PharmaSUG2011 - Paper CD19

Challenges in Implementing ADaM datasets: Balancing the Analysis-Ready and Traceability Concepts

Pushpa Saranadasa, Merck & Co., Inc.

ABSTRACT:
According to the ADaM Implementation Guide Version 1.0, analysis datasets ought to adhere to certain fundamental principles. One such principle is "Analysis datasets should have a structure and content that allows statistical analyses to be performed with minimal programming. Such datasets are described as 'analysis-ready'." Another fundamental principle of the ADaM standards is "Traceability ". This standard is of great importance in ADaM since traceability facilitates transparency, which is an essential component in building confidence in a result or conclusion. A CDISC-compliant submission includes both SDTM and ADaM datasets; therefore, it follows that the relationship between SDTM and ADaM must be clear. This highlights the importance of traceability between the input data (SDTM) and the analyzed data (ADaM). This paper explains the challenges encountered when applying these two principles, e.g., analysis-ready data and traceability, in real life clinical data.

1. Introduction:

The Clinical Data Interchange Standard Consortium (CDISC) has published standards for the organization, structure, and content of clinical data. The Study Data Tabulation Model (SDTM) specifies the structure and metadata for collected data that are to be submitted as a part of an application to regulatory authorities. Analysis (ADaM) datasets, also submitted in support of a drug application, allow statistical reviewers to identify, understand, replicate, explore and confirm the analyses performed and submitted by the sponsor. In many instances analysis datasets should be "one proc away" from the statistical results. The method of implementing CDISC standards is to create ADaM datasets from SDTM domains.

Importantly, according to CDISC standards, analysis datasets and associated metadata must provide traceability to allow an understanding of the path to creation of an analysis value. The metadata should also identify when analysis data have been derived or imputed. Traceability in ADaM permits the understanding of the relationship among the analysis results, the analysis datasets, and the SDTM domain.

Any programmers that have supported clinical trial analysis and reporting understand the complexity of derived datasets that are ready to produce tables and graphs for efficacy and safety analysis. Taking data from several domains, many-to-many merging, transposing data, imputing dates and values, averaging within visits for duplicate results, and creating intermediate datasets for the final ADaM are more complex than we expect.

Though the concept of traceability is very important and appealing, practical applications of transparency with ADaM datasets may lead to many difficulties when dealing with real-life clinical data. This paper explains the challenges in the creation of "analysis-ready" datasets and the associated metadata and illustrates traceability of the "analysis result". In the context of CDISC standards and their implication, the primary objective of this paper is to explain how and where traceability is lacking while we try to maintain the analysis-ready concept. A second aim of the paper is to suggest solutions to keep both principles in harmony.

2. SDTM to ADaM then to Analysis Result: Traceability and One PROC Away

2.1 Definitions

2.1.1 Analysis-Ready Data: The scientific and medical objectives of a clinical trial determine the design of analysis datasets. Analysis data sets should strive to be "One Proc Away" from the statistical result. Creating the analysis dataset program is based on the analysis plans and dataset specifications. The data going into the program are the source data (i.e., SDTM and/or other analysis datasets), and the output is the analysis dataset. Analysis datasets are such that the analysis results can be produced without further data manipulation.

2.1.2 Traceability: The property of traceability means that the genealogy of the relationship between the source datasets (SDTM) and analysis datasets (ADaM) and from there to the results is transparent to the end user of the study.

2.1.3. Statistical Review Aid User Guide: Our company prepares this user guide to assist the reviewer in understanding the directory structure, the organization of the analysis datasets and programs contained in the Statistical Review Aid, pointers to where programs, logs and results may be found, and the step-by-step instructions for how to install the SRA in a local environment and run SAS Analysis programs. The User uide should include the following sections: 1. Introduction; 2. Directory Structure; 3. Datasets; 4. SAS programs; 5. How to Run the SAS Analysis Programs.

2.2 Process flow for Creating Final Analysis Results: The typical flow is that once the Statistical Analysis Plan (SAP) is approved, the table package is created based on the SAP. Using the table package information, the decision is made as to what variables should be included in the datasets. The dataset specifications are then written. The ADaM datasets are then created following the CDISC standards. The process flow is very efficient. Dataset specifications are the key in providing traceability back to the STDM data from ADaM. The specifications identify when and how records should be imputed, what manipulations should be done to capture the correct observations for desired populations, etc.

