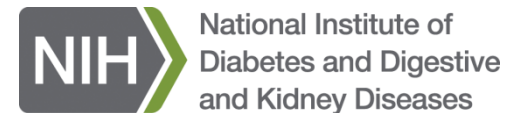


Pre-planning of project management: data and file management

Sungyoung Auh, PhD
Biostatistics Program Office
Office of the Director
NIDDK



Objectives

- To understand and utilize variable naming limitations/conventions and be able to create meta documentation for data
- To understand and utilize data validation tools within Excel/REDCap
- To understand the importance for file organization and uniform naming conventions and be able to implement these for research projects



What is metadata? It is Data about Data

NIDDK Data Management & Sharing (DMS) Webinar 3: Metadata and Data Standards for NIDDK Research Data: <https://www.youtube.com/watch?v=Ps4c6Va7Vlg&t=263s>

Data dictionary:

- At variable level
- *Should be stored in the same directory where data is saved*

Usages:

- Finding/Organizing files and Tracking details
- Managing security
- Critical in data governance and data management
- Help ensure data quality and data integrity.

README file:

- At study/dataset or file level



Data dictionary

- Critical to making your research reproducible because it allows others to understand your data.
- Derive from case report forms (CRFs)
- Contains
 - Name
 - Description
 - Type of variable
 - Unit
 - Allowed Formatting/Range

Variable Name	Variable Description	Type of variable	Unit	Allowed Formatting/Range
ID	Identifier for subject	Categorical		
Sex	Sex	Categorical		(1 for Male or 2 for Female) or (Male or Female)
BMI	Body Mass Index	Continuous	Kg/m ²	18 – 60



Create data structure in spreadsheet, Excel

- Rectangular shape in a way:
 - Each column header for variable name
 - Each observation in a row
 - Each type of observation unit forms a table

