Data Management

Ian Tfirn, MPH

Daniel Norez, MPH

Center for Data Solutions

Objectives

- 1) Understand the methodology for data collection and management
- 2) Reviewing methods for requesting datasets

Introduction

- Straightforward understanding of your project
- What is the outcome you want to measure?
  - does x predict y ?
  - and does z influence this relationship?

Organizing Your Data

General Structure of a Data Dictionary

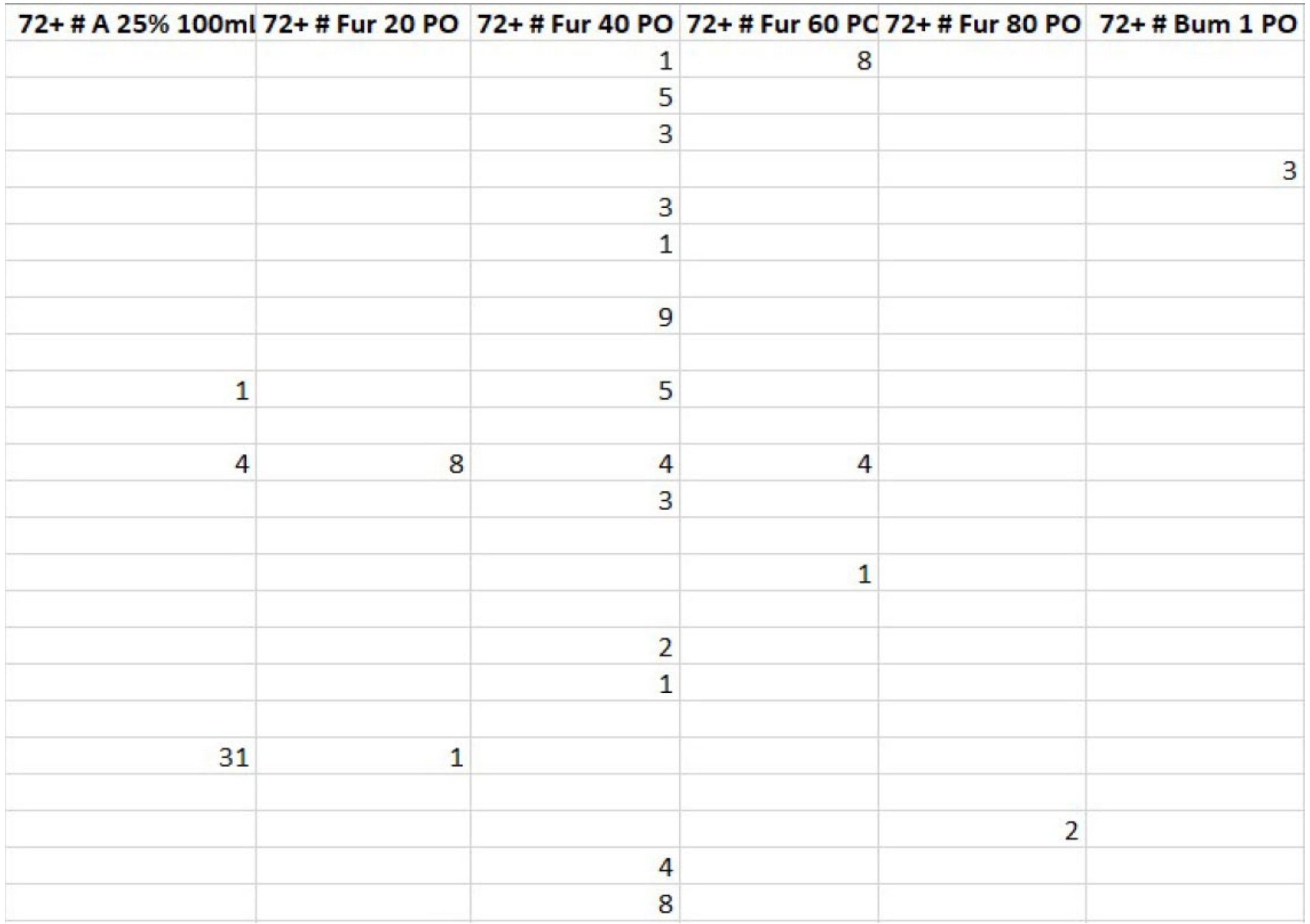
Best Case Scenario

- Some variables are easily understood
  - If variable names are complex, provide resources
    - If your data comes from SEER, let us know
    - Know who collected the data

Data Dictionary?

