PharmaSUG 2018 - Paper DS-17

Creating an ADaM Data Set for Correlation Analyses

Chad Melson, Experis Clinical, Cincinnati, OH

ABSTRACT

The purpose of a correlation analysis is to evaluate relationships between two variables, likely from two different types of measurements (e.g., blood pressure and lab results). As a result, correlation analyses usually require data from multiple source data sets. In this paper, I describe how an ADaM data set using the OTHER class structure (one of the four classes of ADaM data sets: ADSL, BDS, and OCCDS are the others) was constructed from multiple ADaM data sets to produce outputs displaying correlations of actual and change from baseline values at individual visits. Pros and cons of the non-normalized data set structure are also presented, which may give programmers ideas on how to create an ADaM data set for similar types of analyses.

INTRODUCTION

For a recent study that included multiple efficacy parameters, correlation analysis output was requested to quantify the association between the primary efficacy variable and multiple secondary and tertiary efficacy variables. The efficacy parameters were spread across multiple ADaM data sets created using the Basic Data Structure (BDS). Although the regulatory submission of this study was not to be CDISC compliant, we were following CDISC guidelines as closely as possible. To use the individual ADaM efficacy data sets for the correlation analysis would require merging of multiple data sets in the analysis program, which does not adhere to the ADaM fundamental principle of "analysis-ready" data sets. However, standard ADaM data sets structures, ADSL (Subject-Level Analysis Dataset), BDS, and OCCDS (Occurrence Data Structure), do not have the necessary structure and would require significant programming in the output program, which would again violate the fundamental principles. The solution that was implemented was an efficacy correlation data set created using the OTHER class structure.

Beyond presenting the structure and contents of this ADaM data set, all inputs, including the mock outputs and the source efficacy ADaM data sets, are presented. Derivations of the variables in the efficacy correlation ADaM data set are provided as well as annotations of the mock outputs based on this data set. The objective is to provide a complete picture of the process by showing how the data moves from the source data sets to the correlation analysis data set to the final outputs. All data used in this paper is dummy data.

REQUIREMENTS FOR CORRELATION DATA SET

Before an ADaM data set is created, it is necessary to understand what is being analyzed. Output shells (also referred to as mock outputs) are provided prior to programming to identify the variables that will be required to produce the desired output. Initially, tables presenting the correlation and p-values were requested by the team (Table 1). Eventually, they decided on forest plots that present the correlation and 95% confidence interval in tabular and graphical form (Figure 1). In addition, variables and time points were changed in the output specifications prior to the final analysis. The purpose in showing these different output types is to demonstrate different features of the correlation data set and to show that multiple output formats can be created from this data set.

MOCK CORRELATION TABLE

Table 1 displays the mock format of the correlation table. In this table, the correlation (and p-value) of the observed value of the primary efficacy variable with the observed value of each of the other efficacy outcomes are displayed. In this case, the primary efficacy variable and secondary efficacy variable A are collected at Baseline, Week 24, and Week 48, while the tertiary efficacy variables Y and Z are collected at Baseline and Week 48 only.

Table 1 Mock table for correlation analysis between primary efficacy endpoint and other efficacy endpoints based on observed data, intent-to-treat population

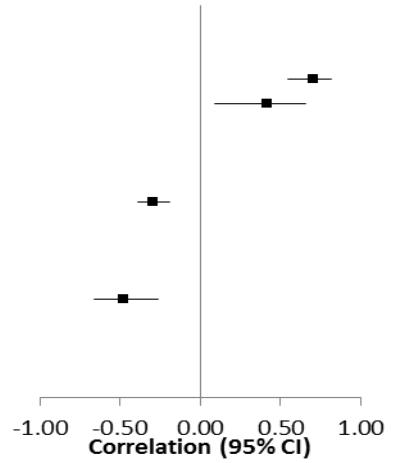
Outcome Visit	Primary Efficacy Variable		
	n	Correlation	p-value
Secondary Efficacy Variable A			
Baseline	xx	x.xx	x.xxx
Week 24	xx	x.xx	x.xxx
Week 48	xx	x.xx	x.xxx
Tertiary Efficacy Variable Y			
Baseline	xx	x.xx	x.xxx
Week 48	xx	x.xx	x.xxx
Tertiary Efficacy Variable Z			
Baseline	xx	x.xx	x.xxx
Week 48	xx	x.xx	x.xxx

