

Streamlining the Integration of SDTM Datasets from Multiple Studies: A Macro Approach

Presented by:

Santosh Reddy Lekkala

Vijay Anuroop Reddy Allam

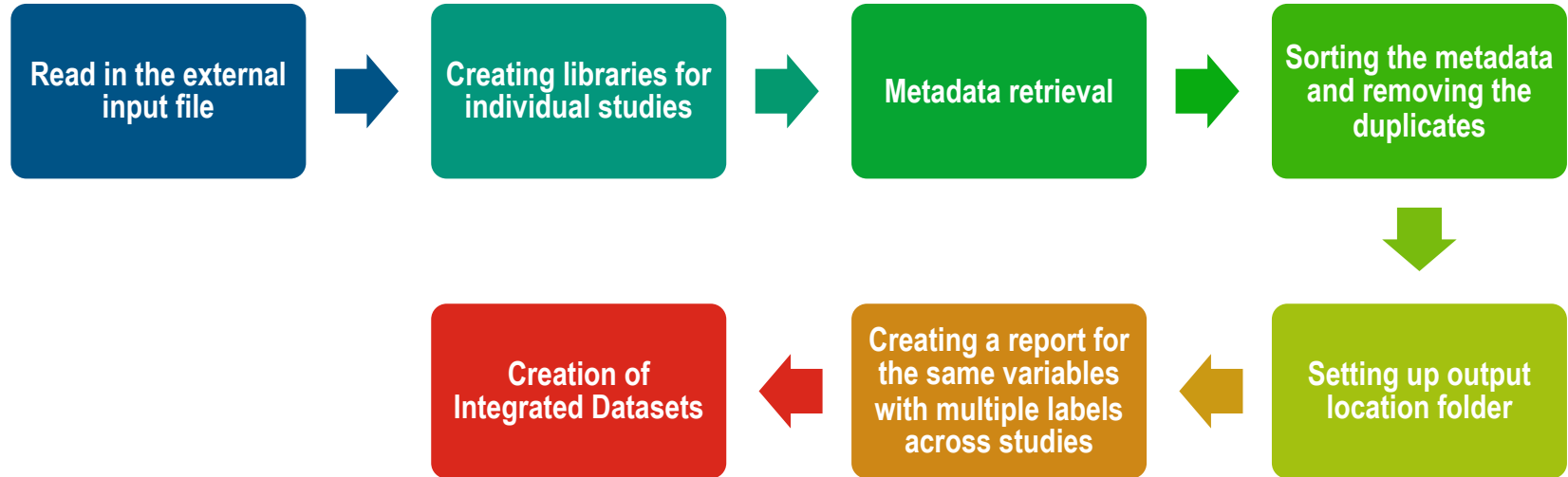
Introduction:

- Building integrated datasets by pooling data from multiple studies is a complex and time-consuming task which is either done at Study Data Tabulation Model (SDTM) or Analysis Data Model (ADaM) stage.
- In this macro we describe a macro designed to streamline the data integration process address common challenges encountered during dataset integration, such as variables with different lengths, missing data and variable label discrepancies.

Advantages:

1. The macro not only performs integration but also serves as an edit check tool when combining SDTM/ ADaM datasets from multiple sources. It generates a report for variables having different labels across studies so that they can be investigated and updated as required.
2. It saves a considerable amount of time, as it auto-adjusts the length of the same variables with different lengths across multiple datasets.
3. The macro can lessen the workload for Programmers by automating tasks and improving data quality, ultimately leading to cost savings

Process Flow:



Read in the External Input File:

The Input for the Macro is an excel file that contains the global macro variables that would be used in the program. This file contains 3 main input columns

Type: The Type column is used to identify if the row relates to Input or Output location.

StudyID: Used to specify the StudyID associated with the location. For this macro the length of the StudyID cannot be more than 7. StudyID column would be blank when type = “Output Location”.

Location: Specifies the location where the datasets to be integrated are available.

Input File Screenshot:

Type	StudyID	Location
Output Location		X:\Path\Output
Input Location	1	X:\Study datasets path
Input Location	2	X:\Study datasets path
Input Location	3	X:\Study datasets path
Input Location	4	X:\Study datasets path

Creating libraries for individual studies:

- After reading in the external file, the macro utilizes SQL queries to extract essential information from the excel sheet and uses SQL queries which serve multiple purposes such as counting the unique number of Studyid's (&N), Extracting Studyid's (&LIB1 - &LIB&N.) and extracting corresponding data locations (&PATH1 - &PATH&N.).
- This information is used to create libraries for individual studies specified in the Excel sheet.

Metadata retrieval:

- The macro retrieves metadata for datasets in each study library. This is achieved using the PROC DATASETS procedure, which provides a detailed overview of the structure and attributes of datasets within the specified libraries.
- The metadata from all the studies is combined in a single dataset and is used in subsequent processing steps.

Sorting the metadata and removing the duplicates:

- The combined metadata obtained from all the studies in the above step is sorted by domain, variable, and Max length for each variable. Duplicates are identified and removed, retaining the maximum variable length. The resulting dataset serves as a clean representation of the metadata.
- The dataset is then sorted by domain and Varnum so that the order of variables remains similar. The resulting dataset contains attributes to include the variable name, label, length, and data type.
- The macro utilizes the CALL EXECUTE routine to dynamically generate data step code for creating individual SDTM data shells based on the metadata. These shells would be later used to append the SDTM datasets from individual studies.

Setting up output location folder:

- To set up the output folder the Macro reads in the location from the Excel file where TYPE = “Output Location” and the program run date is used to create a folder in the provided location.
- The folder would be named in YYYYMMDD format based on the date. Following this step 3 other subfolders are created in the dated folder using the LIBNAME statements –
- INTERIM - to store intermediate datasets
- INTDATA - location to store the final integrated datasets
- INTCSV - location to save the integrated CSV files

Creating a report for the same variables with multiple labels across studies:

- In the step above – “Sorting the metadata and removing the duplicates” when sorted by domain and variable names either in SDTM or ADAM we see that there might be differences in the labels used from one study to another. This might be caused by several factors such as different I.G versions being used, label errors, etc., and needing to be corrected.
- A report is generated for variables with multiple labels using the PROC EXPORT to generate an Excel file. The resulting Excel file, named "Variables with Multiple Labels.xlsx," provides the list of domains and variable labels that needs to be investigated.

Creation of Integrated Datasets:

- In the final step the datasets from individual studies are got into the work library with “_studyid” added as a suffix. The datasets are then set with the shells that were created earlier to get the final Integrated datasets and CSV files.

Conclusion

By using this approach, we could create integrated datasets which would be:

- Readily available when we receive multiple new deliverables for the studies to be integrated
- Use limited resources
- Save time

Thank You

