1

Hello, I am Sawyer Newman, I work at Yale’s Cushing/Whitney Medical Library. During this session, I will discuss research data management and relational databases. The relational database portion will allow me to demonstrate, and for you to think about how data management concepts are well represented and brought to life, if you will, when you understand how databases are designed and utilized.

2

Research data management is, unsurprisingly, the management of research data. But this is not a helpful definition, so let’s break this down a bit. Research data, according to the National Institute of Health, is recorded factual material commonly accepted in the scientific community as necessary to document and support research findings. This does not mean summary statistics or tables; bur rather, it means the data on which summary statistics and tables are based. Research data also includes software you create as a part of a research project. So, if you write a Python script that performs your analysis or data processing, this falls under research data. Data management is the maintenance of research data.

3

Research data maintenance activities can include these core topics -how data is organized, where data is stored, how it is preserved for future use through backups and documentation, and data sharing, which provides access to research datasets for reuse and validation. We will talk about in more detail.

4

Research data management concepts, and how research data management fits into the research process is often represented through a diagram like this, through a research data lifecycle diagram. At different stages of your project, you will make different data management considerations.

In the planning phase, you might create a data management plan to layout where you will consider what kinds of and how much data you will create during your research project, and you might want to consider costs associated with your data access, analysis, or storage that you may be able to write in as line items in a grant application.

In the next phase, as data is collected from secondary sources or created via instrumentation, you will want to find and review available documentation related to your data so you understand and have a record of what exactly your data is, what the scope is, what your variables mean, etc.

Data processing is where you will go from having all of the raw data you might have access to, say an entire database or a full collection of focus group transcripts, and whittle this down to the data you need for analysis in a format that will enable you to conduct your analysis. In the case of a database, this might mean pulling only data from a time period, demographic, or disease of interest (or a combination of these things). In the case of having a full collection of transcripts, you may need to parse and code this qualitative data using a software such as NVIVO. Processed data enables you to run your analysis.

During data analysis, data management considerations include keeping records or documentation on the specifics of your analysis. If you are working with large datasets or performing computationally intensive analysis, as what may be the case with machine learning and natural language processing analyses, your computer may not be capable of operating at a speed that can facilitate this. If this is the case, you will want to leverage high performance computing clusters (HPC). HPC resources at Yale are managed through the Yale Center for Research Computing.

After analysis has been completed, and potentially a paper has been published, you may be required to share your data. Even if you are not required by a funder or a publisher to share data, this is something you may want to consider doing because sharing data supports future research

At this point, you will also need to think about data retention, or long-term storage. At Yale, you are required to retain research data for at least three years after the point of publication.

Within this cycle of research data there are three constant themes, or components, that you should always consider. Data security, documentation, and data storage, operational in the sense that it is where your data exists as you are using it during any phase. Data storage during project planning may be something collaborative and document focused like Box, where data processing may happen on your local machine, data analysis could happen through HPC, and data storage can be handled by the archival tier of Storage @ Yale.

Data needs to be secure in two major ways: secured from accidental destruction, and protected from interference, tampering, or unwanted data capture. In this first sense of being secure, imagine the comparison between a thumb drive carried around in the pocket of a rain coat between lab meetings, to a managed cloud storage system like Box Secure that is backed up across multiple systems and can be accessed from multiple devices. In the second data security concept, where data needs to be kept private, you will need to comply with guidelines prescribed by HIPAA and Yale’s Cybersecurity Group. (Go to cybersecurity.yale.edu/protectyourdata).

Now that we have gone through more specific examples of research data management considerations made during different phases of research, we will return to our four core RDM topics: organization, storage, preservation, and sharing, which represent the bulk of “what you do” as a researcher when you manage your research data.

5

Starting with organization…

6

Organization means that there is a logical, intuitive, repeated structural grouping in place. Good organization should make your files and data easier to understand and work with. Bad organization makes files difficult to find, maybe because it is not clear which folder it is in, and easy to misinterpret, maybe because file names are not descriptive or because version control practices were not put into place.

Organization happens at multiple levels, storage, file folders a.k.a. directories, and the data itself can all potentially be organized. As mentioned previously, data storage locations may vary across the data lifecycle. Depending on how many locations you use, this may make it more difficult to stay organized. Similarly, the more collaborators working on a project together, the more difficult it can be to maintain patterns, such as file and folder naming conventions, that help facilitate good organization. So, here on this slide, you can see these different assessment questions you can ask of these levels of data organization.

Under storage location, how many places do you have data stored? Does having multiple locations improve or hamper your ability to work?

With folder structures, within a storage location, how do you organize your files and folders currently? What, if any, frequent frustrations do you encounter when searching through your computer? Could you lessen these frustrations with better organization practices?

I will talk more about data structures, so for now I will just offer examples of what data structures can be. In the first Python lesson, data structures mentioned included lists and dictionaries. Data structures can also include spreadsheets and databases.

7

These practices and considerations can help you maintain a helpful folder organization structure.