A flow diagram for the process would look like the following:

SAP → Table package → data specification → ADaM data set creation → Reports Creation

2.3 Paper Objective: The primary objective of this paper is to explain how and where traceability is lacking while we try to maintain the analysis ready concept. A second aim of the paper is to suggest solutions to keep both principals in harmony. Due to the time of the presentation only two examples are used to explain the scenarios.

Example 1:
This example shows the tables that need to be created for a Clinical Statistical Report for a cardiovascular study. The study is a parallel study with two phases. In the second phase one third of the patients were switched to another combination of the treatment. This is a common table and most programmers will have already seen this.

In a typical ADaM setting the input data for this table would be taken from ADLB and ADAE datasets, which include records per patient per parameter per time phase. If we are to harmonize the traceability and analysis-ready concepts then the dataset should contain the phasing information and indicator variables in order to capture the correct observations for consecutive elevations while adding observations with the variable "DTYPE". The result is that the " ADLB ' dataset has both the analysis-ready and traceability concepts. When you look at the bottom part of the table you see that the relevant information is coming from ADAE data.

Problem: As with the ADLB dataset , we need to repeat the same process used to create ADAE dataset. With AE domain it was pretty cumbersome to do the same process that was used for ADLB. Therefore we decided to create a dataset that has some of the characteristics in the ADaM data model, but not all. In creation of the ADAE dataset there is no SEQ information since we had to remove a chunk of patient information when they did not fall into any phasing.

Table 1

Summary of Incidence of Adverse Events Through 6 Weeks (Phase 1)
(All-Patients-as-Treated Population)

	Number (%) of AE		Difference in Proportions of AE (TRT1 minus TRT2)
	TRT1 m/n (%)	TRT2 m/n (%)	Difference (95% CI) [†]
ALT and /or AST			
≥ 3xULN, consecutive [‡]	x/xxx (x.x)	x/xxx (x.x)	x.x (x.x, x.x)
≥ 5xULN, consecutive [‡]	x/xxx (x.x)	x/xxx (x.x)	x.x (x.x, x.x)
≥ 10xULN, consecutive [‡]	x/xxx (x.x)	x/xxx (x.x)	x.x (x.x, x.x)
CK			
≥ 10xULN	x/xxx (x.x)	x/xxx (x.x)	x.x (x.x, x.x)
≥ 10xULN with muscle symptoms	x/xxx (x.x)	x/xxx (x.x)	x.x (x.x, x.x)
≥ 10xULN with muscle symptoms that are considered drug-related	x/xxx (x.x)	x/xxx (x.x)	x.x (-x.x, x.x)
Potential Hy's Law Condition [§]	x/xxx (x.x)	x/xxx (x.x)	x.x (x.x, x.x)
Hepatitis-related AEs	x/yyy (x.x)	x/yyy (x.x)	x.x (x.x, x.x)
Gallbladder-related AEs	x/yyy (x.x)	x/yyy (x.x)	x.x (x.x, x.x)
Gastrointestinal-related AEs	x/yyy (x.x)	x/yyy (x.x)	x.x (x.x, x.x)
Allergic reaction or rash AEs	x/yyy (x.x)	x/yyy (x.x)	x.x (x.x, x.x)

%=m/n x 100 = (number of patients within the adverse experience category / number of treated patients) x 100.

For laboratory safety (ALT, AST, CK, Potential Hy's Law Condition), patients must have taken at least one dose of study medication and have at least one postbaseline measurement within 14 days of the last dose of study therapy to be included in the analysis.

[†] Confidence intervals were calculated using the Miettianen and Nurminen method.

[‡] This category includes those patients with (a) two consecutive measurements ≥3xULN,(b) a single, last measurement ≥3xULN, or (c) a measurement ≥3xULN followed by a measurement <3xULN that was taken more than 2 days after the last dose of study medication. For ≥5xULN, consecutive and ≥10xULN, consecutive,substitute ≥5 and ≥10 in the above definition, respectively.

[§] ALT or AST elevations ≥3xULN, with total bilirubin >2xULN. Criteria will be confirmed by alkaline phosphatase measurements and clinical review of medical history and concomitant medications.

