Paper SI05

# Digitization of Specifications - Brave new world for Analysis Definitions

Becky Bates, Saama Technologies, USA
Krishna Gopal, Saama Technologies, USA

## ABSTRACT

Spreadsheets have been the industry standard for capturing SDTM/ADaM specifications. They are used with varying levels of success to partially automate programming. Analysis specifications are still primarily written in Word documents, which are read and interpreted by a human who in turn develops programs to produce the analysis based on their understanding of the written instructions.

With Saama's knowledge-graph based approach to digitization of specifications, analysis definitions are now designed as the foundation standardization and automation of statistical analyses. These digitized specifications dovetail nicely with the Define Analysis Results Metadata (ARM), and will enable the development of robust predictive analytics for specification and code development.

Digitization of analysis definition specifications has the potential to transform analysis and reporting workflows. Key benefits include reducing manual effort, promoting standardization, shortening programming timelines, and improving end-to-end traceability and auditability in order to enable the next generation of clinical trials.

## INTRODUCTION

Analysis specifications are still primarily written in Word documents, which are read and interpreted by a human who in turn develops programs to produce the analysis based on their understanding of the written instructions.

Digitization of analysis specifications is key in the automation of analysis programming as well as the creation of analysis results metadata. Using knowledge graphs greatly reduces the myriad of manually managed documents and spreadsheets currently used to generate the required outputs for clinical trials and greatly increases the transparency and traceability.

To make digitization of analysis specifications even more powerful is the need for standardization in such a way that the analysis specifications can be used to drive autocode generation for the analyses. This paper demonstrates mechanisms on how analysis specifications can be designed so that even complex statistical analyses can be generated using autocode generation.
Many of the concepts presented in this paper do align with CDISC Analysis Results Model (ARM) but this paper is not explaining the model. This paper presents concepts on using standards-compliant analysis results metadata, in a manner where digital specifications drive results.

## DEFINITION OF A SINGLE ANALYSIS

Before any analysis definition specifications can be defined, let alone digitized and standardized, a common definition of what constitutes a single analysis will help simplify the standards. A good example is a summary of treatment emergent adverse events (TEAE) by System Organ Class (SOC) and preferred term (PT) table. How many analyses are included in this one table? One definition could define all results presented in the table as constituting one analysis. Another definition could define each number presented in the table as constituting one analysis. A more practical definition, which will be assumed for this paper, is to define a single analysis as

- using the same dataset or same set of datasets
- having one variable being analyzed
- including the same set of individual observations (i.e. using the same selection criteria)

Another assumption in this paper is that a single analysis uses only one dataset with the exception of when the dataset cannot provide the correct N size for a given analysis (e.g. ADAE requires ADSL to determine the denominator for incidence rates). This assumption supports the principle of analysis datasets being “one proc away”.

Therefore, a Treatment Emergent Adverse Event summary table would be comprised of 4 analyses:

- number of subjects that were at risk of getting a TEAE by treatment group:
    - dataset = ADSL
    - analysis variable = ADSL.SAFFL
    - selection criteria: ADSL.SAFFL = Y
- Number and percent of subjects that experienced at least one TEAE by treatment group:
    - datasets= ADAE and ADSL (supports determination of the denominator)
    - analysis variable = ADAE.AOCCFL
    - selection criteria: ADAE.TRTEMFL = Y and ADSL.SAFFL = Y
- Number and percent of subjects that experienced at least one PT within each SOC by treatment group:
    - dataset = ADAE and ADSL (supports determination of the denominator)
    - analysis variable = ADAE.AOCCSFL
    - selection criteria: ADAE.TRTEMFL = Y and ADSL.SAFFL = Y
- Number and percent of subjects that experienced at least one PT by treatment group:
    - dataset = ADAE and ADSL (supports determination of the denominator)
    - analysis variable = ADAE.AOCCPFL
    - selection criteria: ADAE.TRTEMFL = Y  and ADSL.SAFFL = Y

## ORGANIZATION OF A SINGLE ANALYSIS DEFINITION SPECIFICATION

A single analysis specification can be divided into three sections:

1. Analysis Description which specifies
    - Names and labels of the analyses
    - High level description
    - Documentation for more detailed information if needed
2. Data Requirements which specifies
    - dataset(s)
    - variables needed for the analysis
    - How, or roles, the variables are to be used in the analysis.

