Tool 3.1 – Data Quality Control Checklist

To ensure that data meets certain quality standards before analysis begins, this checklist offers questions to consider when checking data for quality. Once each question is answered or considered, you can check the box on the right.

Depending on the source of your data, some questions you may want to ask the data provider:

Where do the data derive from? □

What are the data used for? □

How will the data be extracted? Is there a standard process for this, or will this need to be developed for this purpose? □

What fields will the data include? Is there any documentation available? □

Is there a specific set of records that will be selected for extraction? How will that be specified? □

Is there a lag in the data entry that would require a wait to get complete data through a particular period? How often and when are the data updated? □

Are there any system conversions that took place during the period these data cover? □

How far back do the data go? □

What are the known weaknesses in the data? □

Prior to processing the data:

What are the criteria for inclusion in the dataset? □

What is the sample size? □

Prior to writing code to process the data:

What data checks should be made to ensure accuracy of the final results? □

While processing the data:

Did I check for missing data? Also, did I check by subgroups, to uncover patterns of missing data? □

Did I check for data conversion issues? □

Did I check for duplicate or partial duplicates? □

Did I check for internal inconsistencies, unexpected values, and outliers? □

Did I check for common programming mistakes while:
reading data files? □
merging files? □
restructuring data? □
creating variables and date values? □
working with arrays? □

Did I select the correct sample(s) for routines that apply to a subset of records? □

Did I apply the correct formats (numeric, character, date) for all variables? □

Did I check sample size for each outcome measure? □

Did I create summary/aggregate checks? □

Did I follow a group of random cases through every programming step? □

Did I check that these data files look consistent over time? □

Do categorical measures add up to 100%? □

Did I follow up on any suspicious findings? □

After processing the data:

If datasets are updated over time, did I review all data checking output with each data update? □

Did I clearly document any data quality issues found and resolved? □