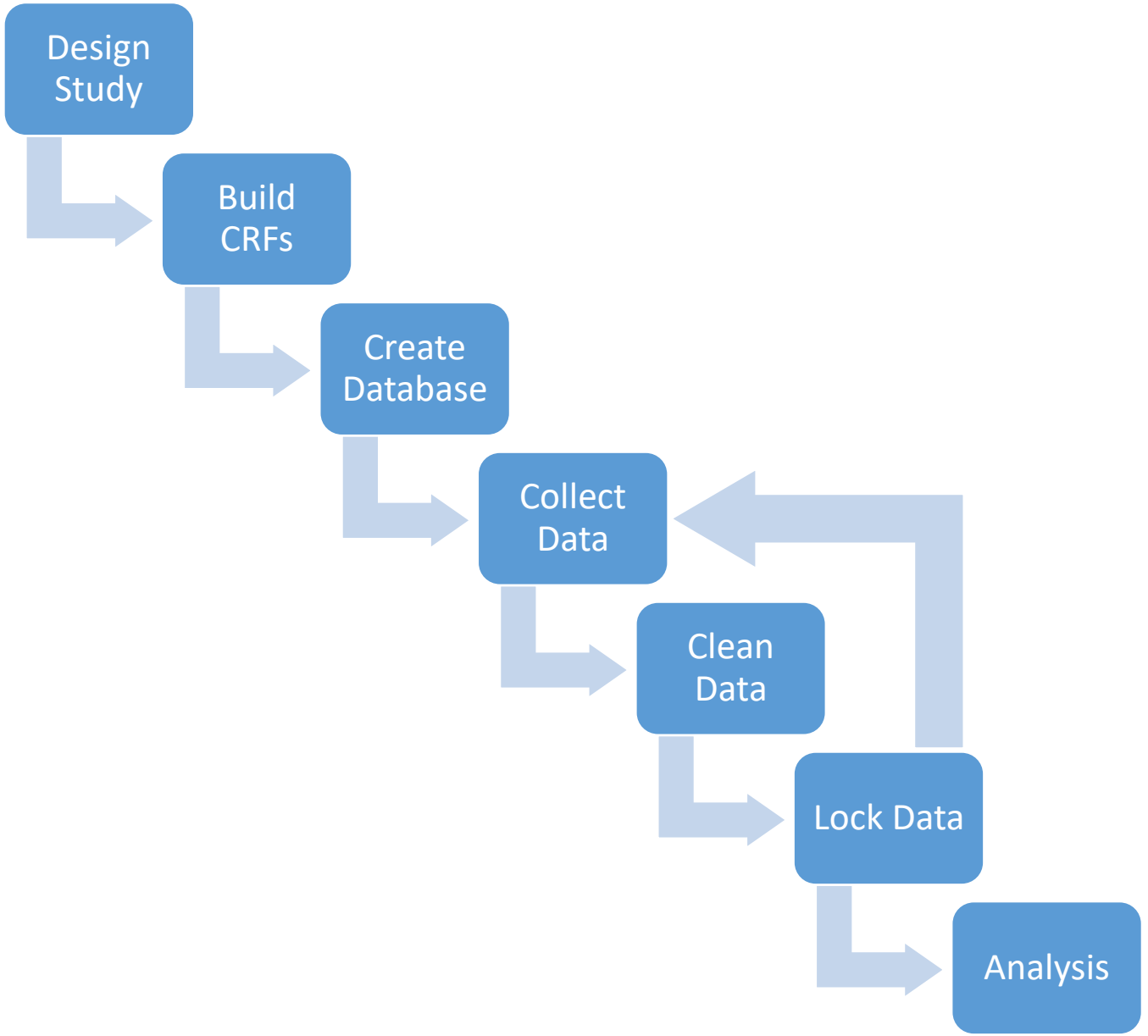
- Technology has made it easy to enter, share and analyze data
- Responsible for protecting the data in our “custody”
- Responsible for ensuring the “best” quality data
- Responsible for ensuring the “best” use of the data



What is Data Management?

- Where data comes from
- Where was it recorded
- When was it collected
- Who collected and entered it
- What information was recorded
- Where was it stored (e.g. hard drive or server)
- How all the information was compiled
- How to get it into statistical software
- What needs to be done to analyze it
- Sharing the data with others

Data Management Cycle



Data Management

Research Coordinators play a critical role in ensuring the quality and validity of the study data, and subsequently the validity of the study results.

