Data Management

An overview

Vahe Khachadourian, MD, MPH, PhD

June 25, 2022

Navetis

What Is Data?

The National Institutes of Health, or NIH, define data as:

“recorded factual material commonly accepted in the scientific community as necessary to validate research findings.”

What Is Data?

The National Science Foundation or NSF considers data to be something,

“….determined by the community of interest through the process of peer review and project management.”

Types of Data

- Numeric or tabular data
- Samples such as biological samples and physical collections
- Software programs and code
- Algorithm
- Geodatabases

Types of Data

There are other research products that need to be considered alongside data in order for the data to be meaningful.

- Questionnaires
- Codebooks
- Descriptions of methodologies

Data are Little Consistent Pieces of Big Things in Reality

Data Management

“Data management” is a relatively new term arising in the mid-2000s with funder requirements for both data management and data sharing.

Data Management

Briefly, data management includes data management planning, documenting your data, organizing your data, improving analysis procedures, securing sensitive data properly, having adequate storage and backups during a project, taking care of your data after a project, sharing data effectively, and finding data for reuse in a new project. Such a wide range of practices means that data management is something you do before the start of a research project, during the project, and after the project's completion.

Data Management

Assuring the protection, integrity and quality, accessibility, usefulness and memory of collected data for the purpose of advancing health sciences, informing the development of health policy, improving health programs and projects, identifying best practices, protocols and treatments, AND to assure the most effective timely responses to outbreaks & emergencies!

DO NOT LET THIS HAPPEN!

Why Data Management: Foundation to Advance Science

Data should be managed to:

- maximize the effective use and value of data and information assets
- continually improve the quality including: data accuracy, integrity, integration, timeliness of data capture and presentation, relevance and usefulness
- ensure appropriate use of data and information
- facilitate data sharing
- ensure sustainability and accessibility in long term for re-use in science

