PharmaSUG 2017 - Paper DS23

The Benefits of Traceability Beyond Just From SDTM to ADaM in CDISC Standards

Maggie Ci Jiang, Teva Pharmaceuticals, Great Valley, PA

ABSTRACT

Since FDA released the Analysis Data Model Implementation Guide (ADaMIG)[1] for public review in 2014, traceability in CDISC ADaM data is well known for its characteristics between SDTM and ADaM. However, in CDISC ADaM development, the ability of traceability is more than just the connection between SDTM and ADaM. When it comes to a more complex analysis, an ADaM data value is not necessarily immediately related to a SDTM source data variable, some or more derivation procedures may be required to achieve the purpose, such as the derivation of an analysis-ready variable, parameter or record, to do it in the ADaM data or to program it in the analytical report instead? It appears that programmers can have an option. The decision often ends up with the one that the programmer favors over the other. Are programmers really free to make this choice? This paper will try to tackle this topic by comprehensive discussion of the ADaM Methodology through some practical examples, and to provide insight on the extensive benefits of traceability beyond just from SDTM data to ADaM data in CDISC ADaM development.

INTRODUCTION

The ADaM methodology is to include all observed and derived rows for a given analysis parameter. The inclusion of all the rows in the ADaM dataset, including those not used in the analysis, requires a way to identify the rows used in the specified analysis. The advantage to this approach is that the inclusion of all rows makes it easier to verify that the selection and derived time-point processing was done correctly, thus providing useful traceability. In addition, the data are also then available to enable other analyses, including sensitivity analyses. [2]

Regulatory reviewers prefer that the path followed in creating and/or selecting analysis rows be clearly delineated and traceable all the way back to the originating rows in the SDTM dataset, if possible and within reason. Simply including the algorithm in the metadata is often not sufficient, as any complicated data manipulations may not be clearly identified (e.g., how missing pieces of the input data were handled). Retaining in one dataset all of the observed and derived rows for the analysis parameter provides the clearest traceability in the most flexible manner within the standard BDS. The resulting dataset also provides the most flexibility for testing the robustness of an analysis (e.g., using a different imputation method). [3]

The design of ADaM datasets and associated metadata should facilitate explicit communication of the content of, input to, and purpose of submitted ADaM datasets. The Analysis Data Model should support efficient generation, replication, and review of analysis results[4]. To satisfy these requirements, the quality of traceability plays an inevitable role.

The two commonly known types of traceability are:

Metadata Traceability: provides the information about data i.e. origin of variable, algorithm used to derive the variable etc. It establishes traceability between ADaM data and SDTM data by describing the algorithm used to derive or populate an analysis-ready variable, a parameter or a record.

Data Point Traceability: enables users (agency reviewers, QC programmers, Biostatisticians etc.) to go directly to the specific SDTM data record(s) used to derive an analysis value. This level of traceability is straightforward when a user is trying to trace a data manipulation path. It can be established by providing clear links in the data to the specific data values used as an input from predecessor to derive an analysis value.

There's the third type of traceability:

Report Traceability: a virtual traceability that connects the analytical results to the ADaM data. This traceability allows users (agency reviewers, QC programmers, Biostatisticians etc.) to trace clearly a

complex analysis report value to the underlying data, to PROC AWAY the analysis result easily, and to be able to reproduce the outputs with less hassle.

The traceability from ADaM datasets to SDTM datasets is measurable as you can see by examining the data; the traceability from analysis results to ADaM datasets is virtual and hard to measure.

This paper first briefly reviews the essence of the Data Point Traceability and the Metadata Traceability, then extends to explore the extensive benefit of Report Traceability in CDISC ADaM standard development.

Metadata Traceability

Metadata Traceability is presented by providing the users a document that can give the reviewers a clear picture of how the analysis data sets have been created. This traceability provides the information about data i.e. the origin of the variable, the algorithm used to derive the variable etc. It establishes traceability between ADaM data and SDTM data by describing the methods used to derive or populate an analysis-ready variable, a parameter or a record.

The most important component is the Defined Definition Table (DDT) in ADaM metadata development. The DDT plays a critical role when programmers start to develop ADaM datasets. The DDT is the actual document that provides the true information regarding the data traceability.

It looks like that the metadata define.xml is the last step of the ADaM development, however, it does not seem as it appears to be. It's highly recommended that statisticians and programmers should always work together to create first the DDT before putting hands on the ADaM datasets development, which is the appropriate way to keep the ADaM traceability on the right track.

There are many types of an ADaM data variable traceability, examples are a simple copy of the variable from a SDTM dataset, a value derivation from the multiple variables of a single SDTM dataset, or the value derivation from the multiple variables of the multiple SDTM datasets. Let's review a few metadata examples.

