SDA-03

A Taste of ADaM

Beilei Xu, Merck & Co., Inc., Rahway, NJ
Changhong Shi, Merck & Co., Inc., Rahway, NJ

ABSTRACT

The Analysis Data Model (ADaM) and the ADaM Implementation Guide (ADaMIG) published by the Clinical Data Interchange Standards Consortium (CDISC) provide the fundamental principles of analysis datasets and specifications for two standard data structures: the subject-level analysis dataset ADSL, and the Basic Data Structure (BDS) which is a general structure that provides "one proc away" readiness for many common analyses. This paper presents the detailed steps used in drug project work to create ADaM BDS datasets illustrated by ADLP, a dataset to support analysis of lipid endpoints. Particular emphasis is given to the following implementation considerations: 1. number of ADaM datasets needed; 2. derivation of analysis endpoints, analysis windows, analysis values, and imputation of missing values; and 3. setup of analysis flags and population flags.

Keywords: ADaM, CDISC, analysis dataset

INTRODUCTION

ADaM is a CDISC standard model that supports efficient generation, replication, and review of analysis results. The CDISC Analysis Data Model Version 2.1 and the CDISC ADaM Implementation Guide (ADaMIG) Version 1.0 are the basic references for this paper. These two documents specify the fundamental principles and standards to follow in the creation of analysis datasets and the associated metadata - "data about the data". This paper focuses on three fundamental features of ADaM analysis data: ADaM Basic Data Structure (BDS), traceability, and "one-proc" away readiness for analysis. The creation of ADaM metadata is out of scope of this paper.

This paper provides step-by-step real-time implementation of the ADaM model in deriving analysis data from SDTM source data and includes the following implementation considerations: 1. number of ADaM datasets needed; 2. derivation of analysis endpoints, analysis windows, analysis values, and imputation of missing values; and 3. setup of analysis flags and population flags. The analysis dataset used for illustration is a lipid endpoint dataset used for longitudinal analysis.

ADaM BASIC DATA STRUCTURE (BDS)

There are two ADaM standard data structures currently available in Version 2.1: 1. Subject-Level Analysis Dataset (ADSL); and 2., Basic Data Structure (BDS). The ADSL dataset structure contains one record per subject and variables such as subject-level population flags, planned and actual treatment variables, demographic information, randomization factors, sub-grouping variables, and important dates. Examples include:

RFSTDTC (Reference Start Date Time) and randomization date. Please refer to the ADaMIG for the details of ADSL.

The Basic Data Structure, on the other hand, contains one or more records per subject, per analysis parameter, and per analysis time point. This structure contains a central set of variables that describe the analysis parameter (e.g., PARAM and related variables) and contains the value being analyzed (e.g., AVAL and AVALC and related variables). Other variables in the dataset provide more information about the value being analyzed (e.g., the subject identification) or describe and trace the derivation of the variable (e.g., DTYPE) or enable the analysis (e.g., treatment variables, covariates). The BDS supports parametric and nonparametric analyses such as ANOVA, ANCOVA, categorical analysis, logistic regression, Cochran-Mantel-Haenzsel, Wilcoxon rank-sum, time-to-event analysis, etc.

In the example of creating the ADLP dataset below, the goal is to set up the lipid endpoint data in BDS format for longitudinal analysis that includes the following different types of BDS variables: subject identifiers, treatment variables, timing variables, analysis parameter variables, analysis visit windowing variables, and flag variables. Table1 and 2 below show a snapshot of the analysis dataset ADLP. It contains all the variables for one patient, USUBJID='1A-4_02', with three lipid endpoints (PARAMCD: LDL, HDL, LDLHDL) and four analysis time points (AVISIT: SCREENING, BASELINE, WEEK 2, WEEK 4). Due to the space limitation, the data are split into two tables. For easy reference, the second table repeats the USUBJID, PARAMCD and AVISIT variables so it can be linked back to the first table. Please refer to the ADaM document for the variable descriptions and labels.

