**Processing Data from Sites**

**IeDEA East Africa**

Following is a description of the process of getting data from the sites and generating master data sets.

**Master Data Sets** consist of two SAS data sets: one containing the cross-sectional data and one containing the longitudinal/visit data. For ease of processing and use, the sites that employ the same data collection and storage methods are grouped together in a single set of master data sets. Such sites are those within AMPATH, FACES, MTCTplus sites and all of the Tanzanian sites. For these programs, the master data sets contain a variable to indicate from which site the patient received services.

1. Formal request for data along with the SOP for Data Quality and Transfer is sent to the site principal investigators and the site data manager. Ideally data should be submitted within 45 days of the request.
2. Data and documentation from each site/program are reviewed to assess completeness of the minimum data set variables and of the visit level information supplied.
3. Comparisons to previous submissions are made.
4. Overall questions about the data files submitted are sent to the site data manager. More often than not, some or all of the files need to be modified and resubmitted.
5. Harmonize data with the other sites. This involves:
	1. renaming variables (ex. gender to male)
	2. reformatting variables (ex. converting dates supplied in text format to numeric date format)
	3. recoding responses so that categories are consistent across sites
	4. derivation of new variables from data supplied (ex. specific variable for ART initiation date)
	5. labeling variables for clarity
	6. formatting responses for consistency across sites
	7. making decisions about how to harmonize, when to collapse categories, and knowing when different terminology has the same meaning
6. Once the master data sets are complete, patient-level queries are generated and sent to the sites. Detailed description of such queries can be found in the SOP for Data Quality and Transfer. Where possible, results of these queries are incorporated into the master data sets.
7. Basic summary statistics are generated and posted to the website.
8. Specific analysis data sets are generated from the master data sets for individual concept proposals depending on site participation and patient inclusion/exclusion criteria.