ICTR Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

Introduction to data management and analysis

Mimi Kim, Sc.D. – Director of BERD Core

Melissa Fazzari, Ph.D. – Associate Director

Division of Biostatistics

EINSTEIN
Albert Einstein College of Medicine

Montefiore

Objectives for this webinar:

1. Understanding rigor and reproducibility in research
2. The importance of proper data management
3. Some best practices in data management and statistical summaries
4. Introduction to BERD House

Example

- Suppose we conducted a clinical trial of a new treatment to reduce blood pressure.
- The goal was to show that the new treatment is more effective in reducing blood pressure compared to standard medication
- We observe a statistically significant difference in systolic blood pressure reduction at 12 months (6 mm Hg in the treatment arm vs -1 mm Hg in the standard arm)

We publish the findings…but do we trust the results?

Scientific Advancement

"The bottom line.....is that when **science is done well**, it produces believable, **replicable**, and **generalizable** findings"

-Jon Krosnick, the Frederic O. Glover Professor of Humanities and Social Sciences in the Stanford School of Humanities and Sciences

Scientific Rigor = science done well

Scientific rigor is the strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation and reporting of results

Scientific Rigor

Some ways to ensure scientific rigor:

- **Randomization and blinding**
- **Adequate power/sample size** – *under reasonable assumptions, allowing us to be confident that if the treatment reduces BP by a certain amount, we will likely be able to detect it*
- **Carefully defined outcome measures** - *exactly how the outcome will be measured, under what conditions, procedures to minimize bias and handle potential data issues like missing data*
- **Proper statistical analysis** – *giving us the best tools to measure and test for an effect*

How does poor data collection impact scientific rigor?

- If data are collected not according to protocol *(ex: no rest period before measuring BP)*.
- Lax data entry can result in inconsistencies and errors. *Data are entered incorrectly, in the wrong units*

What if data collection or entry is different between treatment arms?

Data entry

patient	Systolic BP at baseline	Diastolic BP at baseline	Weight (lbs)
1	121	72	134
2	140	90	152
3	105	120	133
3	105	120	133
4	108	84	104

mix-up

duplicates

wrong units: 104kg = 229lbs

Errors can be systematic (occurring mostly in one group) or random and cause lower power/precision, over/underestimated treatment effects, or incorrect interpretations

Mistakes can cause a treatment effect to appear or disappear

Scientific Reproducibility

- Scientific (computational) **reproducibility** is the ability of independent researchers to obtain identical results when using the same **computational steps**, methods, code, and raw data.

- **Replicability** is consistency with the original results if we follow the same experimental design and methods, obtain new data, and analyze

Note: Often these two terms are used interchangeably

- **Generalizability** is when the results are consistent across different populations (ex: younger patients, rural populations, different countries) and time

Rigor and Reproducibility

When study results cannot be reproduced or replicated, it is not always obvious what is different between the studies and whether it is a consequence of:

- biological variability
- insufficient sample size
- poor generalization to different populations
- experimental and data errors
- failure to implement the same protocol
- incorrect statistical analysis
- p-value hacking
- fraud

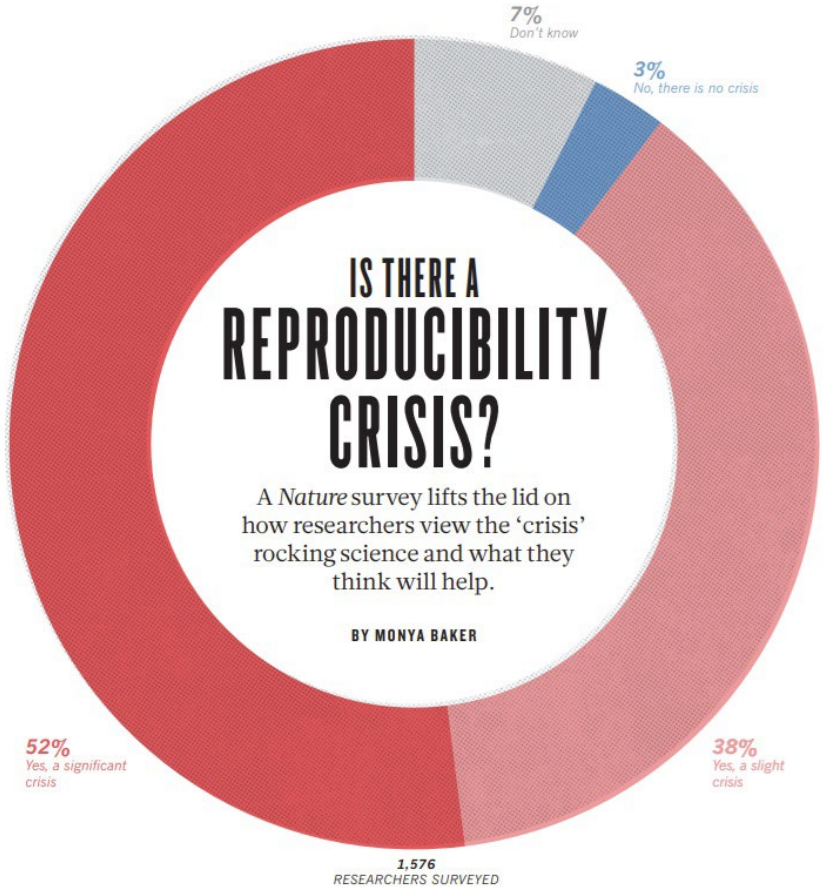
Scientific Reproducibility

As a biostatistician, many times I have received code and data from other biostatisticians or institutions and **not been able to reproduce** tables and figures.

....And to be honest, there have been situations where I have worked on an analysis and then a year later been unable to reproduce my original results easily and spent days tracking down the differences to resolve it.

Reproducibility crisis in medical research?

- In 2016, the journal Nature surveyed 1576 researchers on reproducibility
- 52 percent reported that replicating results is a “significant” problem and another 38 percent believe a “slight crisis” exists.
- More than **70 percent of researchers** have tried and failed to reproduce another scientist’s experiments.
- **More than half of the respondents** reported that they failed to reproduce their own experiments.



IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the ‘crisis’ rocking science and what they think will help.

BY MONYA BAKER

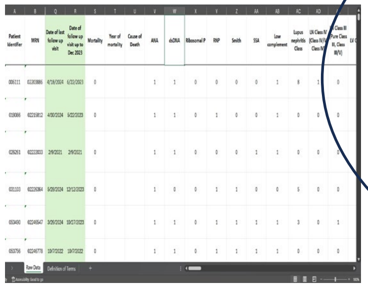
Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).

Rigor and Reproducibility

Study Design and Research Questions

Patient data

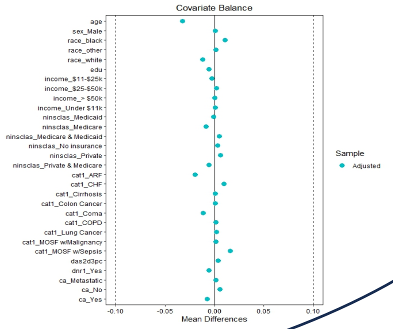
Data entry



A	B	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	
Patient Identifier	MRN	Date of last follow up visit	Date of follow up visit up to Dec 2023	Mortality	Year of mortality	Cause of Death	ANA	dsDNA	Ribosomal P	RNP	Smith	SSA	Low complement	Lupus nephritis Class	LN Class IV (Class IV/V Class IV	N Class III (Pure Class III, Class III/V)
006111	02203886	4/18/2024	6/22/2023	0			1	1	0	0	0	0	1	8	1	0
019066	02215812	4/30/2024	5/22/2023	0			1	1	0	1	0	0	1	0	0	0
026261	02222033	2/9/2021	2/9/2021	0			1	1	0	0	0	1	1	0	0	0
031103	02226364	5/28/2024	12/12/2023	0			1	0	0	1	1	0	0	5	0	0
053490	02246547	3/26/2024	10/27/2023	0			1	1	0	1	1	1	1	3	0	1
053756	02246778	10/7/2022	10/7/2022	0			1	1	0	1	1	1	1	0	0	0

Data management

Analysis/interpretation



Many entry points for error. You can have valid study design and properly collected data and still have a lack of reproducibility and rigor at the end!