The source of the ADLP dataset is the SDTM LB domain which contains patients' lipid measurements at different visits. In ADLP, variables STUDYID, USUBJID, SUBJID, and SITEID are the subject identifiers among which USUBJID is the unique subject identifier. TRTP, TRPA, TRTPN, and TRTAN are the treatment variables. TRTP and TRTA are the planned and actual treatment group. TRTPN and TRTAN are the numeric forms of TRTP and TRTA and are not required variables as are TRTA and TRTP; however, it is useful to have them for analysis programming. Both of the subject identifiers and treatment group variables can be found in the ADSL dataset. Analysis parameter variables include PARAM, PARAMCD, PARAMN, and PARAMTYP. PARAM is the unique identifier for the analysis parameter. It provides the unique description of the parameter with the unit when applicable. PARAMCD is its corresponding short name, and PARAMN is the numeric form of the parameter. Neither PARAMCD nor PARAMN are CDISC required variables like PARAM. If they are present, a 1-to-1 correspondence with PARAM is required. Using PARAM allows the endpoints to be presented in a long and skinny format, which allows one-proc away readiness for analyses.

Note that PARAMTYP is used to indicate the parameter "LDL/HDL ratio," which is derived from the ratio of the LDL and the HDL. There is no data collection for this parameter in the SDTM LB domain; it is derived for analysis purpose. See Step 2 below for more details.

Analysis timing variables include ADT and ADY. ADT is obtained from the observed date LBDTC in the LB domain. The relative analysis day, ADY, is derived based on the statistical analysis plan (SAP). The analysis visit windowing variables, AWRANGE, AVISIT, and AVISITN, are also derived based on ADY and other criteria defined in the SAP. In this example the following four analysis visits are defined: SCREENING, BASELINE, WEEK 2, and WEEK 4.

The analysis value variables include AVAL, BASE, and CHG. AVAL is obtained from the observed value in the LB domain; BASE is the baseline value; CHG is the change from baseline value. ABLFL is a flag variable indicating that the record contains the baseline value. The corresponding AVAL is the value for BASE. Note that DTYPE is used to indicate the record with an averaged value of the multiple records on the same date, and is described in Step 3a and 3b below.

The analysis flag variable, ANL01FL, identifies the record chosen for a specific analysis window. See Step 4 for more details. The parameter population flag, FASPFL, is defined as "patients with at least one post-baseline measurement" per SAP. Using the variables ANL01FL and FASPFL, it is easy to construct a criterion to subset the data for a specific analysis and it enables one-proc away readiness for analysis and is described in Step 6.

The traceability variables, SRCDOM, SRCVAR, and SRCSEQ, provide the SDTM source domain, variable, and value for the analysis record.

Table 1: ADLP Dataset (Part 1)

STUDYID	USUBJID	SUBJID	SITEID	TRTP	TRTPN	TRTA	TRTAN	PARAM	PARAMCD	PARAMN	PARAMTYP	ADT	ADY	AWRANGE	AVISIT	AVISITN
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL-C (mmol/L)	LDL	8001		9/15/2008	-16	< Day 1	Screening	-99
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL-C (mmol/L)	LDL	8001		10/01/2008	1	Day 1	Baseline	0
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL-C (mmol/L)	LDL	8001		10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL-C (mmol/L)	LDL	8001		10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL-C (mmol/L)	LDL	8001		10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL-C (mmol/L)	LDL	8001		10/28/2008	28	>= 22 Days	Week 4	4
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	HDL-C (mmol/L)	HDL	8002		9/15/2008	-16	< Day 1	Screening	-99
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	HDL-C (mmol/L)	HDL	8002		10/01/2008	1	Day 1	Baseline	0
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	HDL-C (mmol/L)	HDL	8002		10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	HDL-C (mmol/L)	HDL	8002		10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	HDL-C (mmol/L)	HDL	8002		10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	HDL-C (mmol/L)	HDL	8002		10/28/2008	28	>= 22 Days	Week 4	4
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL/HDL ratio	LDLHDL	8003	DERIVED	9/15/2008	-16	< Day 1	Screening	-99
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL/HDL ratio	LDLHDL	8003	DERIVED	10/01/2008	1	Day 1	Baseline	0
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL/HDL ratio	LDLHDL	8003	DERIVED	10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL/HDL ratio	LDLHDL	8003	DERIVED	10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL/HDL ratio	LDLHDL	8003	DERIVED	10/15/2008	15	2-21 Days	Week 2	2
1A	1A-4_02	02	4	PLACEBO	1	PLACEBO	1	LDL/HDL ratio	LDLHDL	8003	DERIVED	10/28/2008	28	>= 22 Days	Week 4	4

Table 2: ADLP Dataset (Part 2)