Best Practice #1: Know security requirements

- What you collect will impact how data can be collected, stored and transferred
- Will it contain PHI? [45 CFR section 164.514(b)(2)(i)]
  - Data with identifiers must be stored securely
  - Research - IRB approval of storage plan required
  - Don’t collect PHI if you don’t need it!
  - Minimize who has access to PHI
    - Other than dates, often stored separately / on a separate case report form from the rest of the study data
- Statistical Perspective: We need dates for analyses but don’t want to be sent / responsible for other PHI

What is PHI

Protected health information containing any of the following items:

(A) Names

(B) Postal address information (except town or city, county, health service area, state and zip code, which may be included)

(C) “Day” elements (i.e. except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death

(D) Telephone numbers

(E) Fax numbers

(F) Email addresses

(G) Social security numbers

(H) Medical record numbers

(I) Health plan beneficiary numbers

(J) Account numbers

(K) Certificate/license numbers

(L) Vehicle identifiers and serial numbers, including license plate numbers

(M) Device identifiers and serial numbers

(N) Web Universal Resource Locators (URLs)

(O) Internet Protocol (IP) address numbers

(P) Biometric identifiers, including finger and voice prints

(Q) Full face photographic images and any comparable images

(R) Any other unique identifying number, characteristic or code (except for a study number that allows for re-identification)

Best Practice #2: Consider analysis plan when collecting data

- Need to collect everything needed to answer the research question
- How is each outcome and exposure / covariate measured?
  - Type of variable – continuous, binary, categorical, …
  - Range of valid values

- Statistical Perspective: We don't see the "front end" of data collection, we pull data from the "back end" of the data management system. E.g. OnCore, Advarra EDC, REDCap

WAIS IV		Raw	SS
Block			
Digit Span		27	
	Forward:	Backward:	Sequence:
DS Raw Scores:	10	8	9
DS SS:			
RDS:			