Data management

Rigor and Reproducibility can be impacted when there is:

- **Inconsistent data collection**: Different sites define and measure the outcome differently
- **Lack of standardization**: for example, each site has their own unique EXCEL spreadsheet, resulting in inconsistent data fields and formats across sites.
- **No validation**: no validation rules or error checks in place for the data entry forms, allowing for out-of-range values, typographical errors, and other inaccuracies to go unnoticed.
- **Different versions of the data set** with poor tracking of edits made

Poor data management:

- *How did I get to 84 patients in my analysis set from the original 100 patients on the study?*
- *Which patients did I "correct" BPs on?*
- *Which file represents the final analysis data set?*
- *Why did the resident/study coordinator/PI change these values?*
- *How did I produce this table/figure/summary?*
- *I need to send the study data to a group at Northwestern. It's a mess!!!*

EXCEL

Typical scenario: interviewing patients or extracting data from EHR or chart review and storing in an Excel spreadsheet

Issues with using EXCEL to manage data:

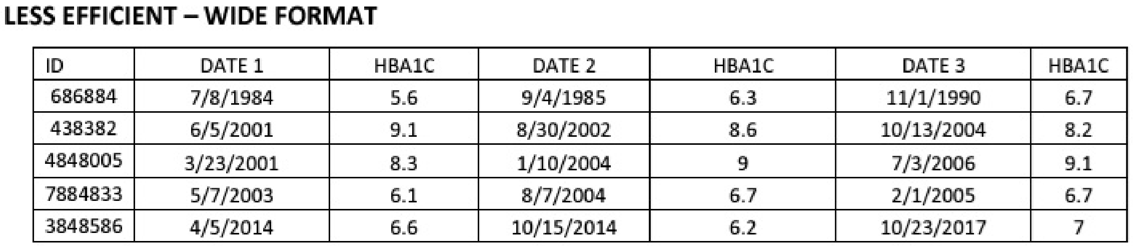
- No explicit version control or date stamp. Once a change is made the previous value is gone forever unless carefully recorded and the previous version is saved
- Wide spreadsheets with lots of variables is difficult to navigate
- Standardization – e.g.: “Black” vs “black” vs “black/other” vs “AA”
- Inconsistent missing codes – “N/A” vs “.” vs “na” vs “missing” vs. “not applicable” vs. “no value” vs. “-”………….
- Dates like *01/18/2204* are seen <u>all</u> the time

Working in EXCEL

- Be consistent with EVERYTHING!
  - naming conventions (DBP_wk1, wk_6_DBP, DBP_1year is not consistent),
  - missing value codes
  - formats for dates – stick to one way
- Use descriptive variable names (but not too long)
- Assume your data sheet will eventually be transported into SAS, R, SPSS for analysis,…highlighting and different colors should not be used to convey information

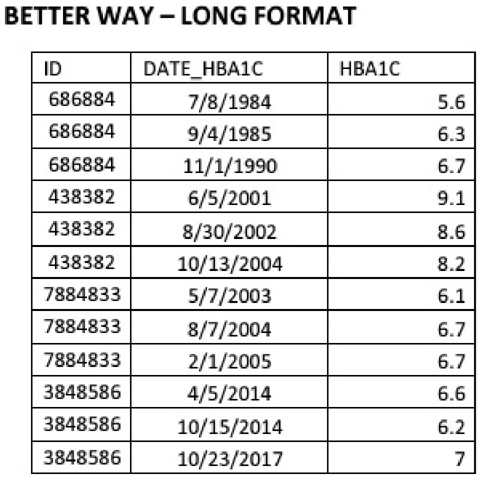
Working in EXCEL

- EXCEL data sheets for data input and clinical use are not formatted in the same way as data sheets for analysis
  - Example long vs wide data



LESS EFFICIENT – WIDE FORMAT

ID	DATE 1	HBA1C	DATE 2	HBA1C	DATE 3	HBA1C
686884	7/8/1984	5.6	9/4/1985	6.3	11/1/1990	6.7
438382	6/5/2001	9.1	8/30/2002	8.6	10/13/2004	8.2
4848005	3/23/2001	8.3	1/10/2004	9	7/3/2006	9.1
7884833	5/7/2003	6.1	8/7/2004	6.7	2/1/2005	6.7
3848586	4/5/2014	6.6	10/15/2014	6.2	10/23/2017	7

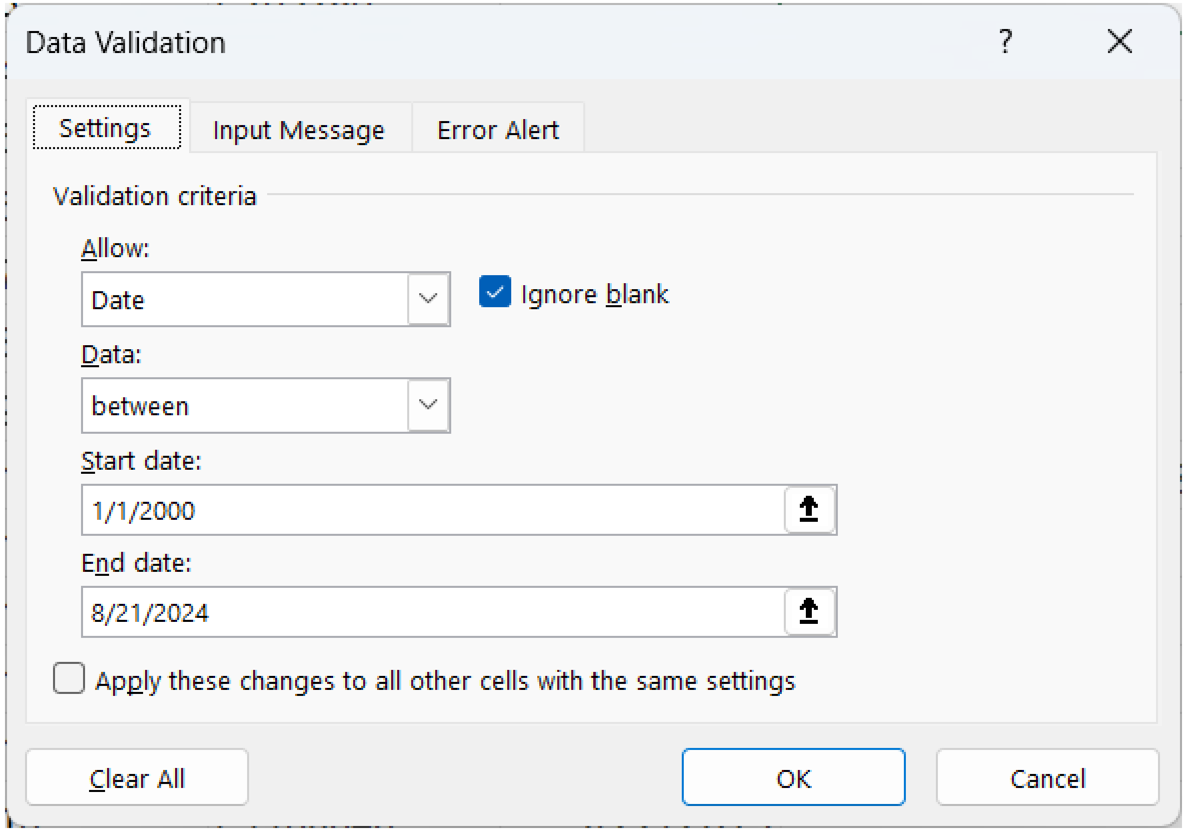


BETTER WAY – LONG FORMAT

ID	DATE_HBA1C	HBA1C
686884	7/8/1984	5.6
686884	9/4/1985	6.3
686884	11/1/1990	6.7
438382	6/5/2001	9.1
438382	8/30/2002	8.6
438382	10/13/2004	8.2
7884833	5/7/2003	6.1
7884833	8/7/2004	6.7
7884833	2/1/2005	6.7
3848586	4/5/2014	6.6
3848586	10/15/2014	6.2
3848586	10/23/2017	7

Working in EXCEL

Use data validation in EXCEL



Data Validation

Settings | Input Message | Error Alert

Validation criteria

Allow:

Date

☑ Ignore blank

Data:

between

Start date:

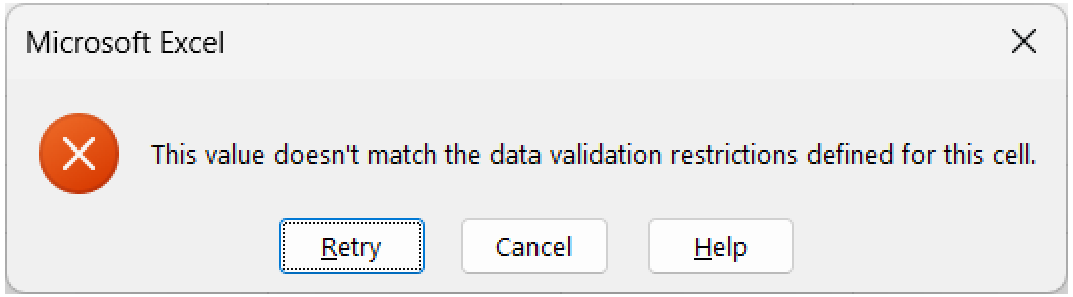
1/1/2000

End date:

8/21/2024

☐ Apply these changes to all other cells with the same settings

Clear All | OK | Cancel



Microsoft Excel

This value doesn't match the data validation restrictions defined for this cell.

Retry | Cancel | Help

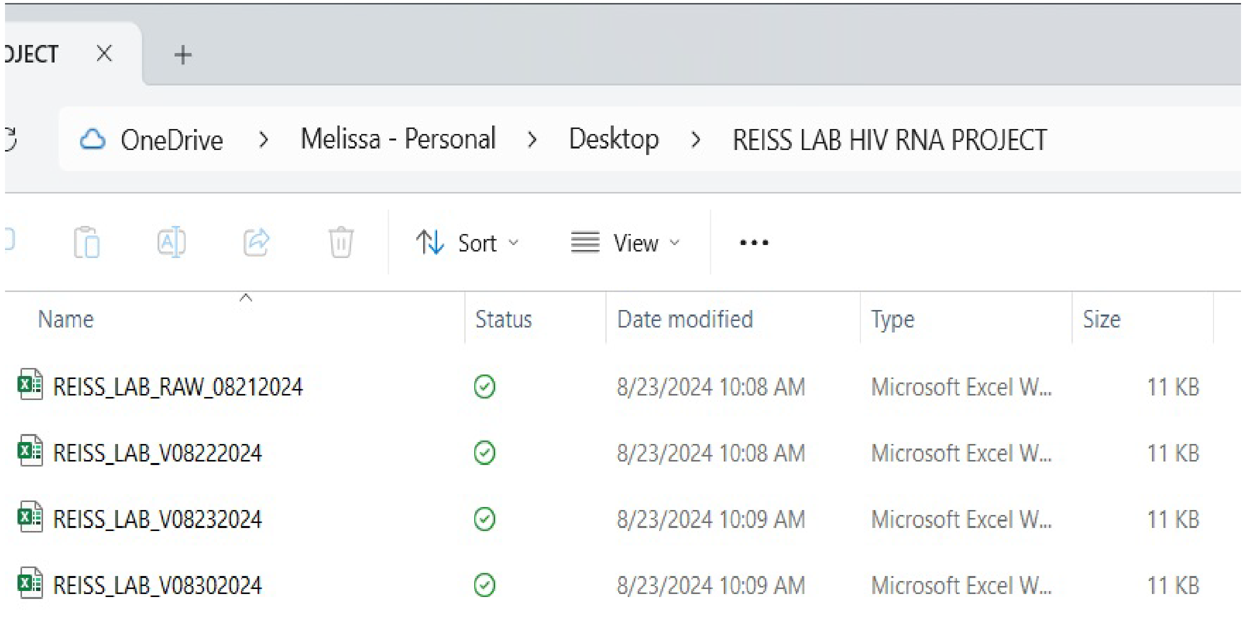
Working in EXCEL

- Start with a copy of the original data that will remain untouched

- When data corrections are made,

  save in a different version

- Create a READ ME file detailing all data corrections

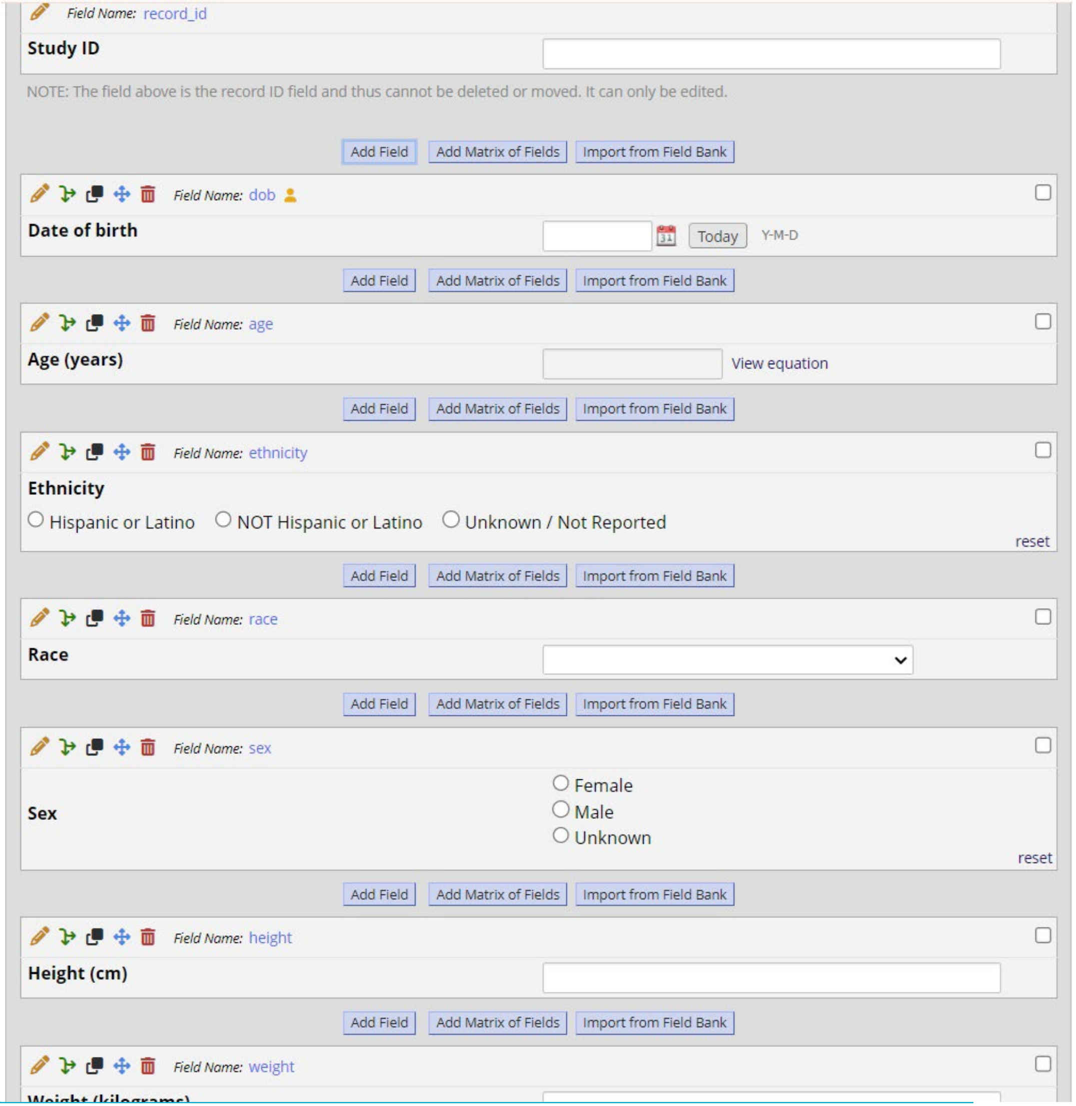


OJECT
OneDrive > Melissa - Personal > Desktop > REISS LAB HIV RNA PROJECT
Sort
View

Name	Status	Date modified	Type	Size
REISS_LAB_RAW_08212024	✓	8/23/2024 10:08 AM	Microsoft Excel W...	11 KB
REISS_LAB_V08222024	✓	8/23/2024 10:08 AM	Microsoft Excel W...	11 KB
REISS_LAB_V08232024	✓	8/23/2024 10:09 AM	Microsoft Excel W...	11 KB
REISS_LAB_V08302024	✓	8/23/2024 10:09 AM	Microsoft Excel W...	11 KB

REDCap
Research Electronic Data Capture

- An efficient electronic data capture software
- Meets HIPAA requirements for collection of protected health information
- Multiple users per project for administrators, analysts, data entry personnel
- **Built-in reports and descriptive statistics**
- Can also be used for longitudinal studies and randomized clinical trials
- Can **build-in controls** for allowable data values and logic



Field Name: record_id
Study ID

NOTE: The field above is the record ID field and thus cannot be deleted or moved. It can only be edited.

