

How Metadata can help drive the creation of Tables, Listings and Figures

Judith Goud, Nurocor, The Netherlands

ABSTRACT

For Regulatory submissions as well as internal data review, it is important to provide consistent representations of the collected data. Many companies have developed Standard Tables, Listings and Figures, and PhUSE also provides Shells that can be used to represent the data.

These Standardized outputs can be generated via pre-defined programs and macros. Two potential approaches for the generation of Listings and Tables are:

1. Variable driven macros utilizing Proc Print, Report or Tabulate
2. Layout instruction and algorithms can be used to create output in a cell by cell fashion

With organizations using a metadata repository (MDR) to manage their standards, this MDR can also be used to manage the content required to generate TFLs, providing full data traceability.

This paper will explore these two methods of Metadata usage in the creation of Tables and Listings. You can then extrapolate from these models with respect to Figures.

INTRODUCTION

To present data collected in clinical trials and their analysis, biopharmaceutical companies generally use tables, listing and figures (TFLs). These representations are based on the Statistical Analysis Plan (SAP) and show important summaries, trends or features of the collected data used to support the conclusions in the Clinical Study Report.

Several industry organizations like the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) and the Pharmaceutical Users Software Exchange (PhUSE) have provided guidance around which data should be represented and the layout of the representations. Regulatory authorities like European Medicines Agency (EMA)¹, the US Food and Drug Association (FDA)², and Pharmaceutical and Medical Device Agency (PMDA)³ in Japan have adopted the ICH guidance and reference PhUSE guidelines.

The ICH E3 guideline 'Structure and Content of Clinical Study Reports'⁴ provides rules for the use of tables, listings, and figures and defines a set of tables and listings that should be part of any Clinical Study Report.

For example, chapter 11 discusses information to be provided regarding analysis populations including specific demographic, disease, and other relevant variables and chapter 12 describes requirements for reporting safety information. Chapter 14 lists the Tables, Figures, and Graphs as referenced throughout the document and paragraph 16.2 includes an overview of patient data listings expected in the Clinical Study Report.

14. TABLES, FIGURES AND GRAPHS REFERRED TO BUT NOT INCLUDED IN THE TEXT

Figures should be used to visually summarize the important results, or to clarify results that are not easily understood from tables.

Important demographic, efficacy and safety data should be presented in summary figures or tables in the text of the report. However, if these become obtrusive because of size or number they should be presented here, cross-referenced to the text, along with supportive, or additional, figures, tables or listings.

The following information may be presented in this section of the core clinical study report:

14.1 Demographic Data Summary figures and tables

14.2 Efficacy Data Summary figures and tables

14.3 Safety Data Summary figures and tables

14.3.1 Displays of Adverse Events

14.3.2 Listings of Deaths, Other Serious and Significant Adverse Events

14.3.3 Narratives of Deaths, Other Serious and Certain Other Significant Adverse Events

14.3.4 Abnormal Laboratory Value Listing (Each Patient)

16.2 Patient Data Listings

16.2.1 Discontinued patients.

16.2.2 Protocol deviations.

16.2.3 Patients excluded from the efficacy analysis.

16.2.4 Demographic data.

16.2.5 Compliance and/or drug concentration data (if available).

16.2.6 Individual efficacy response data.

16.2.7 Adverse event listings (each patient).

16.2.8 Listing of individual laboratory measurements by patient, when required by regulatory authorities.

Figure 1. Excerpts of FDA Guideline for industry/ICH E3: Structure and Content of Clinical Study Reports

PhUSE EU Connect 2019

Based on guidance documents provided by ICH, FDA, PMDA, EMA and the Clinical Data Interchange Standards Consortium (CDISC), and industry best practices, a PhUSE CSS working group has developed further guidance documents regarding common analysis and datasets including shells for tables and listings⁵. These shells developed by PhUSE are being used by an increasing number of biopharmaceutical organizations.