Some programming work is still necessary due to the statistical results reported in the table. In order to use the appropriate statistical method the proportion of patients who had a particular event as well as those who did not have any events also needed to be produced. It is the latter group that is the problem as they are not an intrinsic part of the dataset. As we do these data manipulations in the analysis program and produce the analysis result we lose the analysis-ready concept.

With the ADLB the problem is one of unnecessary variables being included in the dataset. The problem is magnified when the study has two phases. In our study not only did we have to produce reports for phase 1 and phase 2, but also reports for the whole treatment period. When phasing is changed (for example when a patient is discontinued) the definition of day range within the phases is also changed. This adds more indicator variables, lead us to insert more observations for each patient for each phase.

Solution: The solution is to create intermediate datasets necessary to produce analysis results and not pay too much attention to traceability in the one to one mapping from STDM data to ADaM. The trick is to take advantage of existing tools and processes. Some companies have their own macro tools to create standard tables such as AE and concomitant medications. Some tools are designed in such a way that they create an intermediate dataset with all the necessary counts. In our company we have macro program tools that create "count datasets" for the desired subsets. For example the count dataset for body system shown below as Table 2 has all the required information needed to produce the analysis result. The input dataset for the "count dataset" are STDM domains and ADSL subject level datasets. Merging with ADSL is necessary to get the proper analysis population and day ranges. But this is an intermediate data set and does not comply with the ADaM data structure, but it is definitely a very efficient approach. All the counts related to AE can be taken directly from the intermediate datasets. The nice feature is that it can be used to get the number of patients that are not experiencing the particular event in one data step.

Table 2

EXAMPARM	TRT1	TOT_N1	S1	TRT2	TOT_N2	S2
Patients in population	Placebo	120	120	MK020	100	100
With one or more adverse events	Placebo	120	50	MK020	100	65
with no adverse events	Placebo	120	70	MK020	100	35
Gastrointestinal disorder	Placebo	120	10	MK020	100	15
Cardiac disorder	Placebo	120	16	MK020	100	30

Using count datasets to produce the desired analysis results give us a huge efficiency gain in every way and maintains the "analysis -ready" concept. The question may arise about traceability. How can we communicate with the reviewer the role of "count datasets". That is where we can take advantage of section 3 in the SRA user guide where we explain how the input datasets were derived for analysis results.

In this analysis exposure adjusted AE rates and percentage of patients were calculated for special interest AEs and displays the number of patients that are included in each of the associated categories. Clinicians sometimes identify other "defined AE" events known as special interest AEs; i.e., those adverse events which are not usually a part of the standard AE dataset. For example, a special interest AE table may require information from domains not of direct concern to the study. These special interest AE events also can be categorized as Tier 1 AE's; some of these AE's are comprised of one or more associated incidents defined in the protocol.

Problem: Creating the ADaM dataset that is needed for the analysis table was somewhat complex due to the use of many domains. In Table 2 above, "With Diabetes" is an event of interest and event terms are then listed as associated event terms for the main event. In this table the event terms are "anti diabetes medication" taken from the CM (Concomitant Medications) domain, OR "elevations of fasting glucose" taken from the LB (Lab) domain AND "any AE‘s related to diabetes" (these terms should be defined in the SAP) taken from the AE domain. Combining associated event terms from different domains need many to many merging for some special AE's. Capturing the first event occurrence for each unique associated term and each special interest AE for each patient and calculated AE rates requires considerable data manipulation.

Example 2: The table below is a table that also needs to be created for a CSR.

Table 3

Number (%) of Patients with Special Interest AE 's by Treatment Group
All Treated Patients

	Number (%) of AE		Difference in Proportion of AE		AE Rates Per Patient-years of Exposure‖	
	TRT1	TRT2	TRT1 minus TRT2		TRT1	TRT2
	XXX	XXX	Diff. (95% CI†)	p-value‡		
With Diabetes	xx(x.x)	xx(x.x)	x.x(x.x,x.x)	.xxx	.xxx	.xxx
Anti-diabetic medications	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			
Elevations of Fasting Glucose	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			
Diabetes mellitus	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			
type 2 diabetes mellitus	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			
With worsening Glucose Tolerances	xx(x.x)	xx(x.x)	x.x(x.x,x.x)	.xxx	.xxx	.xxx
Fasting Glucose >100	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			
Consecutive Elevations of Fasting Glucose	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			
Hyperglycemia	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			
Blood Glucose increased	xx(x.x)	xx(x.x)	x.x(x.x,x.x)			

†CI = Confidence intervals. Calculated using a method based on Wilson's score method.