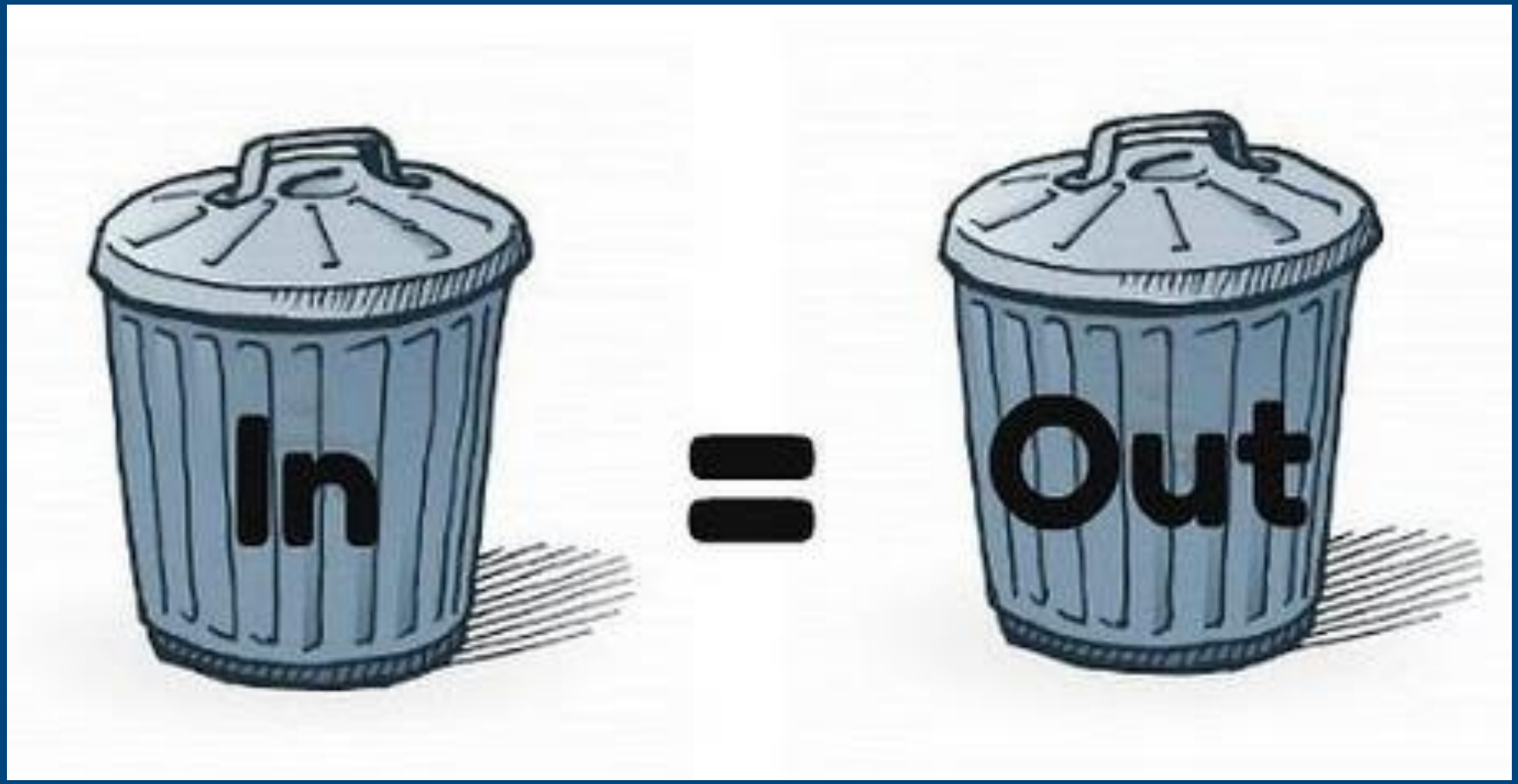
Planning For Data Management

The most import time to address the management of your research data is before you even start to collect that data.

Planning for Data Collection

Good Practice

Develop your hypothesis and define all the primary and secondary measurements and covariates that you need to test your hypothesis.


In
=
Out

Planning for Data Management

Before you directly work with your research data, it helps to think about all of the data management strategies you want to utilize during and after your research project. There are many choices to make, though some have greater consequences than others.

Additionally, the longer the duration of time you need to keep the data and the more people who need to use it, the more important it is to plan ahead how to manage everything.

How to customize data management to your needs

Much of your effort in planning for data management will be spent finding the systems that work best in your research.

Good data management is a balance between best practices and your individual needs, and part of achieving that balance involves understanding those needs.

Documentation

Documentation is one of the most important parts of managing data because data needs context in order to be understood and used.

All details are important. Therefore, you should think of documentation as a letter to your future self so that you understand what your data is and how you acquired it.

Data Dictionaries / Codebooks

Generally, data dictionaries include the following types of information on each variable:

- Variable name
- Variable definition
- How the variable was measured
- Data units
- Data format

Data Dictionaries / Codebooks

- Minimum and maximum values
- Coded values and their meanings
- Representation of null values
- Precision of measurement
- Known issues with the data (missing values, bias, etc.)
- Relationship to other variables
- Other important notes about the data

Codebook

Coding Manual

Variable Name	Q	Description	Codes	Type	Notes
BABYID				I	KEY
REGNUM		The CHD Registry Number	Only for cases	I	
REGMTH		The month the case was registered by CHDR	Only for cases	I	
CASE	2	Case Status:	1=Case 2=Control	I	
BABYSEX		The sex of the baby	1=Male 2=Female		Only for controls Cases will be taken directly from the web
DATE	1	Date of Interview		D	
WHO			1=Wiiam Al Obaid 2=Sarah Al Qatani 3=Lama Sultan 4=Deema Abdulkader 5=Mona Al Otaibi	I	
DOBMM	3	In what month was this child born? Month *Code as 99 if unknown*		I	2 digit Gregorian
DOBYYYY		In what year was this child born? Code as 9999 if unknown			4 digit Gregorian

File Organization

- File organization is simultaneously simple and difficult because it is conceptually easy but takes persistent work to have everything in its proper place.
- The best piece of advice for staying organized is to have a system. The system should be logical, but more importantly, it should work well for you in your everyday research tasks.

