Maximizing Your Time with a CBHDS Biostatistician

Data Preparation

Please consider the following guidelines if you plan to collect and/or organize data in preparation for sharing with a CBHDS biostatistician for analysis.

1. Deciding what to measure
   An early part of any research project is deciding exactly what you want to know about the sample you plan on studying. It can be helpful to organize these characteristics into groups, such as demographic, clinical, medications, laboratory results, etc.

2. Deciding how to measure it
   There are several different types of variables (nominal, ordinal, dichotomous, scale, rate, percent, etc) that can be used to measure the sample characteristics and outcomes for your study. For example, although body mass index (BMI) is typically measured on a continuum, this variable can also be broken up into 4 categories defined by the Center for Disease Control and Prevention (CDC) as: underweight (<18.5), normal weight (18.5-24.9), overweight (25.0-29.9), and obese (>30); an investigator might find a specific variable type more suitable to describe their study sample or specific study aims. We recommend collecting data at its highest level (continuous in the case of BMI), so that you have more flexibility in working with the data. Note that you can reduce a continuous measure to an ordinal or dichotomous measure, but you cannot work in the opposite direction (create a continuous measure from an ordinal measure).

3. Identifying participants/patients
   We recommend identifying participants/patients using unique identification numbers. Numeric IDs are preferred over names or initials to protect patient privacy according to the Health Insurance Portability and Accountability Act (HIPAA) Privacy rule requirements. As such, medical record numbers, health plan numbers, social security numbers, or other unique identifying numbers should not be used as a unique identifier. Please do not use the Excel row number to identify patients. In order to merge your data with other potential data sources, a unique consistent "link" ID is needed to identify individuals.

4. Entering your data into a worksheet
   CBHDS biostatisticians are knowledgeable in several statistical programming languages/packages, including but not limited to SAS, R, Stata, SPSS, etc. As such, we ask that you do not enter your data into a Microsoft Word table. Word tables do NOT import properly into statistical packages. When collecting and organizing your data, please adhere to the following guidelines:
   a. Acceptable file formats for data sharing are REDCap, Excel (.xls or .xlsx), comma-separated values (CSV) file (.csv), SAS (.sas7bdat), SPSS (.sav), or Stata (.dta). Please contact your biostatistician for guidance when using any other format.

   b. The first row of your data file should contain only the column/variable name. All variable labels, units of measurement (e.g., years, days, etc) and coding (e.g., for race, 1=Asian, 2=Black or African American, 3=White, 4=Other, etc) should be included in a data dictionary saved as an Excel (.xls or .xlsx), Word (.doc or .docx), or PDF (.pdf) file.

   c. A data dictionary may contain the following information:

i. The exact variable name as in the data file
ii. The column that the variable can be found (e.g., for Excel files, age at treatment might be in Column B)
iii. A descriptive label explaining what the variable is
iv. The measurement of units or coding of the variable
v. Expected minimum and maximum values

An example of a data dictionary created in Excel has been provided below.

Variable Name	Column	Variable Label	Unit of Measurement/Coding
ID	A	Unique patient identifier	
age_at_tx	B	Age at treatment	Years
sex	C	Patient sex	1=Male, 2=Female
race	D	Race	1=Asian, 2=Black or African American, 3=White, 4=Other
ethnicity	E	Ethnicity	1=Hispanic or Latino, 2=Not Hispanic or Latino
bmi	F	Body mass index (BMI)	

d. For categorical variables, please be sure that all possible responses have been defined in your data dictionary. That is, your data dictionary should help your biostatistician define ALL possible codes in your data for your categorical variables, such as race or sex. Codes that are found in your data, but not defined in your data dictionary, may cause delays in the data cleaning process.

e. When entering data into your file, please make sure that responses for all patients in a single column are either all in a text OR numeric format, but not both. For example, if you collected glucose levels from patients and are providing the actual value in cells, we would expect all entries for this variable to be numeric. If there were issues with this variable for a particular patient, or there are details in the patient's medical record that should be noted, please provide these details in a separate column for your notes. Also, please note that if you only know that a value is above or below a certain value (e.g., >65), please note that in a separate "notes" column and not in the column expected to be entirely numeric.

f. Should any additional columns/variables need to be derived for your data analyses, please provide your biostatistician with specific instructions of values needed to derive any new variables. For example, if you would like to create a dichotomous variable to identify patients that were greater than 65 years old and were overweight, you would need to provide the column/variable name of each characteristic and the corresponding value(s) that will be needed to create the new variable (e.g., age_at_tx > 65 and bmi > 24.9).