MOCK CORRELATION FIGURE

Figure 1 displays the mock output of the correlation analysis in the form of a forest plot. In this figure, the correlation and the 95% confidence interval (CI) are displayed in tabular and graphical form. The correlations are calculated between the change from baseline value of the primary efficacy variable and the change from baseline value of each of the other efficacy outcomes (Variables A, Y, and Z). The substantive difference between the output in Figure 1 and the output in Table 1 (from an analysis perspective) is that change from baseline data is correlated instead of analysis values at the specified timepoint; therefore, no baseline row would be needed.

Figure 1 Mock figure for correlation analysis between change from baseline in primary efficacy endpoint and change from baseline in other efficacy endpoints based on observed data, intent-to-treat population

Outcome Visit	Primary Efficacy Variable	
	n	Correlation (95% CI)
Secondary Efficacy Variable A		
Week 24 Change from Baseline	xx	x.xx (x.xx, x.xx)
Week 48 Change from Baseline	xx	x.xx (x.xx, x.xx)
Tertiary Efficacy Variable Y		
Week 48 Change from Baseline	xx	x.xx (x.xx, x.xx)
Tertiary Efficacy Variable Z		
Week 48 Change from Baseline	xx	x.xx (x.xx, x.xx)



INPUT EFFICACY DATA SETS AND VARIABLES

The source efficacy data sets that are used to create the ADaM correlation data set in the example follow the BDS. The 3 input data sets, along with the efficacy parameters (i.e., PARAMCDs), are:

- ADEFPRIM Data Set - Primary efficacy parameter: PARAMCD = 'PRIMEFF'
- ADEFSEC Data Set - Secondary efficacy parameter A: PARAMCD = 'SECEFFA'
- ADEFTERT Data Set - Tertiary parameters Y and Z: PARAMCD in ('TERTEFFY', 'TERTEFFZ')

Table 2 displays an excerpt of the example primary efficacy dataset. The analysis 01 flag (ANL01FL) is set to 'Y' for AVISIT = 'Baseline' or the ADY closest to the target day for other AVISITs.

Table 2 Input Data Set (ADEFPRIM) containing the Primary Efficacy Variable

USUBJID	PARAMCD	AVISIT	ADY	DTYPE	BASE	AVAL	CHG	ANL01FL
001	PRIMEFF	Baseline	1		5.0	5.0		Y
001	PRIMEFF	Week 24	165		5.0	10.0	5.0	Y
001	PRIMEFF	Week 24	179		5.0	7.5	2.5	
001	PRIMEFF	Week 48		LOCF	5.0	7.5	2.5	Y
002	PRIMEFF	Baseline	1		7.2	7.2		Y
002	PRIMEFF	Week 24	168		7.2	8.1	0.9	Y
002	PRIMEFF	Week 48	334		7.2	6.1	-1.1	Y

For Subject 001, two observations exist for the Week 24 analysis visit. ANL01FL is set to 'Y' for the observation closest to the target visit day for Week 24 (Day 169). Also for this subject, there is no observed data in the Week 48 window, so the last value prior to that window is imputed for Week 48 and identified with DTYPE = 'LOCF'. Although the mock table and figure are based on observed data only, the LOCF observations will be considered in the Other Derivation Type - LOCF section later in this paper.

For Subject 002, all analysis visits have just one observation in each window, so ANL01FL is set to 'Y' for these observations. There are no LOCF observations needed since all analysis visits for the parameter have observed data.

The other ADaM input efficacy data sets (ADEFSEC and ADEFTERT) have a similar structure to the primary efficacy variable data set. Examples are shown in Appendix Table A and Appendix Table B, respectively, in the appendix. Although only 3 efficacy data sets are included as input in this example, you can include as many efficacy data sets and as many efficacy parameters as is needed.

CONTENTS AND STRUCTURE OF CORRELATION ANALYSIS DATA SET

The mock outputs require the following information in the correlation analysis data set:

- Efficacy responses for primary efficacy and other efficacy variables as specified
  - Raw outcome at baseline and post-baseline visits
  - Change from baseline values at post-baseline visits

The resulting efficacy correlation data set is called ADEFCORR. Table 3 presents the contents and format of this data set along with the derivations.

