STEP 4.2: Clean and code your PHDS data set

What is the purpose of this step?

The purpose of this step is to obtain an accurate data set from your survey vendor. This step includes consistent and accurate cleaning and coding of the data set in preparation for constructing PHDS quality measures and the analytic variables needed to report your PHDS project findings.

In this step you will:

- ☑ Specify data files to be created.
- ☑ Obtain and check interim data sets for accurate data labeling and entry.
- ☑ Prepare data files for analysis.

Guidelines and Issues to Consider

While data preparation is often considered part of the analysis, this task is included in the data collection section because it may be completed by the vendor while administering the survey. Additional data preparation and cleaning steps are described in Step 5.

☑ Specify data files to be created.

If you are using a vendor to administer the survey, the vendor should submit a data file that contains the following:

1. Coded responses for all PHDS items, including blank, do not know, refused to answer, and item skipped.
2. Survey disposition, such as if the survey was completed and the reasons for incomplete surveys (see variables noted in Step 4.1).
3. Other descriptive variables identified and collected for the starting sample (e.g., related to enrollment, health care utilization, etc.) that were identified in Steps 2 and 3.
4. Administrative data used for generating the sampling frame.
5. Age of child in months.
6. Any supplemental data linked prior to the removal of identifying information used for survey administration.
7. A data dictionary for the file. An example data dictionary for the ProPHDS survey data file can be found in **Appendix 9**.

You also may choose to have your vendor conduct some initial data preparations, such as:

⇒ Verifying ineligible cases
⇒ Checking for duplicate data records
⇒ Running frequencies on all variables to check for values that are out of range
⇒ Identifying problems with skip patterns

If errors are found, you should have the vendor verify them with the original surveys to ensure that the errors stem from the respondent and not from the administration process. Once these are identified, you will need to make decisions on how they will be handled for the analysis. Refer to Section 5 for more detail on analyzing the results.

☑ Obtain and check "test" and interim data sets for accuracy of data entry and survey administration.

The vendor administering your survey (either internally or externally) should provide you with a test and interim PHDS survey data sets according to a predetermined schedule.

CAHMI recommends that you ask your survey vendor to send a test data set that is based on a handful of mock completed surveys. This data set will test the data entry processes and ensure that the data set you receive matches the data dictionary your vendor is using.

Tips from the Field

- Always label data variables.
- Update your data dictionary with any changes made to data labels or response codes. Good documentation is essential!
- Create a backup of your data set in case of emergency. Also, create temporary and permanent data sets wisely. Think about what you would need to do if you lost the data.
- Always keep a copy of your original data set.

When you receive the test data set you should make sure that your vendor is using the data variable labels agreed upon and that responses to survey items are assigned the agreed-upon values (e.g., 1= "no"; 2 = "yes", etc.). If errors are found, request that they be corrected immediately.

CAHMI recommends that you request at least two interim data sets. The first should include the first 100 surveys entered and the second should include half of your expected final completed survey data set (e.g., if your final complete survey goal is N=2,000, then the second interim data set should be N=1,000). These interim data sets allow you to develop the syntax that you will use to clean and analyze your PHDS data. Therefore, when the final data set is received, you will have already done a significant amount of preparatory work.

☑ Preparing the PHDS data files for analysis.

Valid PHDS findings require careful preparation of your data prior to starting your analysis. The following are necessary steps to prepare the data for analysis. They do not necessarily need to be conducted in the order presented.

Data Prep Step #1: Verify survey completeness.

You should receive the data from the vendor for all of the interviews conducted. However, for your analysis you should limit the data to surveys with at least 80 percent of the items completed.

Data Prep Step #2: Check for ineligible cases**.**

Make sure parents who responded have children who meet the sampling criteria for age and continuous enrollment. (a) Run a frequency on the age variable from the survey responses. Here you should ensure that the age the parent reports in the survey and the age-specific section of the PHDS that the parent completes match the age of child that you have in your administrative data files. Use the parent report as the "gold standard" and exclude cases where the child was erroneously included in the sampling frame. (b) Remove records where the child was found not to be in the health plan, provider, or unit you are sampling.

Data Prep Step #3: Check for duplicate data records.

Make sure every record has a unique identifier.

Data Prep Step #4: Check for out-of-range values.

Run frequencies on all of your variables to check for out-of-range values or odd-looking distributions. At this point, you may not be able to go back and correct the data error. If the error is random and affects only a few cases, then you may want to exclude those cases. However, if the error seems to be systematic and affects a large number of responses, it may be worth finding the source of the error and correcting it.

Data Step #5: Identify problems with skip patterns.

Run frequencies and cross-tabulations to verify that skip patterns were followed correctly. If errors seem random and affect only a small number of records (less than 2%), assume the item stem (the question instructing the respondent to go to a different question) is accurate and then correct the response for the incorrect skip. Systematic errors or problems with a significant number of cases should be verified.

If you want to be absolutely certain that skip patterns were followed, you can require that only the children of parents who responded appropriately to the filter question are included when you create the new variables.

Data Prep Step #6: Assign missing values.

Missing values should be recoded in some way so that you know not to include them in the analyses. You should designate missing values in the data set in a way that ensures they are omitted when calculating measures. Also, recode the response options of "refused" to a missing value. Examine the number of "I don't know" responses that you get. If this total percentage is less than 2 percent, then you should recode them as missing values.