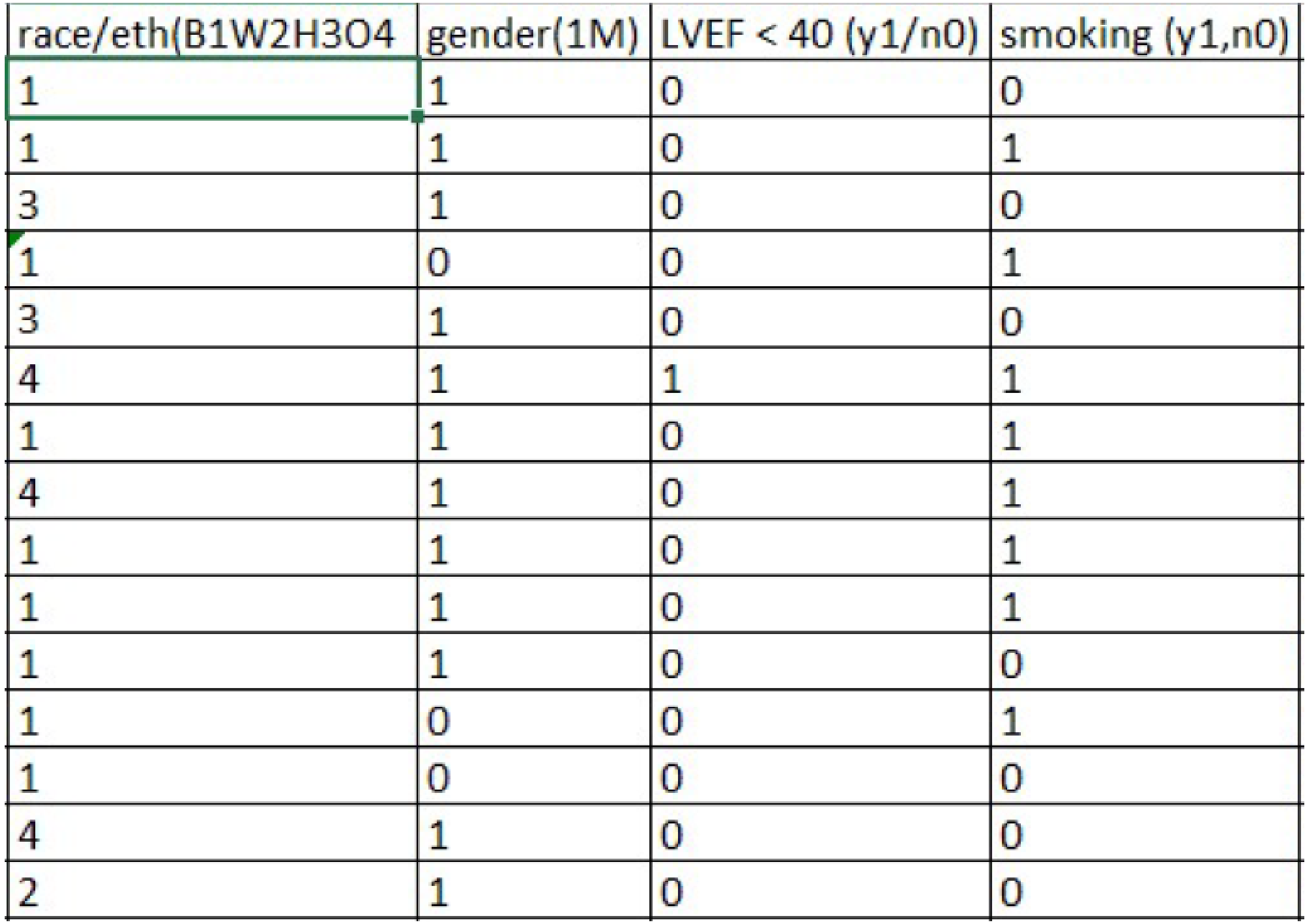
- Do I need to create one?
  - Simple variables? NO
  - Complex and discipline specific? YES
  - National dataset? NO (they provide)

Could a non-physician make sense of the data?