File Organization

In terms of the actual organization of files in your storage system, there are many possible schemes to adopt for your data. A few ideas for organizing folders include:

- By project
- By researcher
- By date
- By research notebook number
- By any combination of the above

Naming Conventions

Using consistent naming is essential for good data management. It takes the benefits of good file organization to the next level by adding order to the actual files within the folder. While these naming conventions nominally apply to files and folders, you can use these principles whenever you need to name groups of objects, such as physical samples. For how simple it is to adopt good naming, there are many rewards for using such conventions.

File Naming

Your biggest consideration for a naming scheme is what information should go in the name. Good names convey context about what the file contains by stating information like:

- Experiment type
- Researcher name or initials
- Analysis type
- Date
- Version number

File Naming

:You have a lot of leeway with naming systems, so long as you follow a few general principles:

- Names should be descriptive
- Names should be consistent
- Names should be short, preferably less than 25 characters
- Use underscores or dashes instead of spaces
- Avoid special characters, such as: “ / \ : * ? ‘ < > [ ] & $
- Follow the date conventions: YYYY-MM-DD or YYYYMMDD

File Versioning

- Include version numbers in your file names.
- A common form for expressing data file versions is to use ordinal numbers such as 1, 2, and 3.
- For major version changes and decimals for minor changes.
- Avoid using confusing labels such as revision, final, final2, or final_final copy.

Documenting Your Conventions

What to document?

- To document your conventions, you should provide a basic framework for how the conventions work and note any information necessary to decoding your system, such as abbreviations used in files names. Documentation should be clear and concise, with as much detail as necessary to make a peer understand what you did.

Documenting Your Conventions

Where to document?

- README.txt files are ideal for documenting conventions, especially those concerning digital data. You should place a copy of your overall conventions in a README.txt file at the top-level of your project folders.

Data Merging and Cleaning

- Merge the databases, check inconsistent variables/values
- Look for any potential pattern, data entry operator pairs, interviewer, variable, values, and etc.
- Majority of the corrections should be based on the original responses
- All decisions made during data cleaning should be documented.

Sensitive Data - Confidentiality and Safety

When working with sensitive data, there are three general recommendations: determine if your data is sensitive, don’t collect sensitive data if you don’t have to, and make a clear security plan with help from security experts and review it frequently.

Keeping Data Secure

- Not collecting sensitive data
- Safe computer practices
- Limiting access to the data
- Encrypting your data
- Properly disposing of data
- Training everyone in security procedures

Destroying Data

- Only by destroying data can you ensure that it is not at risk for loss after that point.

- Additionally, destroying data means that you no longer need to devote resources to maintaining a secure storage environment for that data. It is best practice to destroy sensitive data after it is no longer needed or required to be retained.

Destroying Data

Destroying sensitive data requires more work than simply deleting the file from your computer.

When you delete a digital file from your computer, the file stays on your hard drive and only the reference to the file is removed. This is why it is often possible to recover data from a hard drive after a crash; the files are not truly destroyed even if the operating system does not recognize that they are still there.

Anonymizing Data

- There are lots of reasons to anonymize your data
- Most data privacy laws no longer apply once you remove the personally identifiable information from a dataset, which means that you don’t have to adjust your routine as you would to work with a sensitive dataset

Anonymizing Data

- Types of personally identifiable information
- Direct identifiers are things that identify a person individually, such as
- Name
- Address
- Telephone number
- Email address
- IP address
- Social Security Number, National Insurance number, or other national ID

Anonymizing Data

- Direct identifiers
- Driver’s license number
- Medical record numbers
- Credit card number
- Photographs
- Voice recordings

Anonymizing Data

Unlike direct identifiers, indirect identifiers do not disambiguate with one data point but instead can be used in combination to identify an individual. For example, research showed that 87% of Americans can be uniquely identified by their zip code, birth date, and gender (Sweeney 2000). Alone, none of these data points is a direct identifier but they can be used collectively to identify someone.

