PharmaSUG2010 - Paper CD06

How to build ADaM from SDTM: A real case study

JIAN HUA (DANIEL) HUANG, FOREST LABORATORIES, NJ

ABSTRACT:

Building analysis data based on the ADaM model is highly recommended since ADaM supports a high transparency of derivation. It aims to be ready-for-analysis and is clearly traceable to input data. This is a real application of ADaM on a phase III, randomized, double-blind, placebo-controlled clinical study. This article describes how to: create ADaM data based on four fundamental principles from the ADaM guideline; prepare for ADaM metadata by creating analysis data specifications; understand the SDTM to derive ADaM properly; build ADaM data according to the statistical analysis plan (SAP); and realize differences between the current and future ADaM guidelines to learn from the current study and make improvements for the future studies.

BACKGROUND:

Here described is the first study that we use the ADAM model to create our analysis datasets. We faced a few challenges at the onset of ADaM derivation. First, we have no examples to follow for ADaM derivation in the company. Second, the ADaM Implementation Guide, Version 1.0 (ADaMIG v1.0), which we used intensively at that time, is not finalized yet. It is still collecting comments and being updated. Third, in addition to creating ADaM data from SDTM data, our validation team conducts a cross-validation of ADaM data with CRF data. Fourth, the efficacy analysis in the study is complicated, which includes nearly 30 different efficacy parameters. Therefore, how to apply the ADaM guideline properly, meanwhile the guideline itself is still changing; how to use SDTM data correctly, for our ADaM data to match with our validation team's data; how to prepare the efficacy data to be one step away from analysis results; and how to recognize the differences in the ADaM guideline now and future, to improve the ADaM derivation for future studies. Those are the challenges we face and below are the summaries of our learning experiences from one real case study of ADaM derivation.

ABBREVIATIONS:

ADaM	Analysis Data Model	SDTM	Study Data Tabulation Model
CSR	Clinical Study Report	CRF	Clinical Report Form
SAP	Statistical Analysis Plan	TFL	Tables, Figures & Listings

ADaM DERIVATION:

With all the challenges we faced, we decided to conduct the ADaM derivation in four steps. First, refer to the current ADaM guideline. Second, understand SDTM - the source data used for ADaM derivation. Third, follow the statistical analysis plan. And fourth, realize the differences and revisions needed for future ADaM model.

Step 1. Refer to the ADAM guideline.

As mentioned before, the ADaM Implementation Guide, Version 1.0 (ADaMIG v1.0), which we mainly used as our reference is not the final version. It is still open for comments and some portion will be updated in the future. However, the fundamental principles described in this guideline should remain the same in the future. Therefore, we think it is safe to follow these fundamental principles as much as possible for ADaM derivation. In summary, there are four fundamental principles we identified from the current ADaM guideline, and they are listed as below:

Four Fundamental Principles that we followed when creating ADaM model:
(1) Support traceability
(2) Be analysis-ready
(3) Have machine-readable metadata
(4) Be usable with current tools

Principle (1): Support Traceability.

Traceability is of high importance in ADaM. It permits the understanding of three types of relationships: the relationship between the analysis results and data, the relationship across data, and the relationship within data. I discuss how we support the traceability in this case for each type of relationship from below example:

First, a snapshot of ADEFF (ADaM data for efficacy analysis) is given in next page in order to display the application of traceability. Per Graph 1, the ADEFF snapshot is created in a data structure with one-record per subject, per visit (week), per parameter. Some key variables (USUBJID, TRT1N, AVISITN, and PARAMCD) are included, in addition to the key variables, some analysis variables (BASE, AVAL, CHG) and analysis supportive variables (ABLFL, SRCDOM, SRCVAR, DTYPE) are included as well. Both the data structure and the derived variables support the traceability of ADEFF. In addition to the snapshot, a sample table shell for efficacy result is also given in Graph 2. By comparing ADEFF data with efficacy table, we can see how ADEFF support the relationship between analysis data and results. And by comparing variables within ADEFF data, we can see how it supports the relationship across data and within data.

