Performing Statistical Analysis on 'Big' Project Dataset to Mine Associations and Identify Data Issues on an Ongoing Basis

Kantish Chowdhury, Md. Rajwanur Rahman, Shafi Consultancy Bangladesh
Shafi Chowdhury, Shafi Consultancy Limited

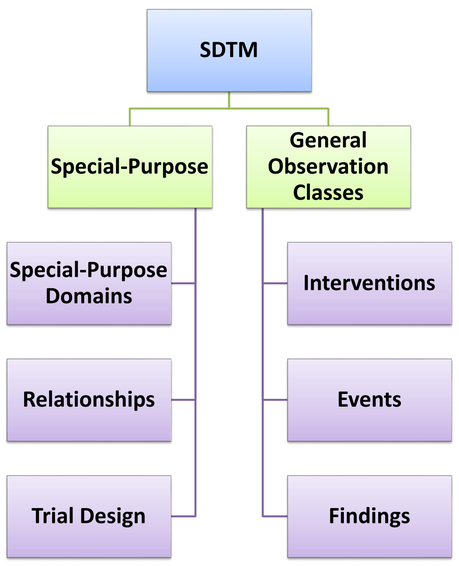
1. Introduction

The use of standard SDTM structured data has provided an opportunity to generate one 'Big' dataset by merging different SDTM datasets together. It is possible to merge all related data together by defining the key identifying variables within those datasets. The advantage of doing this is that the analysis can be performed without further data manipulation to investigate how all the data relate to each other.

Statistical analysis can be performed automatically to check associations of all types of data based on age, sex, race, country and all other covariates specified within a project. Any associations can be highlighted, and potential data quality or trends in the data can be explored on an ongoing basis. Graphical tools can also be used to see how values are changing over time and if there are unexpected events in the target population. Recruitment rates, inclusion exclusion failures, randomization, protocol violations, early terminations and lost to follow up, serious adverse events, prohibited medications, outcomes and endpoint issues can all be analyzed on an ongoing basis to see if there are any relationships that should be explored. Any unexpected results or findings can then be investigated and actions can be taken if required to help raise the quality of the data as soon as any issues are found.

2. The simple 'Big' structure

3. SDTM Domain Overview



4. 'Big' Dataset Structure

Record types for 'Big' dataset.

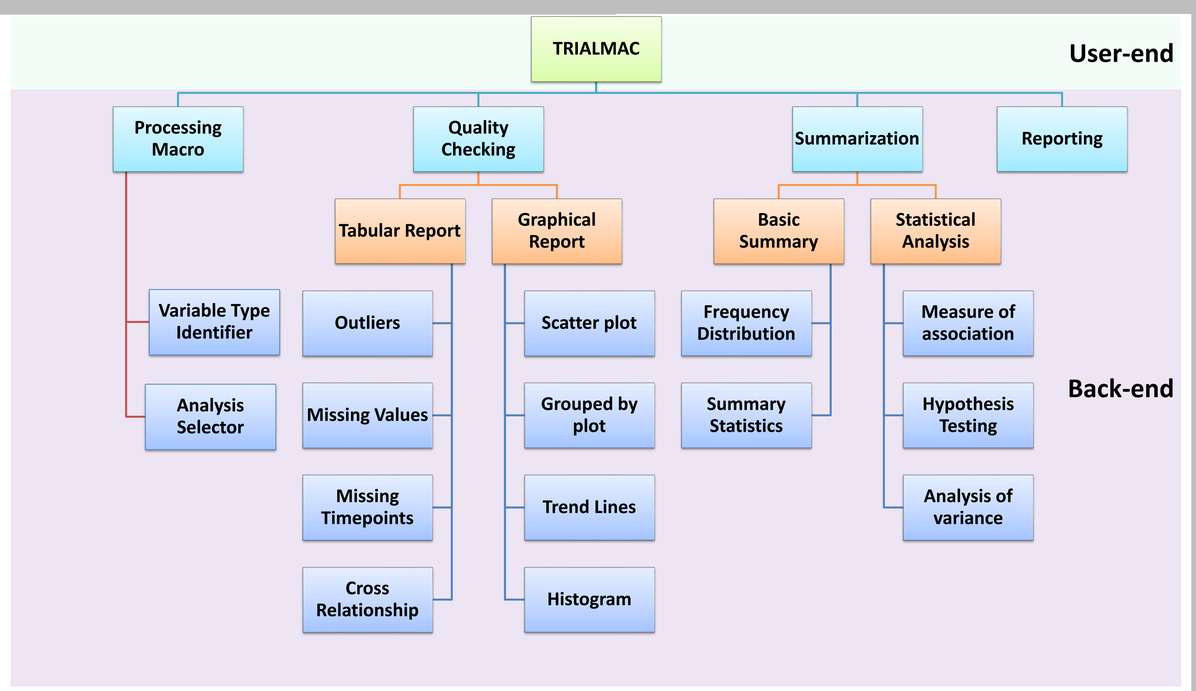
- Record type 1: Date/Time of Informed consent is collected from DM/DS.
- Record type 2: Observational findings from laboratory (LB) and vital signs (VS) domains. In some studies, these may occur before the screening period.
- Record type 3: Demographic records collected during screening period from DM domain.
- Record type 4: Randomization records of randomization period from DM/SE/DS domains.
- Record type 5: intervention records collected throughout the study exposure period, from EX domain.
- Record type 6: Events occurred during/after the intervention from AE and DS domains.
- Record type 7: Post study records (if available) from SE domain.