Follow a consistent organization pattern. Say you have project folders, and each project has associated consent waivers, grant documents, raw, processed, and analyzed data – it would make sense to these folders represented consistently across each of these projects. Whether or not you have consistent patterns, documentation can be created to index a directory’s contents, prescribe a file or folder organization pattern, and describe how files might be interrelated.

Think about what you are organizing or grouping by. Would it be useful to organize by project? By dataset or data type? By grant cycle or academic calendar?

Overall, you should aim to have a folder structure that is user friendly; for yourself, and/or for your collaborators if it is a shared project. In order to do this, think about how it would be most useful to search, browse, and sort through your folders, and create a directory that enables you to do this.

8

File names are a key piece of metadata – metadata, meaning information about our data (for example, a book title is metadata for a book. File names are key pieces of metadata because they are highly visible, and they are the first clue we receive to indicate what a file contains.

9

With data structures, you will want to have a sense of how your data is comprised. Is it multiple spreadsheets? Is it one Excel file with multiple tabs? Is it the result of a database query, like what will be demonstrated in the Jupyter notebook and what you will be doing for this session’s exercise segment? Does the structure of your data allow you to perform the analysis you plan on doing, or does your data need to be reshaped?

10

A common data structure is a spreadsheet. Spreadsheets might also be referred to as tabular data structures. On this slide you can read through some guidelines on how to maintain spreadsheet organization that will enable to keep your data clean – something you will learn more about in a latter segment of this course. In addition to keeping data organized and clean, in order for the data structure to be fully “managed,” you should also create documentation that details variable meaning, units of measurement, acronym meaning, and data creation methods. Spreadsheets and databases will utilize data dictionaries in order to clearly communicate these things.

11

Our next key element of data management is storage. Data storage, here, refers to where your files exist. Data storage choices have implications related to security, cost, and usability.

12

Cloud storage is made up of servers maintained and backed up by a third-party company. If you use a Storage @ Yale Solution that may either be Amazon Web Services (AWS) or Microsoft Azure. The Storage @ Yale team works with these vendors in order to provide these services. Cloud platforms you may also use or be familiar with include Box and Secure Box, Google Drive, and Microsoft Teams. If you have a Yale netID, you will have access to these storage options. Benefits of cloud solutions include: collaborative features because multiple users can access the same files easily; data backup; easy to see version control (at least with Box and Google Drive). Of these three cloud storage solutions, it is important to note that Box Secure is the only one that is cleared for high risk data. In order for you to convert your Box to a Box Secure, you will need to contact IT.

13

Storage @ Yale comes in three different varieties (Standard, Enhanced, and Archive) based on how you will use the files in this directory. Each of these has a different associated cost based on the computational resources associated with the tier. The Storage @ Yale team can help you determine which solution is best for your lab or research effort.

14

Next, I will discuss data preservation and sharing. Preservation overlaps with storage, particularly long term or archival storage because it is where your data can exist for the long term.

15

While preservation and storage can sometimes be conflated with each other. The difference between the two comes from the specification that preserved data remains usable. For data to be usable, the files that make up the data need to be interpretable by your computer and the software available to you, and the context and details of the data need to be understood by the user.

16

Here are the two branches of preservation dependencies. Conceptual preservation and digital preservation. With digital preservation, files are kept in usable formats. Conceptual preservation is achieved by creating documentation including readme files, and data dictionaries that describe your data in enough detail to facilitate future use.

17

Data sharing can be another term used interchangeably with preservation or storage. It can also refer to data sharing as in data transfer – where you might be sending files to a colleague while a project is still ongoing. Data sharing as I am bringing up on this slide is the act of making data available for future research, usually at the point of publication. Data sharing could be a requirement put in place by a grant provider or publisher. Though you do not need to necessarily have a related publication in order to share a data set. In order to share data and make it accessible to others, data might be deposited in a data repository or a research paper with an associated dataset may include a data availability statement saying something along the lines of, “contact the authors in order to access the data.” Data availability statements are common in health science publications due to HIPAA patient privacy regulations. Data types, including omics data types, may have specialized data repositories hosted by an institution. For example, the NIH supports domain-specific repositories such as the Metabolomics Workbench, Genomic Data Commons, The Cancer Imaging Archive, and others. Searching for “Data Sharing Resources NIH” in a web browser will help you find a National Library of Medicine webpage that provides the full list.

18

Now that we have gone through concepts and considerations related to research data management, I will move on to relational databases so you can see some examples of these concepts being applied.

19

Database is a common term, but it can be used in different ways. Libraries provide access to databases that store catalogued data on library resources. Excel spreadsheets are sometimes referred to as databases, though an individual spreadsheet is not a database. At the point of design will also need to plan ahead to think about how you will leverage this database for your data collection, data processing, and data analysis. If you are not creating your own database, but rather using a database created by someone else – understanding the structure and design can help you use it more effectively and determine whether or not this database will suit your needs.

20

Relational databases are data structures that allow you to organize data with more control and ability to query against what are essentially multiple related spreadsheets. These related “spreadsheets” in a database are called tables. Upon creating the tables of a relational database, you declare what datatypes and data sizes will populate the columns of your tables.