Table 1.1 Illustrate a metadata example of ADSL variables from a simple copy of SDTM variables.

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation / Comment
STUDYID	Study Identifier	text	100		Predecessor: DM.STUDYID
USUBJID	Unique Subject Identifier	text	30		Predecessor: DM.USUBJID.
RANDID	Randomization Identifier	text	8		Predecessor: DS.DSREFID
AGE	Age	integer	2		Predecessor: DM.AGE
AGEU	Age Units	text	10	Unit	Predecessor: DM.AGEU

*the copied variables can keep the SDTM variable name and label as is.

Table 1.2 ADSL: Illustrate a metadata example of an ADSL value from the multi-variables of a single SDTM dataset

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
HEIGHT	Baseline Height (cm)	float	3.1		Derived: Set to VSSTRESN where VSTESTCD='HEIGHT' and VSBLFL ='Y'
WEIGHT	Baseline Weight (kg)	float	3.1		Derived: Set to VSSTRESN where VSTESTCD='WEIGHT' and VSBLFL='Y'
BMI	Baseline BMI (kg/m2)	float	2.1		Derived: BMI = WEIGHT/((HEIGHT*0.01)*(HEIGHT*0.01)) Round to one decimal places.

*HEIGHT and WEIGHT have kept their SDTM variable names and labels, they have kept their original values from SDTM though the rules applied to them; BMI is newly calculated variable by using HEIGHT and WEIGHT.

Table 1.3 illustrate an example of an ADaM record from the multi-values of multiple SDTM datasets.

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
PKFL	Pharmacokinetic Population Flag	text	1		Derived: "Y" if EX.EXTRT='DRUG A' and EXCAT = 'STUDY MEDICATION RECORD' and EXSTDTC is not missing; AND PC.PCSTAT is not 'NOT DONE'; "N" otherwise.

*PKFL has been derived by using both SDTM EX and PC data variables.

Data Point Traceability

Data Point Traceability is the fundamental requirement for ADaM to be able to be traced back to SDTM data. This traceability enables users (agency reviewers, QC programmers, Biostatisticians etc.) to be able to trace directly to the specific SDTM variables or data record(s) that have been used to derive an analysis value. This level of traceability is straightforward when a user is trying to trace a data manipulation path. It can be established by providing clear links in the data to the specific data values used as an input from predecessor to derive an analysis value.

ADaM datasets and metadata must clearly communicate how the ADaM datasets were created. The Data Point Traceability should exactly reflect the metadata defined. Per Study Data Technical Conformance Guide, each submitted ADaM dataset should have its contents described with complete metadata in the define.xml file and within the ADRG as appropriate[5].

Table 2.1 Corresponding to Table 1.1 Illustrate an example of ADSL variables from a simple copy of SDTM variables.

Study Identifier (STUDYID)	Unique Subject Identifier (USUBJID)	Randomization Identifier (RANDID)	Age (AGE)	Age Units (AGEU)
C38000-INC-20001	INC_20001_100101		37	YEARS
C38000-INC-20001	INC_20001_100102	151	25	YEARS
C38000-INC-20001	INC_20001_100103		27	YEARS
C38000-INC-20001	INC_20001_100104	152	20	YEARS

In this example, STUDYID, USUBJID, AGE and AGEU are values simply copied from SDTM domain DM; RANDID is a simply copy from SDTM domain DS, which reflect exactly what Table 1.1 metadata have described.

Table 2.2 Corresponding to Table 1.2 Illustrate an example of an ADaM value from the multi-variables of a single SDTM dataset.

Study Identifier (STUDYID)	Unique Subject Identifier (USUBJID)	Baseline Height (cm) (HEIGHT)	Baseline Weight (kg) (WEIGHT)	Baseline BMI (kg/m2) (BMI)
C38000-INC-20001	INC_20001_100101	.	.	.
C38000-INC-20001	INC_20001_100102	180.4	72.4	22.2
C38000-INC-20001	INC_20001_100103	.	.	.
C38000-INC-20001	INC_20001_100104	165.9	70.8	25.7

In this example, HEIGHT and WEIGHT values are the derivation from SDTM domain VS.VSTESTCD and VS.VSBLFL; and the BMI value is then derived from the calculation of HEIGHT and WEIGHT.

Table 2.3 Corresponding to Table 1.3 Illustrate an example of an ADaM record from the multi-values of the multiple SDTM datasets.