Graph1. ADEFF snapshot.

SUBJID	TRT1PN	AVISITN	PARAMCD	BASE	AVAL	CHG	ABLFL	DTYPE	SRCDOM	SRCVAR
001	1	-99	PARAM1	0	0	0	Y		ADXXX	VAR1
001	1	-2	PARAM1	0	0	0			ADXXX	VAR1
001	1	-1	PARAM1	0	0	0			ADXXX	VAR1
001	1	1	PARAM1	0	4	4			ADXXX	VAR1
001	1	2	PARAM1	0	4	4			ADXXX	VAR1
001	1	3	PARAM1	0	4	4			ADXXX	VAR1
001	1	4	PARAM1	0	4	4			ADXXX	VAR1
001	1	5	PARAM1	0	4	4			ADXXX	VAR1
001	1	6	PARAM1	0	4	4		LOCF	ADXXX	VAR1
001	1	7	PARAM1	0	4	4		LOCF	ADXXX	VAR1
001	1	8	PARAM1	0	4	4		LOCF	ADXXX	VAR1
001	1	99	PARAM1	0	4	4			ADXXX	VAR1
002	2	-99	PARAM1	1	1	0	Y		ADXXX	VAR1
002	2	-2	PARAM1	1	1	0			ADXXX	VAR1
002	2	-1	PARAM1	1	1	0			ADXXX	VAR1

Graph2, table shell for the efficacy results

Table 14.X.X.X
Change from Baseline in 12-Week PARAM1 Frequency Rate (OC)
Intent-to-Treat Population

Visit*	Statistics	Placebo (N=XXX)	Treatment 1 (N=XXX)	Treatment 2 (N=XXX)
Baseline	Mean	xx.x	xx.x	xx.x
	SD	xx.x	xx.x	xx.x
	SEM	xx.x	xx.x	xx.x
	Median	xx.x	xx.x	xx.x
	Min, Max	xx, xx	xx, xx	xx, xx
	n	xx	xx	xx
Treatment Period	...			
	...			
Change from Baseline				
	Mean	xx.x	xx.x	xx.x
	SD	xx.x	xx.x	xx.x
	SEM	xx.x	xx.x	xx.x
	Median	xx.x	xx.x	xx.x
	Min, Max	xx, xx	xx, xx	xx, xx
	n	xxx	xx	xx
ANCOVA Results	LS Mean Change from Baseline	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)
	P-value for within-group*	(0.xxxx)	(0.xxxx)	(0.xxxx)
	LS Mean Difference (95% CI)		xx.x (xx.x,	xx.x (xx.x,
	(Lin - Placebo)		xx.x)	xx.x)
	P-value [1]		0.xxxx	0.xxxx

Comments:

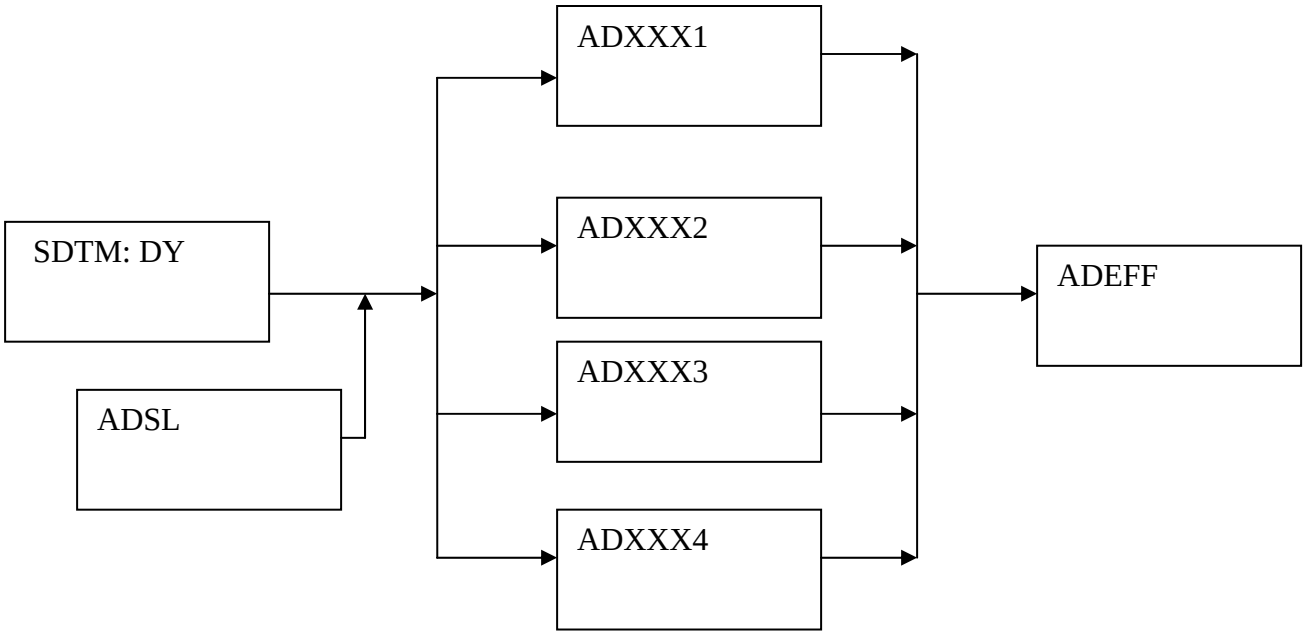
By comparing the table shell and ADEFF data snapshot, the reviewer can easily link the results displayed on the table to the ADEFF data, and understand the relationship between results and analysis data. For example, those three analysis variables (BASE, AVAL, CHG) derived in ADEFF are used for the summary table of efficacy for 'Baseline', 'Treatment Period' and 'Change from Baseline'. The 'DTYPE' variable indicates which records are imputed by the method of last observation carried forward (LOCF). So these records are excluded in the table summary as the table includes observed cases (OC) only. In addition to display the relationship between study results and analysis data, the ADEFF also displays the relationship across data and within data. For example, SRCDOM = 'ADXXX' indicates that the efficacy parameter "PARAM1" is derived from another data called 'ADXXX'. And SRCVAR= 'VAR1' indicates that the value of PARAM1 is related to another variable 'VAR1' which is saved in ADXXX. In addition, ABLFL='Y' indicates which visit is a baseline visit and where does the baseline value come from. All these variables derived in ADEFF permit the reviewer to understand the relationships more clearly.

Principle (2): Be analysis-ready.

Analysis-ready requires ADaM data to be one step away from study results. All variables used for summary tables (or listings) should be available from analysis data, no extra derivation are encouraged in the table programming.

As we mentioned before, ADEFF is created on a one-record per subject, per visit (week) and per parameter structure. This structure builds a close relationship between ADEFF and summary tables, and ensures ADEFF is ready for analysis. However, in addition to create summary tables, we also need to create listings. Some listings are called 'table listings', as they require certain derived information that can not be obtained directly from the raw data, i.e. SDTM or CRF data. Neither can those derived information be obtained from ADEFF. As listings present patient's daily record, they have a different structure with ADEFF, which is created on a weekly visit base. Then, how to make analysis data to be analysis-ready for both tables and listings? How many datasets needed for efficacy analysis? By referring to the ADaM guideline, we conclude with the following solutions. First, create the optimum number of analysis datasets that require minimal processing for analysis and review. Second, closely follow the study protocol, SAP and TFL shell. Third, some analysis datasets could be based on other analysis datasets. Based on these solutions, we created five datasets (including ADEFF) for efficacy analysis. Below is the flow chart to demonstrate all efficacy analysis data we created in this study and how they relate to each other.

