DEPARTMENT OF MEDICINE
Biostatistics Shared Resource

# Collecting Research Data: Data Management Tips

**TIP 1**: Before you actually collect any data, please meet with a biostatistician to discuss your study! Doing so can result in better study designs, better analysis plans, and better analyses, especially when complex analytical methods might be required. Regardless of how much data have been collected, even the most sophisticated statistical analysis cannot overcome the challenges created by a poor research design / data collection process.

Once collected, data for a research project will often be analyzed by a biostatistician. One of the first things that the biostatistician will need to do from an analysis perspective is to read the data into a software package capable of conducting the required statistical analyses. Examples of general, multi-purpose, statistical analysis software include SAS, R, STATA and SPSS. Note that Excel and GraphPad are not generally used by biostatisticians. Each of these programs require data to be in a specific format unique to the software program. The remainder of this handout provides some tips to avoid collecting and recording data that in ways that may negatively impact your study, as well as increase the difficulty for and time required of the biostatistician.

## A Biostatistician's View of Data Structure

Generally, it is helpful to think of a single dataset as a rectangular table with rows and columns. A database is a collection of related datasets. Usually, different datasets within a database relate to a specific aspect of the project.

### Identifying each observation (row):

Each row is a **unique observation**, also called a **"record".** Sometimes this is a patient or an animal. When the same subject is measured repeatedly over time, the "unique" observation, or row, could be identified by **two factors** – the subject and the time of the measurement.

**TIP 2**: It is extremely important to include in every dataset **a unique ID number** for each subject in the project. Equally important is to ensure that the data are **de-identified**; meaning, it is **not** appropriate to include hospital MRNs, social security numbers, names, telephone numbers, exact street addresses, etcetera in any dataset that is shared with the biostatistician. If this information is required for the researchers to obtain data, then a **separate "key"** should be kept by the investigator that allows the non-identifying Study ID to be converted into identifying data.

## Identifying each measurement (column):

Each column is a **measurement**, also called a **"variable or field",** which pertains to that unique observation or observation/time. When data is read into any of these statistical software programs, the program must decide – and define – the **type** of data that is in each variable.

**TIP 3: Keep variable names short**. Most software packages have a limit to how many characters can be included in a variable name. And, the biostatistician does not want to write out very long names when writing code; in addition to being an unnecessary waste of time, it increases the chances of errors (e.g., spelling mistakes). See section on data dictionary below.

**TIP 4: The data type definition must apply to every observation for the same variable.** If the type of data is not consistently recorded for each variable, the data will be entered incorrectly and (at best) lead to errors in the analysis. Meaning, for a given column in the data file, the entries across all rows must be of the same data type; however, columns can certainly have different data types.

Common data types include:

- **Integer** – a number without any decimal places
- **Numeric** – a number which MAY include decimal places
- **Date** – how dates are handled by each program may vary. Usually dates are stored **internally** as the number of days from a specific reference date. This allows the interval between dates to be easily calculated. Dates must be entered in a specific and **consistent format** within one variable. For example, it would not work to have some observations for the variable DOB entered in a MM-DD-YYYY format and other observations for DOB entered in a MM/DD/YY format.
- **Text** – any field that is entered with any letters included. Keep in mind that capitalization matters. For example, you may want to record a binary response as "yes" vs "no". The responses "N", "n", "NO", "No", "no" are all **different** responses due to how the text field is entered. More complicated text fields should have the most common responses assigned to a number code to avoid data entry variation. Then, include an "other" option with an associated text field for a comment. This will produce less text data, and more analyzable data. **TIP 5: Text data is the most difficult data to use in statistical analyses** and should be avoided if possible.
- **Missing** – see below

**TIP 6: ONLY the "raw" data needs to be collected**. For example, if you are interested in the "Time to Discharge" from the hospital, you need to collect the data of admission and the date of discharge. You do NOT need to include the duration in the hospital, as this calculation will be done during analysis. Similarly, if you are interested in the "Age at Surgery", you should collect the date of birth and the date of surgery. This will result in the most exact measurement of the variable of interest.