Study Identifier (STUDYID)	Unique Subject Identifier (USUBJID)	Pharmacokinetic Population Flag (PKFL)
C38000-INC-20001	INC_20001_100101	N
C38000-INC-20001	INC_20001_100102	Y
C38000-INC-20001	INC_20001_100103	N
C38000-INC-20001	INC_20001_100104	Y

*PKFL has been derived by using both SDTM EX and PC datasets.

Report Traceability

This Report Traceability means that the analytical report should be able to be easily connected to the ADaM dataset for reviewers to understand the analysis results without much frustration or confusion. This traceability allows users (agency reviewers, QC programmers, Biostatisticians etc.) to trace clearly a complex analysis report with its value derivation route; more, to allow users to PROC AWAY the analysis result in simple steps, and to help reviewers to reproduce the outputs with less frustration.

Often there are hard times that the reviewers or the QC programmers have difficulty to identify the values in the clinical reports. It happens when the primary programmers or the statisticians have done the complex derivations directly in the programs, and such an approach makes it opaque for the reviewers or the QC programmers to follow.

ADaMIG describes that the Analysis Data Model should support efficient generation, replication, and review of analysis results [6]. However, the challenge is how much is considered sufficient?

The fundamental principles are that ADaM datasets and associated metadata must provide traceability to show the source or derivation of a value or a variable (i.e., the data's lineage or relationship between a value and its predecessor(s)). The metadata must identify when and how analysis data have been derived or imputed [7].

Let's take a look at a report example –

Table 14.3.1.1 provides the overview of treatment-emergent adverse events for a 150-day follow-up study analysis. In addition to the general summarized adverse events information, it asks for the total number of the adverse events by distinguishing between the new from the 150-day follow-up data and the old from the prior FDA submission data.

Protocol Number: xxxxx Page 1 of 1

Table 14.3.1.9
Treatment-Emergent Adverse Events (TEAEs): Overview
Safety Population

		Treatment Drug A		
Time Period	Parameter	NDA (N=)	New Data, n	Complete Safety Update (N=)
Period 1	Number of subjects	xx	xx	xx
	Any TEAEs	xx (xx.x)	xx	xx (xx.x)
	Any SAEs	xx (xx.x)	xx	xx (xx.x)
	Any Severe TEAEs	xx (xx.x)	xx	xx (xx.x)
	Any TEREs	xx (xx.x)	xx	xx (xx.x)
	Any TEAEs Leading to Disc	xx (xx.x)	xx	xx (xx.x)

Note: The New Data column includes incremental data of existing patients in the NDA submission and data of new patients not included in the NDA submission.

The columns "NDA" and "New Data" actually require the comparison between the current ADAE and the prior ADAE used for NDA submission in order to identify which adverse event has been newly collected. To do this, a programmer can have two options, to do it directly in the report program or to do it in the ADAE.

With the 1st option, the programmer derives the columns "NDA" and "New Data" values directly in the table program, "NDA" numbers are from the prior ADAE for NDA, "New Data" numbers are from the current ADAE. The consequence is that it will be tough for the QCer and the reviewer, as it's really hard for them to figure out on their own how the numbers in the two columns come from. In this case, the

traceability from the report to the ADAE data has been lost because the QCer and the reviewer can not identify the “New” adverse events in the current ADAE. This 1st option has actually violated the fundamental “analysis-ready” [8] principle, and is not recommended.

With the 2nd option, the programmer first creates a new variable “D150FL” in ADAE to identify the new adverse events, to document the rules applied to this variable in the metadata DEFINE, the QCer and the reviewer will be given the opportunity to follow the DEFINE to figure out the number of the “New Data” records. The 2nd option is a better option and the correct way of doing it as it can appropriately achieve the ADaM compliance on the data traceability and the “analysis-ready” principles.

Here’s the SAS® code for identifying the new adverse events in ADAE –

```
data adae;
    length D150FL $1.;
    merge ae150(in=a) ae_nda(in=b);
    by usubjid aebodsys aeterm aestdtc aeendtc aeser aesev aerel;
    if b & ^a then flag =1;          /* only in NDA */
    if a & ^b then D150FL = 'Y';    /* flag only new AEs in D150 */
    if flag ne 1;    /* only keep subjects with both ae_NDA and ae150 */
run;
```

The metadata will document this derivation approach:

Table 3.1 Illustrate the metadata example of ADAE variable “D150FL”.

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
D150FL	New Data in Day 150 Update Flag	text	1		Derived: "Y" for the new adverse events since the NDA submission. New adverse event is identified by comparing the 150 Day ADAE to the previous NDA ADAE based on the these variables - USUBJID, aebodsys, aedecod, aestdtc, aeendtc, aeser, aesev, aerel.