Table 3 Variables Included in ADaM Correlation Data Set (ADEFCORR)

Name	Label	Type	Derivation
AVISIT	Analysis Visit	Char	Set to Analysis visit [AVISIT] from the records in the Input Data Set selected to create the analysis variables (PRIMEFF, etc.). Each of the three efficacy data sets (ADEFPRIM, ADEFSEC, ADEFTERT) are transposed so that there is one record per USUBJID, AVISIT and DTYPE selection. Thus, the PARAMCD values for each become a variable that contains the analysis value or the change from baseline value depending on the ENDPOINT. Note that there will be two records per combination to capture analysis value (ENDPOINT = 'Raw') and to capture change from baseline (ENDPOINT = 'Change from Baseline'. After all analysis variables (PRIMEFF - TERTEFFZ) have been created, merge by USUBJID, AVISIT, ENDPOINT, and DTYPE. Exclude records where all analysis variables (PRIMEFF - TERTEFFZ) are missing.
DTYPE	Derivation Type	Char	Set to 'Observed' for records that represents the original (non-imputed) data [DTYPE is null] in the Input Data Set.
ENDPOINT	Endpoint	Char	Set to 'Raw' if origin of the analysis variable is AVAL. Set to 'Change from Baseline' if origin of the analysis variable is CHG.
PRIMEFF	Primary Efficacy Variable	Num	Subset ADEFPRIM where PARAMCD = 'PRIMEFF' and ANL01FL = 'Y' and DTYPE is null. For each USUBJID, AVISIT, DTYPE and ENDPOINT combination, assign accordingly: When ENDPOINT = 'Raw', set PRIMEFF = ADEFPRIM.AVAL When ENDPOINT = 'Change from Baseline', set PRIMEFF = ADEFPRIM.CHG
SECEFFA	Secondary Efficacy Variable A	Num	Subset ADEFSEC where PARAMCD = 'SECEFFA' and ANL01FL = 'Y' and DTYPE is null. For each USUBJID, AVISIT, DTYPE and ENDPOINT combination, assign accordingly: When ENDPOINT = 'Raw', set SECEFFA = ADEFSEC.AVAL When ENDPOINT = 'Change from Baseline', set SECEFFA = ADEFSEC.CHG

Name	Label	Type	Derivation
TERTEFFY	Tertiary Efficacy Variable Y	Num	Subset ADEFTERT where PARAMCD = 'TERTEFFY' and ANL01FL = 'Y' and DTYPE is null. For each USUBJID, AVISIT, DTYPE and ENDPOINT combination, assign accordingly: When ENDPOINT = 'Raw', set TERTEFFY = ADEFTERT.AVAL When ENDPOINT = 'Change from Baseline', TERTEFFY = ADEFTERT.CHG
TERTEFFZ	Tertiary Efficacy Variable Z	Num	Same as TERTEFFY except PARAMCD = 'TERTEFFZ.'

Each parameter from the input data set included in the correlation becomes a variable in the correlation data set. For example, the result (AVAL or CHG) for PARAMCD = 'PRIMEFF' from the input data set becomes the value for variable PRIMEFF in the correlation data set. In addition, variables identifying analysis visit, derivation type, and endpoint are included in the data set. Analysis visit comes directly from the input data set, while derivation type and endpoint (Raw and Change from Baseline) are set based on the type of observations (observed or LOCF) and result type of variables (AVAL or CHG) selected from the input data set.

Table 4 displays the resulting ADEFCORR observations for Subject 001 using observed data. Week 24 data for variables TERTEFFY and TERTEFFZ are missing because these parameters are not collected at Week 24. There are no PRIMEFF and SECEFFA values for Week 48 because this subject does not have any observed data for Week 48; the LOCF data for this visit will be discussed later. Even though the Week 48 observations will not be used in the outputs (since the primary efficacy variable is missing), these observations in the ADEFCORR should remain in the data set in case correlations are later requested for TRTEFFY vs TRTEFFZ. Only analysis visits that have values in at least one source data set are included in the correlation data set. If a visit has missing values for all efficacy variables, then this observation is removed from the correlation data set.