Add Field | Add Matrix of Fields | Import from Field Bank

Field Name: dob
Date of birth Today Y-M-D

Add Field | Add Matrix of Fields | Import from Field Bank

Field Name: age
Age (years) View equation

Add Field | Add Matrix of Fields | Import from Field Bank

Field Name: ethnicity
Ethnicity
○ Hispanic or Latino ○ NOT Hispanic or Latino ○ Unknown / Not Reported
reset

Add Field | Add Matrix of Fields | Import from Field Bank

Field Name: race
Race

Add Field | Add Matrix of Fields | Import from Field Bank

Field Name: sex
Sex
○ Female
○ Male
○ Unknown
reset

Field Name: height
Height (cm)

Field Name: weight
Weight (kilograms)

Best practices in data management

- Always have a time-stamped **raw, untouched** data set. Never manipulate this version so you can always go back to it.
- Keep a careful **data flow chart** – when patients are removed from a data set of values are changed you should have a record of when and why
- Work with **“frozen” data files** by date
- **Clean and analyze data in SAS/R/Stata** if possible (code based)
- Create code that can be **re-run easily** – include lots of comments and all data changes coded

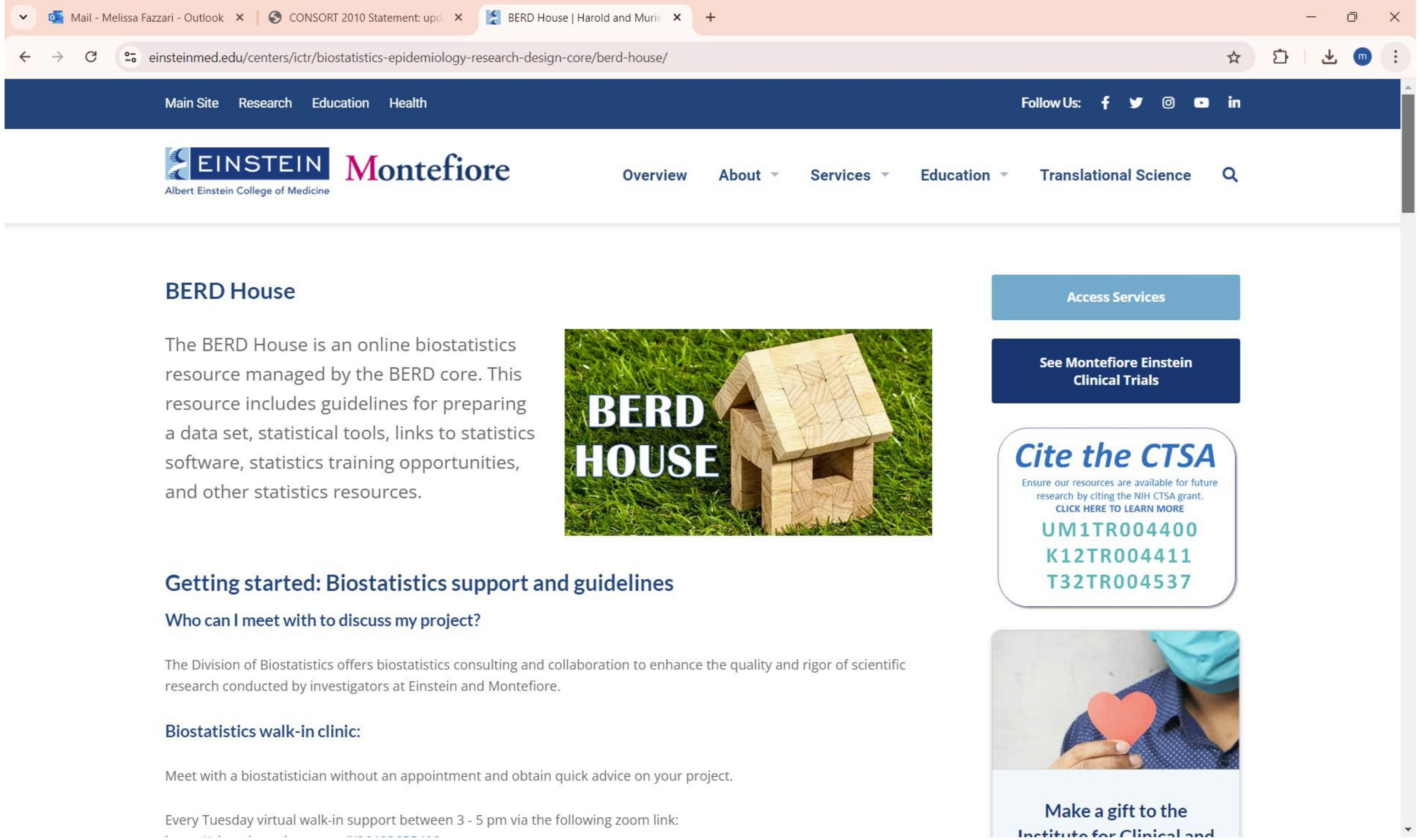
Helpful resources

- Published guidelines for clinical trials (e.g. CONSORT, PCORI)
- **Research Data Management Planning Checklists**

https://researchdata.wisc.edu/resources/research-data-management-planning-checklist/

  - Data collection – types, formats
  - Data storage
  - Data security
  - Data retention

BERD House


Mail - Melissa Fazzari - Outlook
CONSORT 2010 Statement: upd
BERD House | Harold and Murie
einsteinmed.edu/centers/ictr/biostatistics-epidemiology-research-design-core/berd-house/
Main Site
Research
Education
Health
Follow Us:
EINSTEIN
Albert Einstein College of Medicine
Montefiore
Overview
About
Services
Education
Translational Science

BERD House

The BERD House is an online biostatistics resource managed by the BERD core. This resource includes guidelines for preparing a data set, statistical tools, links to statistics software, statistics training opportunities, and other statistics resources.

Getting started: Biostatistics support and guidelines

Who can I meet with to discuss my project?

The Division of Biostatistics offers biostatistics consulting and collaboration to enhance the quality and rigor of scientific research conducted by investigators at Einstein and Montefiore.