72+ # A 25% 100ml	72+ # Fur 20 PO	72+ # Fur 40 PO	72+ # Fur 60 PC	72+ # Fur 80 PO	72+ # Bum 1 PO
		1	8		
		5			
		3			
					3
		3			
		1			
		9			
1		5			
4	8	4	4		
		3			
			1		
		2			
		1			
31	1				
				2	
		4			
		8			

Could a non-physician make sense of the data?



race/eth(B1W2H3O4	gender(1M)	LVEF < 40 (y1/n0)	smoking (y1,n0)
1	1	0	0
1	1	0	1
3	1	0	0
1	0	0	1
3	1	0	0
4	1	1	1
1	1	0	1
4	1	0	1
1	1	0	1
1	1	0	1
1	1	0	0
1	0	0	1
1	0	0	0
4	1	0	0
2	1	0	0

Data Dictionary

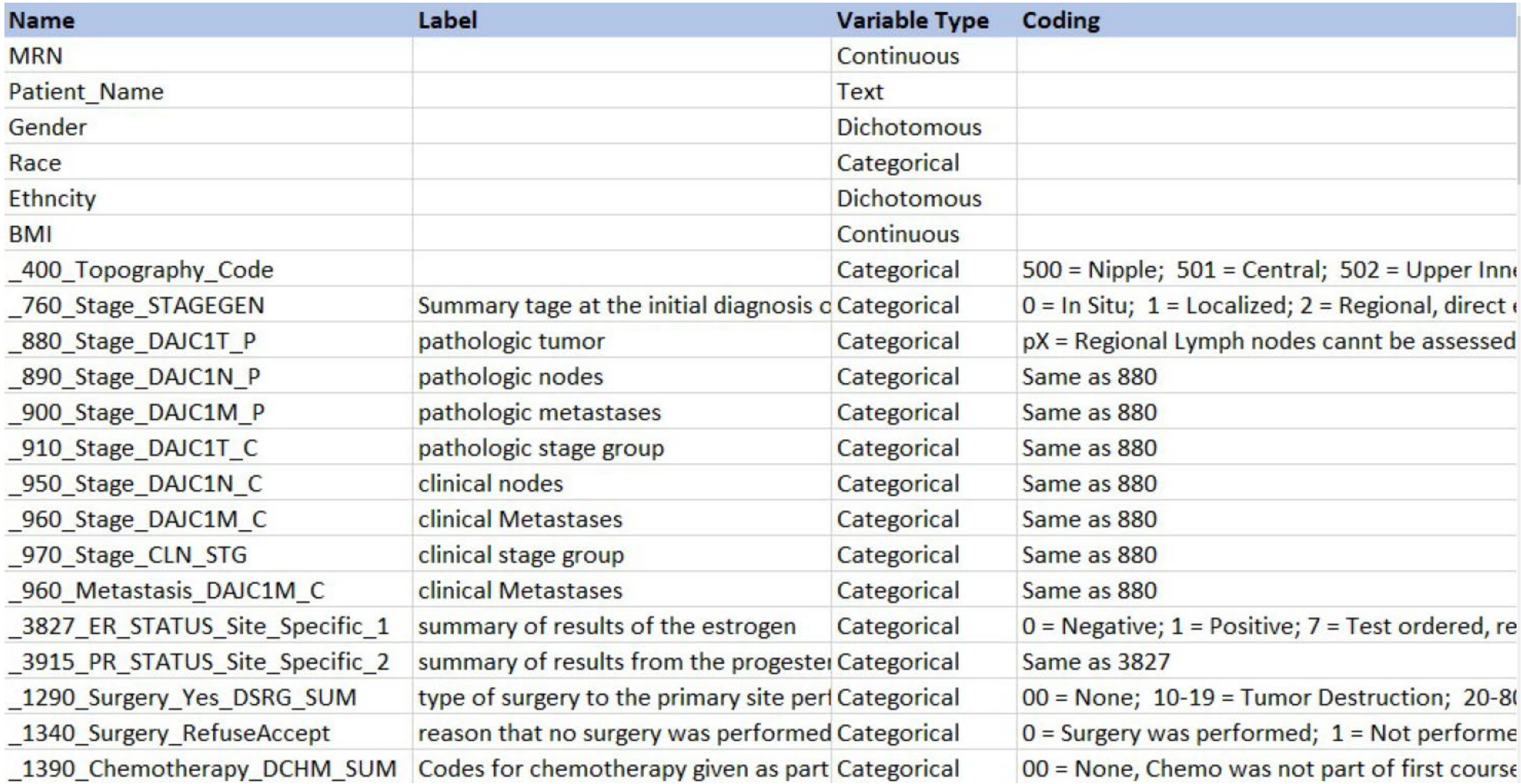
(Surveys, Chart Reviews)