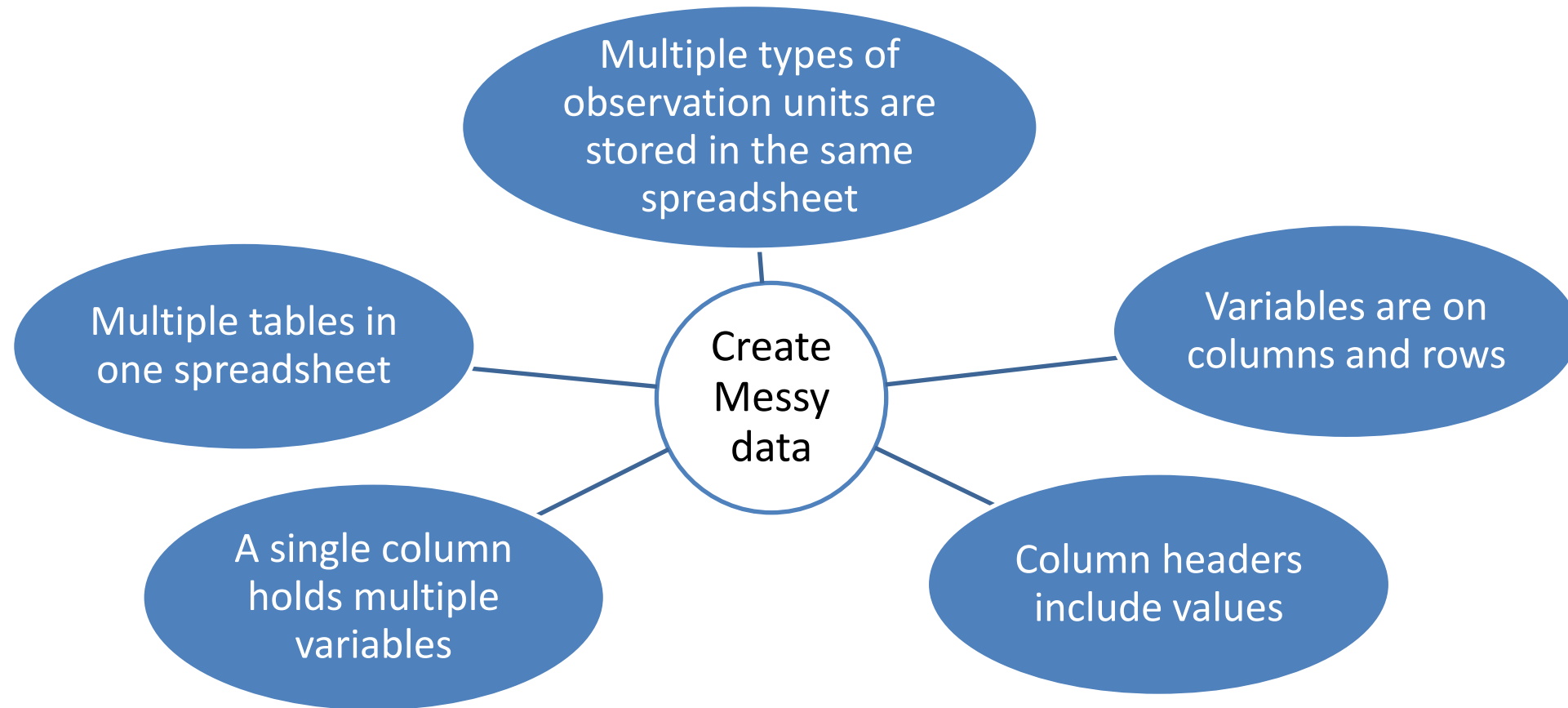


Create “Tidy” dataset, a data structure organized to facilitate data analysis

	A	B	C
1	subject	age	response
2	1		
3	2		
4	3		
-			



Create data structure in spreadsheet, Excel



Multiple types of observation units

	A	B	C	D	E	F
1	subject	GTT date	GTT weight	time	glucose mg/dl	Insulin ng/ml
2	321	2/9/2015	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	266.1	0.129
5				3	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/2015	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/2015	24.7	0	198.5	0.151
15				5	530.6	off curve lo

- “subject” is an observation unit
- “GTT date” and “GTT weight” are measured for each “subject”

- “time” is another observation unit
- Multiple observation units for “glucose” and “insulin”

- Here are multiple types of observation units are stored in the same table



Tidy or Messy?

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447



National Institute of
Diabetes and Digestive
and Kidney Diseases

Tidy or Messy?

Multiple tables in one spreadsheet

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447



Tidy or Messy?

Multiple tables in one spreadsheet

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

- Variables are stored in both columns and a row
- A single Column holds multiple variables



Tidy or Messy?

Multiple tables in one spreadsheet

Column headers include values

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

- Variables are stored in both columns and a row
- A single Column holds multiple variables



Tidy data

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

- Rectangular shape
- Column header in a single row
- Each observation in a row



National Institute of
Diabetes and Digestive
and Kidney Diseases

Multiple types of observation units

	A	B	C	D	E	F
1	subject	GTT date	GTT weight	time	glucose mg/dl	Insulin ng/ml
2	321	2/9/2015	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	266.1	0.129
5				3	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/2015	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/2015	24.7	0	198.5	0.151
15				5	530.6	off curve lo

- “subject” is an observation unit
- “GTT date” and “GTT weight” are measured for each “subject”

- “time” is another observation unit
- Multiple observation units for “glucose” and “insulin”

- Here are multiple types of observation units are stored in the same table



Tidy data?

	A	B	C
1	subject	GTT date	GTT weigh
2	321	2/9/2015	24.5
3	322	2/9/2015	18.9
4	323	2/9/2015	24.7

Per subject

Per "GTT time"
within subject

- Separated into two spreadsheet:
 - Variable in a single row
 - Each observation in a row
 - Each type of observation unit forms a table

	A	B	C	D
1	subject	GTT time	glucose mh/dl	Insulin ng/ml
2	321	0	99.2	lo off curve
3	321	5	349.3	0.205
4	321	15	286.1	0.129
5	321	30	312	0.175
6	321	60	99.9	0.122
7	321	120	217.9	lo off curve
8	322	0	185.8	0.251
9	322	5	297.4	2.228
10	322	15	439	2.078
11	322	30	362.3	0.775
12	322	60	232.7	0.5
13	322	120	260.7	0.523
14	323	0	198.5	0.151
15	323	5	530.6	off curve lo



National Institute of
Diabetes and Digestive
and Kidney Diseases

Tidy data?