Graph 3, flowchart of efficacy analysis datasets



Comments:

DY is an SDTM dataset which contains raw efficacy information. Its structure is one-record per subject, per question, per day. ADEFF is an ADaM dataset which contains average weekly values for all efficacy parameters. To be analysis-ready for certain listings, more importantly, to improve transparency of the derivation from DY to ADEFF, we created four other ADaM datasets to support efficacy analysis in this study. ADXXX1 contains derived data for study visit and period. ADXXX2 contains derived data for patient's gastrointestinal (GI) assessment. ADXXX3 contains derived data for patient's rescue medication. ADXXX4 contains derived data for all other efficacy information. These four ADaM datasets are created on a daily record base, so they match the listing structure and are ready for listing output. In addition, the derivations of efficacy parameters are complicated in this study. For example, in order to derive the primary efficacy parameter, we need to consider information from multiple fields, including patients' GI assessment, rescue medication, and study duration. Therefore, those four extra efficacy datasets not only permit the ready-to-analysis for listings, but also support the traceability for ADEFF derivation.

Principle (3): Have machine-readable metadata

Metadata can be defined as: "a structured, encoded data that describes characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described

entities" *(William R. Durrell, Data Administration: A Practical Guide to Data Administration, McGraw-Hill, 1985).* A one sentence summary of metadata is "the information about data".

Four types of ADaM metadata have been defined. They are 'analysis dataset metadata', 'analysis variable metadata', 'analysis parameter value-level metadata', and 'analysis results metadata'. In our study, we created data specifications for each ADaM data to support the generation of ADaM metadata. Our data specification includes 3 parts: 'the heading', 'the content', and 'the notes'. Below are examples of how we use data specification to prepare for ADEFF metadata.

1. Use 'heading' to prepare for 'analysis dataset metadata':

The 'heading' part of ADEFF specification is listed as below:

1. Purpose

The derived data set ADEFF is created to store values for efficacy parameters, including: primary, secondary and additional efficacy parameters. In addition, information of "Patients having XXXXX" is also included. ADEFF is created on a per-subject, per-parameter and per-analysis-week base. The data is built in the way according to ADaM implementation guideline 1.0, which is aimed to be ready-to-analysis, traceability to input data, and transparency of derivation.

2. Source SAS Data Set

The sources SAS data sets for ADEFF are the subject-level data: ADSL and derived patient diary data: ADXXX1, ADXXX2, ADXXX3, ADXXX4 and SDTM: QS.

3. Derived Data Set Structure

The structure of the derived data set ADEFF will be vertical. The variables USUBJID, PARAMCD and AVISITN are the key variables to identify each unique record.

A sample 'analysis datasets metadata' for ADEFF is given as below:

Dataset Name	Dataset Description	Location	Structure	Class	Key Variables	Documentation
ADEFF	Contains values for efficacy parameters	xxxx/adeff.xpt	One record per subject per parameter per analysis visit	BDS (Basic Data Structure)	USUBJID, PARAMC, AVISITN	SAP section 14.X.X for detailed efficacy parameter definitions
....						

Comments:

As shown by the 'heading' of data specification, the key information, required by metadata such as: dataset name, description, structure and key variables, are clearly defined in the heading part of our data specification. Therefore, when the data specification is created, the information for ADEFF metadata is ready. We just need to copy and paste the information into ADEFF metadata.

2. Use 'content' to prepare for 'analysis variable metadata':

Part of the 'content' of ADEFF specification is copied as below:

VARIABLE NAME	**VARIABLE LABEL**	**TYPE**	**CONTROLLED TERMINOLOGY**	**LENGTH**	**COMMENTS**
SUBJECT IDENTIFIER VARIABLES					
STUDYID	Study Identifier	char		32	Copy from ADSL: STUDYID
USUBJID	Unique Subject Identifier	char		32	Copy from ADSL: USUBJID
TIME VARIABLES					
PHASE	Study Phase Description	char	PRETREATMENT PHASE TREATMENT PHASE	40	Merge from ADXXX1 ADXXX2 ADXXX3 ADXXX4: PHASE
PHASEN	Study Phase Number	num	1=PRETREATMENT PHASE 2=TREATMENT PHASE	8	Merge from ADXXX1 ADXXX2 ADXXX3 ADXXX4: PHASEN
.....					