- Variable name
  - SAS has naming rules
  - Begin all variable names with a letter or underscore. After the first character, numbers are fine.
  - No Spaces, underscores instead
  - Remove all special characters and punctuation
  - Keep variable names relatively short (under 32 characters)

- Label (question associated with variable)
  - Full survey question
  - Units of measurement
  - Explanation of acronyms (e.g. variable name “SBP” should say “Systolic Blood Pressure” in the label)

- Variable type
  - Categorical
    - male or female, state, tobacco use
  - Continuous
    - age, weight, SBP/DBP
  - Other
    - Text – Patient names, patient ID, etc.

- Coding
  - Categorical
    - Male= M, Female= F, Florida= FL, Tobacco = Y/N
    - No= 0, Yes= 1, Don't know= 7, Missing= .
  - Continuous
    - List range of acceptable values

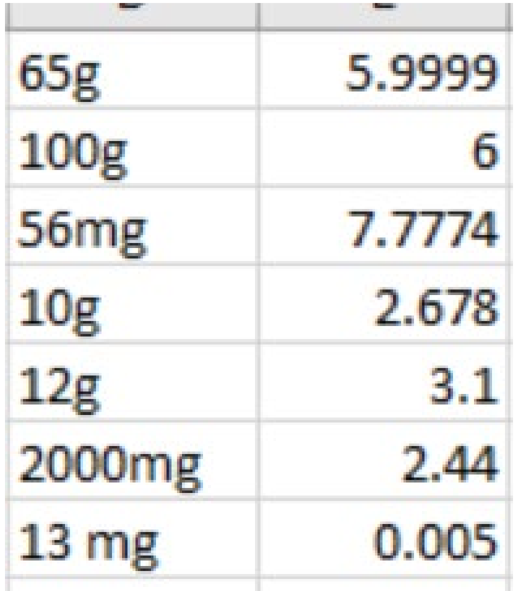


Name	Label	Variable Type	Coding
MRN		Continuous	
Patient_Name		Text	
Gender		Dichotomous	
Race		Categorical	
Ethncity		Dichotomous	
BMI		Continuous	
_400_Topography_Code		Categorical	500 = Nipple; 501 = Central; 502 = Upper Inne
_760_Stage_STAGEGEN	Summary tage at the initial diagnosis o	Categorical	0 = In Situ; 1 = Localized; 2 = Regional, direct e
_880_Stage_DAJC1T_P	pathologic tumor	Categorical	pX = Regional Lymph nodes cannt be assessed
_890_Stage_DAJC1N_P	pathologic nodes	Categorical	Same as 880
_900_Stage_DAJC1M_P	pathologic metastases	Categorical	Same as 880
_910_Stage_DAJC1T_C	pathologic stage group	Categorical	Same as 880