5. Why use 'Big' Dataset?

'Big' dataset contains all of the observational data in a study from the beginning to the end. All observational records are merged together by corresponding key variables and sorted by date/time variables along with other key variables.

All SDTM domain information can be easily found in one place. Special-purpose domains information (e.g. demography, subject visits, subject elements), intervention taken (e.g. exposures), baseline and related findings (e.g. vital sign and laboratory results) before and/or during and/or after intervention taken, adverse and dispositional events of the subjects in a clinical study are all present. It is even possible to find where the subjects discontinued for a trial, related endpoints, outcomes, recruitment rates, randomization, early terminations and lost to follow up, serious adverse events, prohibited medications and further discrepancies.

A Statistician may need to see all related information with Laboratory results to ensure there were no issues with collected data. In this scenario, the Statistician can use the 'Big' dataset to summarize the data. They can also review the relationship with other variables and assess any patterns they may come across.



A user can choose to go fully automatic to get quality checks and summary results for pre-defined variables, or they can choose their own variables and report configuration. In the back-end, the TRIALMAC system consists of several different types of modular macros.

There are four types of modular macros: processing macro, quality checking macro, summarization macro and reporting macro. Processing macros have two separate modules, one is for variable type identifier and another is automatically selecting appropriate analysis based on variable types. Variable type identifier macro uses the information from SDTM variable names and assigns pre-defined attributes, which are used in analysis selector macro to select appropriate analysis.

'BIG' dataset Basic Structure (DM, SE, LB VS, EX, DS, AE)

DM, SE - Special Purpose; LB, VS - Findings; EX - Intervention; AE,DS – Events

	Observational Records	Study Period
DM/DS	Record Type-1	Start of Study (IC Occurs)
Findings (LB,VS) records before screening	Record Type-2 (Multiple records)	
DM	Record Type-3	Screening Period
Findings (LB,VS)	Record Type-2 (Multiple records)	
DM/SE/DS	Record Type-4	Randomisation Period
Findings (LB,VS)	Record Type-2 (Multiple records)	Pre-Treatment Period
EX	Record Type-5	Treatment Period (1)
Findings (LB,VS)	Record Type-2 (Multiple records)	
Events (AE, DS)	Record Type-6 (Multiple records)	
EX	Record Type-5	
Findings (LB,VS)	Record Type-2 (Multiple records)	Rest Period
Events (AE, DS)	Record Type-6 (Multiple records)	
EX	Record Type-5	Treatment Period (n)
Findings (LB,VS)	Record Type-2 (Multiple records)	
Events (AE, DS)	Record Type-6 (Multiple records)	
EX	Record Type-5	
Findings (LB,VS)	Record Type-2 (Multiple records)	Follow-up Period
Events (AE, DS)	Record Type-6 (Multiple records)	
DM/SE	Record Type-4	End of Study
Findings (LB,VS)	Record Type-2 (Multiple records)	End of Study and Follow-up
Events (AE, DS)	Record Type-6 (Multiple records)	
Post Study	Record Type-7	Post Study

Statistical analysis will be more convenient with respect to 'Big' dataset from SDTM domains rather than having to use the independent SDTM domains and prepare them to create individual analysis datasets in the later stage of a clinical study.

6. Continuous Quality checking with 'Big' dataset

To implement data quality checks and statistical analysis, a system was designed to reduce the user workload, increase the consistency and efficiency, and to remove the need for additional programming when performing complicated data checks or generating a basic summary.

There are two types of quality checking modular macros available with TRIALMAC. One is tabular type and another is graphical type. Tabular type macro provides outputs as a traditional report while graphical type macros generate different charts for the provided variables from the 'Big' dataset.

To identify data issues on an ongoing basis, traditional methods are used. Four different types of macros are used to identify outliers, missing values, missing timepoints and check the cross relationship between variables. Different types of macros are used to produce graphical output from the 'Big' dataset. Graphical outputs include scatter plot, grouped by plot, trend lines and histograms.

Statisticians and programmers can get basic summary from summarization macros, which includes frequency distribution and summary statistics. While for an advanced statistical summary, they can rely on the measure of association, hypothesis testing (mean/ median test), and the analysis of variance results.

A reporting macro is used to collate all output produced by the quality checking and summary macros to generate jargon-free and reusable reports.

7. What type of data quality checks and statistical analysis may be done with 'Big' Dataset?

There is an endless possibility of using the 'Big' dataset for quality checks as all of the domains' information are combined here. Following are a few of the basic checks which can be done on 'Big' dataset:

- Identify extreme values for different variables, especially demographic and findings variables.
- Missing value check for relevant variables.
- Potential data quality or trends checks before database lock.
- Identify endpoint issues.
- Find issues with timepoints missing.
- Recognize unexpected results or findings for further investigation.

8. Conclusion

The 'Big' dataset can be created on an ongoing basis with minimal effort, which then allows the study team to quickly perform data quality checks, ongoing summary and continuous issue tracking without wasting important resource for programming and generating outputs one by one. As 'Big' dataset is generated from structured SDTM datasets, it can always be updated without any additional programming when the source data is updated.

'Big' dataset from SDTM can give the statistician and programmer a painless gateway from traditional quality checks and ongoing summary. Allowing automation of analysis to lead the way in improving QUALITY and saving time and resource.

Please feel free to contact us for further information.

www.shaficonsultancy.com