	A	B	C
1	subject	GTT date	GTT weigh
2	321	2/9/2015	24.5
3	322	2/9/2015	18.9
4	323	2/9/2015	24.7

Per subject

Per "GTT time"
within subject

- Separated into two spreadsheet:
 - Variable in a single row
 - Each observation in a row
 - Each type of observation unit forms a table

	A	B	C	D
1	subject	GTT time	glucose mh/dl	Insulin ng/ml
2	321	0	99.2	to off curve
3	321	5	349.3	0.205
4	321	15	286.1	0.129
5	321	30	312	0.175
6	321	60	99.9	0.122
7	321	120	217.9	to off curve
8	322	0	185.8	0.251
9	322	5	297.4	2.228
10	322	15	439	2.078
11	322	30	362.3	0.775
12	322	60	232.7	0.5
13	322	120	260.7	0.523
14	323	0	198.5	0.151
15	323	5	530.6	off curve lo

Variable names include:

1. Space
2. Measurement unit



Choose good variable name in tidy data

- Maximum length of variable name is 32 characters.
- Be short but meaningful
- Case sensitive in R/Python

- Avoid spaces, numbers in first place and special characters
- Do not include measurement unit
- Do not include legend
- Use underscore (_) instead of space
- Avoid special characters

Good name	Avoid
glucose	glucose (mg/dl)
Sex	M/F
genotype	genotype (1=A, 2=B, 3=C, 4=D,5=E)
weight	w.
day_01	1st day
diff_ISHAK	ΔISHAK



Tidy data?

	A	B	C
1	subject	GTT_date	GTT_weigh
2	321	2/9/2015	24.5
3	322	2/9/2015	18.9
4	323	2/9/2015	24.7

	A	B	C	D
1	subject	GTT_time	glucose	Insulin
2	321	0	99.2	lo off curve
3	321	5	349.3	0.205
4	321	15	286.1	0.129
5	321	30	312	0.175
6	321	60	99.9	0.122
7	321	120	217.5	lo off curve
8	322	0	185.8	0.251
9	322	5	297.4	2.228
10	322	15	439	2.078
11	322	30	362.3	0.775
12	322	60	232.7	0.5
13	322	120	260.7	0.523
14	323	0	198.5	0.151
15	323	5	530.6	off curve lo



Variables and Entering data

	A	B	C	D	E	F
1	Insulin	Insulin	note		sex	sex
2	to off curve		Insulin below curve		female	female
3	0.205	0.205			Female	female
4	0.129	0.129			F	female
5	to off curve		Insulin below curve		Male	male
6	0.251	0.251			m	male
7	2.228	2.228			M	male
8	0.523	0.523				
9	0.151	0.151				
10	off curve to		Insulin below curve			
11						

Continuous: numbers only

Categorical: be consistent with format

Missing values: Blank is best even though it is problematic



National Institute of
Diabetes and Digestive
and Kidney Diseases

Tidy data

	A	B	C	D	E
1	subject	GTT_time	glucose	Insulin	note
2	321	0	99.2		Insulin below curve
3	321	5	349.3	0.205	
4	321	15	286.1	0.129	
5	321	30	312	0.175	
6	321	60	99.9	0.122	
7	321	120	217.9		Insulin below curve
8	322	0	185.8	0.251	
9	322	5	297.4	2.228	
10	322	15	439	2.078	
11	322	30	362.3	0.775	
12	322	60	232.7	0.5	
13	322	120	260.7	0.523	
14	323	0	198.5	0.151	
15	323	5	530.6		Insulin below curve



More principles for Tidy data

Put one thing in a cell

Do not use color as a data

Do not include calculations and graphs

	A	B
1	plate-well	value
2	1-A1	1.2
3	1-A2	0.3
4	1-A3	2.6
5	1-A4	7.8
6	1-A5	1.3
7		
8	avg	2.64

	A	B	C	D
1	plate	well	value	outlier
2	1	A1	1.2	FALSE
3	1	A2	0.3	FALSE
4	1	A3	2.6	FALSE
5	1	A4	7.8	TRUE
6	1	A5	1.3	FALSE
7				
8				



Data dictionary

- Critical to making your research reproducible because it allows others to understand your data.
- Derive from case report forms (CRFs)
- Contains
 - Name
 - Description
 - Type of variable
 - Unit
 - Allowed Formatting/Range

Variable Name	Variable Description	Type of variable	Unit	Allowed Formatting/Range
ID	Identifier for subject	Categorical		
Sex	Sex	Categorical		(1 for Male or 2 for Female) or (Male or Female)
BMI	Body Mass Index	Continuous	Kg/m ²	18 – 60