_950_Stage_DAJC1N_C	clinical nodes	Categorical	Same as 880
_960_Stage_DAJC1M_C	clinical Metastases	Categorical	Same as 880
_970_Stage_CLN_STG	clinical stage group	Categorical	Same as 880
_960_Metastasis_DAJC1M_C	clinical Metastases	Categorical	Same as 880
_3827_ER_STATUS_Site_Specific_1	summary of results of the estrogen	Categorical	0 = Negative; 1 = Positive; 7 = Test ordered, re
_3915_PR_STATUS_Site_Specific_2	summary of results from the progester	Categorical	Same as 3827
_1290_Surgery_Yes_DSRG_SUM	type of surgery to the primary site perf	Categorical	00 = None; 10-19 = Tumor Destruction; 20-80
_1340_Surgery_RefuseAccept	reason that no surgery was performed	Categorical	0 = Surgery was performed; 1 = Not performe
_1390_Chemotherapy_DCHM_SUM	Codes for chemotherapy given as part	Categorical	00 = None, Chemo was not part of first course

Data Entry

Consistency

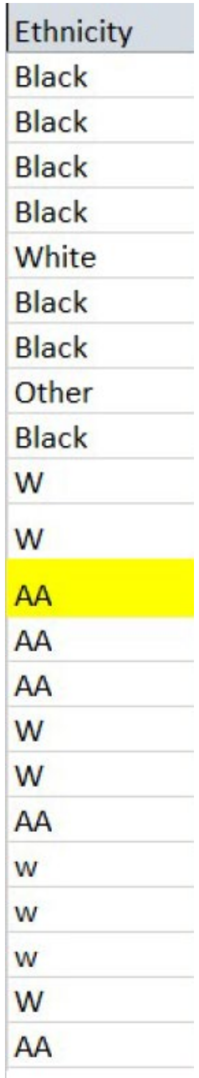
- Continuous variables
  - If rounded, round for all
  - Go to the same decimal place
  - Only use a single unit (e.g. use hours or minutes, not “1H 16M”
  - For missing values, leave the cell blank or use a period



65g	5.9999
100g	6
56mg	7.7774
10g	2.678
12g	3.1
2000mg	2.44
13 mg	0.005

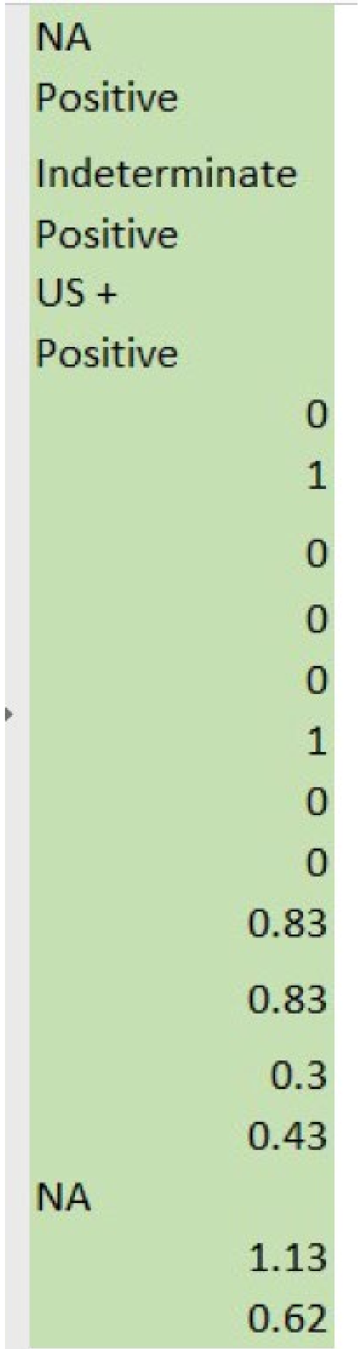
Consistency

- Categorical Variables
  - Yes/No, Y/N, 1/0
  - Variables should have consistent formatting
    - W or White, not both
    - Keep capitalization consistent
  - Spelling



Ethnicity
Black
Black
Black
Black
White
Black
Black
Other
Black
W
W
AA
AA
AA
W
W
AA
w
w
w
W
AA

What is wrong here?