21

These two tables represent tables in a hypothetical database. The first contains subject data, and the second contains data related to admissions.

Here you can see how the data in this database representation appears to be standardized. The subject\_id column always contains a five-digit number. The date of birth column always contains a date, and this date is always formatted in the year, month, day style. Insurance type includes what appears to be categorical data. A column like this one would most likely has a set number of options that can be input in order to maintain constancy. For example, not all private insurance providers are captured. If you were hoping to perform an analysis comparing different private insurers, this database would not be suitable for you.

In these tables, you can also see how these tables are related to each other as indicated by the red lines. Each admission has an associated subject and these are linked by way of the subject\_id identifier. The first subject has had two separate admissions, which can be seen in the first two rows of the second table.

In database design, these visual representations of database tables and how they are related are incredibly helpful.

22

Another vital tool that will help you understand database design is the documentation that contains details about the data captured in the database. These details can be captured in a data dictionary, which you will see on the next slide.

23

This is a data dictionary for the subject table we looked at two slides ago. For each column name, the data dictionary details what data type, what format of the data is acceptable, and now to handle missing data or Nulls. In this case, none of these fields can be left blank if a new record is created – or in essence, if a new person is added as a patient or subject. At the data collection point, these fields would be mandatory to fill out in a patient intake form.

24

We interact with databases regularly. If you fill out a form that someone else enters into a database, you are being catalogued in a database, and the cataloger is adding content to the database. If you are looking at patient representation within a hospital database system, or you are querying a research database, these are both more front-end database interactions. The user interface of the CDC Wonder Cancer Incidence database follows the ideas of an SQL query but provides a Graphical User Interface. SQL queries, which I will discuss further, and what you will be using in the exercises are text commands you can use to interact with relational databases.

25

There are multiple ways in which you might create a relational database. Here are a few listed. MySQL is an open source option. Microsoft Access was commonly used at one point because of the prevalence of Microsoft products. Python and R have the ability to create and query databases using SQL libraries. RedCap is a software used in clinical research. With RedCap, you are able to deploy surveys, similar to what you would do with Qualtrics or Google Forms. However, the added value of RedCap is that since survey and form entry results are stored in a relational database, you are able to link across all forms and data ingested from other sources like activity tracking instruments, or any number of data sources.

26

I have referenced “SQL” several times now. SQL stands for Structured Query Language. As I mentioned before, it is a language used to build, query, and modify relational databases.

27

The demo and exercises that accompany this session provide examples of SQL queries run through Python. Read through the demo Jupyter notebook after you are finished with these slides to learn the “how-to’s” of SQL queries.

28

This is a big picture view of what a SQL query looks like. This presumes there is already a database that has been established, and an SQL query like this one will pull data from this database from specified columns (established with the SELECT column\_name FROM table\_name), where certain parameters are met (established with the WHERE statement). GROUP BY and ORDER BY affect how data is presented in the results, not which data is pulled.

With a query like this, we could say SELECT patient\_id from admissions\_table WHERE admission\_date is between a certain range AND a patient is within a particular age range, and order by admission date. This would result in a subset of our database.

SQL is not case sensitive, but it is common to see SQL terms in all caps as an easy way to differentiate what is an SQL term (SELECT, FROM, GROUP BUY, etc.) and other terms that point to database specifics like column names, and constraint parameters.

29

Here is another example of an SQL query. This example brings in additional elements including JOINing, which allows you to link one table to another by a common identifier, like what you saw on slide 21 in the example patient and admissions tables. The other new element in this example is GROUP BY. Like ORDER BY, GROUP BY affects the result appearance. GROUP BY statements allow you to see summary results. If you group by age demographic, you will see a summary number based on that grouping. This could be a count, an average, a max value, etc. You will see this in the demo workbook.

30

Before you move onto the demo workbook, here is a quick list of Yale-based services that can help you with the research data management and database topics discussed during this session.

31

If you are not sure how to get started or where to go, you can always as the Cushing/Whitney Medical Library. I am available to answer questions or direct you to more specific help related to RDM, computational strategies for working with data, finding datasets for reuse, and data visualization. Our library also has a Bioinformatics Hub, where we have two experts in omics data analysis.

32

Maybe you have a floppy disk or a CD ROM, or the file type itself is inaccessible to you. The Library’s digital emulation services can assist you in cases where you have file types that you are unable to open on your computer.

33

Yale IT can help with database configuration and hosting, security assurance, and data storage solutions.

34

The Yale Center for Research Computing (or YCRC) can help in cases where large datasets need to be transferred, code optimization, and with high performance computing (or HPC).

35

The last group I will mention is the Joint Data Analytics team (a.k.a. JDAT). JDAT provides curated access to electronic health records by working with researchers to create custom queries to the Epic health database.

36

This concludes the presentation on research data management and relational databases. Next, read through the jupyter notebook for this session and ask questions during the morning office hour session. In the afternoon, you will complete some exercises based on what you will learn in the demo notebook. Any time after this lesson, feel free to email [medicaldata@yale.edu](mailto:medicaldata@yale.edu) with your RDM questions. Thank you for your attention during this time.