‡ p-values are from Fisher's Exact test.

‖ AE rates per patient-years of exposure = (number of patients with AE/sum of days at risk for AE) x 365.25days/year.

The dataset being created is no longer able to keep all the observations related to patients who finally end up in the ADaM data. There will be some instances in which domain information is completely dropped for some patients due to the way special interests were defined. In this sort of set-up, when you derive a new parameter from another parameter, it is no longer possible to trace back to the source STDM domains. In this case our main goal is to have an ADaM dataset that adheres to the "analysis-ready" concept.
Solution: Do not create an intermediate dataset. It is not efficient and doesn't serve our purpose. It inevitably leads to massive duplication and unnecessarily increases the complexity of the datasets. For the example in Table 2 we considered our analysis data specification to be our metadata. It contains thorough descriptions regarding the domains that should be used to capture the special interest AE's and how to derive the number of events per patient. We followed a different approach. Instead of tracing back to STDM data, we traced back to the data specification. As the time to submission approaches we can follow the same method we followed in Example 1. That is, include a clear specification document in section 3 of the Statistical Review Aid User Guide. It allows the reviewer to understand what the definitions are and what process was taken to create the Analysis dataset, which is ultimately used to produce the analysis result.

The final dataset with selected variables should appear as shown in Table 4.

Table 4

SUBJID	EVENT	EVNTDT	EVNTTERM	EVNTCODE	EVNTVALU	EVNTFL	TERMFL
101	Diabetes	01/04/10	Type 2 Diabetes mellitus	1	1	1	1
101	Diabetes	01/05/10	Diabetes mellitus	1	1	0	1
101	Diabetes	01/10/10	Increasing Glucose	1	1	0	1
101	Diabetes	01/12/10	Anti diabetes medication	1	1	0	1
101	Diabetes	02/12/10	Insulin resistant diabetes	1	1	0	1
102	Diabetes			1	0	1	0
103	Diabetes			1	0	1	0
104	Diabetes	05/09/10	Increasing Glucose	1	1	1	1
104	Diabetes	06/03/10	Insulin resistant diabetes	1	1	0	1
104	Diabetes	06/10/10	Latent autoimmune diabetes	1	1	0	1
101	FG Tolerances	03/08/10	FG >100	2	1	1	1
101	FG Tolerances	04/11/10	Hyperglycemia	2	1	0	1
102	FG Tolerances	06/20/10	Blood Glucose Increase	2	1	1	1
102	FG Tolerances	06/27/10	Hyperglycemia	2	1	0	1
103	FG Tolerances	01/22/10	Consecutive elevation of FG	2	1	1	1
103	FG Tolerances	02/02/10	Blood Glucose Increase	2	1	0	1
103	FG Tolerances	03/12/10	Hyperglycemia	2	1	0	1
104	FG Tolerances			2	0	1	0

3. Conclusion: There are clear benefits of adopting an industry standard for analysis datasets. Traceability and Analysis-Ready concepts are very important principles of the ADaM guidelines. In real life clinical data balancing both principles is very challenging except for the direct and simple cases which involve only one domain or the cases in which the analysis is straightforward.

In more complex studies the most important thing is to have a detailed and clearly written data specification. These data specifications ultimately serve as metadata. When the submission is ready, it is an easy task to cut and paste items from the data specs into Section 3 of the Statistical Review Aid User Guide. If more companies were to adopt the use of a Statistical Review Aid User Guide, then the reviewer's job will be much easier.

References

ADaM Implementation Guide Version 1.0 (V1.0)
Analysis Data Model Version 2.1

Acknowledgments

Thanks to John Troxell and Rob Hoffman for reviewing this work and providing constructive feedback.

AUTHOR CONTACT

Your comments and questions are valued and welcome. Contact the author at:
Pushpa Saranadasa
Phone: (267) 305-5385
E-mail: Pushpa_saranadasa@merck.com