Table 4 USUBJID = 001 Observed Data for ADaM Correlation Data Set (ADEFCORR)

AVISIT	DTYPE	ENDPOINT	PRIMEFF	SECEFFA	TERTEFFY	TERTEFFZ
Baseline	Observed	Raw	5.0	71	1.73	6.3
Week 24	Observed	Raw	10.0	74		
Week 24	Observed	Change from Baseline	5.0	3		
Week 48	Observed	Raw			2.01	6.3
Week 48	Observed	Change from Baseline			0.28	0.0

Table 5 displays the resulting ADEFCORR observations for Subject 002 using observed data. Similar to Subject 001, Week 24 data for variables TERTEFFY and TERTEFFZ are missing because these parameters are not collected at Week 24.

Table 5 USUBJID = 002 Observed Data for ADaM Correlation Data Set (ADEFCORR)

AVISIT	DTYPE	ENDPOINT	PRIMEFF	SECEFFA	TERTEFFY	TERTEFFZ
Baseline	Observed	Raw	7.2	66	1.92	4.8
Week 24	Observed	Raw	8.1	68		
Week 24	Observed	Change from Baseline	0.9	2		
Week 48	Observed	Raw	6.1	65	1.89	7.2
Week 48	Observed	Change from Baseline	-1.1	-1	-0.03	2.4

ANNOTATED MOCK OUTPUTS

Next, annotations and programming details for Table 1 and Figure 1 are shown based on the correlation data set (ADEFCORR). These provide the specifications needed by the output programmer.

Table 6 Annotated mock table based on Table 1

Outcome Visit	Primary Efficacy Variable **PRIMEFF** n	Correlation, r	p-value
Secondary Efficacy Variable A **SECEFFA**			
Baseline	xx	x.xx	x.xxx
Week 24	xx	x.xx	x.xxx
Week 48	xx	x.xx	x.xxx
Tertiary Efficacy Variable Y **TERTEFFY**			
Baseline	xx	x.xx	x.xxx
Week 48	xx	x.xx	x.xxx
Tertiary Efficacy Variable Z **TERTEFFZ**			
Baseline	xx	x.xx	x.xxx
Week 48	xx	x.xx	x.xxx

Population: ITTFL = 'Y' (Variable ITTFL is not shown in sample data since all subjects have ITTFL = 'Y')
Selection clause for entire output: DTYPE = 'Observed'
Selection clause for visit rows:
Baseline: AVISIT = 'Baseline' and ENDPOINT = 'Raw'
Week 24: AVISIT = 'Week 24' and ENDPOINT = 'Raw'
Week 48: AVISIT = 'Week 48' and ENDPOINT = 'Raw'

For example, the full selection criteria for the Week 24 Visit in Table 6 is ITTFL = 'Y' and DTYPE = 'Observed' and AVISIT = 'Week 24' and ENDPOINT = 'Raw'.

The annotations for Figure 1 would be very similar to Table 6, except for the following selection clauses for visit rows:
Week 24 Change from Baseline: AVISIT = 'Week 24' and ENDPOINT = 'Change from Baseline'
Week 48 Change from Baseline: AVISIT = 'Week 48' and ENDPOINT = 'Change from Baseline'

For example, the full selection criteria for the Week 24 Visit in Figure 1 is ITTFL = 'Y' and DTYPE = 'Observed' and AVISIT = 'Week 24' and ENDPOINT = 'Change from Baseline'.

OTHER DERIVATION TYPE - LOCF

As mentioned, the structure of the mock outputs changed several times over the course of the study. Originally, there was a mock table exactly like Table 1, except that LOCF data was to be used instead of observed data. In order to create records in the correlation analysis data set that can be used for the LOCF analysis, both observed and LOCF data must be read from the input data set to create a complete set of observations for LOCF data in the correlation data set. Observed and LOCF data for analysis cannot be included on the same record in the correlation data set because we would need a variable on the record for each endpoint variable to indicate whether it was observed or LOCF in order to provide a straightforward selection clause. That is not practical, nor would it provide an "analysis-ready" data set.

