

Who will win the Masters this year?

- Jordan Spieth
- Dustin Johnson
- Jon Rahm
- Justin Timberlake
- I don't know
- I don't care



Data Management

Prepping Data for Analysis

Heather Tavel, MPH

Evaluation Team: Kaiser Permanente Colorado
Institute for Health Research

Agenda

- ❑ What is data management and preparation?
- ❑ Why is it necessary?
- ❑ What tools can be used to do this?
- ❑ What are the steps/best practices?
- ❑ Challenges
- ❑ Examples

What is Data Management

“An administrative process by which the required data is acquired, validated, stored, protected, and processed, and by which its accessibility, reliability, and timeliness is ensured to satisfy the needs of the data users.”

- www.businessdictionary.com

Data Preparation

“Data Preparation is the process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis.

- Handling messy, inconsistent, or un-standardized data
- Trying to combine data from multiple sources
- Reporting on data that was entered manually”

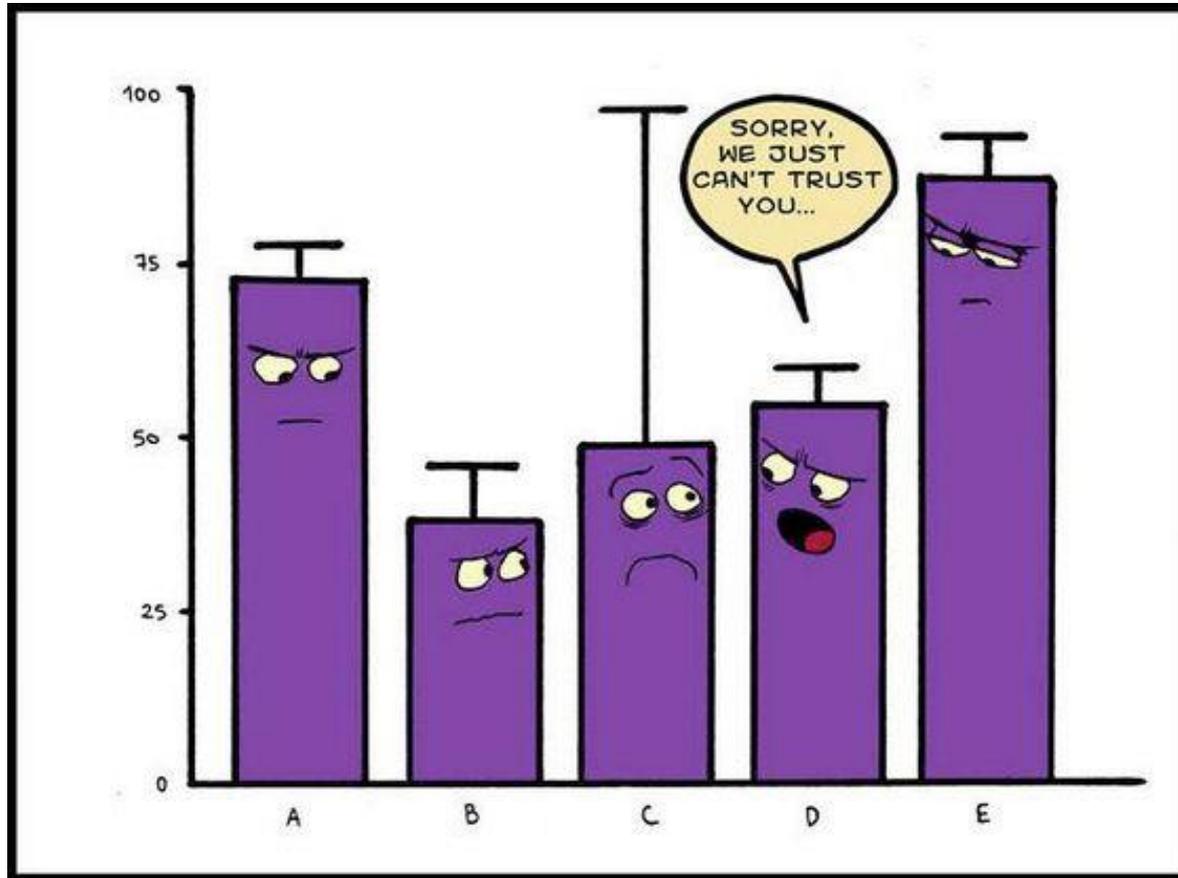
-www.datawatch.com

Why data management is necessary

“I created an awesome survey because of your amazing seminar last year. Why do I need to do more work on prepping the data?”