g. Keep column/variable names short ($\leq$ 12 characters) and make sure that they are unique. All variable names should contain no spaces and should not start with a number or symbol. In lieu of spaces, underscores ("_") can be used. For instance, "Age at treatment" can be named "age_at_tx" to be under 12 characters with no spaces.

h. Please be consistent in the codes for categorical values. For example, for a categorical variable like patient sex, use a single common value for males (e.g., "male"), and a single

common value for females (e.g., “female”). Do not at times write “F”, and other times “Female”. Pick one code and stick to it throughout the data file. Also note that numeric values are preferred (e.g., 1=Female and 2=male).

i. For character variables, please also be consistent with the letter case. For example, ‘female’, ‘Female’, and ‘FEMALE’ are all considered different responses in statistical software. Letter case should be consistent throughout your data file for each variable.

j. For variables with the same response options (e.g., yes/no) for coding, do not code one variable as ‘1=yes, 0=no’ and another variable as ‘1=no, 0=yes’ or ‘1=yes, 2=no’ as this may cause confusion. Please note that the convention for coding yes/no variables is: 0=no and 1=yes.

k. For missing data, please leave these cells empty. If there are special notes on why the data is missing, please include this is a separate “notes” column (e.g., N/A, patient died, etc).

l. Be careful about extra spaces within cells. A blank cell is different from a cell that contains a single space. That is, “male”, “ male”, and “male ” are different.

m. Avoid the use of commas in both text and numeric fields. For example, if transferring 4-digit (or more) numeric data, please use ‘1234’ instead of ‘1,234’. For text fields, please use a ‘/’ to separate items rather than using a comma (e.g., for medications, please use “Acetaminophen/Amoxicillin” instead of “Acetaminophen, Amoxicillin”).

n. Dates should be formatted as MM/DD/YYYY. If the exact time is important, please include as a separate column/variable in your data file.

o. Use 24-hour clock to input time variables. For example, for 1:45pm, please use 13:45 instead.

p. Do not indicate any distinguishing patient group or characteristic with colored font or highlighted cells as these will not be transferred when data is imported into our statistical software. Instead, please include a separate column/variable to indicate the patient meets specific criteria. For example, if you want to identify patients who were older than 65 years and use a wheelchair, please create a separate column/variable indicating 1=yes or 0=no if the patient meets that criteria instead of highlighting them in pink.

q. Please do not include any blank rows or columns, as these will be read as observations/variables when importing the data into our statistical software.

r. Please do not include hidden rows or columns of data in the file you will be sharing.

s. If patients have more than one observation for the same variable, include multiple rows for that patient, identified by the patient identifier. As shown in the sample data file below, **PatientID** is the unique patient identifier, **Timepoint** is the assessment time of BMI, and **BMI** is the actual BMI value.

PatientID	Timepoint	BMI
1	1	26.9
1	2	25.5
1	3	25.3
1	4	24.9
2	1	22.3
2	2	23.0
3	1	24.2
3	2	24.2
3	4	24.5

t. Do not include comments or explanations of variable names, study design, data collection, or any irregularities that occurred during the study or data collection in the actual data file. These should be provided in a separate Word (.doc or .docx) or PDF (.pdf) file.

   For all other guidelines for organizing data to be shared with your biostatistician, please refer to Broman and Woo's (2018) article entitled "Data Organization in Spreadsheets".

5. Sharing data with your biostatistician
   a. Please check your data for any typos before sharing with your biostatistician.

   b. Shared data with the CBHDS team should not contain any protected health information (PHI), including but not limited to: patient name, date of birth, phone number, address, email address, medical record number, health plan number, social security number, or other unique identifying number, characteristic or code. Please remove all PHI before sharing. For a full listing of PHI, please refer to What is Considered Protected Health Information Under HIPAA?

   c. Please note that for human research studies, before allowing CBHDS biostatisticians access to the data, you will need to add them as 'key personnel' in the IRB study protocol.

   d. If there are corrections to your data, it is the responsibility of the investigator to provide the CBHDS biostatistician with an updated data file as soon as possible.

For grants, analyses, and papers/presentations that we support outside of direct grant funding or contractual agreements with CBHDS, please acknowledge that your work was supported by core grant funding through the iTHRIV CTSA using the following language:

"Research reported in this publication/presentation/work was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number

UL1TR003015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health."

These guidance documents are modified versions of documents prepared by our colleagues at the Biostatistics Collaboration Center (BCC) at Northwestern University. We are grateful for Dr. Leah Welty and her team for their guidance and input.