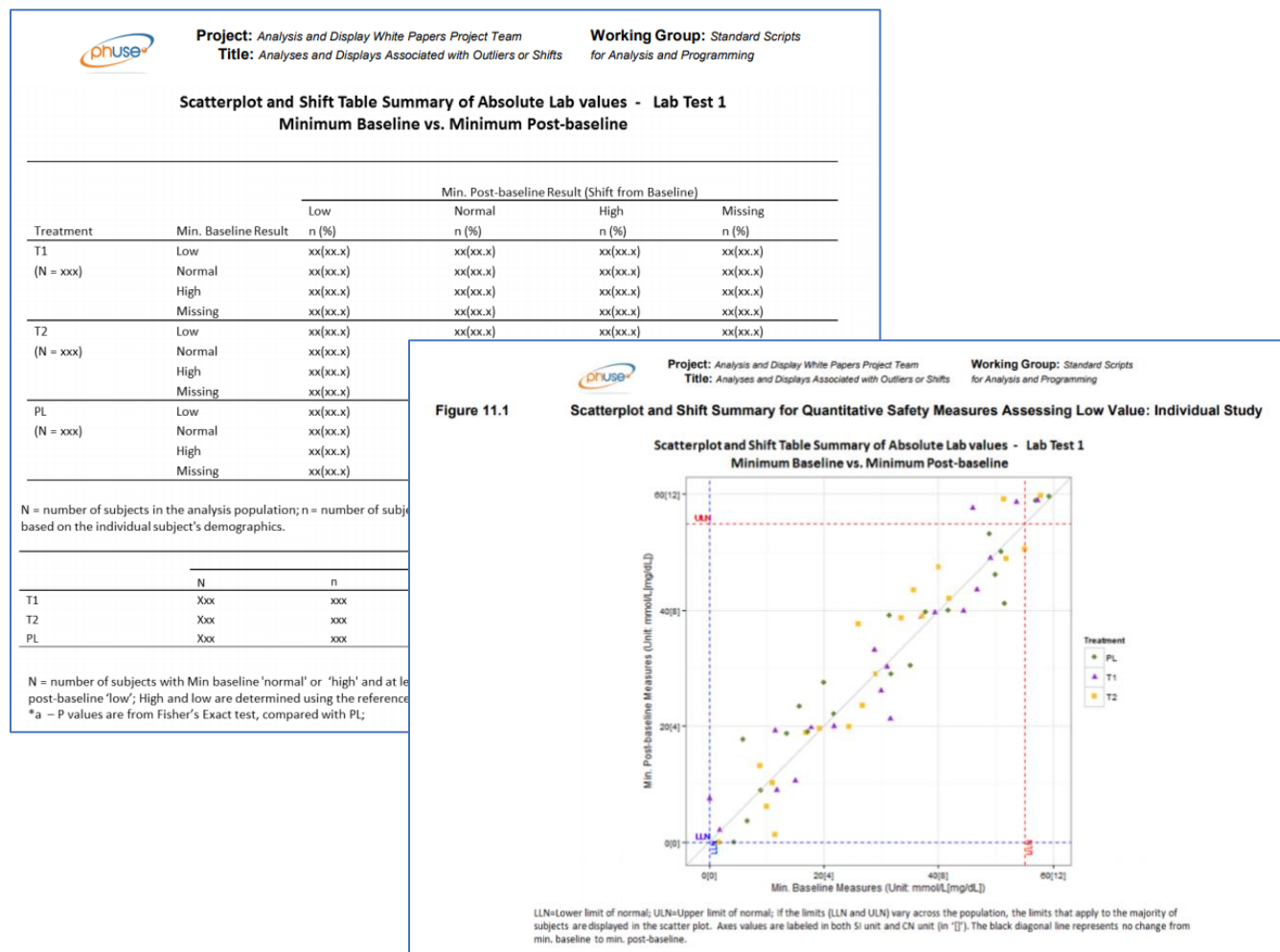


Figure 2. Example Table & Figure Shells from 'Analyses and Displays Associated with Outliers or Shifts from Normal to Abnormal' white paper

CURRENT PROCESS EVALUATION

Whether or not using the PhUSE shells, many biopharmaceutical organizations have developed Standard Tables, Listings, and Figures to represent the data for their clinical studies with associated standard programs and processes. However, the process to produce these TFLs is not as standardized as the output.

The following questions were used to analyze the current processes to produce TFLs practiced within biopharmaceutical organizations:

- Does the organization maintain a repository of TFL specific programs?
- Does the organization support a Macro library?
- Do the Macros cover all TFL type outputs?
- Is any "custom code" required for any of the outputs?
- How are Titles and Footnotes applied? Is it parameter driven, a lookup table, defined within the program(s)?

The analysis revealed there are many approaches to delivering Tables, Listings and Figures (TFLs). The most commonly used methods are (a combination of):

- Head's down programming, utilizing old code from past projects.
- Maintaining a library of programs dedicated to each of the table outputs.
- Macro libraries that support different table styles is another approach to producing Standardized outputs.

These most common processes are highly manual and require substantial interpretation of documents and datasets. Typically, the process starts with a Statistical Programmer reading the SAP, reviewing source data and any TFL standards. This programmer then develops programs or looks for programs that can be reused from previous studies, applies changes to functions, procedures, & controlled terminology, and if Macros are used, assigns TFL parameter values.

PhUSE EU Connect 2019

The next step is the development and testing of algorithms and procedures, e.g. SQL – Merge – Macro – etc. Finally, the Statistical Programmer self-tests the produced output before handing off to a second Statistical Programmer for quality control. The 2nd Statistical Programmer checks the results by performing double programming or other QC steps. This process then leads to validated, or approved, output. Figure 3 shows a representation of this process.