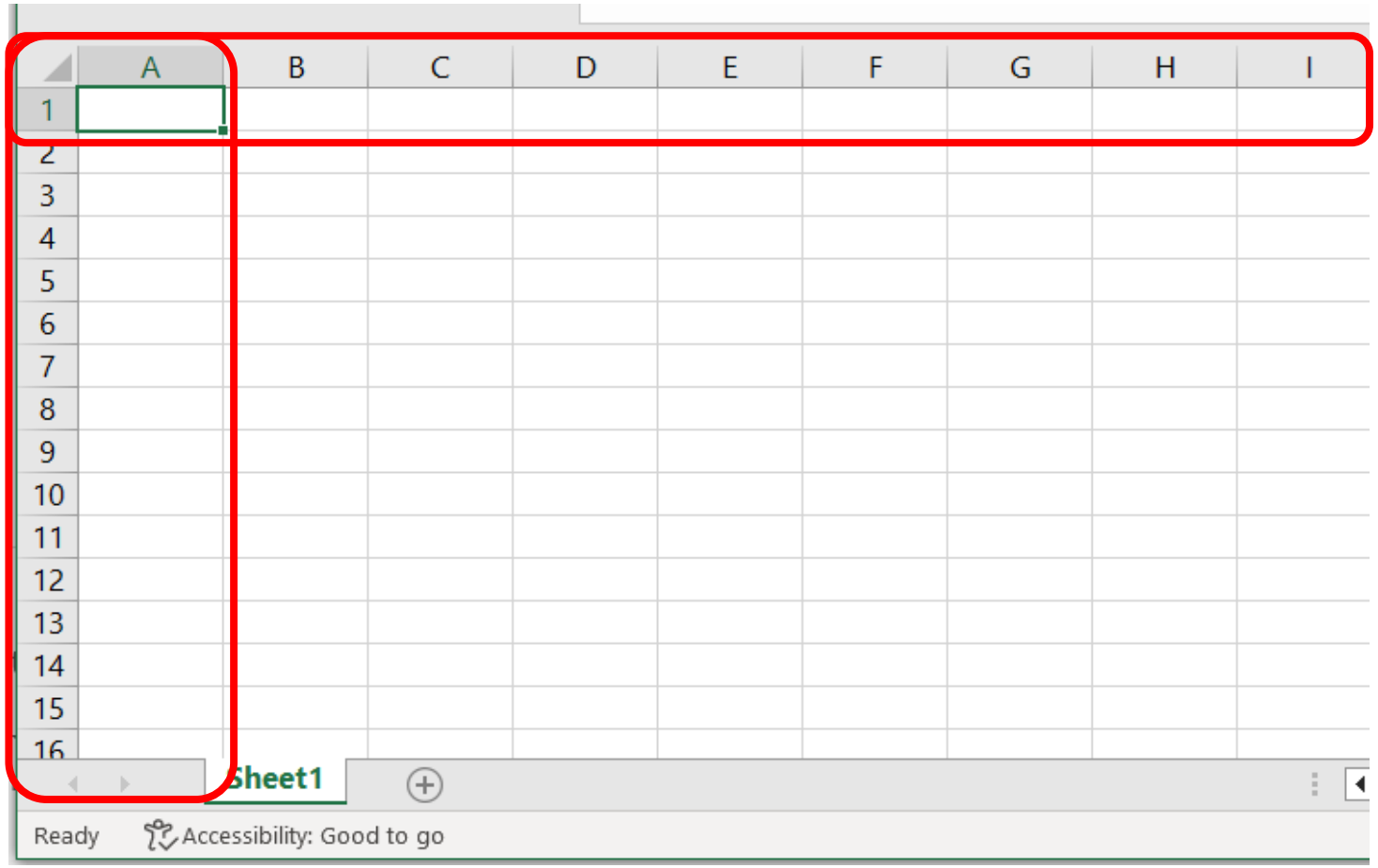
Front End Data Entry

	A	B	C	D	E	F	G
1	block	digitspan	digitspanss	dsraw_fwd	dsraw_back	dsraw_seq	dsss_fwd
2		27		10	8	9	

Statistical Back End Data View

Data Structure

- Data is arranged in ROWS and COLUMNS
- Cells should ONLY contain a single piece of information



Data Structure

- *Unit* or *Level* of Analysis
  - Patient/individual/animal
  - Group – e.g site, clinic, experiment
- *Record*
  - One instance of the unit – e.g. one patient, one respondent
  - In repeated measures / longitudinal studies, each “contact” is a record. E.g. each visit, each replicate

➢ ROWS = Records

Data Structure

- *Variable*
  - A factor / measure of interest in your study
  - Each variable should capture **a single** piece of information
  - E.g. BP = systolic / diastolic is 2 pieces of information
    = 2 variables

➢ COLUMNS = Variables

Data Structure: Wide Format

Student ID	Baseline	Midterm	Final
001	50	75	100
002	60	65	70
003	53	63	73
004	70	71	85