- Survey results may come back in different formats
- Different software packages may require different format/layout for analytic process
- Different analysts may prefer different format/layout
- Humans create, respond to, and enter data for surveys.
- Humans are flawed.

Why do we have to make sure the data are correct?



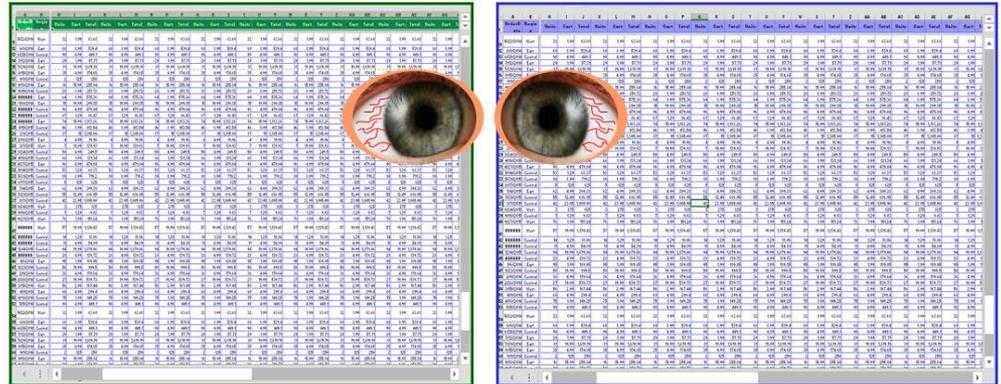
Tools/Methods for Data Management

- Manual

- e.g. Excel
- For smaller data sets

- Programmatic

- e.g. SAS, SQL, SPSS?
- Ideal for larger datasets
- Eyeballing the data only goes so far



Steps of Data Preparation and Management

- Data access and acquisition
- Editing & reviewing data
- Data entry planning
 - Table structures
 - Coding Variables
 - Preparing Codebook/Data Dictionary
- Data entry
- Data cleaning
- Data modification
- Data documentation

Data Acquisition

- EMR or other administrative data systems
- Paper forms
- Survey downloads
- Excel Spreadsheets
- Text files
- Secure FTP
- Analytic files (e.g. SAS datasets)

Editing & reviewing data

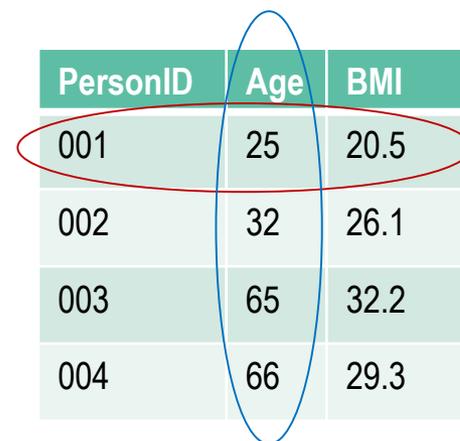
- Part of the quality control process
 - Process to ensure that information on a questionnaire/form is ready to be entered for analysis
 - During data collection, interviewer and supervisor check completed questionnaires for errors
 - Check for missing values that could potentially be followed up on
- You may be getting the data after it is past this point...
 - Entered into the system through survey tool, or provided by a 3rd party
 - Have to take a look to know what's in there!

Planning for data management

- Table structures
- Coding variables
- Codebook preparation

Tables

- Data are stored in a computer file (data file) that is arranged in a certain format, typically called a 'Table'
- Rows are the horizontal elements within a table, also called a 'record'
 - Subjects
 - Respondents
- Columns are the vertical elements of a table, also called 'fields'.
 - Physical or conceptual attributes
 - Questions (answers) on a survey
 - ... etc...that may vary by case/respondent/subject



PersonID	Age	BMI
001	25	20.5
002	32	26.1
003	65	32.2
004	66	29.3

Common Table formats

■ Wide

- Attributes are 'side by side'
- Easier for analyzing within and across participants
- Easier for manual data entry

PersonID	Age	BMI
001	25	20.5
002	32	26.1
003	65	32.2
004	66	29.3

■ Tall-thin

- Attributes are stacked
- Space efficient
- Easier to add attributes
- Requires more data manipulation for analysis

PersonID	Var	Value
001	Age	25
001	BMI	20.5
002	Age	32
002	BMI	26.1
003	Age	65
003	BMI	32.2
004	Age	66
004	BMI	29.3

Column Names

- Sometimes restricted in length by software package
- Avoid special characters (underscore_OK)
- Do not start with a number
- Keep short (but informative!!)