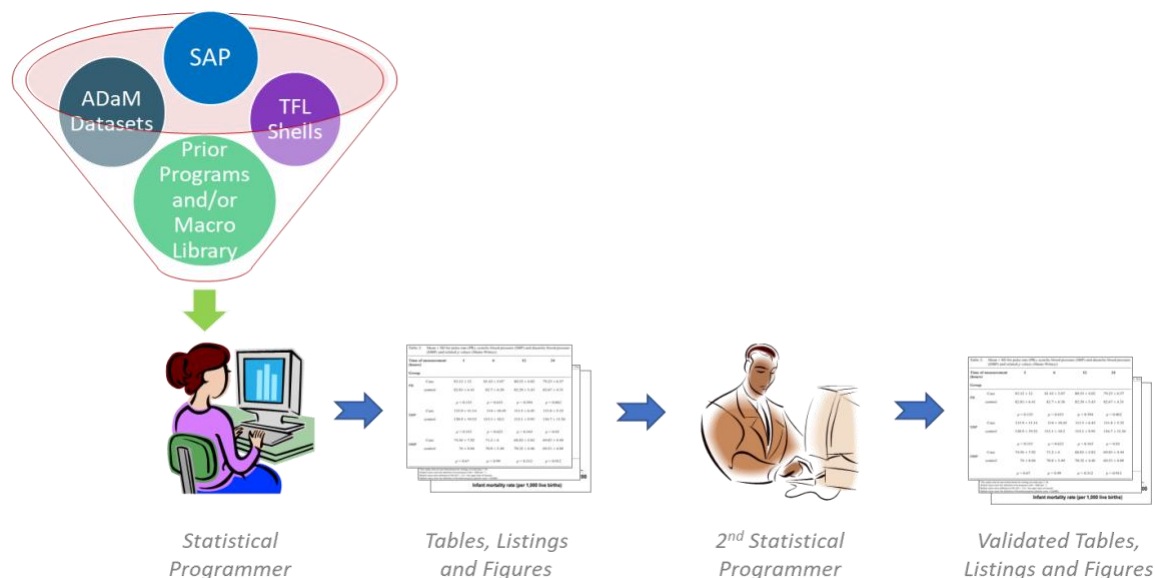


Figure 3. Current TFL generation process using QC/Double programming

The information used in the current process may reside in different systems and locations like documentation systems, file drives, macro libraries, in the statistical computing environment, etc.

The common current process is labor intensive and error-prone (hence the double programming requirement). Organizations have attempted to improve the efficiency by creating and maintaining libraries of programs and/or macros that can be reused between studies. These libraries have provided improved output quality and process efficiency, but they still need to be manually added for each study.

UPDATED PROCESS USING SUPPORTING METADATA

Using metadata to drive the TFL generation process provides further quality and efficiency improvements, data traceability, and interoperability.

If this metadata is managed in a central metadata repository, and program generation is applied, the process becomes much more automated (see figure 4)

