Creating a Data Quality Plan

Accurately answering an evaluation question about your site depends on the availability of complete and accurate data. While it is unrealistic to expect your data to be 100% accurate, a benchmark for defining quality data depends on the circumstance. This can't be generalized, but a basic rule of thumb is that you need complete and accurate data for all data elements involved in a given calculation required for at least 80% of the participants. This is usually not a problem for an analysis involving only one data element, but data quality issues are sometimes prohibitive in analyses reflecting multiple data elements.

Given enough time and funding, an evaluator should run quality assurance tests and then ask each site to correct as much missing and incomplete data as they can until data quality reaches acceptable standards. Unfortunately, depending on timeframes this is often not possible and the evaluator may need to exclude a certain portion of the data.

Assessing Data Quality

One of the first steps in a data quality assessment should be to identify which specific data elements you will need to include in the study. For each data element, you will want to identify what values are valid. This will help you focus your data quality review, as a comprehensive data quality review might be too time-intensive. The following example shows a chart of data elements needed for evaluation questions A, B, and C.

	A	B	C	Criteria
Assessment				
Does not have multiple intakes based on the same assessment		x		
Family Stress Checklist score		x		Not null
Mother's birth date		x	x	Not null
Mother's ethnicity		x	x	Not null
Marital status			x	Not null, <>255(unknown)
Mother's age		x	x	>=12
Intake Record				
Parity (previous deliveries)		x		Not null
High-risk pregnancy		x	x	Not null and not unknown
Trimester of first pre-natal care			x	Not null
Home Visits				
Date of first home visit in Home Visit Log				Not null
Monthly Contact Logs exist for x% of participant's service period				
Birth Record				
Birth Weight			x	Not null
Gestation Age			x	Not null

Identifying Data Quality Issues

You can use a number of tools to identify data quality issues. Some examples include:

1. **QAMP**. The majority of PIMS sites have found this to be useful. If you are unfamiliar with this program, you should sign up for the QAMP webinar on the PIMS website.
2. **PIMS Quality Assurance** reports. Beginning in PIMS 6, a few quality assurance reports were added directly in PIMS in the Quality Assurance category. Also, many of the prepackaged reports highlight **missing data**.
3. **PIMS Custom Query** tool. Use this tool to explore your data with your own queries. One useful method for finding missing data is to set a criteria for a given field as "is blank". This will help you find any missing data. This tool is fairly complex, and it is recommended that you attend the **Custom Queries** webinar to learn this.
4. **Microsoft Access**. Even without substantial expertise in Microsoft Access, it is fairly easy to sort data and look for outlier data, i.e. data that falls above or below an expected range. First, sort the data in a given column in ascending order, then sort the data in descending order. It will be easy to identify some of the worst data entry errors.

Also, check with your evaluator to see what sort of tools their analysis software (e.g. SPSS) has built-in for handling data quality issues. Though you'll want to provide them with the best quality data possible, they may be able to handle some of this themselves.

Cleaning Up Incorrect or Suspect Data

If you have enough good data, the easiest way to handle suspect data is to exclude it from the analysis. However, in some cases you may want to consider cleaning some data to improve your pool of available data. This can easily become very time-consuming and should only be done as needed. You should always document your work, highlighting any decisions you make about which participants and/or sites were excluded from the analysis, and specifically noting any data that you edited. Note that data editing should only be done when you have compelling evidence of a data entry error. When in doubt, do NOT edit data. Following are some methods by which an evaluator or data quality specialist may want to consider cleaning PIMS data.

Correct Outlier Data

Once you've identified the erroneous data, you can null or (sometimes) edit the data. It's important to keep a log of all data corrections you make.

Nulling Data
Any data that obviously makes no sense should be nulled out. E.g.:

- a baby with birthweight = 0
- a participant with a parity of 99

Correcting Data
In some cases you can and should be able to justify manually modifying data. This should only be done in cases when you have clear evidence of the actual data.

Reviewing Dates

Particularly if you are studying retention or anything else involving time durations, you will want to take a closer look at dates- not just outliers, but consistency of existing dates. For example, it is not unusual to see the following pattern:

- January 2003 MCL (monthly contact log)
- October 2003 MCL
- November 2003 MCL
- December 2003 MCL
- February 2004 MCL
- March 2004 MCL
- April 2004 MCL

In the above pattern, it seems likely that the 1/03 Monthly Contact Log should have been 1/04. It is not unusual for even the most detail-oriented and attentive data entry staff to enter the incorrect year at the beginning of a new year. In order to justify correcting such a mistake, you should look for other supporting evidence- for example, maybe the participant above had an assessment in 9/03, which would support an assumption that the January 2003 Monthly Contact Log was a data entry error.

