Insights into ADaM

Matthew Becker, PharmaNet, Cary, NC, United States

ABSTRACT

ADaM (Analysis Dataset Model) is meant to describe the data attributes such as structure, content, and metadata that are typically found in clinical trial analysis datasets. The ADaM models are built from the CDISC SDTM baseline. In this workshop, we will look into the latest DRAFT ADaM Implementation Guide (v1.0) that is out for comment. The goal is to provide the attendees knowledge of the ADaM model, how it relates to the CDISC SDTM (study data tabulation model) base, and how it may help in reducing FDA review time. Depending on the timing of release of the final ADaM 1.0 implementation guide, the workshop will incorporate that guide.

INTRODUCTION

In 2005, the Analysis Dataset Model (ADaM) v0.7 was released for comment. It was created by a large number of reviewers experienced in clinical trials and regulatory submissions. The ADaM model is based upon the nomenclature of CDISC SDTM domains with additional attributes, variables and data structures. The goal of the model is to be used as a guideline for the creation of analysis datasets that will be used to generate statistical analysis for submissions. In 2008, an updated implementation guide was released – ADaM IG v1.0.

Over the past year, many companies have instituted committees and/or task forces to make operational tasks more efficient by using standards. By instituting CDASH for standard case report forms and CDISC SDTM for our clinical database standard, it was obvious to us that having a standard analysis dataset structure was a natural progression. In order to take advantage of the standards that experienced personnel had developed, we have moved towards ADaM as our standard for analysis dataset content.

In this paper, we will (1) review the structure of the ADaM v1.0 models, (2) delve into the relationship between CDISC SDTM domains and ADaM and (3) discuss our experience in using ADaM as the structure for analysis datasets.

PRINCIPLES

The main principles of ADaM include:
1. "Standardize" delivery to regulatory agencies
2. Provide clear documentation of the content, source and quality of the analysis datasets
3. Provide clear documentation of the results of a clinical trial (statistical methods, transformations, assumptions, derivations, imputations)
4. Provide a "roadmap" of how metadata, programs and documentation translate the SAP to the statistical results
5. ADaM datasets are usable by current tools (e.g. SAS®)
6. Provide XML metadata for future analysis tool development
7. Analysis-ready or "one proc away."

ADaM model datasets are based upon the nomenclature of CDISC SDTM domains with additional metadata. Prior to having a standard structure for metadata, we took the CRF database and utilized it as the baseline for our analysis datasets. A typical diagram of data flow from source to analysis is shown in Figure 1. We can assume that previously the "Data Sources" were the CRF data structures defined within data management. Now, the "Data Sources" are assumed to be CDISC SDTM domains.
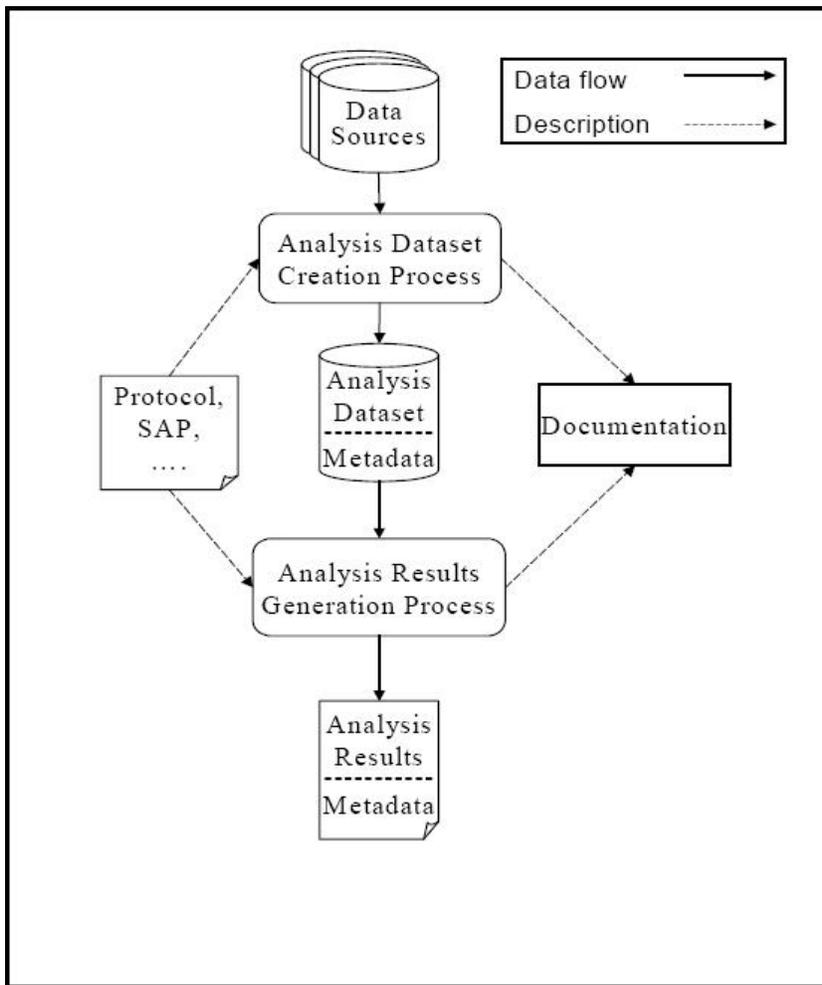
Figure 1. Data Flow From Source to Analysis