Storage and Backups

The motto for storage and backups is that “lots of copies keep stuff safe” (LOCKSS), not feasible for the average researcher.

Storage and Backups

- For most research data, a good rule of thumb is to follow the 3-2-1 backup rule. This guideline recommends maintaining three copies of your data on at least two different types of storage media with one offsite copy. Three copies is a good balance between having enough copies and having a manageable number of copies.
- The offsite copy is one of the most important copies as it protects against fire, natural disaster, and any other local event that might cause you to lose your data.

Backups

Imagine losing your research data. You go into work and find that everything is gone: your computer, your notes, and all of your work.

- Would you be able to recover?
- Could you recreate all of the lost work?
- Do you have the money to recreate all of the work?
- Do you have backups in place to restore from?
- How much would losing the main copy of your data set you back?

Backups

Update your back up copies periodically.

On top of frequency, you may need to decide on the type of backup to regularly perform: full or incremental.

Full backups copy all of your files for every backup.
Incremental backups only copy files that have been added or changed since the last backup.

Backups

Test your backups:

One important aspect of keeping backups is that you must test them. This consists of two parts. First, periodically checking your backups ensures that they work properly and that your data remains safe even if you lose the main copy of your files. The second reason to test your backups is so that you know how to restore data from them. You don’t want to be learning how to recover data from your backup when you’ve just lost the primary copy of your files.

Long-term Storage and Preservation

Keeping files readable

One of the biggest challenges of keeping data in the long term is the rapid evolution of technology.

Long-term Storage and Preservation

Keeping files readable
- File formats
- Hardware

Keeping datasets interpretable
- Improving your documentation

Good Practice

Store raw data fields rather than calculated values.

BMI→ Height + Weight

Hypertension: Y/N → SBP & DBP

- You can always calculate BMI from Height and Weight later, but you cannot calculate Height/Weight from BMI
- DO NOT code continuous variables ---- store them intact.

Good Practice (Miss ng D ta)

Allow for the possibility of missing data fields
Never code a zero or other possible values as default.

Develop codes for missing values if needed.

- Patient does not remember (888)
- Patient will not disclose (999)

Dates can be particularly problematic:
Date of diagnosis:
October 1 2021, Fall 2021, 2021, In the past 3 years

Improving Data Analysis

- Documenting the analysis process
- Such information includes everything from analysis procedures to the version of software (custom made or proprietary/branded) used. Keeping track of your analysis makes it easier to repeat, easier to correct errors, and easier to describe to others when publishing your results.

Preparing Data For Analysis

Often, half of the challenge of analysis is preparing the data for analysis. This means cleaning the data, performing error checking and quality control, and making the data more consistent.

Preparing Data For Analysis

One of the best things you can do to streamline your data analysis is ensure that your data is consistent, meaning your data should have the same units and format across each variable. Consistency prevents errors by making sure your values are all correct.

Preparing Data For Analysis

Beyond absolute correctness, consistent formatting makes for easier analysis because related values are all represented in the same way.

Error checking

Error checking means scanning through your data to look for improbable and missing values, though there are other methods for checking. Your preferred method will depend on the form of your data and the tools you have available, but here are two powerful ways to perform error checking.

Error checking

- One of the easier ways to check for errors is to make a simple plot of your data (or subsets of your data) along some logical coordinate.
- Outliers and artifacts in your data are often easier to identify visually than when buried within a large dataset.

Thank You

Contact
Name
Organization
Contact Info

Data Management Planning

- What types of data do I have? How much do I have?
- Do I use any third-party data?
- What data tools and technology are readily available to me?
- How long must I keep my data?
- Will I share my data?

Data Management Planning

- Does my data have security concerns, such as personally identifiable information?
- What does my funder/institution/employer require?
- Is there anything particular in my research workflow that might affect my data management?

Data Management Planning

Data management plan addresses the following major topics:

- What data will you create?
- How will you document and organize your data?
- How will you store your data and, if necessary, keep it secure?
- How will you manage your data after the completion of the project?
- How will you make your data available for reuse, as necessary?

Thank You

Contact
Name
Organization
Contact Info