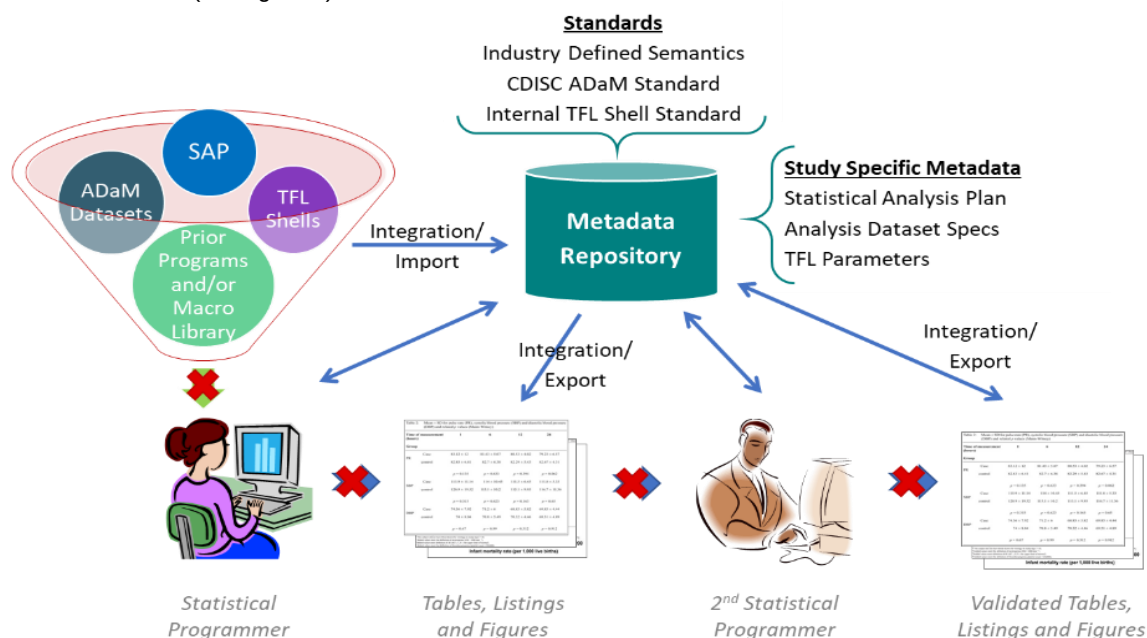


Figure 4. Future TFL generation process using a metadata repository

PhUSE EU Connect 2019

Instead of the Statistical Programmer interpreting the SAP, the input datasets, and search for TFL shells and prior programs or macros, she or he can collect all validated content in the metadata repository to create a study specific package which can be used to generate the output. The 2nd Statistical Programmer can review the study package on an ongoing basis and approve the output.

Besides the input dataset standards, SAP and TFL parameters, the metadata used in the updated process is more elaborate:

- Default values can be assigned based on metadata defined earlier in the study development cycle, for example biomedical concepts including associated collection, tabulation and analysis implementations
- Information about which data from the dataset can directly be mapped into a table, figure or listing
- Controlled Terminology to be applied for each variable
- Standard Algorithms and SQL Procedures can also be defined as metadata to allow reuse across studies
- Table, figure or listing layout information (templates) such as Fonts and Margins.

Several industry sources are available that provide guidance for organizations around the metadata mentioned above. For example, CDISC has defined standards for various clinical lifecycle stages as well as pointers for transformations between the stage specific standards.

Recently the CDISC 360 Project started with the mission to enhance standards “to automate preparation of study specification metadata and end-to-end study data processing”. Starting ‘with the end in mind’ first an end-to-start specification is created, based upon which start-to-end metadata will be defined. This metadata is then used to demonstrate the ability to process data and execute data transformations. The expectation is that the CDISC 360 Project will provide metadata that can be used to further implement the begin-to-end strategy within biopharmaceutical organizations.

The US National Cancer Institute (NCI), in collaboration with CDISC, provides controlled Terminology for many standard variables through its Enterprise Vocabulary Services. Further PhUSE or internal standard shells can be used to standardize TFL layout and parameters.

To store, manage, and govern the multitude of metadata standards leading up to the TFL generation process, organizations can use a Metadata Repositories (MDR). Originally MDRs were set up to support the governance of SDTM and ADaM dataset standards, while these standards are required by the FDA² and PMDA³. Over time the use of the MDR has expanded to include Collection (CDASH) and sometimes even protocol standards on the front end. Extending on the downstream end, it makes sense to utilize the MDR to also support standardized Tables, Figures and Listings.