Table 7 displays the resulting ADEFCORR observations for Subject 001 using LOCF data. The highlighted cells indicate "new" data that is included in the LOCF observations for this patient. There are still no Week 24 data for variables TERTEFFY and TERTEFFZ because these parameters are not collected at Week 24. It is not typical to impute to visits at which the data was not originally collected. All other efficacy data exactly matches the data in the DTYPE = 'Observed' observations. If the correlation data set includes both observed and LOCF records, data from Table 4 would be set with data from Table 7.

Table 7 USUBJID = 001 LOCF Data for ADaM Correlation Data Set (ADEFCORR)

AVISIT	DTYPE	ENDPOINT	PRIMEFF	SECEFFA	TERTEFFY	TERTEFFZ
Baseline	LOCF	Raw	5.0	71	1.73	6.3
Week 24	LOCF	Raw	10.0	74		
Week 24	LOCF	Change from Baseline	5.0	3		
Week 48	LOCF	Raw	7.5	74	2.01	6.3
Week 48	LOCF	Change from Baseline	2.5	3	0.28	0.0

There are two features of the creation of LOCF records that require additional discussion. First, the duplication of data in the resulting correlation data set may cause some apprehension. Admittedly, it is not optimal, but it does provide very simple selection criteria to produce the required outputs. The second item is the use of DTYPE = 'LOCF'. This is misleading since it implies that all results on the record are LOCF values. A better alternative would be to use a different variable name, perhaps ANLTYPE, so it is not misinterpreted. The syntax used to execute programs to create output with meaningful file names prevented us from using a different variable name in this study.

The LOCF data for Subject 002 would look exactly like the data shown in Table 5, except that the DTYPE variable would all be set to 'LOCF'.

CHANGES TO THE SPECIFICATIONS WHEN LOCF DATA ADDED

Addition of records to perform analysis of LOCF data would require changes to the variable derivations and the annotated mock outputs. If LOCF observations are included in the ADaM correlation data set, updates to the variable derivations previously shown would need to be made to DTYPE and each of the efficacy parameters.

- The following would need to be added to the DTYPE derivation:
  Set to 'LOCF' for records that combine the original (non-imputed) data and the data created based on the last observation carried forward method [DTYPE is null or 'LOCF'] from the Input Data Set.
- The derivation of efficacy parameters would be updated as follows (example shown for PRIMEFF):
  Subset ADEFPRIM where PARAMCD = 'PRIMEFF' and ANL01FL = 'Y' and specified DTYPE selection.
  For each USUBJID, AVISIT, DTYPE and ENDPOINT combination, assign accordingly:
  When ENDPOINT = 'Raw', set PRIMEFF = ADEFPRIM.AVAL
  When ENDPOINT = 'Change from Baseline', set PRIMEFF = ADEFPRIM.CHG

For the annotated output in Table 6, the only change would be to replace DTYPE = 'Observed' with DTYPE = 'LOCF'.

EVALUATING DATA SET AGAINST ADAM FUNDAMENTAL PRINCIPLES

Let's review the fundamental principles of ADaM data sets and evaluate whether the correlation data set meets them, as described in the Analysis Data Model document.

- Traceability: "The overall principle in designing analysis datasets and related metadata is that there must be clear and unambiguous communication of the content and source of the datasets supporting the statistical analyses performed in a clinical study."

  Evaluation: In simplest terms, the creation of the efficacy variables in the resulting correlation data set is a transpose of the AVAL or CHG variable by analysis visit from the source efficacy data set for observations based on derivation type and analysis flag. The variable metadata associated with ADEFCORR will explain the transformation and how each variable is populated and the source of each variable.

- Analysis Ready: "Sponsors should strive to submit 'analysis-ready' datasets, i.e., analysis datasets that have a structure and content that allows statistical analysis to be performed with minimal programming"

  Evaluation: Because the variables being correlated are on the same record, observations used in the analysis program can be created by using a WHERE clause. No merging of data sets is required.

- Metadata: "Metadata and other documentation should provide clear and concise communication of the analyses, including statistical methods, assumptions, derivations and imputations performed."

  Evaluation: The annotations provided for the tabular and graphical correlation outputs describe the data being used, while the variable metadata explains how the data set is created.

- Software: "Analysis datasets must be readily usable by commonly available software tools"

  Evaluation: ADEFCORR is a data set that is usable by SAS®, which is software that is readily available and commonly used.