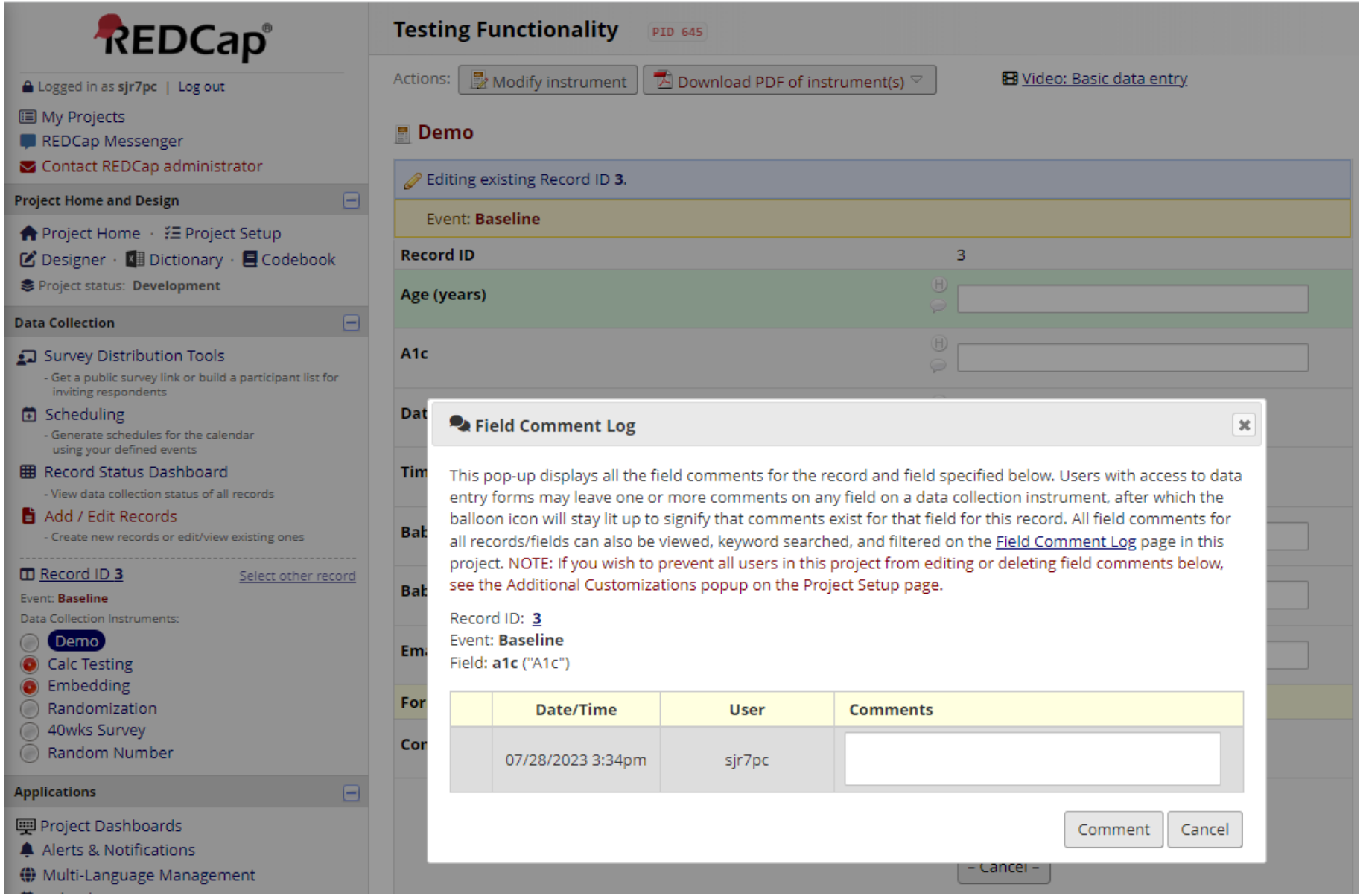
Data Structure: Long Format

Student ID	Time	Score
001	1	50
001	2	75
001	3	100
002	1	60
002	2	65
002	3	70

How Statistical Software Reads Data

- Most software expects each row is a record and each column is a variable
- Functions / algorithms are programmed to expect this structure
- Expects all the data to be analyzed to be coded as numbers (not text)
- Colours don’t exist
- Comments added through special function (vs. a field in the database) don’t exist

Comments added through special buttons (vs. a field in the database) don't exist in "back end" data pulls



Best Practice #3: Don't Use Excel to Collect Data

- Excel is NOT a data management system
- Will not prevent you from setting up data incorrectly
- Will not prevent "out of range" entries
- Will not ensure consistent coding
- No way of tracking changes to the data
- Can create problems with variable formats
- Statistical formula are known to give wrong answers in certain structures / instances!
- **Statistical Perspective: Data Integrity and Security are too easy to compromise in Excel!**

Best Practice #4: Variable Naming Standards

- Variable names = first row of the data
- Keep them short and meaningful
  - 8-12 characters as target; 32 characters maximum
- No spaces: use “_” if gap is needed
- No special characters: *$%&@!
- Start with a letter
- Use all lower or all upper cases letters
  - Upper and lower case are different in some packages!
- Match standards as much as possible: e.g. sex & gender
- Statistical Perspective: we type these a lot in our code, check they are short and informative enough for us to work with

Best Practice #4: Variable Naming Standards

Table 1. Examples of good and bad variable names.

good name	good alternative	avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

Broman KW & Woo KH. Data Organization in Spreadsheets. *The American Statistician*, 72(1):2-10, 2018. https://doi.org/

Best Practice #4: Variable Naming Standards

Dates are particularly hard to understand with good names

Front End

- Date of Birth
- Date of Screening
- Date of Randomization
- 6 Month follow-up Visit Date

I	J	K	L
dob	date1	date2	date3
01-02-84	03-12-23	04-12-23	10-12-23

Best Practice #5: Keep it Numeric

- Data *values* should be numbers when possible
  - Break this rule if you must
- Continuous numeric data can be compressed / categorized later, if appropriate.
- Categorical data should be stored in numbered categories
  - Can use “other” category, with an other text variable, to capture rarer responses
  - Can add numbered categories if “other” option becomes common
  - Save the key / coding in the data dictionary
- Statistical Perspective: statistical software always needs numbers

Best Practice #6: Keep it Consistent

- Use consistent units (mmHG, grams, degrees F)
  - Add units to labels, variable names, and data dictionary
- Use the same date format for all dates
  - REDCap requires M-D-Y for some calculations, others assume Y-M-D unless told otherwise. All dates must be in same format for calculations on dates to work.
- Use consistent coding
  - No=0 in all yes/no variables, then Absent=0 in an absent/present variable
- Text / string variables must be identical for statistical software
  - “male”, “ male”, “Male”, “M”, “M “ will all be read as different
  - Multiple Dx in one field means re-coding it all
    - E.g. history of “diabetes, MI” will not be counted in either “diabetes” or “MI”.
- Statistical Perspective: The more text, the harder it is to clean the data and to run analyses.