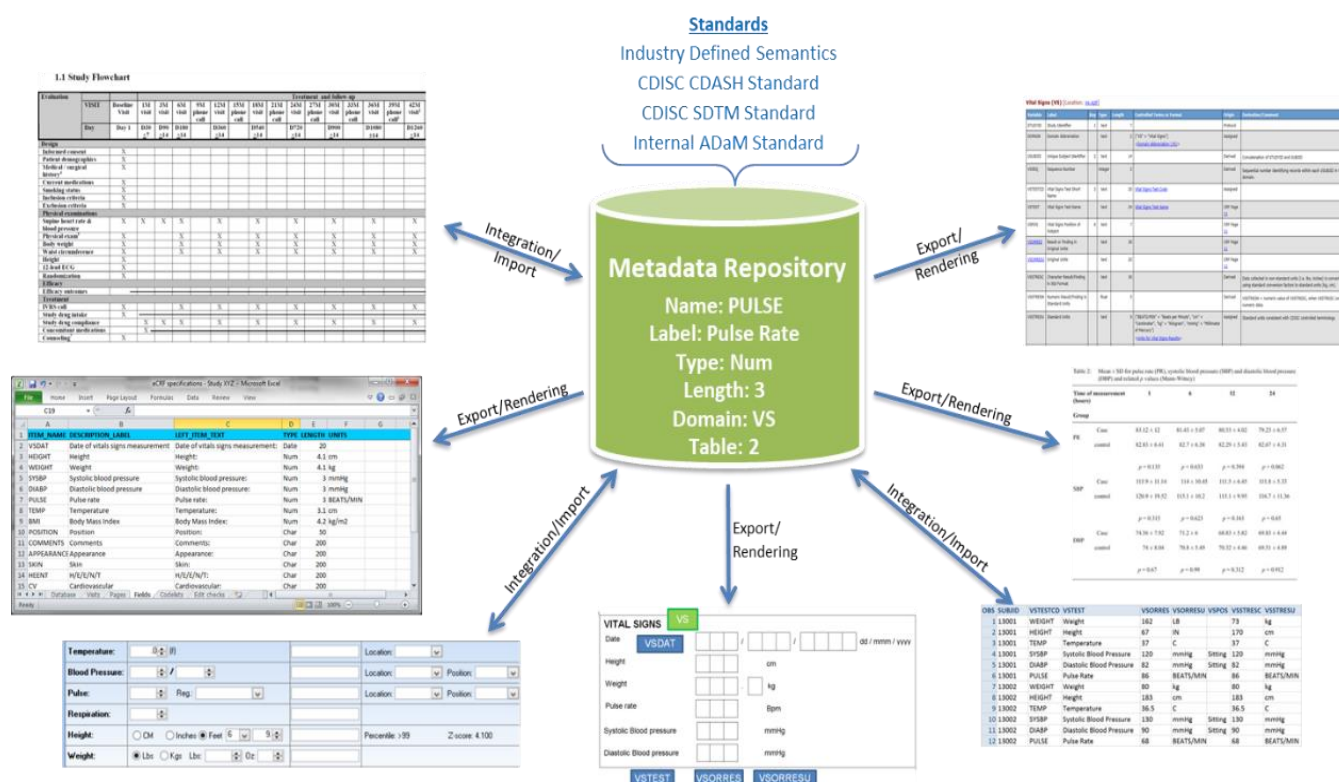


Figure 5. Metadata Repository enabling metadata-driven processes

PhUSE EU Connect 2019

To make full use of the metadata standards for the different clinical lifecycle stages in the MDR, it should be possible to link them to each other, thus providing information where and how the collected data ended up in the TFLs. This linking leads to improved data traceability which is important for regulatory compliance.

Figure 6 shows the example for the full data lineage of the age group (AGEGR1) field in the ADSL Listing. It is derived from AGE as present in the study SDTM specification which in turn comes from the birth and informed consent dates as collected.

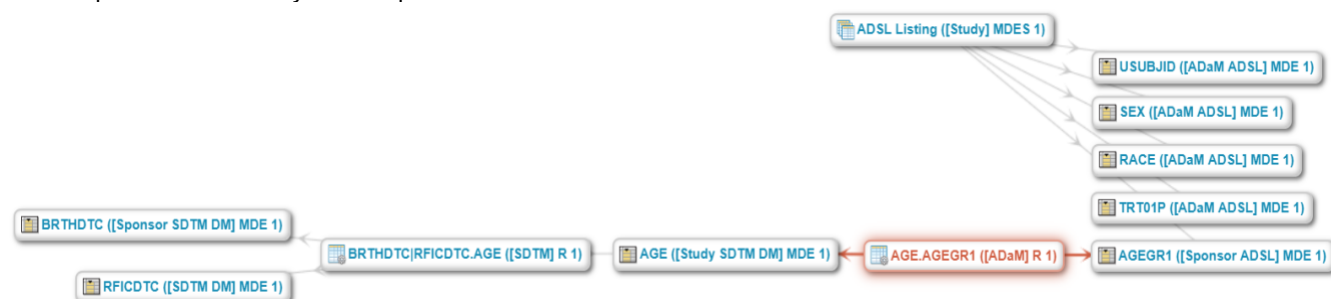


Figure 6. Traceability example from Age Group in ADSL listing back to collection clinical lifecycle stage

The 'BRTHDTC|RFICDTC.AGE' and 'AGE.AGEGR1' elements in figure 6 represent the transformational metadata for the derivations. This metadata typically includes the computational method which can be used in the define.XML and can further contain validated code in any required language. The example below shows 2 commonly used languages, SAS and R.

Published Asset: AGE.AGEGR1 ([ADaM] R 1)	
Asset Type: Rule	
Classifiers	
Name	Value
Lifecycle Status	Draft
Endorsement Level	Global
Clinical Lifecycle Stage	Data Tabulations
Rule Type	Computation
Artifacts	
Name	Value
Computational Method	Assign age groups for "5 and under", "6 to 11" and "12 and over"
SAS Derivation	"If Age <= 5 then AGEGR1 = ""5 and under""; Else if Age > 5 and Age < 12 then AGEGR1 = ""6 to 11""; Else if Age >= 12 then AGEGR1 = ""12 and over"";
R Derivation	AGEGR1 <-ifelse(AGE <= 5, "5 and under", ifelse(AGE <= 12, "6 to 11", no = "12 and over"))
Relationships	
Relationship	Asset Type Asset Name
Input Elements	Metadata Element AGE ([Study SDTM DM] MDE 1)
Produces	Metadata Element AGEGR1 ([Sponsor ADSL] MDE 1)