CONCLUSION

Changes to the output specifications made during the course of the study provided an opportunity to test the robustness of this data structure. We did not have to change the general format of the efficacy correlation analysis data set when the team decided to change from a table output to a forest plot display. When the correlations outputs based on LOCF were removed, we simply removed the DTYPE = 'LOCF' observations from the data set. This had no impact on the correlations based on observed data. When efficacy variables or visits were added or removed, we only needed to modify the data sets and parameters (PARAMCDs) included in the program. It was helpful to have the variable names in the correlation data set match the PARAMCDs from the input efficacy data sets.

The fact that the variables to be correlated are on the same record would be beneficial if correlations other than the primary endpoint vs. other efficacy endpoints are needed. This could be completed with the existing correlation data set without any further modifications. However, this structure will not work if you want to correlate different endpoints (e.g., raw results with change from baseline results). In that case, to follow the structure presented in this paper, the variable names for the efficacy parameters would need to indicate whether raw or change from baseline is represented.

One of the other challenges was writing derivations for the correlation data set. This was especially true in the early stage of the project when the team was still discussing the format of the outputs. This is a case when starting early on the programming activities caused a significant amount of work on specifications that was not needed in the final version. At one point, there were rows on the correlation outputs for average of parameters across multiple time points (e.g., average value of Week 24 and Week 48). There were also correlations of percent change from baseline data as well. These complicated the variable derivations a great deal, as did the inclusion of LOCF data. The final data set specifications were more streamlined.

While the solution provided in this paper used the OTHER class, this structure should only be used after the existing structures (e.g., BDS, OCCDS) have been evaluated and deemed not appropriate. If in doubt about whether you should use the OTHER class structure, discuss with an ADaM consultant. For the correlation analysis ADaM data set, the OTHER class structure was the only alternative. While not perfect, the OTHER data structure provided the foundation for a robust solution for producing a variety of correlation analysis outputs.

REFERENCES

CDISC Analysis Data Model Team, 2009, "Analysis Data Model (ADaM) v2.1", available on CDISC website at <http://www.cdisc.org/standards/foundational/adam >

CDISC Analysis Data Model Team, 2009, "Analysis Data Model (ADaM) Implementation Guide v1.0", available on CDISC website at <http://www.cdisc.org/standards/foundational/adam >

ACKNOWLEDGMENTS

The author would like to acknowledge:

- Richann Watson and Oleh Kulyk for providing comments.

CONTACT INFORMATION

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

Chad Melson
Experis Clinical
4445 Lake Forest Drive, Suite 470
Blue Ash, OH 45242
Phone: 513-808-9078
chad.melson@experis.com

APPENDIX

INPUT DATA SET FOR SECONDARY EFFICACY PARAMETERS (ADEFSEC)

The following table displays the contents of the data set containing the secondary efficacy parameters.

Appendix Table A

USUBJID	PARAMCD	AVISIT	ADY	DTYPE	BASE	AVAL	CHG	ANL01FL
001	SECEFFA	Baseline	1		71	71		Y
001	SECEFFA	Week 24	165		71	74	3	Y
001	SECEFFA	Week 48		LOCF	71	74	3	Y
002	SECEFFA	Baseline	1		66	66		Y
002	SECEFFA	Week 24	168		66	68	2	Y
002	SECEFFA	Week 48	334		66	65	-1	Y

INPUT DATA SET FOR TERTIARY EFFICACY PARAMETERS (ADEFTERT)

The following table displays the contents of the data set containing the tertiary efficacy parameters.

Appendix Table B

USUBJID	PARAMCD	AVISIT	ADY	DTYPE	BASE	AVAL	CHG	ANL01FL
001	TERTEFFY	Baseline	0		1.73	1.73		Y
001	TERTEFFY	Week 48	340		1.73	2.01	0.28	Y
001	TERTEFFZ	Baseline	0		6.3	6.3		Y
001	TERTEFFZ	Week 48	340		6.3	6.3	0.0	Y
002	TERTEFFY	Baseline	0		1.92	1.92		Y
002	TERTEFFY	Week 48	334		1.92	1.89	-0.03	Y
002	TERTEFFZ	Baseline	0		4.8	4.8		Y
002	TERTEFFZ	Week 48	334		4.8	7.2	2.4	Y