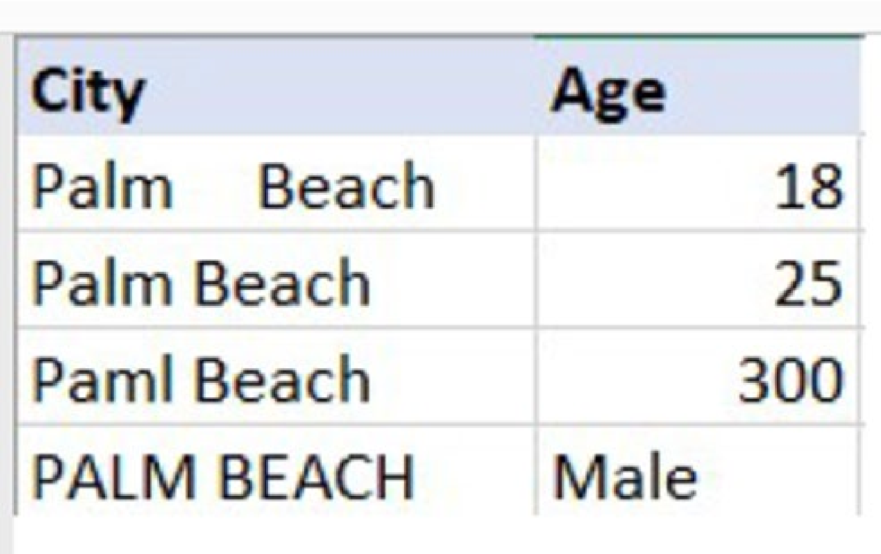
NA
Positive
Indeterminate
Positive
US +
Positive
0
1
0
0
0
1
0
0
0.83
0.83
0.3
0.43
NA
1.13
0.62

- Numeric and character values
- N/a
- Is US+ the same as Positive
- Rounding: 0.83 vs. 0.3 vs. 1

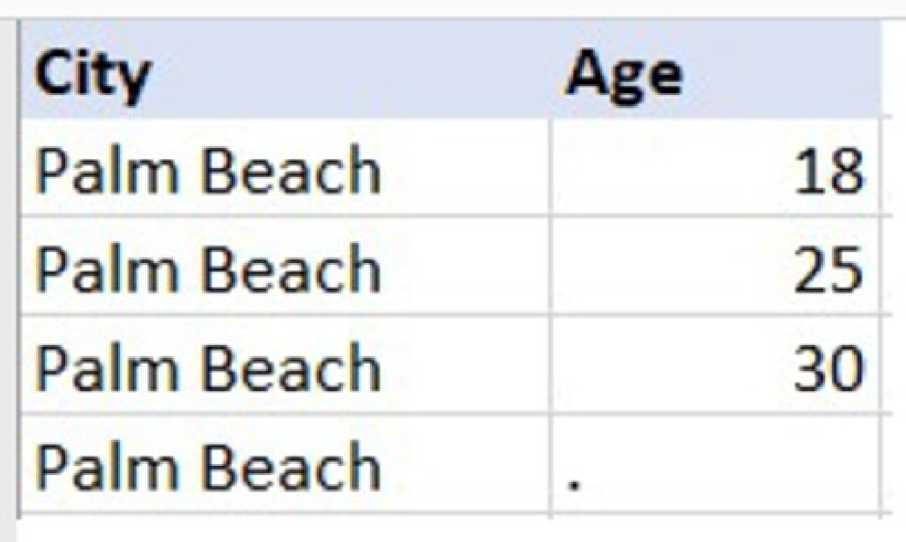
Data Cleaning

Correcting Errors

- Typographical
  - Extra   Spaces
  - Mispeled character data
  - Case SeNsiTiViTy
- Numeric Errors
  - Irrational numbers
  - Characters where numbers belong



City	Age
Palm Beach	18
Palm Beach	25
Paml Beach	300
PALM BEACH	Male



City	Age
Palm Beach	18
Palm Beach	25
Palm Beach	30
Palm Beach	.