The **analysis dataset creation process** includes the steps taken to create our analysis datasets from the source data. In our case, the source data is a clinical database in SDTM format. The steps include definition of metadata at the variable and value level for the analysis datasets and programming to create the analysis dataset. Metadata at this step may also include links to specific sections of the protocol and/or SAP.

Analysis datasets generated from the analysis dataset creation process must:
1. Include a subject-level analysis dataset named ADSL
2. Consist of the optimum number of analysis datasets needed to generate the statistical results for a clinical trial with little or no programming or data manipulation
3. Must maintain the link with any base SDTM datasets used (i.e. SDTM variables carried into the analysis dataset must keep their attributes)

The **analysis results generation process** includes the steps taken to generate the analyses (e.g. tables, listings, figures). The steps include definition of the results metadata and the programming to create the required output. Metadata at this step may also include links to specific sections of the SAP and/or statistical appendix in the clinical study report.

SDTM RELATIONSHIP
CDISC SDTM is becoming the standard structure for collected CRF data. In contrast to analysis datasets, it contains a relatively low number of derived fields. Derivations such as baseline flags, study day and subject reference start date are a few of the common derived fields found in certain SDTM domains. As Biostatisticians and Statistical Programmers, we need to create a number of derived fields to perform our analysis.

In the early stages of CDISC SDTM, there was a pilot project to attempt to put all required derived fields for analysis into the SDTM structure. There were a number of issues that were reviewed and discussed. The resulting decision was to create a separate structure, the ADaM structure, as the final source for the analysis metadata.

But how are SDTM and ADaM used together? There must be a relationship between SDTM and ADaM. In order to support the documentation of traceability from analysis (statistical output) back to the analysis datasets back to the RAW data, a relationship is imperative. A number of SDTM variables are directly carried into the ADaM dataset(s) used for analysis. When this is done, the SDTM variable must not change –the variable name, the variable attributes nor the variable value. It must be copied without modification. ADaM builds on the nomenclature of SDTM. The main difference is that ADaM adds additional attributes and variables to conduct the statistical analyses.

Some of the differences between SDTM and ADaM include:
1. ADaM datasets may not always be vertical (especially in ADAMIG v0.7); less so in ADAMIG v1.0
2. ADaM datasets use redundancy for easy analysis – common variables are found across all analysis datasets (e.g. population flags, subject identifiers, etc.)
3. ADaM datasets have a greater number of numeric variables (e.g. SAS formatted dates, numeric representation of a character grouping variable from SDTM)
4. ADaM datasets may combine variables across multiple domains
5. ADaM datasets are named AD<xxxxxx>

METADATA COMPONENTS
The documentation of the analysis datasets provides a concise link from the CRF data to the analyses defined within the statistical analysis plan. To assist all pertinent parties for a clinical trial (e.g. internal project team members, client team members, agency reviewers), the sources of each analysis dataset should be clearly documented. Two key pieces of this documentation are the analysis dataset metadata (Figure 2) and the analysis variable metadata (Figure 3).