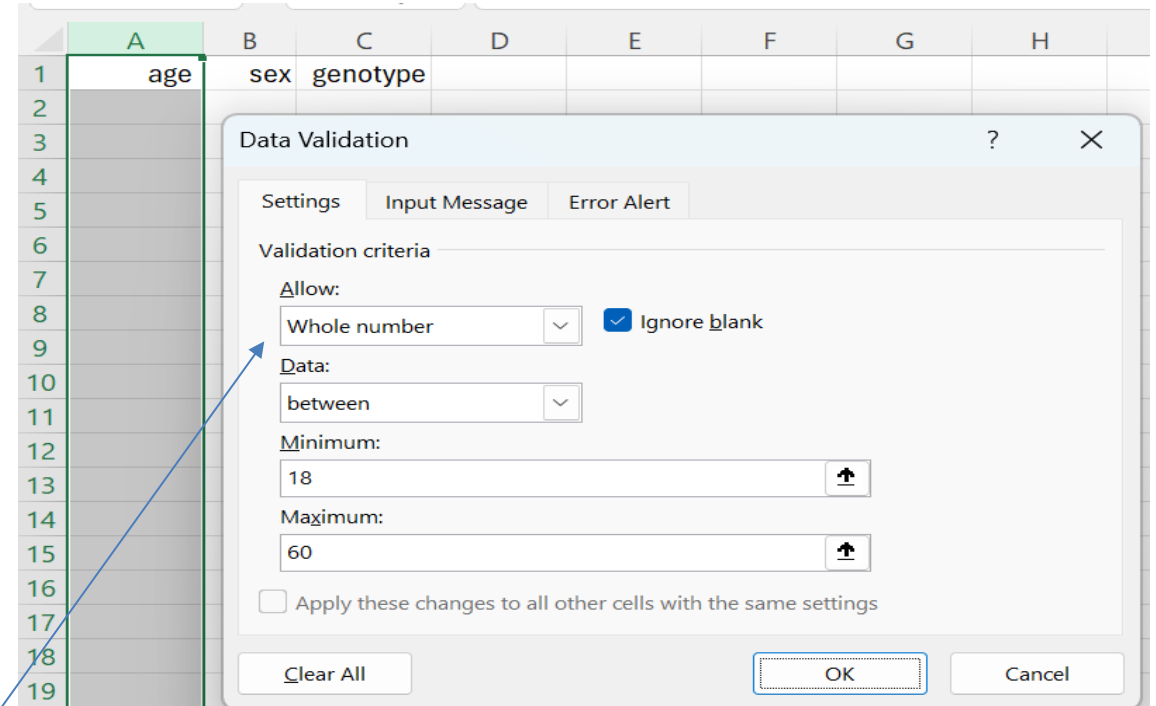
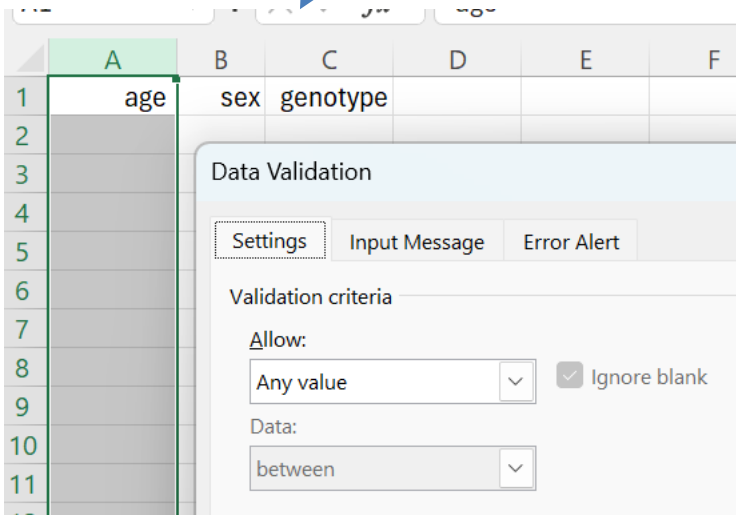


Data Validation Tool in Excel for Continuous

- Select column

	A	B	C
1	age	sex	genotype
2			
3			
4			

- In the 'Menu' bar, choose 'Data'
- In the 'Data Tools', choose 'Data Validation'



- Choose "Whole number"
- Assign minimum and maximum of "age"