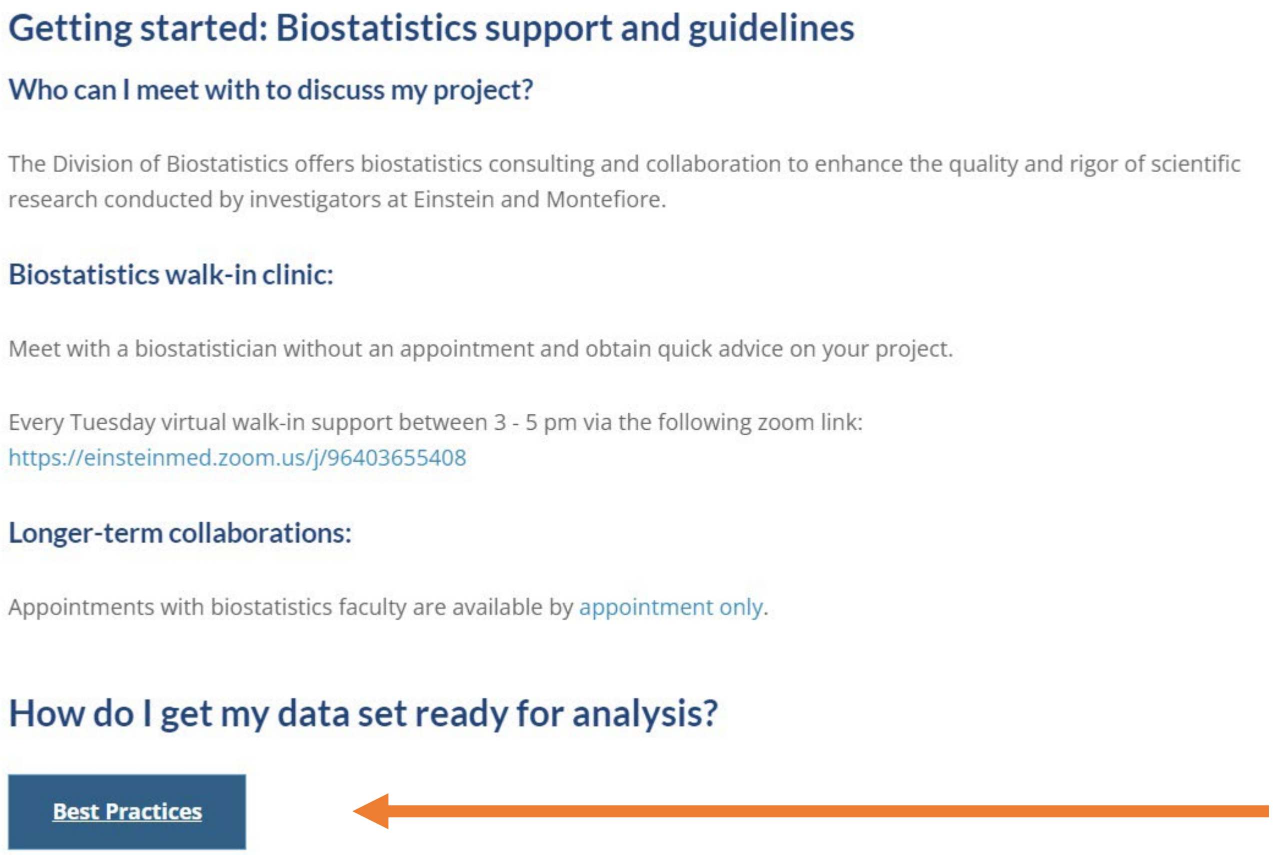
Biostatistics walk-in clinic:

Meet with a biostatistician without an appointment and obtain quick advice on your project.

Every Tuesday virtual walk-in support between 3 - 5 pm via the following zoom link:

Access Services
See Montefiore Einstein Clinical Trials
Cite the CTSA
Ensure our resources are available for future research by citing the NIH CTSA grant.
CLICK HERE TO LEARN MORE
UM1TR004400
K12TR004411
T32TR004537
Make a gift to the
Institute for Clinical and

BERD House



Getting started: Biostatistics support and guidelines

Who can I meet with to discuss my project?

The Division of Biostatistics offers biostatistics consulting and collaboration to enhance the quality and rigor of scientific research conducted by investigators at Einstein and Montefiore.

Biostatistics walk-in clinic:

Meet with a biostatistician without an appointment and obtain quick advice on your project.

Every Tuesday virtual walk-in support between 3 - 5 pm via the following zoom link:
https://einsteinmed.zoom.us/j/96403655408

Longer-term collaborations:

Appointments with biostatistics faculty are available by appointment only.

How do I get my data set ready for analysis?

Best Practices in summarizing data

- Find a statistical software/data management package that works for you
- Create a **standard template** for reporting results – what variables are most important to summarize? What kinds of summaries are expected/appropriate?
- Avoid cut and paste or manual entry of summaries

Types of Data

Important to understand the form of your data before trying to summarizing it

1. Continuous Data

   Examples: Vitamin D levels, gene expression, weight,….

   A continuous variable can take any value within an interval

2. Categorical (discrete) data

   Examples: gender, event occurrence (yes/no) pain score (1-5)…..

   A categorical variable can only take discrete values

Types of Data

- We may further break down discrete data into:
  - **Nominal data**

    For example, gender

  - **Ordinal data**

    For example, pain scores are between 1 (no pain) and 5 (severe pain).

    The order matters, but not the difference between values.

Measures of Central Tendency

Describes the central position within a set of data – **which value the data tend to cluster around**

Common statistics:

- Arithmetic Mean – impacted by outliers (for example, the average net worth in the US is highly impacted by Bill Gates and Jeff Bezos)
- Median – the middle number in a ranked list
- Mode – the most commonly occurring number

Measures of Spread

- **Range**
  - ▻ Interval between the highest and lowest values in a sample
  - ▻ Heavily influenced by extreme values
- **Percentiles**
  - ▻ Order data and divide into 100 separate cut-points (percentiles)
  - ▻ 25th percentile is 13.2 -> that 25% of observed values are ≤ 13.2
  - ▻ **IQR – 75th -25th percentiles**
- **Standard deviation –** how much a “typical” observation varies from the mean. Assumes a symmetrical distribution

$$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}}$$

Summarizing Discrete data types

- For nominal data types (e.g.: race/ethnicity, biological sex)
  - Counts (n) and %
- For ordinal data (e.g.: pain scale 1-5)
  - Counts (n) and %
  - Median to reflect the middle value
  - Range and IQR to reflect the spread of values
  - The “average” value makes less sense here unless we can assume equal distances between values

Summarizing Discrete data types

Example: Suppose we survey 1000 people on their yearly income to examine the association with health outcomes. We code three levels of income:

1 = low	$0 - $30,000/year
2 = moderate	$30,001-100,000/year
3 = high	>$100,000/year

Observed incomes (n=7): 25k, 30k, 35k, 36k, 195k, 250k, 720k

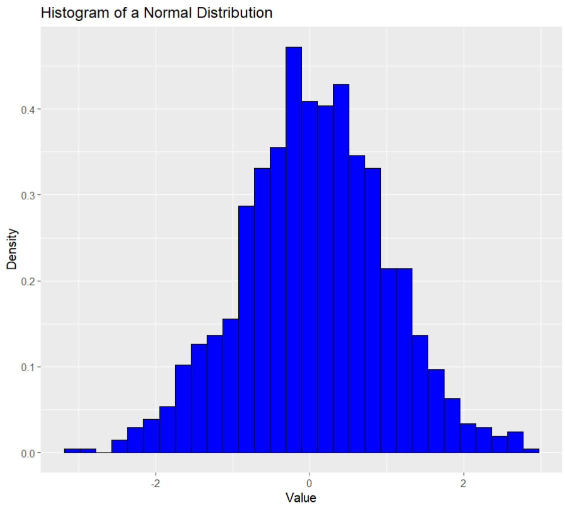
Average income level: 2.14. *What does this mean?*

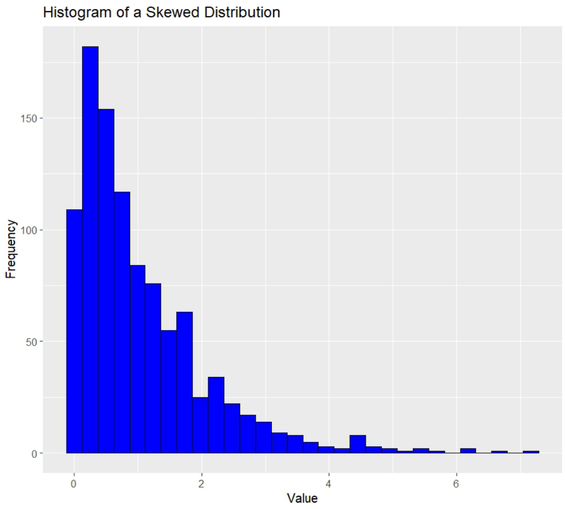
The mean treats income levels as if they are points on a **continuous** scale with **uniform spacing,** which does not make sense here

Summarizing Continuous data types

- If bell-shaped, present mean and standard deviation

- If skewed, present median, IQR


Histogram of a Normal Distribution


Histogram of a Skewed Distribution

Statistical Analysis

The distribution of your outcome variable **completely dictates** the type of statistical analysis or model or machine learning algorithm you will use:

- Continuous, normally distributed -> t-tests, regression, ANOVA
- Continuous, non-normally distributed -> Wilcoxon, Kruskal-Wallis test
- Binary -> chi-square tests, logistic regression, decision trees, XGB
- Count -> Poisson regression, ordinal logistic regression