Figure 7. Example Transformational metadata to derived AGEGR1

METADATA DEFINITION AND IMPLEMENTATION TO DRIVE OUTPUT GENERATION USING A MACRO LIBRARY

Viewing the two processes to generate TFL output outlined above, different amounts and types of metadata are involved. In this paragraph the metadata required to support a process using TFL Macros are described. Dependent on your internal procedures they may include:

- CDISC ADaM standards metadata
- Global level metadata representative of each of your TFL shells and macro calls including content such as:
 - Macro Names and Macro Calls
 - Table Names and IDs
 - Descriptions
 - Input Data
 - Filters
 - Column Variables
 - Row Variables
 - Statistics (i.e. n, Mean, Min, Max, SD...)
 - Titles and Footnotes
- At the Global level, default parameters are assigned for each TFL in production

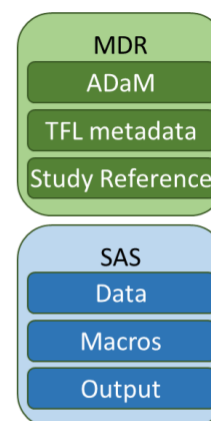
PhUSE EU Connect 2019

- At the Study level, users can copy from the Global level and make any necessary changes to the default parameters. Bear in mind that Standard TFLs should require few to no parameter changes, except for Title numbering and Footnotes
- Additional Study Level Reference metadata may include:
 - Protocol number
 - Study Number
 - Output Presentation Format (i.e. PDF)
 - Any additional content that might be common across all TFLs

Note: All computations, variable formats, and display properties are handled within the Macros.

When all metadata mentioned above is managed in an MDR, the content of below table example will be dictated by the indicated sources (see figure 8). The MDR provides the headers, titles, column and most row descriptions, and footnotes based on the input dataset and standard TFL definitions and study parameters. The table cell content is produced by the statistical analysis software (SAS® or other language) using the input data and standard macros.

Company Header(s)					
Titles		Column Variable			
Description		PBO (N=xxx)	Treatment 1 (N=xxx)	Treatment 2 (N=xxx)	Treatment 3 (N=xxx)
Row Variable 1	N	xx	xx	xx	xx
	Variable Value 1	xx(%)	xx(%)	xx(%)	xx(%)
	Variable Value 2	xx(%)	xx(%)	xx(%)	xx(%)
	Missing	xx	xx	xx	xx
Row Variable 2	N	xx	xx	xx	xx
	Mean	xx.X	xx.X	xx.X	xx.X
	SD	xx.XX	xx.XX	xx.XX	xx.XX
	Min	xx	xx	xx	xx
	Max	xx	xx	xx	xx
	Missing	xx	xx	xx	xx



Footnote(s)

Figure 8. Example Table Shell using Metadata defined Macro Calls

Though this approach offers efficiency in generating the TFL output, it warrants management of standards and metadata in different system and does not easily provide the full data traceability that regulatory authorities are looking for. The next paragraph will detail how this data traceability can be enabled by capturing further metadata about TFL content in a metadata repository.

METADATA DEFINITION AND IMPLEMENTATION TO DRIVE OUTPUT GENERATION BASED ON THE METADATA

The metadata required in the production of TFLs without TFL macros is more complex. In this case a table is built on a Cell by Cell basis. Cell specific content metadata contains information on how the value in the cell is derived, such as:

- percentage value for lines for Variable Value 1 and 2 is derived from the total number of subjects in the category divided by the grand total of subjects
- the SD for Row Variable 2 is derived from the mean and each individual value within the category

If metadata is defined at this level, each value in the Table can be traced back to data collection, providing full transparency for regulatory authorities.



Figure 9. Traceability example from Adverse Event Severity table back to collection clinical lifecycle stage

PhUSE EU Connect 2019

In the table below the MDR provides the headers, titles, footnotes, column and all row descriptions, as well as algorithms for each cell in the table. The table cell content is produced by the statistical analysis software (SAS® or other language) using the input data and metadata from the MDR.