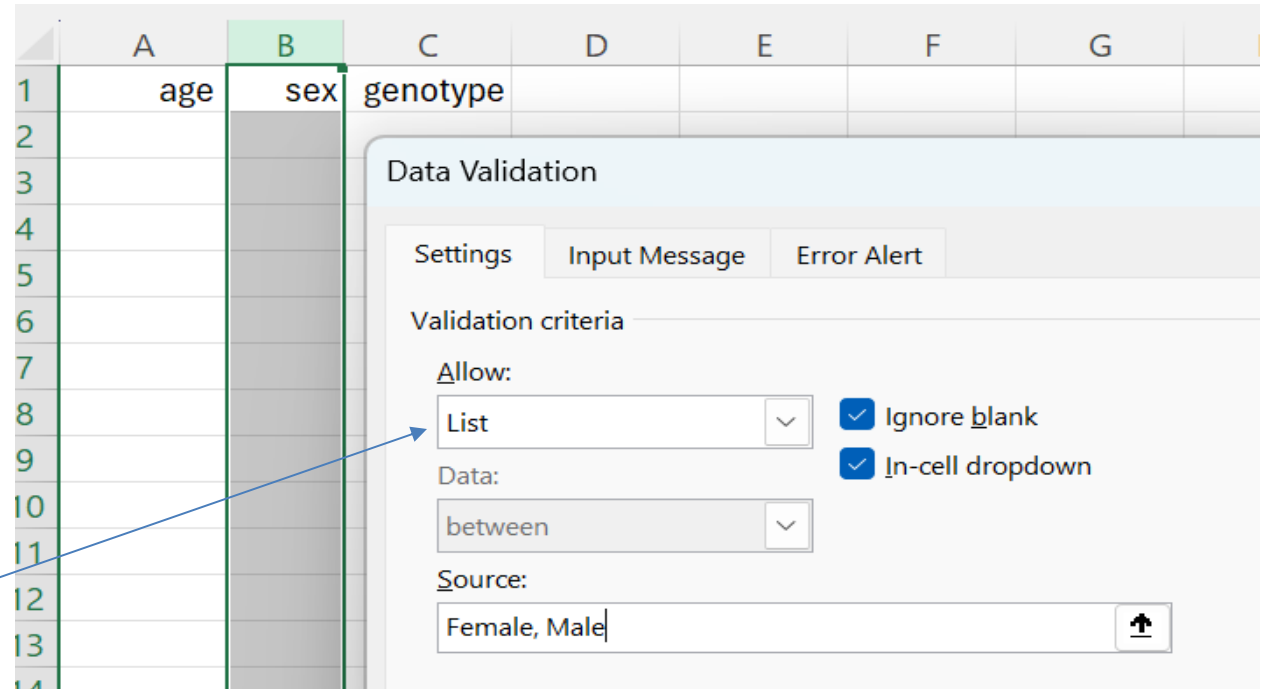
Data Validation Tool in Excel for Categorical

- Select column

	A	B	C
1	age	sex	genotype
2			
3			
4			

- In the 'Menu' bar, choose 'Data'
- In the 'Data Tools', choose 'Data Validation'

- Choose "List"
- Type "Female" and "Male" under Source



The screenshot shows the 'Data Validation' dialog box in Excel. The 'Settings' tab is selected. Under 'Validation criteria', 'Allow:' is set to 'List'. The 'Data:' dropdown is set to 'between'. The 'Source:' field contains 'Female, Male'. The 'Ignore blank' and 'In-cell dropdown' checkboxes are both checked. An arrow points from the 'List' dropdown in the dialog to the 'sex' column in the spreadsheet shown in the previous block.

	A	B	C
1	age	sex	genotype
2		Female	
3			
4		Female	
5		Male	
6			

- Prevent typos for sex when you enter data.