USUBJID	PARAMCD	AVISIT	AVAL	DTYPE	ABLFL	BASE	CHG	ANL01FL	FASPFL	SRCDOM	SRCVAR	SRCSEQ
1A-4_02	LDL	Screening	2.36			2.8	-0.44	Y	Y	LB	LBSTRESN	1001
1A-4_02	LDL	Baseline	2.8		Y	2.8	0.00	Y	Y	LB	LBSTRESN	1002
1A-4_02	LDL	Week 2	2.73			2.8	-0.07		Y	LB	LBSTRESN	1003
1A-4_02	LDL	Week 2	2.68			2.8	-0.12		Y	LB	LBSTRESN	1004
1A-4_02	LDL	Week 2	2.71	AVERAGE		2.8	-0.09	Y	Y			
1A-4_02	LDL	Week 4	2.74			2.8	-0.06	Y	Y	LB	LBSTRESN	1005
1A-4_02	HDL	Screening	1.06			1.08	-0.02	Y	Y	LB	LBSTRESN	1006
1A-4_02	HDL	Baseline	1.08		Y	1.08	0.00	Y	Y	LB	LBSTRESN	1007
1A-4_02	HDL	Week 2	1.12			1.08	0.04		Y	LB	LBSTRESN	1008
1A-4_02	HDL	Week 2	1.09			1.08	0.01		Y	LB	LBSTRESN	1009
1A-4_02	HDL	Week 2	1.11	AVERAGE		1.08	0.03	Y	Y			
1A-4_02	HDL	Week 4	1.15			1.08	0.07	Y	Y	LB	LBSTRESN	1010
1A-4_02	LDLHDL	Screening	2.226			2.593	-0.367	Y	Y			
1A-4_02	LDLHDL	Baseline	2.593		Y	2.593	0.000	Y	Y			
1A-4_02	LDLHDL	Week 2	2.438			2.593	-0.155		Y			
1A-4_02	LDLHDL	Week 2	2.459			2.593	-0.134		Y			
1A-4_02	LDLHDL	Week 2	2.448	AVERAGE		2.593	-0.145	Y	Y			
1A-4_02	LDLHDL	Week 4	2.383			2.593	-0.210	Y	Y			

ANALYSIS DATA SETUP

When planning the analysis data setup, first consider the number of ADaM datasets needed for a study. ADSL is required for all the studies and is usually set up first. Also, it is often optimal to have more than one BDS analysis dataset.

Although analyses are conventionally referred to as efficacy and safety analyses, ADaM does not support the concept of one big dataset for efficacy such as ADEFF, or one big dataset for safety such as ADSAF. Rather, ADaM datasets are specific to analysis endpoints. For example, ADLP is used for lipid endpoints; ADGL for glucose endpoints; ADLB for regular LB safety endpoints; ADVS for vital sign endpoints; and ADCAH for hypoglycemia endpoints. Furthermore, one can set up analysis datasets by populations, or keep all populations in one dataset using different population flags or analysis flags. The challenges in planning ADaM datasets are when one endpoint can be for both efficacy and safety analysis, or when the balance between the number of datasets and convenience of analysis cannot be easily achieved. We cannot make one analysis dataset to fit all analyses, nor would we make one dataset per analysis.

The following are the setup steps using the ADLP data shown above.

Step 1. **Obtain Variables from Source SDTM LB Domain**

The ADaM variables for LDL measurements are obtained directly from the source SDTM LB domain. Lipid measurements for LDL-C collected in the LB domain are shown below. Due to space limitations, only records for LDL measurements are shown; HDL measurements are of a similar structure:

USUBJID	LBSEQ	LBTESTCD	LBTEST	LBSTRESN	LBSTRESU	VISIT	VISITNUM	LBDTC	LBDY
1A-4_02	1001	LDL	LDL Cholesterol	2.36	mmol/L	SCRNING	1	2008-09-15	-16
1A-4_02	1002	LDL	LDL Cholesterol	2.8	mmol/L	DAY 1	2	2008-10-01	1
1A-4_02	1003	LDL	LDL Cholesterol	2.73	mmol/L	DAY 15	3	2008-10-15	15
1A-4_02	1004	LDL	LDL Cholesterol	2.68	mmol/L	DAY 15	3	2008-10-15	15
1A-4_02	1005	LDL	LDL Cholesterol	2.74	mmol/L	DAY 28	4	2008-10-28	28