Note that this sort of review can be very time consuming. One PIMS report that might be very helpful in this assessment is **QA03: Service Level History Reconciliation**. This will not only help you identify gaps in Monthly Contact Log completion, but also help you compare home visit counts between the home visit log and the monthly contact log, if you use both.

Another useful analysis would be to look at the sequence of key engagement events. For most participants, dates should occur in the following chronological order:

1. screening
2. assessment
3. intake
4. first home visit
5. termination

You can compare these dates by creating your own queries, or by using the **Export Stats** function in the Reports menu to get a spreadsheet of all PIMS dates.

Some examples:

- An immunization date of 2/20/1004. Depending on context and actual due dates, you might be able to justify changing this to 2/20/2004.
- A mother's birthdate may be 12/7/06 and the assessment date is 12/12/06- this would indicate the mother would have been 5 days old at assessment! Lacking other evidence, it may be best to null out the birthdate.
- A home visit date appears out of range- a home visit is entered for 1/10/01, when the participant wasn't assessed until 2/12/05. In this case, you may want to look at the "created timestamp" for a possible clue.

Choosing a Data Analysis Cohort

Exclude Participants with Unusual Case Histories

You may want to exclude some unusual cases (even without data quality issues) from retention studies if not all analyses. Some examples would include:

- A transfer from another HF program outside of the evaluation area
- A participant who enrolled 3 years after target child's birth (with no evidence of incorrect birth date)
- A participant with several consecutive missing monthly contact logs before resuming service
- A participant with duplicate intakes based on the same assessment

Choose an Acceptable Date Range

For every site, you will need to look at various records to determine a viable cutoff date for that site (i.e., the last date for which they have valid data). You should first talk to the program manager at each site to get an idea for how up-to-date they think their data is. However, you should always check the data for yourself in at least a few tables. As an example, you might want to look for the most recently entered record in the assessment, intake, and home visit tables.

Problem Areas in PIMS

Following is by no means a comprehensive overview of common data quality problem areas in PIMS, but is a sample of a few.

Gravida and Parity

The PIMS manual defines **Parity** and a related data element, **Gravity**, as follows:

- *Gravida* – Enter the number of pregnancies, including current target pregnancy, the mother has had.

- *Parity* – Enter the number of completed deliveries the mother has had. This may be zero if the target pregnancy is the initial pregnancy and it has not been completed by the time this information is recorded.

In one study, 13% of participants with a target birth after the date accepted services (i.e. postnatal enrollments) had a parity equal to or greater than gravida. Due to this inconsistency, the measure of # of previous pregnancies/births was considered to be too unreliable to be included in the study.

The reason for this inconsistency may be due to a lack of understanding of these two technical terms.

Immunizations

It is not unusual for sites to have difficulty obtaining up-to-date and accurate information about immunization completion. You should work with sites to understand how regularly they get this information, and where they get it from (directly from doctors, participant's word, etc)- there is often a significant time lag between the time an immunization is completed and the program finds out about this. It may be possible to accommodate this time lag by moving up your cutoff date for any immunization studies.

Immunization Schedules

While sites may have different immunization schedules in place, this doesn't have to be problematic if you had expected all sites in an aggregated data set to follow the same schedule. You may have to avoid using the PIMS reports (which look to each site's schedules which may or may not be correct) and conduct any calculations relating to immunization schedules using a single universal immunization schedule table and comparing this to each child's birth date and immunization completion dates.

Race Category

As Hispanic was removed as a "race" in the national census, it can be confusing to conduct any analysis referencing Hispanic as a race. You may need to look at a combination of race category and race subcategory to make this determination.

Service Level Definitions

Data analysis for aggregated data can be very confusing if sites don't have the same service level definitions. Service level definitions are mapped to a service_level_code in PIMS, but it is important to note that PIMS automatically assigns service_level_codes and this number has no inherent meaning. For example, for one site Level I might be service_level_code = 1 and for another site service_level_code = 3. An evaluator would have to create an additional cross-reference table, and use this to map the service level codes for each site to a universal code.

Such a table might look like the following:

Site ID	Service Level Name	Local Service Level Code	Universal Service Level Code
MO002	Level I	1	1
MO002	Level II	3	2
MO002	Level III	2	3
MO003	Level 1	1	1
MO003	Level 2	2	2
MO003	Level UE	3	4

Next Steps for Improving this Document

- Include notes from 2004 HFA Implementation Study
- List reading material related to data quality checks
- Include examples for QAMP, Custom Reports, data sorting, creating service level matrix, etc.