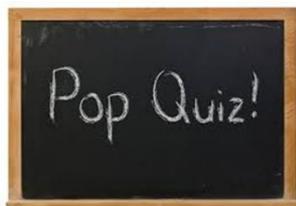
ID	Q1	Q2	Q3
001	25	20.5	120
002	32	26.1	105
003	65	32.2	115
004	66	29.3	132

Vs.

PersonID	Age	BMI	Systolic_BP
001	25	20.5	120
002	32	26.1	105
003	65	32.2	115
004	66	29.3	132

Rows - Uniqueness of observations

- Rows/observations should be identifiable as unique entities/cases
- Duplicate rows can:
 - indicate a data entry error
 - cause confusion
- A unique key variable or combination of variables can help tease apart true duplicates from identical responses
 - Respondent Ids (001, 002, 003) reused for different sites
 - Survey responses over time from same respondents



What would help make observations unique in these cases?

Data Types

- Categorical
 - Race
 - Sex
- Continuous
 - Age
 - BMI
- Some of this is pre-determined by the data coming in. However, continuous variables can be transformed into categorical.
 - Age categories (18-24, 25-44, 45-64, 65+)

Data Types, con't

- Numeric (consists of only numbers)
 - Continuous: 0, 1, 2, 6.4, 1356, etc
 - Ordinal: 1,2,3,4,5, etc
 - Boolean: 0,1
- Character (alpha-numeric)
 - “881CMP”
 - “321WC”
 - “Strongly Agree”
- Numeric values can be used in mathematical computations OR be used categorically

Data Coding

- Coding allows large amounts of information into a form that an analytic tool can handle more easily.
 - Typically thought of as assigning a numeric value to represent a text response
 - Also can be a 'grouping' variable
- Not all data necessarily need to be coded

Examples of coding

- Likert/Ordinal scale

Strongly disagree =1

Disagree =2

Neutral =3

Agree =4

Strongly Agree =5

- Boolean Values/True-False

Male = 0

Female = 1

- Character coding:

- Bill type codes of xx, xx ,xx ,xx = 'OP'

- Bill type coes of xx, xx, xx = 'IP'

Coding missing data

- Why does this happen?
 - Question left blank
 - Respondent chose 'N/A'
 - Data entry errors

- Will need to know –
 - How software package handles missing data analytically
 - '99' in SAS might work for categorical, but not for a numeric field
 - Some software packages (e.g. SAS, STATA) require blanks
 - Whether 'N/A' means something different than an actual missing value
 - Can they be coded the same during cleanup?

Codebooks

- Data should be prepared or entered identically for all participants/respondents/sites
- Indicates how the questionnaire data should be transformed
 - Manual entry or programmatic transforming
 - Computer packages have limits (variable name lengths, etc)
- At the minimum, specifies the valid coding options for the each field
 - Related Questionnaire question
 - Type of variable
 - Length of variable
 - Variable name
 - Coding/translation instructions

Examples

- Pre-planned data entry with the codebook being embedded right in template
- [2016 DPRP and CDPHE Spreadsheet Template.xlsx](#)

Examples, cont'd

	A	B	C	D	G	H
	Variable Name	Ordinal Position	Data Type	Data Type Length	Variable Description	Valid Values
1	ATIME	17	num	8	Start time of an encounter. - Use admission time for inpatient, emergency or institutional encounters. - For other encounters (such as ambulatory visits), use check-in time if this field is populated, otherwise use appointment time. If unknown, specify as null	SAS Time value
19	ADATE	3	num	8	Encounter or admission date from source system. OE/HH and HS adates will be based on the billing cycle, since these represent ongoing home visits. All others are based on the admission or visit date on claim or source encounter.	SAS Date Values between January 1, 1998 and the last VDW load
20	DEPT	19	char	6	The department where the encounter took place as documented in the source data. This is not necessarily the specialty of the clinician providing services.	See "DEPT Codes" tab in this workbook.
22	SOURCE_DATA	20	char	1	Classification of the source database that was used to create this record.	<p>E = Your site's EHR (Electronic Health Record) table operated by your health care organization: Excludes claims data. This category includes the following:</p> <ul style="list-style-type: none"> - Direct extract from your site's EHR - Data that interfaces with your site's EHR (such as a separate lab or radiology system that interfaces with your site's EHR) <p>C = Source is from claims or pre-authorized referrals database</p> <p>L = Local data source but unrelated to your site's EHR</p> <p>M = Multiple, (typically hospital encounters with claims and EHR rounding)</p>
23	ELECTRONIC_CHART_REV	21	char	1	This encounter can be reviewed in Kaiser Colorado's electronic medical record system.	<p>Y=Yes</p> <p>N=No</p> <p>P=Partially (part of the record can be reviewed electronically)</p> <p>U=Unknown</p> <p>Can not be null</p>
24						

Data entry

- True data entry required for written forms
- Care should be taken to follow codebook to a 'T'
- Sometimes data comes to you already entered - skip to data cleaning/modifying

Checking the data

- **Visual:** Scan raw data pages for blatant errors or inconsistencies
- **Verification:** Enter data twice and compare (double data entry)
- **Validity:** verify that only valid codes are used for each question
- **Consistency:** Verify that responses to certain questions are related in reasonable ways to responses to other questions. e.g. birth date and age
- **Usability:** If you did not get to plan for data entry, is the data usable for analysis?