The ADaM variables for LDL measurements are shown below:

USUBJID	PARAM	AWRANGE	AVISIT	AVISITN	ADT	ADY	AVAL	SRCDOM	SRCVAR	SRCSEQ
1A-4_02	LDL-C (mmol/L)	< Day 1	SCREENING	-99	9/15/2008	-16	2.36	LB	LBSTRESN	1001
1A-4_02	LDL-C (mmol/L)	Day 1	Baseline	0	10/01/2008	1	2.8	LB	LBSTRESN	1002
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	15	2.73	LB	LBSTRESN	1003
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	15	2.68	LB	LBSTRESN	1004
1A-4_02	LDL-C (mmol/L)	>= 22 Days	Week 4	4	10/28/2008	28	2.74	LB	LBSTRESN	1005

PARAM is created to contain the parameter name with the unit for LDL endpoint. Numeric parameter name, PARAMN, and short name of the parameter can also be created to facilitate the analysis programming. AWRANGE and AVISIT are based on the analysis window defined in Statistical Analysis Plan (SAP). ADT directly corresponds to LBDTC and ADY directly corresponds to LBDY. AVAL is taken from the variable LBSTRESN for the corresponding record. For traceability purpose, the three variables SRCDOM, SRCVAR, and SRCSEQ are used to keep the source domain, source variables, and the source variable value of the records in SDTM.

Step 2. Derive New Analysis Endpoints (PARAMTYP)

Sometimes it is necessary to derive analysis endpoints based on the collected endpoints. For example, the HDL/LDL ratio can be derived as a new parameter. In this case, the PARAMTYP variable is added to indicate it is a derived parameter.

USUBJID	PARAM	AWRANGE	AVISIT	AVISITN	ADT	ADY	AVAL	PARAMTYP
1A-4_02	LDL/HDL ratio	< Day 1	SCREENING	-99	9/15/2008	-16	2.226	DERIVED
1A-4_02	LDL/HDL ratio	Day 1	Baseline	0	10/01/2008	1	2.593	DERIVED
1A-4_02	LDL/HDL ratio	2-21 Days	Week 2	2	10/15/2008	15	2.438	DERIVED
1A-4_02	LDL/HDL ratio	2-21 Days	Week 2	2	10/15/2008	15	2.459	DERIVED
1A-4_02	LDL/HDL ratio	>= 22 Days	Week 4	4	10/28/2008	28	2.383	DERIVED

Step 3. Handle Negatives (or under detection) and Multiple Records on the Same Date (DTYPE)

a. When there is more than one measurement on the same date, the average of the multiple measurements is taken to represent the measurement for that day as defined in the SAP. See the record highlighted in blue below. On 10/15/2008, there are two LDL measurements. One record was added as the average of the two measurements that occurred on 10/15/2008. In order to differentiate the derived records that have all the source variable values missing, the DTYPE variable is added to the ADLP dataset from the collected records.

USUBJID	PARAM	AWRANGE	AVISIT	AVISITN	ADT	AVAL	SRCDOM	SRCVAR	SRCSEQ	DTYPE
1A-4_02	LDL-C (mmol/L)	< Day 1	SCREENING	-99	9/15/2008	2.36	LB	LBSTRESN	1001	
1A-4_02	LDL-C (mmol/L)	Day 1	Baseline	0	10/01/2008	2.8	LB	LBSTRESN	1002	
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	2.73	LB	LBSTRESN	1003	
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	2.68	LB	LBSTRESN	1004	
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	2.71				AVERAGE
1A-4_02	LDL-C (mmol/L)	>= 22 Days	Week 4	4	10/28/2008	2.74	LB	LBSTRESN	1005	

b. DTYPE can also be used to indicate the derivation of a missing analysis value, for example, if Week 8 is another analysis time point not collected. Based on analysis need, the record for Week 8 can be derived using a method defined in the SAP such as LOCF (Last observation carried forward) or WOCF (worst observation carried forward). In this situation, the DTYPE variable will have values of 'LOCF' or 'WOCF', respectively.

Usually, negative lab values will be converted to positive values using some formula defined as per the SAP. For example, when the formula is AVAL=abs(negative values)/2, one more record will be added with DTYPE='ABS/2'.