ANALYSIS DATASET METADATA

Figure 2. Example Analysis Dataset Metadata

Analysis dataset metadata provides key pieces of information describing each analysis dataset, including documentation and/or analysis dataset creation programs. Figure 2 contains an example of how analysis dataset metadata might appear. Key components of the analysis dataset metadata are:

1. Dataset: the dataset name – should always begin with "AD" as a prefix
2. Description: details on what the analysis dataset contains
3. Location: location of the dataset in the project directory (or submission path)
4. Structure: usually outlines if the analysis dataset is multiple records per subject or a single record per subject
5. Purpose: provides information about why the dataset was created and how it will be used
6. Key Variables: common variables that distinguish each record in the analysis dataset
7. Documentation: Defines the trace-back information on how this dataset was defined (i.e. algorithms used, analysis dataset creation program, etc).

ANALYSIS VARIABLE METADATA

Figure 3. Example Analysis Variable Metadata

Analysis variable metadata describes the variables within the analysis dataset. In addition, it may include links to relevant documentation. Common fields within the analysis variable metadata include:
1. Dataset: analysis dataset name
2. Variable Name: name of the variable
3. Variable Label: description of the variable
4. Variable Type: either Char or Num
5. Variable Length: length of the variable value
6. Decodes: format name and values (if applicable)
7. Origin: how this variable was derived or the source dataset.variable it came from
8. Role: variables role in the analysis (SELECTION, ANALYSIS, SUPPORT).

ANALYSIS RESULTS METADATA

Analysis Name	Description	Reason	Dataset	Documentation
Table 5.1 – Demographic Data (Safety)	Summary of demographic data for the Safety Set	Analysis pre-specified in SAP	ADSL select records where SAFETY=1	SAP Section XX
Table 5.2 – Demographic Data by Gender (Safety)	Summary of demographic data by gender for the Safety Set	Analysis pre-specified in SAP	ADSL select records where SAFETY=1	SAP Section XX

Table 1. Example Analysis Results Metadata

The analysis results metadata (Table 1) describes the attributes of the important analysis results for the study. It may include statistical statements for treatment effect, p-values, tables and/or figures. It provides metadata information in a standard format. The main goal of the analysis results metadata is to provide a link to reviewers from a result in the report to metadata describing the analysis, reason for the analysis, analysis dataset(s) and program(s) used.

Attributes of the analysis results metadata include:
1. Analysis Name: unique identifier for the analysis (table number, figure number, etc)
2. Documentation: text description of the analysis performed
3. Reason: why the analysis was performed
4. Dataset: name of the analysis dataset used
5. Program: programs using the analysis dataset defined above to produce this analysis
    a. Is NOT the program that created the analysis dataset

ADSL

The minimum requirement for ADaM is that a subject-level dataset named ADSL exist. The purpose of ADSL is to provide a single location for key information for each subject in the trial. ADSL is always a single record per subject dataset. Table 2 provides a list of required and common variables found in the ADSL dataset for the majority of clinical trials. *Please note that this is not a complete list of potential variables you may need in your trial's ADSL dataset.*

Variable Name	Variable Label	Type	Required/Conditional
Study Identifiers			
STUDYID	Study Identifier	Char	Required
USUBJID	Unique subject identifier	Char	Required
SITEID	Site identifier	Char	Required
Subject Demographics			
AGE	Age	Num	Required
SEX	Gender	Char	Required
RACE	Race	Char	Required
Population Indicators			

FASFL	Full Analysis Set Flag	Char	Conditionally Required
SAFFL	Safety Population Flag	Char	Conditionally Required
ITTFL	Intent-to-Treat Population Flag	Char	Conditionally Required
PPROTFL	Per-Protocol Population Flag	Char	Conditionally Required
COMPLFL	Completers Population Flag	Char	Conditionally Required
Treatment Variables			
ARM	Description of Planned Arm	Char	Required
TRTxP	Planned Treatment for Period x	Char	Required
TRTxA	Actual Treatment for Period x	Char	Conditionally Required
TRTSEQP	Planned Sequence of Treatments	Char	Conditionally Required
TRTSEQA	Actual Sequence of Treatments	Char	Conditionally Required
Trial Dates			
RANDDT	Date of Randomization	Num	Conditionally Required (Required in randomized trials)
TRTSTDT	Date of First Exposure to Treatment	Num	Required
TRTENDT	Date of Last Exposure to Treatment	Num	Required
TRTxSTDT	Date of First Exposure in Period x	Num	Conditionally Required
TRTxENDT	Date of Last Exposure in Period x	Num	Conditionally Required

Table 2. Common ADSL Variables

ADAM BASIC DATA STRUCTURE

The ADaM basic data structure is a multiple-record-per-subject dataset structure. The ADAMIG developers believe the majority of analysis can be done using analysis datasets in the ADaM basic data structure. It is a normalized design with one of more records per subject per analysis parameter per time point. While other variables may be created, the most important variables in the ADaM basic data structure are those describing the subject, the analysis field and the time point. Table 3 displays the groups into which ADaM variables can be categorized.