**TIP 7**: Please do **not** include multiple responses in a single data field.

For example, if you want to collect the reasons for surgery and valid choices are 1=medically necessary, 2=patient requested, 3=family decision, 4=other, then it would not be helpful to list all reasons in one

text field like “1, 3” or “2 and 4” when multiple reasons may exist. First, the inclusion of a comma or the word “and” converts the entire variable to a text field, even those with a single number response. Second, if you would like to know how many subjects had surgery because of “patient request” (response 2), the analyst needs to combine “2”, and “1,2” and “2 and 3” and (however many possibilities are in the data by writing code). A better way to collect this data would be into 5 separate Yes/No fields plus a text field to collect the “other reason” that would rarely be needed:

| ID | Srg Reason Medical Nec. | Srg Reason Patient Requ. | Srg Reason Family Decis. | Srg Reason Other | Surg Reason Othr Comment |
|---|---|---|---|---|---|
| 1 | Yes | | Yes | | |
| 2 | | Yes | | Yes | Study Particip. |
| 3 | | Yes | | | |

Now, it is a simple matter to determine how many patients had surgery due entirely or in part to their own request by looking at one column, or variable.

**TIP 8:** Create a Data Dictionary

Ideally, **before** any data collection begins, the investigators will create a data dictionary for the project. This will be a **separate document from the file containing the data**. If the data are provided in Excel, the dictionary can also be in Excel, but must be placed on a separate sheet. For complicated projects or for inexperienced researchers, it is advisable to involve the biostatistician (again) at this time so they can verify the data will be in the most relevant and usable format for the planned analysis.

The advantages of creating a data dictionary are many:

- Allows confirmation that all relevant pieces of information (variables) required to assess each of the research hypotheses have been obtained and will be available for planned analyses.
- Encourages one to decide on the **name** for each measurement field and to **provide a description** of the variable along with any relevant units of measurement or study time point information.
- Determines “how” each variable will be stored in the dataset by selecting the most appropriate **data type**, the appropriate choice being dictated by how it is measured or observed. For example, using data collected from the electronic health record means that the type of data available is already fixed. This process encourages data entry to be consistent between people doing data entry, now and in the future.
- Allows one to design how datasets within a database will relate to each other (e.g., by choosing which data fields will be used to connect or link the data in different datasets together).
- When the data is given to the biostatistician, the data dictionary provides a road map for what is included in the database. This improves project **efficiency**!

As part of the data dictionary, it is also important to think about issues that may create data that are missing (either by happenstance or by design), and to ensure the biostatistician understands the coding used to denote such missingness. Specifically:

- **Missing data** for a given variable should ideally be coded as the same data type as the non-missing data. For example, for a text field, you might select "NA" or "UNK", depending on whether the missing data is not applicable due to a "skip pattern" (see below) or the data could have been obtained and was not (UNK). For nonnegative integer or numeric data, you might choose "-99" or some other value that is guaranteed to be out of range of valid responses. Similarly, for dates, you could choose 01/01/1900 – or any date that is known to be out of range of valid responses (note: use the same choice for all missing dates). It can be problematic to indicate a missing date with "NA" because it forces the analysis program to read the dates as TEXT and we lose the nice properties for calculating intervals between dates. For each variable where missing data occurs, the code used to identify the missing information should be provided as part of the data dictionary.
- **Skip patterns.** Sometimes you only want to collect certain variables ONLY if a specific event has happened. For example, if the patient has had prior chemotherapy, then we collect the starting date and the regimen. All subjects would answer an initial question "has the subject had prior chemotherapy" (response: Yes or No). Only those with a Yes response, are expected to have data in the two follow-up questions for start date and regimen. These types of situations should be noted in the data dictionary.

## Database Creation

In addition to the data dictionary, you will typically record and store your data in either an **Excel** Spreadsheet(s) or in a **RedCap** database.