Statistical Analysis

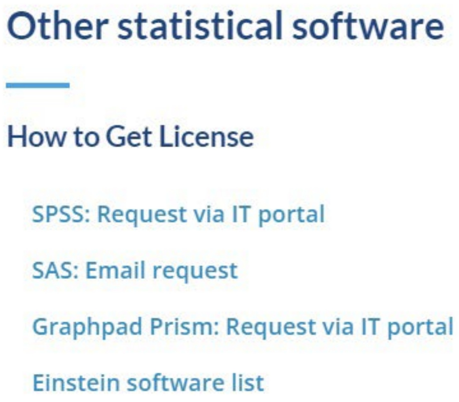
- Data management usually takes a lot of time and is a very important part of the research

- Sophisticated modeling or machine learning is meaningless if the experiment and data are not correct

- Remember – a logistic regression model, Random Forest, neural network, clustering algorithms,…. will take your variables as inputs whether they are right or wrong and find patterns in the data

Introduction to BERD House

- Using statistical packages can increase rigor and reproducibility
- If you are interested in obtaining a statistical software package, BERD House has resources listed



Other statistical software

How to Get License

SPSS: Request via IT portal

SAS: Email request

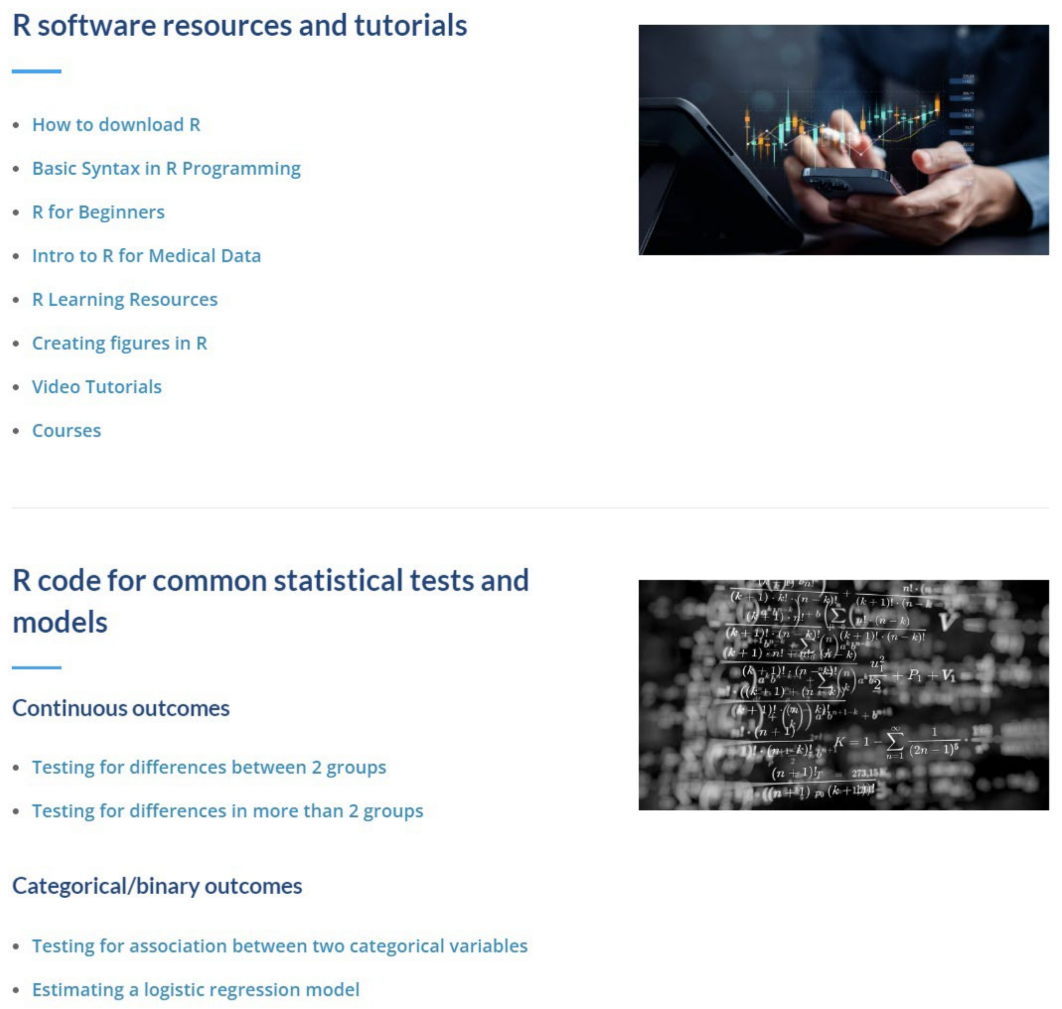
Graphpad Prism: Request via IT portal

Einstein software list

https://einsteinmed.edu/centers/ictr/biostatistics-epidemiology-research-design-core/berd-house

- R is freely available for download online. BERD House has many tutorials for learning R, and we now offer introductory R workshops

Introduction to BERD House



R software resources and tutorials

- How to download R
- Basic Syntax in R Programming
- R for Beginners
- Intro to R for Medical Data
- R Learning Resources
- Creating figures in R
- Video Tutorials
- Courses

R code for common statistical tests and models

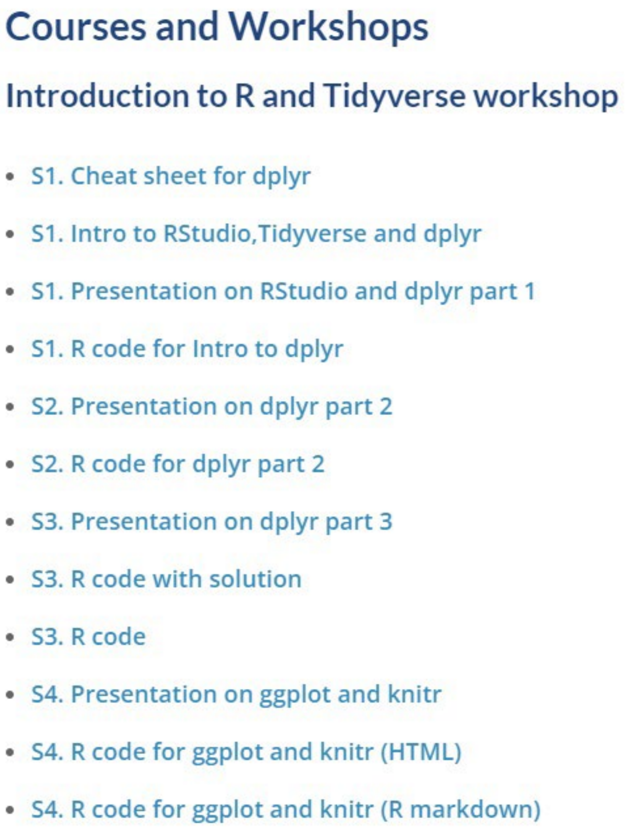
Continuous outcomes

- Testing for differences between 2 groups
- Testing for differences in more than 2 groups

Categorical/binary outcomes

- Testing for association between two categorical variables
- Estimating a logistic regression model