Table 3.2 SAS dataset corresponding to Table 3.1 Illustrate the example of an ADAE variable "D150FL".

Unique Subject Identifier (USUBJID)	Body System or Organ Class (AEBODSYS)	Dictionary-Derived Term (ADDECOD)	Start Date/time of Adverse Event (AESTDTC)	End Date/time of Adverse Event (AEENDTC)	Serious (AESER)	Severity (AESEV)	Causality (AEREL)	New Data in Day 150 Update Flag (D150FL)
INC_20001_100101	Nervous system disorders	Poor quality sleep	6/13/2014	7/1/2014	Y	MILD	RELATED	
INC_20001_100101	Nervous system disorders	Headache	7/1/2014	7/5/2014	N	MILD	RELATED	Y
INC_20001_100103	Gastrointestinal disorders	Dysphagia	12/1/2014	3/1/2015	N	MODERATE	NOT RELATED	
INC_20001_100104	Psychiatric disorders	Depression	1/10/2015	3/12/2015	Y	MILD	NOT RELATED	Y

Table 3.3 The example of the final report.

Protocol Number: xxxxx Page 1 of 1

Table 14.3.1.9
Treatment-Emergent Adverse Events (TEAEs): Overview
Safety Population

		Treatment Drug A		
Time Period	Parameter	NDA (N=80)	New Data, n	Complete Safety Update (N=100)
Period 1	Number of subjects	80	100	100
	Any TEAEs	32 (40.0)	23	47 (47.0)
	Any SAEs	0 (0.0)	1	1 (1.0)
	Any Severe TEAEs	0 (0.0)	0	0 (0.0)
	Any TEREs	15 (18.8)	5	18 (18.0)
	Any TEAEs Leading to Disc	0 (0.0)	0	0 (0.0)

Note: The New Data column includes incremental data of existing patients in the NDA submission and data of new patients not included in the NDA submission.

Table 3.1 and table 3.2 have clearly provided the data lineage of the Table 3.3, i.e. the relationship between the report and its source ADAE. This approach has ensured the report traceability which facilitates transparency to the users.

CONCLUSION

The traceability from analysis results to ADaM datasets can be achieved though it is virtual and hard to measure. When it comes to the situation that the programming derivation can be done in either ADaM data or in the report, it's proved to be a better option to do it in ADaM data. To do it in ADaM data will help relate counts from tables, listings and figures in a study report to the underlying data[9].

The benefits:

- Provides the traceability of the analysis result lineage to ADaM, ADaM to ADaM, ADaM to SDTM.
- One DEFINE document can serve all, the programmer, the QCer and the reviewer.
- Give the QCer and reviewer an easier life to replicate the report numbers without doubts or confusion.
- The resulting dataset also provides the most flexibility for testing the robustness of an analysis (e.g. using a different imputation method or an automation tool).
- Significantly reduce the analysis reports maintenance time, if a derivation rule requires a change, only the variable in the ADaM dataset requires an update instead of touching the complex report program.
- Help reduce the reviewing time and the approval length of time.

You're strongly recommended to adopt this approach if your answer is a "no" to a question like "can the reviewer understand my analysis result without looking at my programming code" or "can the reviewer understand my analysis result with the help of metadata"?

There's a drawback to this approach. The development of ADaM data is often ongoing until the data lock time. Maintaining the DEFINE document up to date is crucial. It's a challenge for the developer to ensure the DEFINE document is a real working one for all users.

REFERENCES

[1] ADaMIG Version 1.1 < CDISC Analysis Data Model Team >, 2016-02-12.
[2] ADaMIG Version 1.1 < CDISC Analysis Data Model Team >, page 71.
[3] ADaMIG Version 1.1 < CDISC Analysis Data Model Team >, page 71.
[4] ADaMIG Version 1.1 < CDISC Analysis Data Model Team >, page 04.
[5] Study Data Technical Conformance Guide v3.2.1, General Considers, page 13
[6] ADaMIG Version 1.1 < CDISC Analysis Data Model Team >, page 04
[7] ADaMIG Version 1.1 < CDISC Analysis Data Model Team >, page 10
[8] ADaMIG Version 1.1 < CDISC Analysis Data Model Team >, page 10
[9] FDA Techinical Conformance Guide v3.3, 8.3 Study Data Traceability

CONTACT INFORMATION

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

Name: Maggie Ci Jiang
Enterprise: Teva Pharmaceuticals
Address: 2 West Liberty Blvd
City, State, ZIP: Malvern, PA 19355
Work Phone: 610-893-1201
E-mail: Maggie.jiang@tevapharm.com