In addition, a sample 'analysis variable metadata' for ADEFF is given as below:

Dataset Name	**Variable Name**	**Variable Label**	**Variable Type**	**Format**	**Codelist/Controlled Terms**	**Source/Derivation**
ADEFF	STUDYID	Study Identifier	Text	$32		ADSL: STUDYID
ADEFF	PHASEN	Study Phase Number	Integer	8	1=PRETREATMENT PHASE 2=TREATMENT PHASE	Merge from ADXXX1 ADXXX2 ADXXX3 ADXXX4:PHASEN
ADEFF	....					

Comments:

By comparing the 'content' of the data specification with the sample metadata, the key information, required by ADEFF variable metadata such as: variable name, label, type, format and derivation, are specified in the content part of data specification. Some information could be generated by the SAS function 'proc content' as well. However, the 'source/derivation' of metadata could only be copied from the 'comments' column of data specification. It is good to note that, in the future, we will rename variable types from 'Character' to 'Text' and from 'Numeric' to 'Integer', and add a few more types, such as 'time', 'datetime' and 'float'. This makes our data specification be more consistent with ADaM metadata terminology.

3. Use ‘notes’ to prepare for ADEFF ‘parameter value-level metadata’:

Multiple analysis parameters can be saved in analysis datasets. As mentioned before, our study is complicated for its efficacy analysis, and our ADEFF contains nearly 30 efficacy parameters. Each parameter is assigned a different name, code and id number, the algorithm for deriving each parameter is different as well. In order to prepare the parameter value-level information for the metadata, we saved the corresponding information in the ‘notes’ part in ADEFF data specification. Below is a sample of the notes.

5. Notes

5.1 Parameter Assignment

PARAMCD (PARAM SHORT NAME)	**PARAM (PARAM DESCRIPTION)**	**PARAMN (PARAM NUM)**	**RECORDS WITH LOCF?**	**PARAMTYP (RELATED TO OTHER PARM)**	**SRCDOM (SOURCE DOMAIN)**	**SRCVAR (SOURCE VARIABLE)**
PARAM1	PARAM1 12-Week Overall Responder	32	Y	PARAM3	ADXXX1 ADXXX2	
PARAM2	PARAM2 12-Week Overall Responder	22	Y	PARAM4	ADXXX3 ADXXX4	

...

5.2 Derivation of Efficacy Parameter Values in Baseline.

For parameter (PARAMCD=PARAM1,PARAM2), the baseline PARAM1 and PARAM2 weekly rates will be derived as the corresponding overall weekly frequency rates based on the number of PARAM1 and PARAM2 a patient had in the Pretreatment Period. (Refer to Section 15.X.X of SAP for the algorithm.)

....

Comments:

ADaM is still developing a standard method for how to display the parameter level metadata in define.xml. Therefore, we don’t have a sample of parameter level metadata in hand. However, we believe that the key information needed for parameter level metadata has been defined and saved in the notes of our data specifications for ADEFF.

4. Analysis result metadata.

The result metadata tells how tables are created and what needs to be known about table programming. Using the previous efficacy table shell (table 14.4.2.1) as an example, a corresponding result metadata for that table could look like:

DISPLAY IDENTIFY	DISPLAY NAME	RESULT IDENTIFY
Table 14.4.2.1	Change from Baseline in 12-Week PARAM1 Frequency Rate	Pair wise treatment comparison

PARAM	PARAMCD	ANALYSIS VARIABLE	REASON	DATASET	SELECTION CRITERIA
PARAM1 Weekly Observed Frequency Rate	PARAM1	AVAL, CHG, BASE	Conduct efficacy analysis as specified in protocol	ADEFF	ittfl='Y' and paramcd eq 'PARAM1' and dtype in ('')