Data Sources

Options for getting data

- 3 data sources
  - Primary - The investigator gathered the data him/her self
  - Publically available State/National – Data is available via a database or annual survey
  - UF – Resources that give data on UF Health patients

- Primary Data
  - Surveys
  - Interviews
  - Chart review

- Publically available data
  - SEER – National Database
  - NHANES – National Survey
  - Have their own data dictionary

Options for getting UF patient data

Integrated Data Repository (IDR)

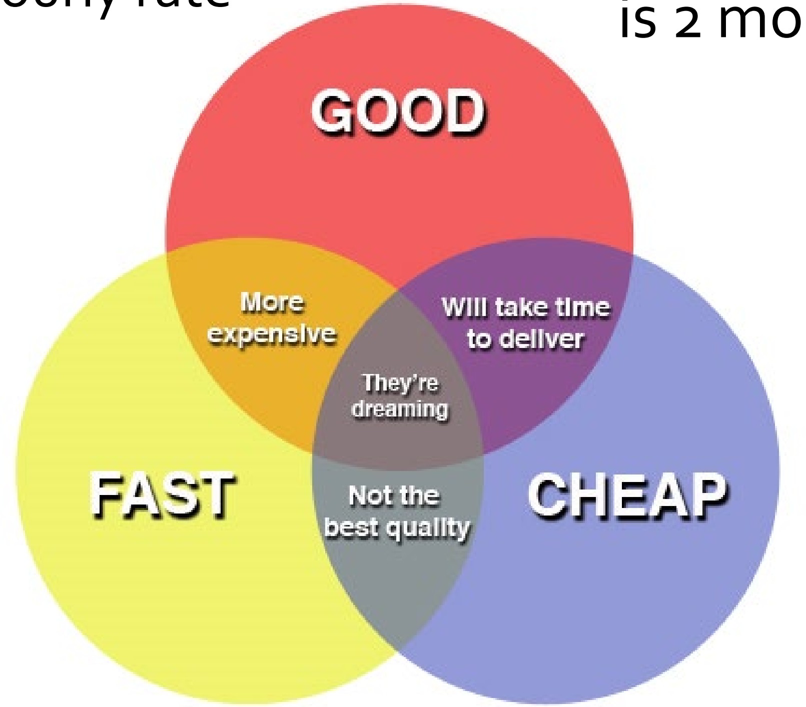
Fee for service, but quick

First four hours free, charges an hourly rate ($90/hour)

Data Analytics and Reporting (DARC)

No cost, but takes time

Average turnaround time is 2 months



Links for DARC and IDR

- DARC - http://1b-esx-infonet.umc.ufl.edu/Data-Analytics-and-Reporting/Pages/Request-a-New-Report.aspx
- IDR - https://idr.ufhealth.org/services/analyst-data-support-services/idr-data-request-form/