3. Statistical Method Requirements which specifies
    - the statistical method being used
    - the statistical parameters needed to conduct inferential analysis (when applicable),
    - statistical tests to be performed (when applicable),
    - the statistics to be generated from the analysis.

### ANALYSIS DESCRIPTION

The Analysis Description is meant to provide the user a high level understanding of the analysis being defined. The following parameters can be used to provide unique identification of each analysis and provide contextual information as well.

**ANALYSIS DESCRIPTION PARAMETERS**

| Attribute | Description |
|---|---|
| Analysis Name | Unique name for analysis for a given reporting event |
| Analysis Label | Label to provide some insight to the name |
| Analysis Description | Description of analysis |

| Attribute | Description |
|---|---|
| Analysis Endpoint | Endpoint that is being analyzed.<br>This is to support the Define ARM specification needs |
| Reason for analysis | Why the analysis is being performed.<br>Example of reasons is Primary objective, regulatory reporting requirement. This is to support the Define ARM specification needs |
| Document | Document to either provide the user with more information or could be used for review/QC purposes. Ultimate goal is to have the digitized specifications to be single source and use as the source for other documents. |

The parameters mentioned above can be found in the CDISC Analysis Results Model as well.

**DATA REQUIREMENTS**

The dataset(s) and variables that are used and how they are used (i.e. roles), uniquely defines a single analysis once coupled with a statistical method.

Parameterizing datasets and variables into codified analysis roles enables the digitization process. The following are key analysis roles.

**ANALYSIS ROLES**

| Analysis Role | Description |
|---|---|
| Primary Analysis Dataset | This dataset contains the analysis variable and all other variables needed for the analysis with the following exceptions<br>a) The full compliment of subjects is not contained within the dataset. Example being ADAE.<br>b) Common variables that may not be permanently stored on the primary analysis dataset. Examples: SEX, AGE, and RACE. |
| Supporting Analysis Dataset | This dataset is used when the full complement of subjects are not included in the primary analysis. Or common variables are not included in the primary analysis dataset. Typically ADSL would be the supporting analysis dataset.<br>However, if ADSL is truly the only logical choice, this parameter may not be needed. |
| Analysis Variable | The variable that is being summarized or analyzed. For example:<br>● SEX and AGE would be analysis variables for a demographic table.<br>● AVAL would be for analysis using Basic Data Structure (BDS) datasets.<br>● AOCC flag variables for Occurrence datasets. |
| Comparator Variable | The variable that is used to define the groups to be compared. Typically these would be TRTA, TRTP, or ARM. |
| Population Variable | The variable that is used to select the appropriate subjects within a population being analyzed |
| Group By Variables | These variables are used to perform the same ‘analysis by’ variable levels. Typically these would be PARAMCD, AVISIT, or ATPT. |

| Analysis Role | Description |
|---|---|
| Selection Variables | These variables are used to select the appropriate observations within the analysis dataset for a given analysis. Typically this would be ANLxFL, CRITyFL, PARAMCD when only specific parameters are to be included in the analysis. |

Some of the roles mentioned above can be found in the CDISC Analysis Results Model, such as the primary analysis variable, data references (subset criteria- Population and selection variables) and parameters. The other elements are part of the “Documentation” and “Programming Notes” section of the current ARM documentation standard (CDISC ARM v1.0 for Define-XML v2.0). Parametrizing these elements allows for Automated generation of the description section.

**STATISTICAL METHOD REQUIREMENTS**

The statistical method requirements are separated from the dataset requirements so the same statistical method can be associated with more than one analysis specification. An example of this is change from baseline. The same exact statistical methods can be applied to labs, vitals, and ECGs. The only difference is the datasets and variables being associated with the statistical methods.