**Excel** may be a better choice for a **single, simple dataset** that will only be used for a **one** project. The rows and columns of a spreadsheet corresponds to the format of a single dataset. It is possible to think of an Excel **Workbook** as a database with several related datasets on separate spreadsheets.

### Advantages:

- Almost everyone has access to and some familiarity with Excel spreadsheets.
- Most statistical analysis software programs can directly import Excel spreadsheets.

### Disadvantages:

- There are no restrictions (guard rails!) as to the **types of data that can be used within the same variable**. Even if Excel does not care, the statistical analysis software WILL care!
- It is not easy to use and see skip patterns in data collection. This may result in spreadsheets that look like there is a lot of missing data when in fact such data was not meant to be recorded.
- Data dictionary file must be produced "by hand".

- There is no security of the data in an Excel spreadsheet unless the sheet is password protected (REVIEW → PROTECT SHEET → select password) or the Excel spreadsheet is kept in a secure location like Box.

**Redcap** may be a better choice for a more **complex database** that will be added to by several users and/or added to over time on a continuing basis. A complex database will typically involve collecting several "forms" from each patient, potentially at different time points in the study.

**Advantages** (copied from https://redcap.urmc.rochester.edu/redcap/):

- **Build online surveys and databases quickly and securely in your browser** - Create and design your project using a secure login from any device. No extra software required. Access from anywhere, at any time.
- **Fast and flexible** - Go from project creation to starting data collection in less than one day. Customizations and changes are possible any time, even after data collection has begun.
- **Advanced instrument design features** - Auto-validation, calculated fields, file uploading, branching/skip logic, and survey stop actions.
- **Data quality** - Use field validation, branching/skip logic, and Missing Data Codes to improve and protect data quality during data entry. Open data queries to automatically identify and resolve discrepancies and other issues real-time.
- **Custom reporting** - Create custom searches for generating reports to view aggregate data. Identify trends with built-in basic statistics and charts.
- **Export data to common analysis packages** - Export your data as a PDF or as CSV data for easy analysis in SAS, Stata, R, SPSS, or Microsoft Excel.
- **Secure file storage and sharing** - Upload and share any type of file with anyone in the world through the File Repository feature or Send-It tool. Also works with exports and other built-in file uploading features.
- **Data-based triggers and alerts** - Send real-time alerts and notifications to your team or other stakeholders via email, text, or phone based on certain data being entered or specific questions having a particular answer.
- **RedCAP** cannot randomize patients but **CAN** accept an external file with randomized assignments (for example produced in PASS) such that randomized designs happen seemlessly
- Learn more about REDCap by watching video tutorials of REDCap in action and an overview of its features, please see the Training Resources page.
- If you require assistance or have any questions about REDCap, please contact REDCap Administrator.

***NOTICE: If you are collecting data for the purposes of human subjects research, review and approval of the project is required by the Institutional Review Board.***

**Disadvantages:**
Despite how "easy" they make Redcap look in the videos linked above, there is a "learning curve" to build a RedCap database! However, there are also experts on campus that can be contacted either for advice (contact the Redcap administrator) or for actually building a database to specification for a fee.

Redcap Support Request form is here:
https://www.urmc.rochester.edu/smd/it/project-request.aspx
(see link on page for Redcap Project Requests)

## Data Management

After the data is read into the statistical analysis software, the biostatistician will confirm the accuracy and integrity of the data, as they best understand it should be. Data queries will be sent back to the investigator for review and verification. Less data checking is necessary when data is imported from Redcap compared to Excel because the data entry into Redcap has already done some of this work.

- Validity of codes using the data dictionary. If all responses should be "A" or "B", then any other codes should be referred back to the investigator for clarification.
- Range checks of all variables to verify that no data entry errors have been made.
- Calculating intervals known to be always positive such as age and survival (etc) can identify if dates have been entered incorrectly.
- Basic summaries of main variables should be examined by the investigator to verify it is as they expect in terms of scale. For example, was subject height recorded in inches or meters?