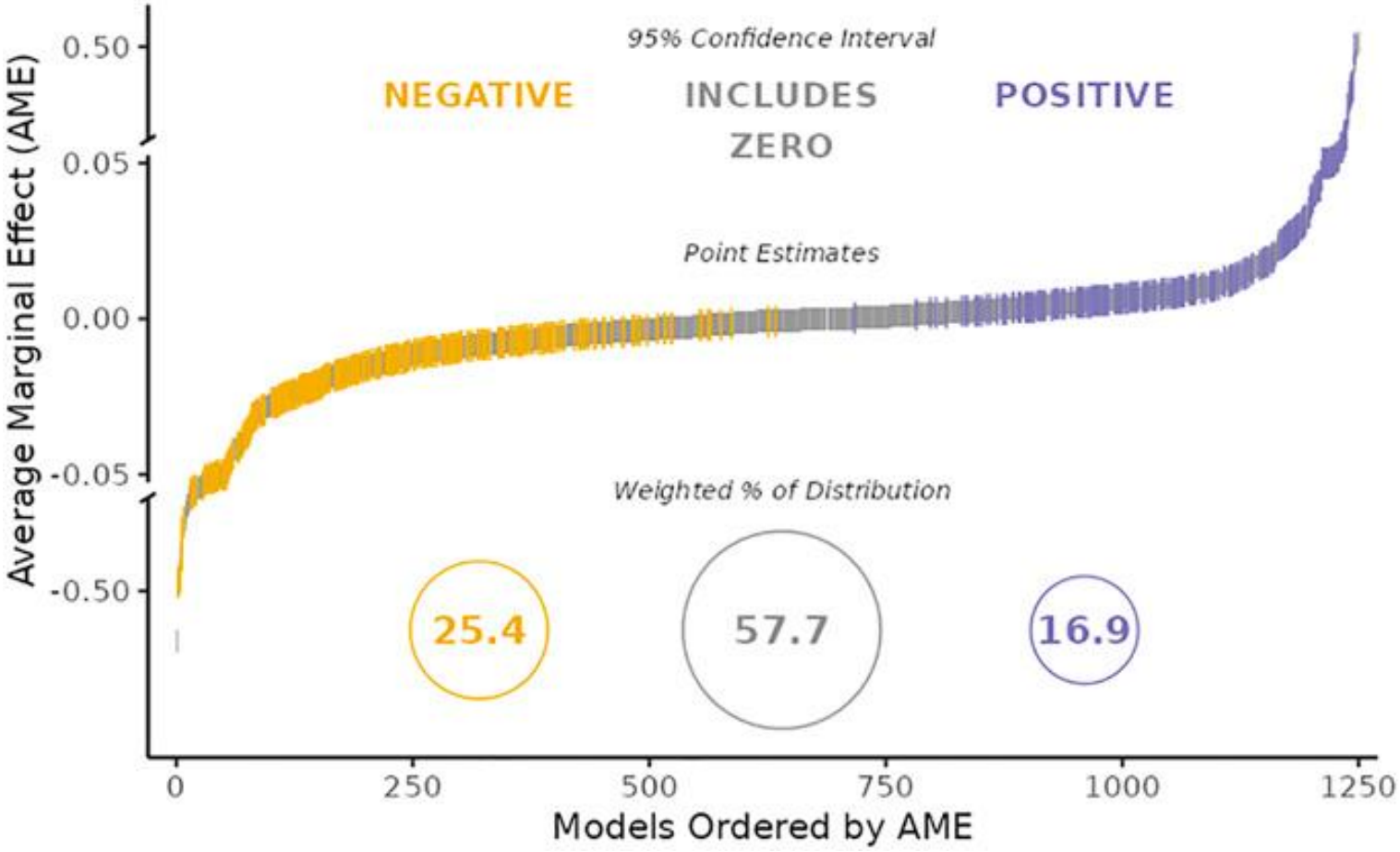
Best Practice #7: Don't change meanings

- Different people make different decisions with the same data
- Subjective decisions on what to enter where will impact the study results and conclusions
- If data entry is unclear – there is a lack of documentation that needs to be addressed
  - Ask someone how and where to enter the data
  - All CRCs on the study need to have the same understanding of the data