Data Cleaning/Cleansing

Where possible,

- Eliminate duplicate data entries
- Ensure uniqueness of key fields/row identifiers, such as patient ids
- Make sure that the raw data were accurately entered into a computer readable file
- Check for invalid coded values
- Code values and missing values appropriately
- Follow up on missing values for variables where complete data is necessary (ie recruitment date, ID, important dependent variables)
- Verify that more complex multi-file rules have been followed



"This is not what I meant when I said 'we need better data cleansing!'"

Data Validation

- Process by which coding rules can be enforced during data entry
- Employing built-in data validation into your process can make data preparation easier
 - Excel has great tools (selections from lists, error checking)
- Programmatic languages allow for ‘defensive programming’ and automated ways to code variables

Data Modification

Raw data: Measurements that have not been organized, summarized or otherwise manipulated.

- Modify by:
 - Creating new variables
 - Recoding (reverse coding, collapsing categories)
 - Transformation (obtain a normal distribution)
 - Composite variables (scales, indices)

Challenges

- Surveys not always planned with the analysis in mind
 - May need to retrofit data
 - Codebook may be great but not necessarily the best for statistical software
- Data provided from downloads can be very messy, requiring quite a bit of cleanup
- Several different ways to do the same thing
- Changing analyses
- Can't see what you don't know to look for

Examples

- [Revised DPRP and CDPHE Spreadsheet.xlsx](#)

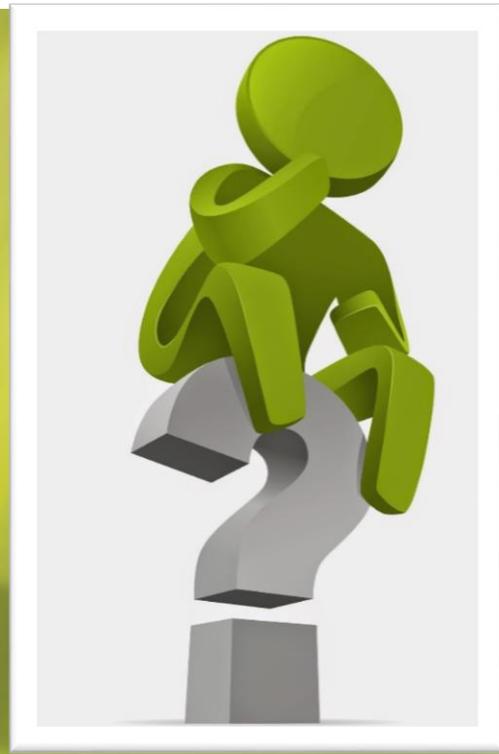
- Coalition Survey
 - [CCATSurvey_PrintVersion.pdf](#)
 - [Survey Monkey results Coalition Effectiveness Survey_2016 Sept.xlsx](#)
 - [Revised Coalition Effectiveness Survey_2017_V1.xlsx](#)

Data Documentation

- In addition to the codebook/data dictionary
- Best practice to document all changes
 - Easy to forget
 - Helps another analyst understand why/what you did
 - Makes it easier to repeat same next time
- Example – changes made to data for this class
 - [Data Steps for Diabetes Data.docx](#)

References

- ❑ McKenzie, J. F., Neiger, B. L., & Thackeray, R. (2013). *Planning implementing & evaluating health promotion programs: A Primer (6th Edition)*. San Francisco, CA: Pearson Benjamin Cummings.
- ❑ California State University, Long Beach, PPA696 Research Methods: <http://web.csulb.edu/~msaintg/ppa696/696codes.htm>



Questions?

THANK YOU!

The Evaluation Team

**Please complete the last poll (to the right)
before exiting the webinar**

Evaluation Team

- ❑ Cheryl Kelly (Evaluation Investigator)
- ❑ Marisa Allen
- ❑ Brooke Bender
- ❑ Bre Barela
- ❑ Lisa Harner
- ❑ Denise Hartsock
- ❑ Bonnie Leeman-Castillo
- ❑ Carmen Luna
- ❑ Shayla Perkins