# Data Checks

## Purpose

Data checks are an important quality control measure in maintaining a clinical dataset. Frequent and regular data checking identifies potentially problematic or missing values in close enough proximity to data collection that the discrepancy can be rectified more easily than at the conclusion of the study. The appropriate interval for performing data checks will vary by project, but typically data checks occur weekly, biweekly or monthly. In addition to regular checks, periodic checks may be requested by study investigators for specific data. In these cases, data queries are forwarded to members of the Research Informatics group or other study staff.

## Scope

This SOP is intended for use by any clinical research staff who may be involved in maintaining and data checking a clinical dataset.

## Responsibilities

#### Data Manager

* To generate data checks
* To address discrepancies when necessary

#### Project Coordinator

* To oversee data checking
* To address discrepancies when necessary

## Procedure

1. Data Check Programs

* Custom programs are written to analyze electronically stored data for discrepancies including outlier values, missing or unverified forms and data fields, and the overall state of data collection and entry
* The programs are typically authored by members of the Research Informatics department but may also be written by other study staff
* After a program is developed, study staff can run the program with little to no programming experience
* Reports are generated from these programs and then distributed to the appropriate study staff members, who then correct or verify the error
* Reports are then regenerated to verify the correction

1. Address Missing Data

* Data Manager should run frequent data checks that can aid in identifying missing data and differentiate between data that is expected to be missing and data that should have been collected
  + For data that should have been collected, staff check to see if the data was collected but not entered electronically, or if the data was not collected
  + In many cases, data that was not collected may still be obtainable. For example, if the data being checked is from a questionnaire, and a missing value is found, staff may follow-up with the study participant to obtain a response

1. Address Outliers and Implausible Data

* Outliers or data that may seem implausible often occur in clinical datasets. These data may represent real values, but sometimes they signify errors in data entry or collection
* Programs may flag potentially implausible data by applying simple logic (e.g. a variable measuring “hours/day” of an activity has some value > 24 hours) or using a prior knowledge of the acceptable range of values (e.g. oxygen saturation = 0% for living participant; participant height = 170 and the unit of measure is inches)
* In addition to implausible values checks previously described, or when such checks are not feasible, plausibility of data may be evaluated by identifying outliers based on data distributions
* Outliers and potentially invalid data should be checked against the data source to confirm accurate data entry.
  + Otherwise, the manner with which outliers and implausible data are handled will vary by the project and data type
  + The Data Manager should dictate specific methods at his or her discretion and study staff should document any changes made to the raw data

1. Data Check Reporting

* Results of data checks can be easily exported from program software into useful formats, such as pdfs or spreadsheets
* If not performing the checks and generating the reports directly, the Data Manager should appoint an individual staff member to be responsible for running data checks and should detail a procedure for handling data discrepancies.

## References

#### Software

There are many statistical software packages available to aid in data checking. These include:

1. SAS©. In the Division of Sleep Medicine Epidemiology, SAS© is the software package used most frequently in data checking. SAS© has strong data handling capabilities and is able to process large, rectangular datasets efficiently. Because data checking often involves checking the values of many variables across multiple observations (or subjects), SAS© is effective in this process. SAS© software is maintained by and available for purchase through the SAS Institute.
2. R. R is open source software for statistical computing and graphing. Maintained by a growing user community and available for free online through [r-project.org](http://www.r-project.org/), R provides a variety of software packages and statistical functions that can aid in data management as well as analysis. Its flexibility can be an asset especially for unique or unusual data types.

=============== Previous: [Data Entry and Verification](https://github.com/sleepepi/data-management/blob/master/300-data-entry-and-verification.md)

[Back to README](README.md)