Options for getting UF patient data

- I2b2
    - Cohort discovery tool – does not provide PHI

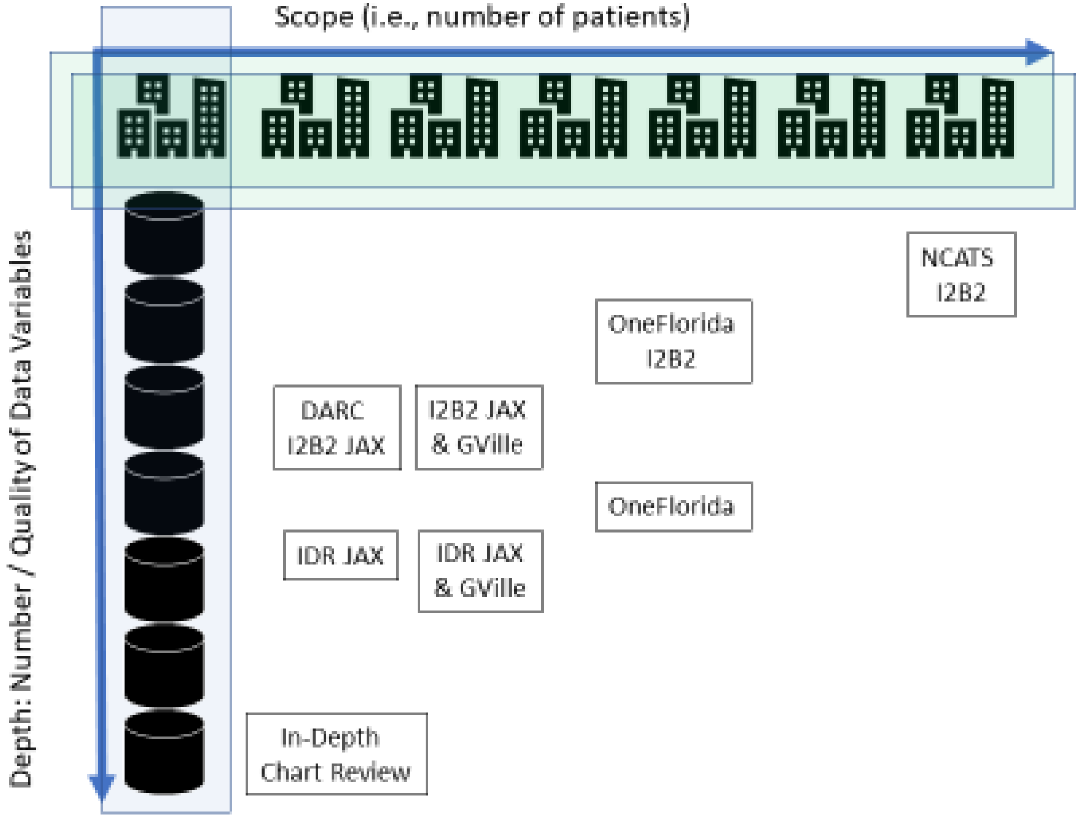
- When is it useful?
    - Getting a sample size estimate for your inclusion criteria
    - Understanding the demographic make-up of your potential sample
    - Determining whether or not a study is feasible

Options for getting UF patient data

A Snapshot of Available UF Data Sources

- In-Depth Chart Review
- DARC – EPIC reporting from JAX
- I2B2 – Cohort Discovery Tool via Integrated Data Repository (IDR); JAX & /or Gville
- IDR – Integrated Data Repository, EPIC data from JAX and/or Gville
- OneFlorida – Multicenter Collaborative, cohort Discovery via its I2B2 or in-depth data*
- NCATS I2B2 – Cohort discovery from CTSAs nationwide

*I2B2, OneFlorida I2B2, and NCATS I2B2 can be accessed without IRBs, as can cohort discovery through DARC. All other Inquiries require IRB approval.



Scope (i.e., number of patients)

Depth: Number / Quality of Data Variables

NCATS I2B2

OneFlorida I2B2

DARC I2B2 JAX

I2B2 JAX & GVille

OneFlorida

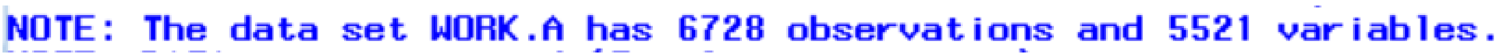
IDR JAX

IDR JAX & GVille

In-Depth Chart Review

How do you choose what and how much data to collect

- Pull data that is immediately useful
  - Do you need 5,000 variables to answer your question?
  - Will 50 variables be enough to answer your question?
  - You save time and money, we save time
  - More data than necessary is cumbersome
- Future Projects



```
NOTE: The data set WORK.A has 6728 observations and 5521 variables.
```

Questions?

"I don't understand your question.
Could you restate it as an answer?"