MODEL SPECIFICATION

```
/*-LSMENAS from proc mixed paired-wise test-*/
proc mixed data=adeff order=internal;
     by avisitn;
     class sitegrpn trt1pn;
     model chg=trt1pn sitegrpn base;
     lsmeans trt1pn /pdiff cl;
     estimate 'Placebo vs. Treatment 1'  trt1pn -1 1 0  / cl;
     estimate 'Placebo vs. Treatment 2'  trt1pn -1 0 1  / cl;
     ods output lsmeans=lsmean  diffs=diffs estimates=estimates;
run;
```

Comments:

The result metadata includes information of table programming, such as: table number, table name, input data, variables used for analysis, the selection criteria, the reason of creating the table, and the analysis model used in the table. It gives detailed information about how table is created. Currently, this type of information is not available in our data specification, as data specification is designed for each ADaM data, not for each table. Since it is important for the reviewer to know how tables are created, especially for some key efficacy tables, we are thinking of a proper way to create the result metadata in future studies.

Principle (4): Be usable with current tools.

In our company, we created certain tools, i.e. macro programs, to generate table, listings and figures more efficiently *(Reference: EZTABLE Macros: Jian Hua (Daniel) Huang, Forest Laboratories Inc, PharmaSUG Paper Year 2008)*. We also have a tool to create define.xml for analysis datasets. After we apply the ADaM model to our analysis datasets, we make sure that those existing tools could be applied to ADaM data as well.

Step 2. Understand of SDTM.

Since SDTM data are used as the source data for ADaM, it is critical to understand SDTM data correctly and comprehensively before deriving ADaM. The standard structure of SDTM is described in *"Study Data Tabulation Model version1.1"*, prepared by CDISC submission data standards team. The implementation of SDTM is explained in *"Study Data Tabulation Model Implementation Guide version 3.1.1"*. In addition to refer to these documents for a comprehensive understanding of SDTM, we summarize a few more comments by learning from our use of SDTM. First, it is important to understand the derivation of SDTM. Although SDTM are considered as the source data for ADaM, they are still derived from the raw data which are collected from CRF pages. Therefore, not only we need to know how to use SDTM, but also we need to understand where SDTM come from. Second, for certain key variables used in ADAM, it is encouraged that we derive them right in ADaM rather than copy them from SDTM. For example, in our study, we notice that the treatment emergent adverse event (TEAE) flag is created by SDTM already, however, it dose not apply the 30 days window to the TEAE flag, which is required in the statistical analysis plan (SAP). To be consistent with the SAP, we decided to derive the TEAE flag again in ADaM and drop the same variable from SDTM. Another example is the baseline record flag. By default, SDTM derive baseline record flag as the last non-missing record before the first dose of double-blind treatment. However, in our study, we have triplicate ECG records and we use the average value of triplicate ECG as the baseline. So, again, we derive the baseline record flag in ADaM. This time, we can not drop the ECG baseline record flag from SDTM, as it is a required variable in SDTM. Therefore, we keep the baseline record flag in SDTM and explain the differences between SDTM and ADaM in the comments column accordingly. Third, if any variables used in ADAM are copied from SDTM, we keep the same attributes, such as variable name, label, length, type and format. Fourth, SDTM usually keeps the character values only. While, sometimes, the numeric values are needed for the convenience of ADaM derivation. The way, we handle this difference, is to save the numeric values in SDTM+, and use SDTM+ to derive ADaM. During the submission, the numeric values are excluded from SDTM in order to keep SDTM simple and slim. Because SDTM keeps character values, the exclusion of numeric values do not affect the traceability between SDTM and ADaM.

Step 3. Follow the statistical analysis plan (SAP).

The SAP is the fundamental document used for clinical study analysis. It describes study design, defines study population, explains analysis model, and discusses methods for data imputation. Because SAP plays such an important role, we have to follow the SAP closely in ADaM derivation. We already mentioned that, according to the SAP, we derived five different ADaM data to conduct our efficacy analysis. We also mentioned that certain key variables, such as TEAE and baseline record flags, are derived differently in ADaM than in SDTM, as the derivation in ADaM follows more accurately with the algorithm specified in the SAP. These are examples of how we follow the SAP to create ADaM. One more example is how we derive the last dose date of treatment in this study if the date is missing from CRF termination page. SDTM creates a similar date called reference end date (RFENDTC), which could be used as the last dose date. It collects the last date available from multiple domains, including demographic, subject visit, lab, ECG, vital signs. However, this is not an accurate way to impute the last dose date in this study. The SAP defines the last

calling date from IVRS to be imputed as the last dose date if it is missing from CRF termination page. Therefore, we did not use RFENDTC from SDTM, but derive this key variable based on SAP in the ADaM.