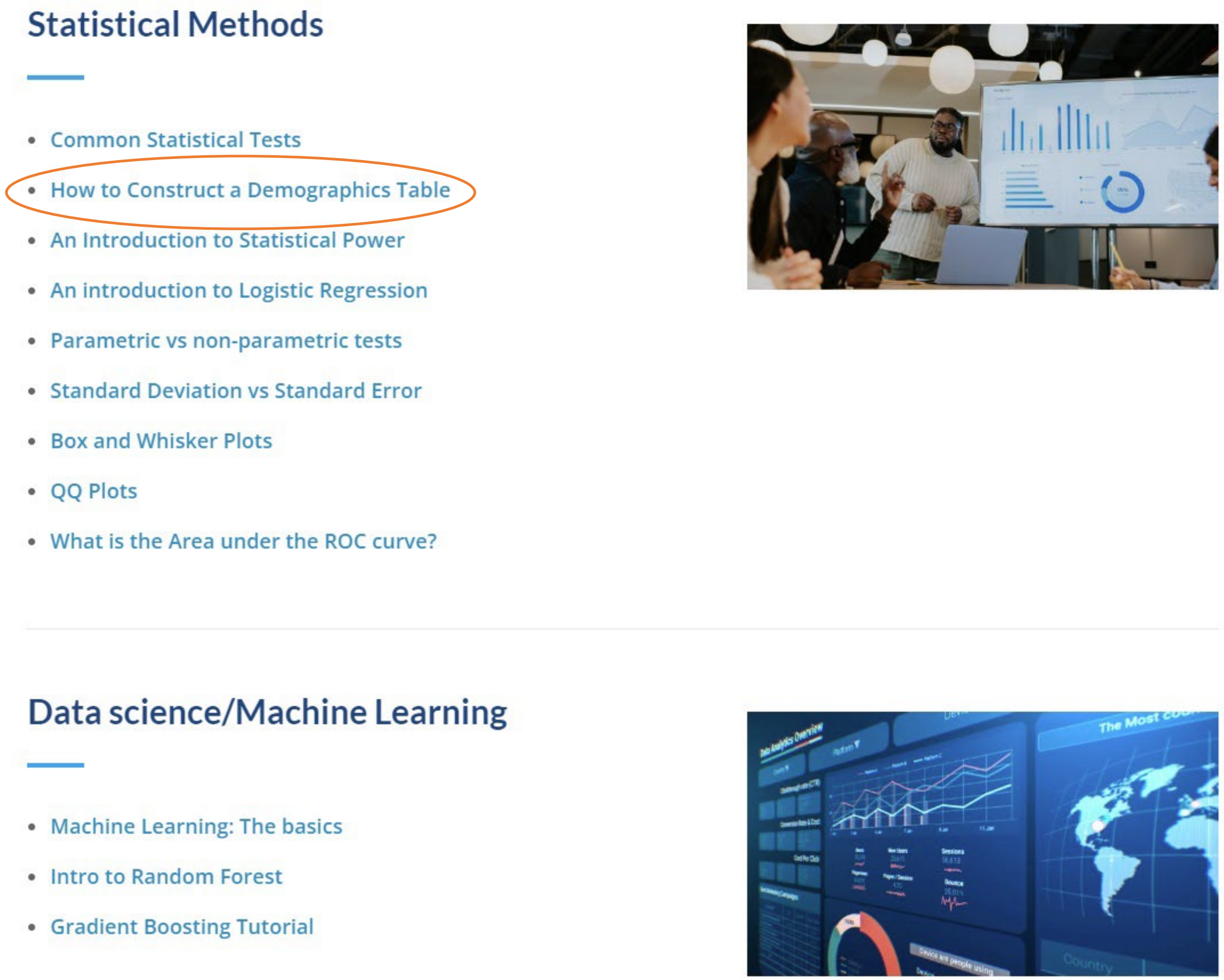
Introduction to BERD House



Courses and Workshops

Introduction to R and Tidyverse workshop

- S1. Cheat sheet for dplyr
- S1. Intro to RStudio,Tidyverse and dplyr
- S1. Presentation on RStudio and dplyr part 1
- S1. R code for Intro to dplyr
- S2. Presentation on dplyr part 2
- S2. R code for dplyr part 2
- S3. Presentation on dplyr part 3
- S3. R code with solution
- S3. R code
- S4. Presentation on ggplot and knitr
- S4. R code for ggplot and knitr (HTML)
- S4. R code for ggplot and knitr (R markdown)



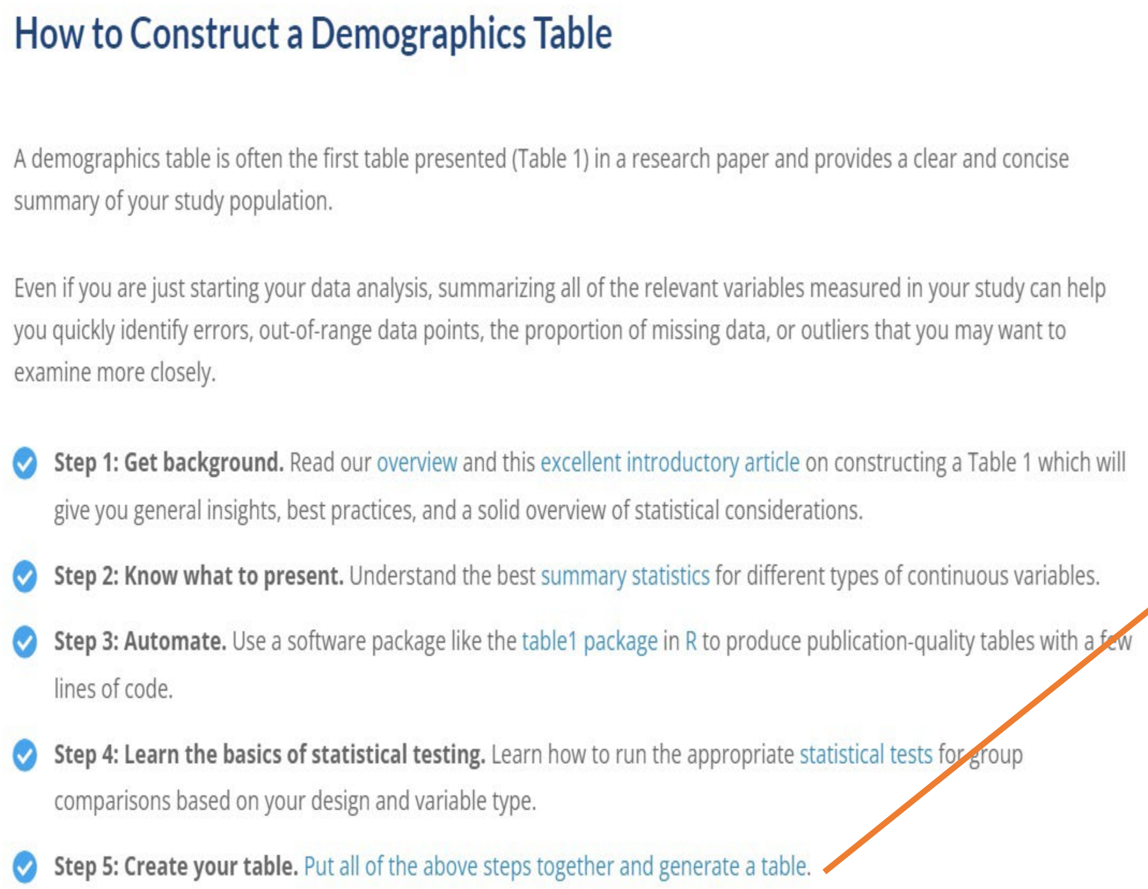
Statistical Methods

- Common Statistical Tests
- How to Construct a Demographics Table
- An Introduction to Statistical Power
- An introduction to Logistic Regression
- Parametric vs non-parametric tests
- Standard Deviation vs Standard Error
- Box and Whisker Plots
- QQ Plots
- What is the Area under the ROC curve?

Data science/Machine Learning

- Machine Learning: The basics
- Intro to Random Forest
- Gradient Boosting Tutorial

Introduction to BERD House

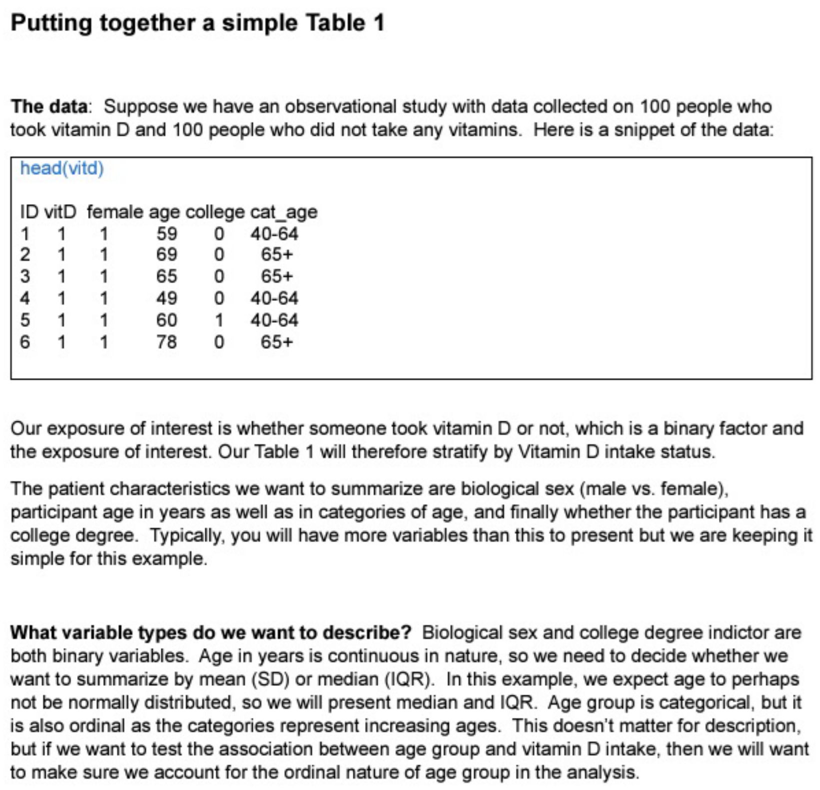


How to Construct a Demographics Table

A demographics table is often the first table presented (Table 1) in a research paper and provides a clear and concise summary of your study population.