Data Validation in REDCap

Full support Common Data Elements (CDEs)

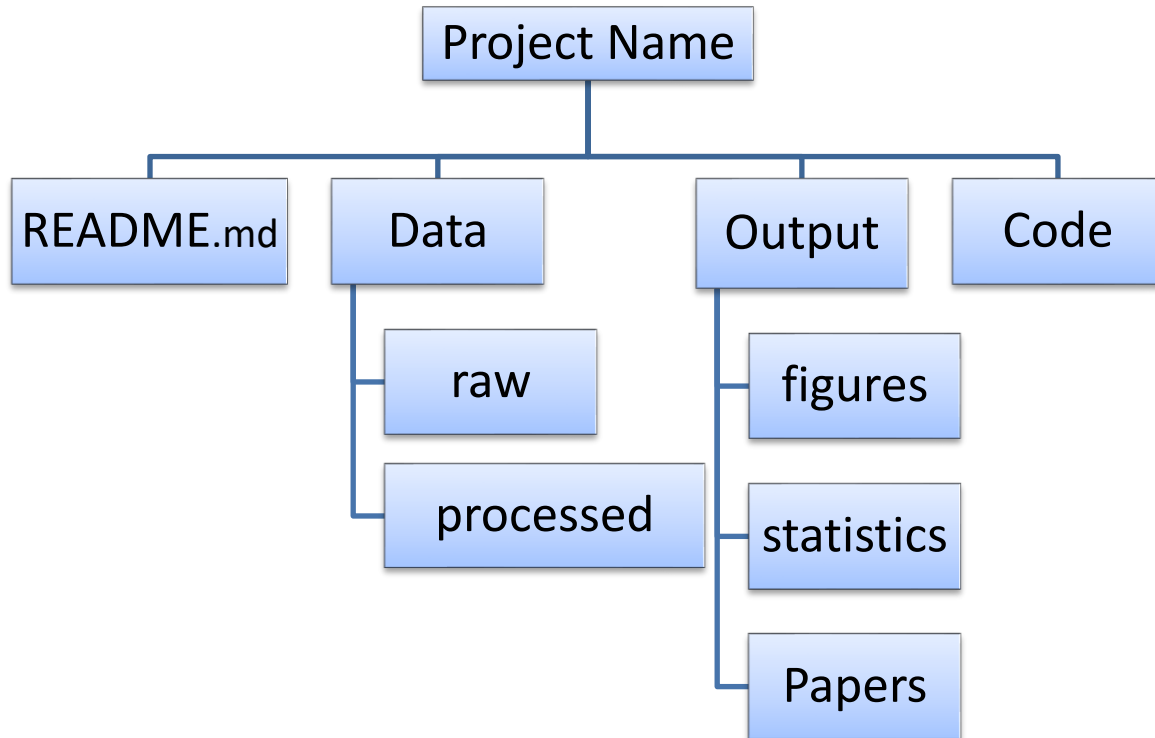
Enforces data integrity at entry point

Numeric constraints

Categorical rules



File Management principles



- Develop an organizational structure
- Prepare a README file
- Adhere to it consistently

Caltech Library File Naming Convention worksheet:
<https://authors.library.caltech.edu/records/mmnpf-cez11>

- This worksheet help you create file naming convention by:
 - Choosing keywords to use in file names
 - Encoding and ordering the keywords
 - Adding version control information



<https://authors.library.caltech.edu/records/mmnpf-cez11>



National Institute of
Diabetes and Digestive
and Kidney Diseases

File Management principles: File Naming Conventions

Human Readable

- Clear and descriptive
- Use YYYY-MM-DD to convey history
- Avoid vague terms such as "final"

Use:

- 01_Data_Cleaning.r
- 02_Figure1.r
- 2026-8-20_data_group1.txt
- 2026-8-21_data_group2.txt

Machine Readable

- Limit to 32 characters
- Use underscore (_) or hyphen (-) instead of space
- Use leading zero like "01"
- Use version number "v01" or version date "YYYY-MM-DD"

Avoid:

- 01.r
- 02.r
- 1-Data_Cleaning.r
- 08-24-2026_data_group1.txt



National Institute of
Diabetes and Digestive
and Kidney Diseases

References/Resources

- Broman, K.W., and Woo, K.H. (2018) Data Organization in Spreadsheets, *The American Statistician*, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989)
- Wickham, H. (2014) Tidy Data. *Journal of Statistical Software*, 59:10, 1-23, DOI: [10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)
- <https://www.nih.gov/heal/heal-initiative-requirements/data-sharing-policy/common-data-elements-cdes-program>
- <https://support.microsoft.com/en-us/excel/get-started/apply-data-validation-to-cells>
- <https://guides.lib.virginia.edu/RDM/file-management>



National Institute of
Diabetes and Digestive
and Kidney Diseases



**National Institute of
Diabetes and Digestive
and Kidney Diseases**



National Institute of
Diabetes and Digestive
and Kidney Diseases