- Statistical Perspective: Make sure the statistician is aware of anything unusual to ensure correct cleaning and analyses

Subjective Decisions

- Breznau et al. (2022). doi:10.1073/pnas.2203150119



Best Practice #8: Data Dictionary

- All variables should be documented
- Allows for easier re-use of data
- REDCap will produce this for you
  - code book or data dictionary formats
- The key to all your short-hands
  - 0=No, 1=Yes
  - 0=Placebo, 1=Treatment
  - CAD = coronary artery disease, PVD = peripheral vessel, MI = myocardial infarction (heart attack)
- Gives how variables were defined if there might be any subjectivity / different definitions used
- Statistical Perspective: Without a data dictionary, your statistician will not be able to work with the data. We only see the variable names, not the descriptions or original questions, on the “back end”.

Best Practice #8: Data Dictionary

	B	C	D	E	F	G	H
1	CRF Name or Calculated Field	Variable Name	Variable Label	Description	Data Type/Format (e.g. numeric, date, string)	Data Codes (for numeric data that represents categories)	Missing Allowed
2	BASELINE	date0	Date at Baseline	Date patient completed baseline assessment	date m-d-yyyy		n
3	BASELINE	sex	Sex	Biological sex of patient	numeric	0, Male 1, Female	n
4	BASELINE	gender	Patient Gender Identity	An individual's personal sense of being a man, woman, or other gender, regardless of the sex that person was assigned at birth (LOINC 76691-5)	numeric	0, Identifies as male 1, Identifies as female 2, Female-to-male transsexual 3, Male-to-female transsexual 4, Identifies as non-conforming 5, Other 8, Asked but unknown 9, Unknown	n
5	BASELINE	race	Race	Self-reported patient race	numeric	0, White 1, Black 2, American Indian or Alaskan Native 3, Asian 4, Native Hawaiian or Other Pacific Islander 5, More than One Race 9, Decline to answer / Unknown	n
6	BASELINE	ethnicity	Ethnicity	Self-reported patient ethnicity	numeric	0, Non-Hispanic 1, Hispanic	n
7	BASELINE	dob	Date of Birth	Patient date of birth	date m-d-yyyy		n
8	calculated	age	Age at baseline	Calculated age in years as: date0 - dob	numeric		n

Best Practice #9: Missing Data

- Best practice for research: Code why it is missing
  - Not applicable
  - Not measured / not asked / respondent refused
  - Below limit of detection
- Best practices for coding:
  - Avoid entering “NA” (or any character data) in a numeric variable
  - Use a numeric code that is NOT a valid answer
    - E.g. Do not use 99 as the missing code in age
  - Use codes that are obviously wrong
    - E.g. 999 is clearly not a valid age
- Statistical Perspective: If a variable has no value, we assume it’s a data entry error and we will query the CRC to fix it.

Best Practice

- #10: Do not use font color or highlighting as data
- #11: Make backups, or ensure it's stored where automatic backups are done
  - Oncore, Advarra EDC, REDCAP, will be backed up automatically
- #12: Use data validation to avoid errors
  - Controls the type or values that can be entered into a variable

Best Practice #13: Don’t assume the data is clean

- Periodically, a study’s data should be monitored / checked for potential errors.
- Involved looking errors such as:
  - missing data,
  - out of range entries,
  - two-way entries that don’t make sense. E.g. pregnant male
- Data queries will be sent to CRCs to correct
- Critical that queries are dealt with quickly
- Statistical Perspective: Anything that doesn’t make sense to the statistician can make the analysis x10 harder

Best Practice #14: Lock Data for Analyses

- Once the data has been cleaned, the database should be locked or frozen for analyses.
- Corrections in the database will not make it into the analyses after data lock, unless the statistician is informed of the change

"*Freezing the data means that no data will be added or changed during the next phase of the analysis. Freezing the data saves time and money, but freezing uncleaned data obviously is a waste of time.*"

Conclusion

- Data Management is a critical part of any study
- Thoughtful data entry form design and careful data entry can “make or break” a study
- CRCs are the “front line” for data quality