Step 4. Set Analysis Flag Variables (ANLzzFL)

ANLzzFL is a character variable indicating whether a record was used for the zzth analysis or not. 'zz' represents an index for a record selection algorithm such as 01, 02, etc. The analysis flag shows the representative record for a specific analysis time point. For example, in the ADLP dataset above in Step 3, the highlighted record with DTYPE='AVERAGE' will have variable ANL01FL equal to 'Y'; the collected two records on 10/15/2008 highlighted in green will have a blank value for ANL01FL. Therefore the derived average record is the representative record of the Week 2 analysis. For SCREENING, BASELINE, and WEEK 4, since there is only one record for each time point, that record is automatically flagged as 'Y' for ANL01FL.

USUBJID	PARAM	AWRANGE	AVISIT	AVISITN	ADT	ADY	AVAL	DTYPE	ANL01FL
1A-4_02	LDL-C (mmol/L)	< Day 1	SCREENING	-99	9/15/2008	-16	2.36		Y
1A-4_02	LDL-C (mmol/L)	Day 1	Baseline	0	10/01/2008	1	2.8		Y
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	15	2.73		
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	15	2.68		
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	10/15/2008	15	2.71	AVERAGE	Y
1A-4_02	LDL-C (mmol/L)	>= 22 Days	Week 4	4	10/28/2008	28	2.74		Y

Step 5. Compute Change, Percent Change from Baseline (BASE, CHG, PCHG)

Once data for all the analysis time points are collected or derived, one can calculate the BASE value for each time point. To achieve the BASE value for each time point, take the baseline analysis time point and merge the value as BASE to each time point, then calculate change from baseline (CHG) as AVAL-BASE, and percent change from BASELINE(PCHG) as (AVAL-BASE)/BASE*100.

Step 6. Set Population Flag Variables

Population flags can have different levels. For example, at the patient level the Full Analysis Set Flag, FASFL, resides in ADSL since it is a subject level population flag. At the parameter level, the Full Analysis Set Param-Level Flag, FASPFL, has one unique value per parameter per patient. Finally, at the record level, the Per-Protocol Record-Level Flag, PPROTRFL has one unique value per record, i.e. among different records for the same parameter, some will be used and some will not be used per protocol analysis.

For example, a patient with USUBJID "1A-4_02" is considered as a FAS patient per SAP because the patient has at least one post baseline measurement for parameter "LDL-C (mmol/L)". When analysis is required for LDL at Week 2 for the FAS population, the subsetting criterion is used: where ANL01FL='Y' and FASPFL='Y' and AVISIT='Week 2" and Param=' LDL-C (mmol/L)', i.e., the row highlighted below, and the ADaM dataset below is ready for SAS® procedures for analysis.

USUBJID	PARAM	AWRANGE	AVISIT	AVISITN	AVAL	BASE	CHG	DTYPE	ANL01FL	FASPFL
1A-4_02	LDL-C (mmol/L)	< Day 1	SCREENING	-99	2.36	2.8			Y	Y
1A-4_02	LDL-C (mmol/L)	Day 1	Baseline	0	2.8	2.8			Y	Y
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	2.73	2.8	-0.07			Y
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	2.68	2.8	-0.12			Y
1A-4_02	LDL-C (mmol/L)	2-21 Days	Week 2	2	2.71	2.8	-0.09	AVERAGE	Y	Y
1A-4_02	LDL-C (mmol/L)	>= 22 Days	Week 4	4	2.74	2.8	-0.06		Y	Y

CONCLUSION

This paper provides a basic programming flow to implement CDISC Analysis Data Model Version 2.1. The program steps enable the creation of the ADaM Basic Data Structure (BDS), traceability between analysis data and source data. In addition, it enables the "one-proc" away readiness for analysis. Further development can be made to standardize the programs for analysis data setup.

REFERENCES

CDISC Analysis Data Model Version 2.1.

CDISC Analysis Data Model Implementation Guide Version 1.0

RECOMMEDNED READING

A taste of SDTM, Northeast SUG Proceedings 2009, Changhong Shi and Beilei Xu

ACKNOWLEDGEMENTS

The author would like to thank the management team for their review of this paper.

CONTACT INFORMATION

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

Beilei Xu
Merck Co. & Inc.
RY34-A320
P.O. Box 2000
Rahway, NJ 07065
(732)-594-9980
beilei_xu@merck.com

Changhong Shi
Merck Co. & Inc.
RY34-A320
P.O. Box 2000
Rahway, NJ 07065
(732)-594-1383
changhong_shi@merck.com