Company Header(s)					
Titles		Column Variable			
Description		PBO	Treatment 1	Treatment 2	Treatment 3
		(N=xxx)	(N=xxx)	(N=xxx)	(N=xxx)
Row Variable 1	N	xx	xx	xx	xx
	Variable Value 1	xx(%)	xx(%)	xx(%)	xx(%)
	Variable Value 2	xx(%)	xx(%)	xx(%)	xx(%)
	Missing	xx	xx	xx	xx
Row Variable 2	N	xx	xx	xx	xx
	Mean	xx.x	xx.x	xx.x	xx.x
	SD	xx.xx	xx.xx	xx.xx	xx.xx
	Min	xx	xx	xx	xx
	Max	xx	xx	xx	xx
	Missing	xx	xx	xx	xx
Footnote(s)					

MDR
 ADaM
 TFL metadata
 Study Reference

SAS
 Data
 Macros
 Output

Combined
 TFL metadata
 Output

Figure 10. Example Table Shell using Metadata defined table content

The content required to succeed in the preparation of cell by cell-based outputs includes:

- CDISC ADaM standards metadata
- Global level metadata representative of each of your TFL shells, like
 - Output Type (Table, Listing, Figure)
 - Parameters
 - Output Name
 - Dataset
 - Top Rows & Columns

1_1	Description	e.g. "Summary of ..."
1_n	Variable	Formats are derived from the Metadata
2_1	Filter	Defined Rule / Algorithm from the metadata
2_n	Value, Count, %	Data Displayed
 - Table Group Row & Column
 - **Group_1**

1_1_1	Variable	e.g. SEX
1_1_2	Variable Label	e.g. "Sex" Assigned via Metadata
1_n_2	Include "n" row	Yes/No
1_n_n	Variable Format	e.g. \$Gender. Assigned via Metadata
1_n_n	Count %	e.g. Include Count, Percent
1_n_n	Variable Values	Group_1 Variable Value by Top_1_n & Filters
 - **Group_2**

1_n_n	Missing Count	Yes/No
2_1_1	Variable	e.g. HEIGHTBL
2_1_2	Variable Label	e.g. "Baseline Height cm" Assigned via Metadata
2_1_n	Include "n"	Yes/No
2_1_n	"n" Counts	Group_2 Count of Variable 1 by Top_1_n & Filters
2_2_2	Include "Mean"	e.g. Yes, "Mean"
2_2_n	"Mean" Values	Group_2 Mean of Variable by Top_1_n & Filters
		e.g. Mean HEIGHTBL > . by TRT01N & Filter
2_3_2	Include "SD"	Yes/No, "SD"
2_3_n	"SD" Values	e.g. SD HEIGHTBL > . by TRT01Nn & Filter
2_4_2	Include "Min"	Yes/No, "Min"
2_4_n	"Min" Values	e.g. Min HEIGHTBL > . by TRT01Nn & Filter
2_5_2	Include "Max"	Yes/No, "Max"
2_5_n	"Max" Values	e.g. Max HEIGHTBL > . by TRT01Nn & Filter
2_6_2	Include "Missing"	Yes/No, "Missing"
2_6_n	"Missing" Counts	e.g. Count HEIGHTBL . by TRT01Nn & Filter

PhUSE EU Connect 2019

- All computations, formats, and display properties are assigned within the Metadata Repository
- The number of Columns in Top Rows & Columns are derived and assigned based on the metadata
- At the Global level, default parameters are assigned for each TFL in production.
- At the Study level, users can copy from the Global level and make any necessary changes to the default parameters. Bear in mind that Standard TFLs should require few to no changes except for Title numbering and Footnotes.
- Additional Study Level Reference metadata may include:
 - Protocol number
 - Study Number
 - Output Presentation Format (i.e. PDF)
 - Layout content such as spacing and margins
 - Any additional content that might be common across all TFLs

Note: All algorithms, variable formats, and display properties are defined within the metadata.

Alternative approaches, or a combination of the above approaches, are available dependent upon your ability to standardize, automate, and implement. But no matter the approach, metadata can be used to support the methodology.

CONSUMING METADATA AND CREATING OUTPUTS

In either approach the definition of all the metadata above can lead to a dynamic, automated, TFL generation process. A process where approved metadata promotes the generation of validated code, which can then be run to produce the desired outputs. The governed metadata can manage and produce code in any language used in the organizations statistical computing environment. An example process is shown in figure 11.

