2.3: Guide to Data Cleaning

Data cleaning monitors and processes data to ensure reports are accurate. It typically takes places in two primary stages.

1st Stage: Pre-Report Data Cleaning

- Pre-report data cleaning refers to data cleaning and maintenance steps taken before any critical reports are due. This involves the ongoing monitoring of critical data elements by staff and supervisors to take steps to ensure data is accurate.

2nd Stage: Peri-Report Data Cleaning

- Peri-report data cleaning is done right after a report is generated and before it is distributed, when it's no longer feasible to go back and clean data. The goal is to minimize the damage of the inaccurate and missing data. Cleaning is focused on removing any records from a report that inaccurately reflect services due to errors or absence in the data.

Steps for Pre-Report Data Cleaning

Home Visitor Tasks

Home visitors are the first line of defense against missing or inaccurate data. Home visitors can review their family case files on a routine basis (e.g., weekly) to ensure each case file is accurate and current. These checks can include a scan of home visit forms, screenings, and referrals for missing and incomplete data. Exhibit 1 is one example depicting how home visitors can review their case files and summarize the information in a table.

Exhibit 1. Family Missing Data Summary (Completed by Home Visitors)

ID	Name	Parent DOB	Child DOB	Income	Race	Ethnicity
1	Jane Smith	X	X	X	X	0
2	Ashley Waller				X	X
3	Sara Prince					
4	Barbara King	X	X	X	X	X
5	Linda Goose		N/A	0	0	0
6	Seana Smith		N/A	X	X	X

X – Missing 0 – Refused N/A – Not Applicable

Supervisor Tasks

Supervisors can assist home visitors in ensuring case files are being accurately and completely maintained in the data system, and can add an additional review of entries.

Supervisors will typically check case file data less frequently (e.g., monthly) and look across all the home visitors. Exhibit 2 provides an example of a supervisor level summary.

Exhibit 2. Home Visitor's Missing Data Summary (Completed by Supervisor)

Home Visitor	Number of cases late/ missing depression screening	Number of cases missing referral	Number of cases missing DV screening
Anna S.	5	3	0
Barba G.	0	0	1
Wanda S.	6	6	8

Steps for Peri-Report Data Cleaning

Data Manager Tasks

Data managers or evaluators typically run descriptive reports on data elements from the final report to identify outliers, unusual values, or unintended data trends (e.g., large number of families with no income). Once this process is complete, the data manager develops a list of potential issues for the supervisor. Exhibit 3 illustrates an example of a monthly service report a data manager could run. In this example, the number of pregnant caregivers served was larger than the number of total visits. This may require a followup with a supervisor to determine if this is a data error or a situation in which two caregivers were enrolled but missed a monthly visit.

Exhibit 3. Descriptive Data Report (Completed by Data Manager)

Monthly Service Numbers

	Number Served	Total Visits
Pregnant Caregivers	5	3
Female Caregivers	3	3
Children	3	3

Supervisor Tasks

A supervisor works with both the data manager and the home visiting staff to better understand and address emerging data issues. To ensure this process is consistent year to year, a supervisor should decide and document the elements the report needs to reflect, what is considered accurate, and how best to reflect data quality issues in the notes.

Helpful Tools for Data Cleaning

Missing Data Reports

Missing data reports are summaries of key data elements, also known as the numbers that make up a report. They indicate the number of cases with the element, without the element, and sometimes those with an element that is out of the range of what is expected. Missing data reports can be generated at the case file level, the home visitor level (across all his/her case files), the supervisor level, or even the overall program level. They can be used for pre-reporting data cleaning by a supervisor or peri-reporting by a data manager. Exhibit 4 provides an example of a missing data summary. The table is organized by each of the data entry tables.

Exhibit 4. Missing Data Summary

Missing Data	Number of Cases Missing
Table 1	5
Table 2	0
Table 3	6
Table 4	7

Descriptive Data Reports

These reports contain the descriptive data information for each data element in a report. This is typically the mean, median, range, and number of missing values. These are helpful for spotting values that may inaccurately be reflected in the report, such as large numbers of "outliers" or unexpected values. Descriptive data reports are typically run by data managers or evaluators as part of the peri-reporting data cleaning process. Exhibit 5 is an example of a descriptive report that can be used to monitor data quality. In the example table, unusual or inaccurate values such as a 95-year-old pregnant caregiver or a 5-year-old female caregiver are highlighted for a followup.

Exhibit 5. Table 4 Descriptive Data Summary

Table 4	Mean Age	Min/Max
Pregnant Women	21.5	13/95
Female Caregiver	22.4	5/45
Male Caregiver	27.8	21/47