Variable Groups	Description
Subject Identifiers	Variables that uniquely identify a subject (e.g. USUBJID)
SDTM Identifiers	Variables from the RAW SDTM structure that can be used to trace data in the analysis dataset back to SDTM (e.g. –SEQ, VISIT)
ADaM Timing Identifiers	Variables describing the time point/timing with respect to the analysis parameter (e.g. AVISIT, ADY)
ADaM Parameter Identifiers	Variables describing the analysis field (e.g. PARAM, PARAMCD)
ADaM Analysis Values	Variables containing the analysis data for the analysis parameter (e.g. AVAL, AVALN)
Analysis Enabling Variables	Variables required for the statistical analysis [subset, grouping, etc.] (e.g. SAFFL, TRTP, ACAT)
Supportive Variables	Variables provided to support the traceability from the analysis dataset back to SDTM

Table 3. ADaM Variable Grouping

For a full list of required, conditionally required and permissible variables in the ADaM basic data structure, please read the "CDISC ADaM Implementation Guide v1.0." Below, we will be summarizing some of the key variables commonly seen.

VARIABLE STANDARDS

Similar to the variable metadata for past submissions, all ADaM variable names must be less than or equal to 8 characters in length, begin with a letter or underscore and be comprised of letters, underscores and digits. All ADaM variable labels must be less than or equal to 40 characters in length.

SUBJECT IDENTIFIER AND TREATMENT

All ADaM datasets must contain the SDTM STUDYID and USUBJID variables. This is a MINIMUM requirement. Additionally, the SDTM SITEID and SUBJID may be included.

The planned treatment for a period must be included (TRTxP, where x is the period). For a single period randomized trial, the variable would be named TRT1P. In addition, the numeric representation of the treatment must also be included (TRTxPN). Conditionally required variables include the actual treatment for period x (TRTxA, TRTxAN).

TIMING VARIABLES

Any SDTM timing variables should be carried into the ADaM datasets if they are used for analysis or would help in traceability (e.g. EPOCH, --DTC, --DY, VISITNUM, VISIT). Character dates and/or times carried over from SDTM should be also be reproduced in a numeric date/time variable with an appropriate SAS format.

ANALYSIS PARAMETER

One of the key pieces of data in our analysis datasets are variables that define the analysis parameter. Required variables in ADaM include:

1. PARAM – Parameter Description: description of the analysis parameter (e.g. "Supine Systolic Blood Pressure (mmHg)")
2. PARAMCD – Parameter Code: short name of the analysis parameter in PARAM (e.g. SYSBP)
3. AVAL – Analysis Value: numeric analysis value described by PARAM
4. AVALC – Character Analysis Value: character analysis value described by PARAM

Other conditionally required variables seen are BASE/BASEC (baseline value for PARAM), CHG/CHGC (change from baseline for PARAM), PCHG (percent change from baseline).

ANALYSIS DESCRIPTOR

Analysis descriptor variables are conditionally required and document target days, relative ranges, and imputed/derived methods. DTYPE (derivation type) could hold values such as "LOCF," "WOCF," "AVERAGE" and is used to denote when the analysis value (AVAL/AVALC) has been derived or imputed from another record.

CATEGORICAL VARIABLES

The following are permissible categorical analysis variables for analysis datasets:

1. ACAT – Analysis Category: Categorical representation of AVAL (e.g. may be used to categorize AVAL into 'Low' 'High' 'Normal' for a lab value in a range)
2. CRIT – Analysis criterion: A string identifying criterion (e.g. SYSBP > 90)
3. CRITx – Analysis Criterion x: If more than one criterion for a parameter to evaluate

INDICATOR VARIABLES

Conditionally required indicator variables on analysis datasets include population flags, imputation flags, baseline record flags, etc. Some examples of common indicator variables in ADaM are:

1. ABLFL – Baseline Record Flag: character indicator to identify the baseline record for a parameter [Y/N/null]
2. ITTFL – Intent-to-Treat Population Flag: character indicator to identify whether the subject is in the ITT population [Y/N]
3. ITTFN – Intent-to-Treat Population Flag (numeric): numeric indicator to identify whether the subject is in the ITT population [1=Yes, 0=No]

SUPPORTIVE VARIABLES

Any variable from SDTM that will help with traceability from the analysis dataset back to the RAW SDTM data should be included. Common variables from SDTM that support this traceability include –SEQ, VISIT and VISITNUM.