Even if you are just starting your data analysis, summarizing all of the relevant variables measured in your study can help you quickly identify errors, out-of-range data points, the proportion of missing data, or outliers that you may want to examine more closely.

- **Step 1: Get background.** Read our overview and this excellent introductory article on constructing a Table 1 which will give you general insights, best practices, and a solid overview of statistical considerations.
- **Step 2: Know what to present.** Understand the best summary statistics for different types of continuous variables.
- **Step 3: Automate.** Use a software package like the table1 package in R to produce publication-quality tables with a few lines of code.
- **Step 4: Learn the basics of statistical testing.** Learn how to run the appropriate statistical tests for group comparisons based on your design and variable type.
- **Step 5: Create your table.** Put all of the above steps together and generate a table.



Putting together a simple Table 1

The data: Suppose we have an observational study with data collected on 100 people who took vitamin D and 100 people who did not take any vitamins. Here is a snippet of the data:

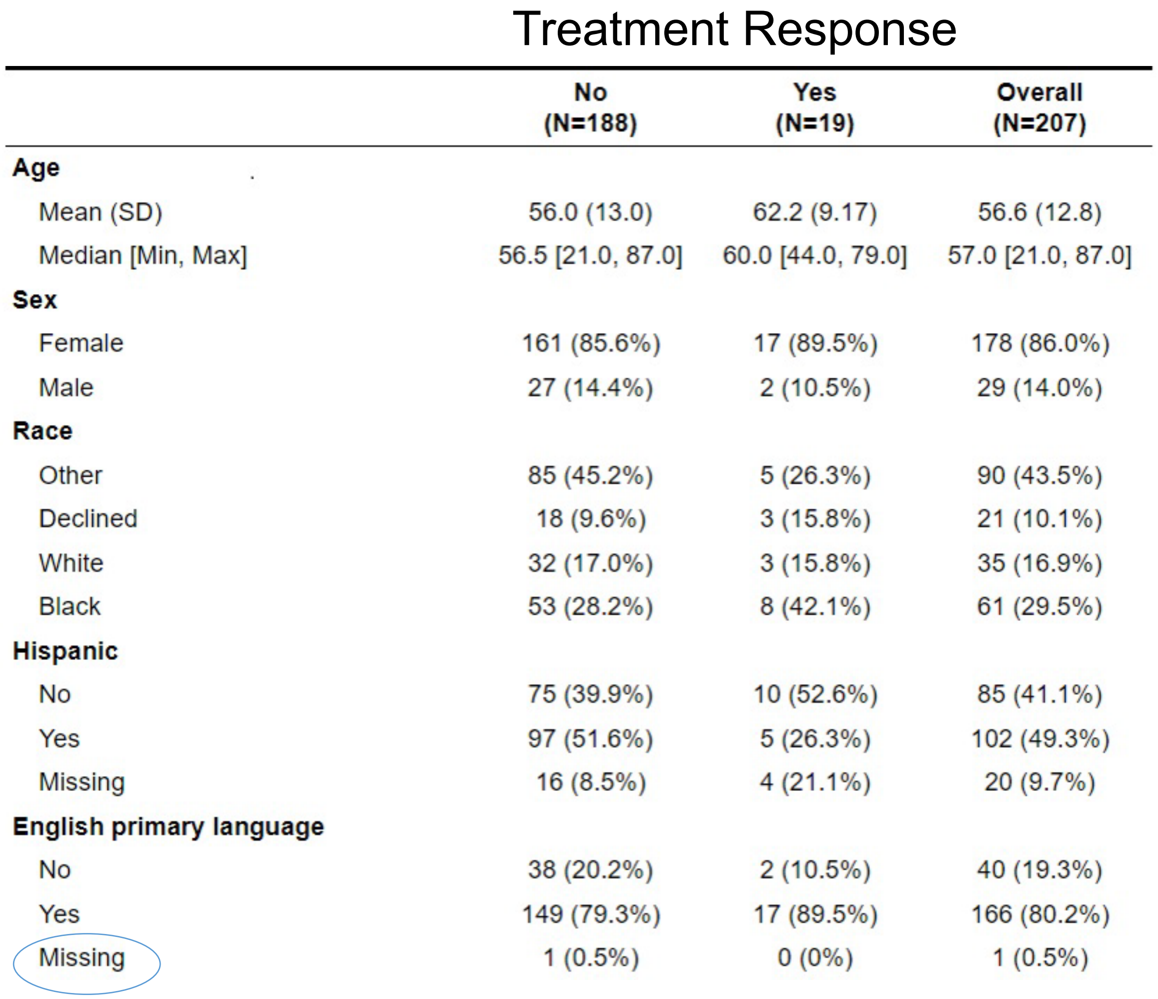
```
head(vitd)

ID vitD  female age college cat_age
1    1     1      59     0    40-64
2    1     1      69     0     65+
3    1     1      65     0     65+
4    1     1      49     0    40-64
5    1     1      60     1    40-64
6    1     1      78     0     65+
```

Our exposure of interest is whether someone took vitamin D or not, which is a binary factor and the exposure of interest. Our Table 1 will therefore stratify by Vitamin D intake status.

The patient characteristics we want to summarize are biological sex (male vs. female), participant age in years as well as in categories of age, and finally whether the participant has a college degree. Typically, you will have more variables than this to present but we are keeping it simple for this example.

What variable types do we want to describe? Biological sex and college degree indictor are both binary variables. Age in years is continuous in nature, so we need to decide whether we want to summarize by mean (SD) or median (IQR). In this example, we expect age to perhaps not be normally distributed, so we will present median and IQR. Age group is categorical, but it is also ordinal as the categories represent increasing ages. This doesn't matter for description, but if we want to test the association between age group and vitamin D intake, then we will want to make sure we account for the ordinal nature of age group in the analysis.

Table 1 (demographics table)



Treatment Response

	No (N=188)	Yes (N=19)	Overall (N=207)
Age			
Mean (SD)	56.0 (13.0)	62.2 (9.17)	56.6 (12.8)
Median [Min, Max]	56.5 [21.0, 87.0]	60.0 [44.0, 79.0]	57.0 [21.0, 87.0]
Sex			
Female	161 (85.6%)	17 (89.5%)	178 (86.0%)
Male	27 (14.4%)	2 (10.5%)	29 (14.0%)
Race			
Other	85 (45.2%)	5 (26.3%)	90 (43.5%)
Declined	18 (9.6%)	3 (15.8%)	21 (10.1%)
White	32 (17.0%)	3 (15.8%)	35 (16.9%)
Black	53 (28.2%)	8 (42.1%)	61 (29.5%)
Hispanic			
No	75 (39.9%)	10 (52.6%)	85 (41.1%)
Yes	97 (51.6%)	5 (26.3%)	102 (49.3%)
Missing	16 (8.5%)	4 (21.1%)	20 (9.7%)
English primary language			
No	38 (20.2%)	2 (10.5%)	40 (19.3%)
Yes	149 (79.3%)	17 (89.5%)	166 (80.2%)
Missing	1 (0.5%)	0 (0%)	1 (0.5%)

Discussion

Questions?

Any topics of interest for future webinars?

Feedback?