To make a diverse range of statistical methods more amenable to digitization, the statistical methods are divided into following three parts:

- Statistical Method Parameters: parameters needed to define a statistical method
- Statistical Test Parameters: statistical test(s) to be performed and all the parameters associated with those tests
- Statistical Output Parameters: specific list of statistics needed to be generated from the statistical analysis

The parameterization presented below is programming language  agnostic. For the purpose of illustration, the following SAS Mixed Procedure® pseudo-code is used to demonstrate the concepts above.

```
proc mixed method=TYPE2;
class TRTP;
model aval = AGE TRTP;
repeated / type=UN subject=USUBJID;
lsmeans TRTP/ alpha=.05 adjust=DUNNETT CL PDIFF;
run;
```

**STATISTICAL METHOD PARAMETERS** (see red text in sample SAS Mixed Procedure code)

| Statistical Method Parameter | Statistical Method Parameter Value |
|---|---|
| Method | TYPE2 |
| Covariate | AGE |
| Fixed Effect | TRTP |
| Covariance Structure | Unstructured |
| Repeated Variables (subject=) | USUBJID |

This list of parameters defines the statistical methodology that is to be used for this method. The list of statistical method parameters and their values can be changed to accommodate any statistical methodology.

Another important note is that the parameter values that are pointing at variables are not dataset dependent variables. AGE and TRTP can exist on multiple datasets. This allows the statistical methods requirements to be applied to multiple analysis definition specifications with the caveat that the primary or supporting analysis datasets associated within the analysis definition must contain these variables.

**STATISTICAL TEST PARAMETERS** (see purple text in sample SAS proc mixed code)

| Statistical Test | Apha Level | Sided-ness | Adjustment |
|---|---|---|---|
| T-test | 0.05 | 2-sided | Dunnett |

For this specific example, the LSMEANS statement performs the t-test and there is no other option. To remain code agnostic and to prospectively define the statistical method, the specification still needs to identify what statistical test(s) is (are) to be conducted. Additionally, this parameterization also allows multiple statistical tests to be performed if needed.

**STATISTICAL OUTPUTS** (see orange text in sample SAS proc mixed code)

| Statistical Output Parameters |
|---|
| Dunnett p-values |
| LS Means |
| 95% Lower Confidence Limit for LS Means |
| 95% Upper Confidence Limit for LS Means |
| LS Mean Difference |
| 95% Lower Confidence Limit for LS Mean Differences |
| 95% Upper Confidence Limit for LS Mean Differences |

The list of statistical outputs are endless. However, at least one statistical output is required per statistical method.

## CONCLUSION

The key to digitizing specifications is dividing the specifications into logical sets of information and then parameterizing the information within each set. Once digitized, the analysis definitions and statistical method specifications can be ingested systematically to auto generate code that produces the very outputs that the specifications are defining.

To further the impact of digitization is the standardization of not just the organizational aspect but the controlled terminology. The data requirements discussed in this paper already have a strong set of well-defined standards and controlled terminology through the use of CDISC defined - SDTM, ADaM, and Define standards.
The statistical methods are just at the beginning of that journey. The suggested parameterization of the statistical methods presented in this paper lends itself very readily to standardization from both the organizational and controlled terminology aspects.

Standardizing digitized analysis definitions and statistical method specifications using knowledge graphs even further enhances the usability, traceability, and transparency, which are the foundation principles that the current data standards are built on. Using such digitally connected information leads to a single source of truth driving both documentation and programming activities. Now is the time to build equivalent strong standards for analysis definitions and statistical method specifications.

## CONTACT INFORMATION

Author Name: Becky Bates
Company: Saama Technologies

Address: 900 E Hamilton Ave Suite 430, Campbell, CA 95008
Work Phone: 321-720-3838
Email: rebecca.bates@gmail.com
Website: https://www.saama.com

Co-Author Name: Krishna Gopal
Company: Saama Technologies
Address: 900 E Hamilton Ave Suite 430, Campbell, CA 95008
Work Phone: 919-636-3001
Email: krishna.gopal@saama.com
Website: https://www.saama.com