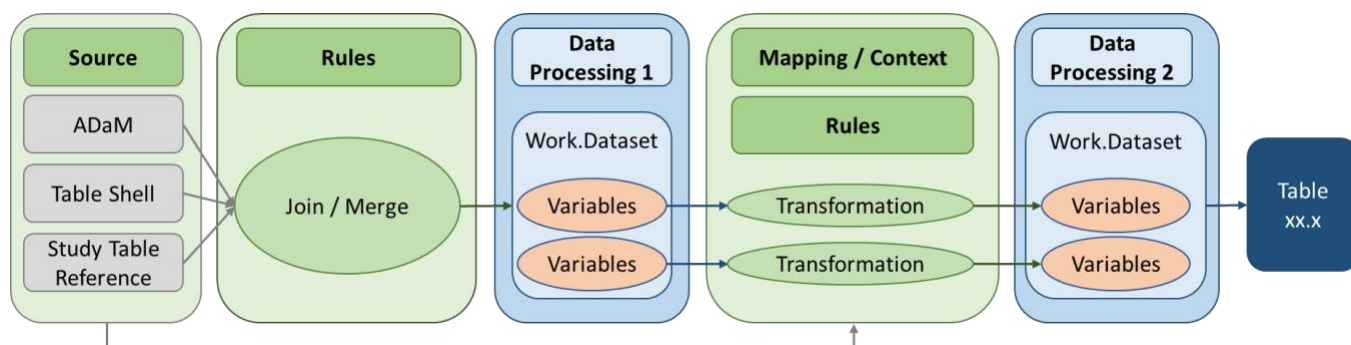


Figure 11. Process flow for creating a table

The depicted process shows a two-step approach where first Rules are applied to join/merge all necessary source datasets and then the variable level mapping and transformation is performed. Within this process the source metadata as well as the Rules, Mapping, and Context information are managed in the MDR and this metadata is consumed by a data processing system, generating and executing the programs to produce the output.

Access to the Metadata from the consuming systems will depend on internal processes and system capabilities and may use:

- API Calls to extract the metadata view for program consumption
 - Get <Study> TFL Metadata <Filter> (Filter on Tables, or a specific Table)
- MDR Exports to provide the metadata view for program consumption
 - Export <Study> TFL Metadata <Filter> (Filter on Tables, or a specific Table)
- SQL can also be used to extract the metadata view for program consumption
 - Select ... TFL Metadata ... Where ... <Filter> (Filter on Tables, or a specific Table) <Study>

Once metadata are available, usage will be dependent on the consuming systems.

- Will the consuming system code read the metadata to create and/or supplement programs?
- Are programs in place to support metadata driven TFL outputs?
- Will program code be maintained in the MDR and/or in an external library?

Regardless of the behavior of the consuming system, utilizing the metadata approach implies less efforts in validating study results as the metadata and external programs have been validated for usage. This leads to process efficiency and shortened timelines for submission.

CONCLUSION

There may be many ways to produce TFLs, and the decision for a “best” approach can only be taken by each organization individually. The current tool set needs to be evaluated and will determine which approach will fit their model with minimal interruption, unless the decision is that disrupting the current process will, in the long run, provide the ultimate benefits of automating and simplifying the process.

PhUSE EU Connect 2019

In any case, when metadata is used to drive the creation of TFLs, it provides greater change control, affords less QC effort for each study, and allows quicker time from the SAP to validated outputs.

The advantages of a process flow fully based on governed/approved metadata are:

- An accelerated Path for Submissions by reduction in Programming and reduction in QC efforts
- Enhanced Validation Efficiency: Verify once, use multiple times
- Improved Traceability from TFL to Collection

REFERENCES

1. EMA, Regulation (EU) No 536/2014 of the European Parliament and the of the Council on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC, 16 April 2014, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0536&from=EN>
2. FDA, Study Data Technical Conformance Guide v.3.3, 23 March 2017, <https://www.fda.gov/downloads/forindustry/datastandards/studydatastandards/ucm384744.pdf>
3. PMDA, Technical Conformance Guide on Electronic Study Data Submissions, 27 April 2015, <https://www.pmda.go.jp/files/000206449.pdf>
4. ICH, ICH Harmonised Tripartite Guideline: Structure and Content of Clinical Study Reports E3, 30 November 1995, https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E3/E3_Guideline.pdf
5. PhUSE, <https://www.phuse.eu/white-papers>:
 - i. Analyses & Displays Associated with Demographics, Disposition, and Medications in Phase 2-4 Clinical Trials and Integrated Summary Documents: Version 2.0, 02-Mar-2018
 - ii. Analysis & Displays Associated with Adverse Events: Focus on Adverse Events in Phase 2-4 Clinical Trials and Integrated Summary Documents: Version 1.0, 03-Feb-2017
 - iii. Analyses and Displays Associated with Thorough QT/QTc Studies: Version 1.0, 11-Mar-2016
 - iv. Analyses and Displays Associated with Outliers or Shifts from Normal to Abnormal: Focus on Vital Signs, Electrocardiogram, & Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Summary Documents: Version 1.0, 10-Sept-2015
 - v. Analyses and Displays Associated to Non-Compartmental Pharmacokinetics - with a Focus on Clinical Trials: Version 1.0, 25-Mar-2014
 - vi. Analyses and Displays Associated with Measures of Central Tendency - Focus on Vital Sign, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Submission Documents: Version 1.0, 10-Oct-2013

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Judith Goud
Nurocor, Inc.
106 East 6th Street
Suite 900
Austin, Texas 78701
Judith.Goud@nurocor.com
www.nurocor.com

Brand and product names are trademarks of their respective companies.