Step 4. Realize the differences between current and future ADAM.

During the time when we created ADaM in this study, we used the '*ADaM Implementation Guide, Version 1.0 (ADaMIG v1.0)'* as the main reference guideline. Since then, many comments have been collected and ADaMIG v1.0 has been continually updated. Therefore, we refer to another important ADaM document '*Updated ADaM Slides for CDISC Interchange training course, 9-13 November 2009, Baltimore, MD'* to update ourselves. We realize some differences in the new ADaM document, and we think they are important for any improvements of ADaM derivation in the future. Below are a few cases of ADaM updates summarized from our study.

Case 1, some SDTM variables are suggested to be kept in ADSL, such as: RFSTDTC, RFENDTC and ARM. RFSTDTC (the reference start date) and RFENDTC (the reference end date) and ARM (treatment arm) were not kept in our ADSL before. At the time when we derived ADSL, these were not required variables. Now, when I review the updated slides for CDISC interchange training course for ADaM, these variables are recommended to be kept in ADSL.

Case 2, change the naming of certain variables. Following are certain treatment related variables we created in ADaM: TRTxP, TRTxPN, TRTxSDT, TRTxSDTM, TRTxEDT, TRTxEDTM. They represent the planned treatment and date and time of treatments. The small 'x' stands for the number of treatment phases, i.e. TRT1P stands for planned treatment for phase1, and TRT2P stands for planned treatment for phase 2. Now, the naming of these variables is slightly changed. In new ADaM document, two spaces are reserved for the treatment phases, i.e. TRTxxP, TRTxxPN, TRxxSDT, TRxxSDTM, TRxxEDT, TRxxEDTM. Under the new naming, the variable of planned treatment for phase 1 is like: TRT01P, this gives more flexibility to assign treatment for multiple phases study, even for studies with 10 or more phases.

Cases 3, derive new ADaM variables. We realize that certain new variables are suggested to be derived in ADaM from CDISC interchange training course. These variables include: TRTSDT (start date of first exposure to any treatment), TRTEDT (end date of exposure to any treatment) and ONTRTFN (On treatment record flag).

For the past 2 years, CDSIC team has been working on the development of ADaM structure and contents continually. We think it is important to follow up with the development of ADaM guideline and keep us updated. However, it is not necessary to wait until the final version of ADaM guideline is ready to start creating ADaM data. As mentioned before, those key principles described in ADaM guideline shall remain unchanged regardless of the updates of content or structure.

CONCLUSIONS:

In conclusion, we think the following steps are important when applying ADaM models to analysis data. First, refer to the most updated ADaM guideline. Second, apply ADaM foundational principles. Third, understand SDTM domains and variables. Fourth, follow the study protocol, SAP and TFL shell. And fifth, realize that ADaM guidelines are still changing.

REFERENCES:

1. The ADaM Implementation Guide, Version 1.0 (ADaMIG v1.0).
2. The Analysis Data Model, Version 2.1 (ADaM v2.1).
3. Updated ADaM Slides for CDISC Interchange training course, 9-13 November 2009, Baltimore, MD.
4. Study Data Tabulation Model, Version 1.1 (SDTM v1.1).
5. SDTM Implementation Guideline, Version 3.1.1.

CONTACT INFORMATION:

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

Jian Hua (Daniel) Huang, MSc,
Forest Research Institute
Harborside Financial Center, Plaza V
Jersey City, NJ 07311
Tel: 1-201-427-8291
Jian.huang@frx.com