EXPERIENCE WITH ADAM

Our experience with ADaM has been increasing over the past year. With more clients requesting SDTM as the CRF data structure, we have moved towards standardizing our analysis dataset structure as well. Utilizing ADaM, we have been able to take the time and thought of others to use that structure as our standard. Doing so has not come without its share of issues.

When we first set-up a study, we create an Excel file containing the metadata for our analysis datasets. This metadata includes columns for the derived dataset name, variable name, variable label, variable length, variable type, format, controlled term, and description. An example of our Excel DDT is shown in Figure 4.

Figure 4. Example of Internal DDT

As you can see, the description column contains either the SDTM domain.variable that is copied over or the algorithm used to derive that field.

One of the key items to keep in mind in moving to ADaM is the learning curve. Specifically with ADAMIG v1.0, there will be time spent developing code to fit your SDTM data to the new standards. The goal is to set-up standard programs that will take your SDTM structure into the minimum ADaM ADSL or ADaM basic data structure. For some, the structures defined in the ADaM implementation guide may be quite daunting.

Another experience we have had is entirely dependent upon your client. A client required that ADSL contain every piece of key information for each subject. Any variable that defined that subject had to be put on ADSL, including overall rating values for a standard test. In this case, our ADSL had over 100 variables and became quite unwieldy to view in the SAS Viewer.

We had both positive and negative comments on the vertical structure for the ADaM basic data model datasets. For statistical programmers well versed in SAS, whether the analysis dataset was a horizontal or vertical structure made no difference. Writing code doing summary statistics "by" the PARAMCD and timing variable were trivial. However, for QC programmers not as well versed in SAS, they preferred a horizontal analysis dataset so their code was more obvious and simple.

As mentioned above, a largely positive experience has been the standardization of code. In the long run, being able to have standard programs that will take an SDTM AE domain and create the minimum required AE analysis dataset in the ADam basic data structure will help in efficiency. Moreover, if we can standardize to a certain level, we

8

envision standard code that will generate standard AE tables. In a simple view, we regard the goal as: standard input -> standard analysis dataset -> standard table -> standard QC programs/processes.

Lastly, we can use our Excel file from Figure 5 to generate define.pdf. We have even begun creating an Excel file template that incorporates all the required ADSL and ADaM basic data structure variables.

CONCLUSION

With the DRAFT ADaM v1.0, the CDISC ADaM group has defined the variables expected in ADSL and the majority of other analysis datasets. At PharmaNet, we are using this documentation in coordination with CDASH and CDISC SDTM to begin standardizing our processes from data collection to analysis dataset structures. While we are still in the process of creating standard programs to support the process, we have become more experienced in the use of ADaM derived datasets and pitfalls to avoid.

For companies considering standardization, it makes sense to use the thoughts, processes and documentation created by industry experts in your review. With clinical trials moving towards standard CRFs and standard CRF data domains, having a standard analysis data structure makes sense. The ADaM team has taken the time to document an analysis dataset standard for us, and it should definitely be considered for your efficiency committees.

REFERENCES
1. "Analysis Data Model Version 2.1", CDISC Analysis Dataset Modeling Team, 18DEC2007
2. "ADaM Implementation Guide Version 1.0", CDISC ADaM Team, 30MAY2008
3. "Analysis Data Model (ADaM)", Annamaria Muraro, Helsinn Healthcare, CDISC Italian User Group 2007
4. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration

RECOMMENDED READING
1. "Analysis Data Model: Version 2.1", CDISC Analysis Dataset Modeling Team, 18DEC2007
2. "ADaM Implementation Guide Version 1.0", CDISC ADaM Team, 30MAY2008

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:
>Matthew Becker, Sr. Director, Statistical Programming
>PharmaNet
>1000 Centregreen Way Suite 300
>Cary, NC United States 27513
>Work Phone: 00 + 1 + 919 388 4031
>Fax: 00 + 1 + 609 452 5537
>Email: mbecker@pharmanet.com
>Web: http://www.pharmanet.com

Brand and product names are trademarks of their respective companies.