PharmaSUG 2018 - Paper DS-01

Traceability: Some Thoughts and Examples for ADaM Needs

Sandra Minjoe, PRA Health Sciences; Wayne Zhong, Accretion Softworks;
Quan (Jenny) Zhou, Eli Lilly and Company; Kent Letourneau, PRA Health Sciences;
Richann Watson, DataRich Consulting

ABSTRACT

One of the fundamental principles of ADaM is that datasets and associated metadata must include traceability as a link between analysis results, ADaM datasets, and SDTM datasets. The existing ADaM documents contain some examples of simple traceability, such as variable derivations and inclusion of the SDTM sequence number, but what about more complex examples?

An ADaM sub-team is currently developing a Traceability Examples Document, showing how traceability can be employed in a wide variety of practical scenarios. Some of these examples contain content from other CDISC documents, modified to focus on the traceability aspects. Others are being developed specifically for the Traceability Examples Document. As members of the Traceability Examples ADaM sub-team, we are including in this PharmaSUG paper and presentation a selection of examples that demonstrate the power of traceability in complex analyses.

INTRODUCTION

Traceability is mentioned throughout the published ADaM documents. It is, in fact, one of the fundamental principles of ADaM. Section 2.1 of the CDISC ADaMIG v1.1 states “ADaM datasets and associated metadata must provide traceability to show the source or derivation of a value or variable (i.e., the data’s lineage or relationship between an analysis value and its predecessor(s)).”

ADaM documents include examples that demonstrate this principle of traceability. These examples don’t imply a rule or required process, just one possible way to handle specific analysis needs that abide by the fundamental principle of traceability. The ADaM Traceability Examples document, currently in development, focuses on examples that demonstrate traceability in both data and metadata.

At the time of this writing, there were fifteen examples being developed as part of the ADaM Traceability Examples document. In order to share with the community, even before the ADaM Traceability Examples document is published for public review, the following set of examples are described here in this paper by the ADaM authors who are developing them:

- General ADSL Traceability
- Traceability with Parameters from Multiple Input Datasets
- Traceability when Creating Rows in BDS
- Traceability When Multiple Analysis Variables are needed on the Same Row

To fully understand this paper, it is expected that the reader be fluent in the CDISC standards ADaM, SDTM, and Define-XML. Documents about these and other CDISC standards can be downloaded from the CDISC website https://www.cdisc.org/. See the Recommended Reading section for a list of CDISC standards documents most pertinent to this paper.

EXAMPLE 1: GENERAL ADSL TRACEABILITY

Common ADSL variables include variables copied from SDTM and derived within the ADSL dataset. This somewhat basic example includes variables copied unchanged from SDTM domains, as well as those derived. It is included here as a reminder and refresher before stepping into more complicated traceability needs.

DATA FLOW

Data for ADSL in this example is coming directly from SDTM domains DM, DS, and EX.

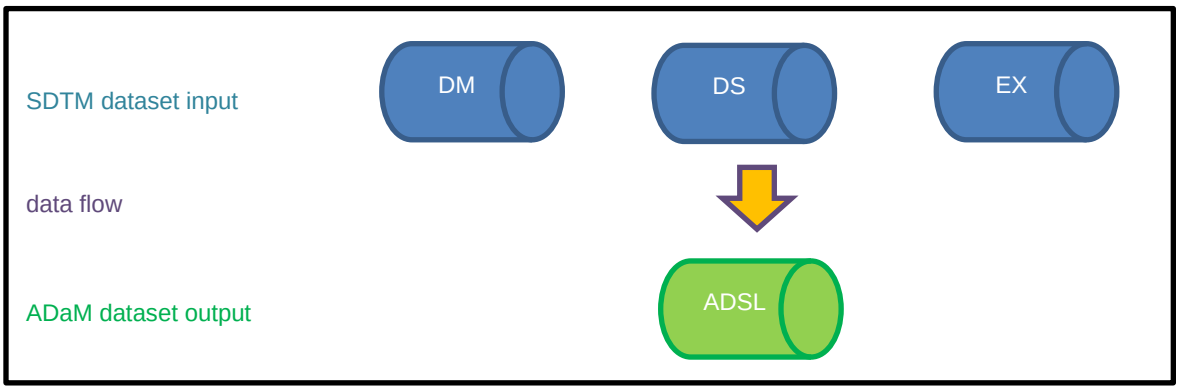

Figure 1: Example ADSL Data Flow

TRACEABILITY NEEDS

The following metadata table shows an example of variable-level traceability for each ADSL variable. In this example, AAGEGR1 was created to serve the analysis of grouping subjects by age categories. The traceability from AAGEGR1 all the way back to the SDTM variable DM.BRTHDTC is demonstrated with the following:

- Deriving AAGE as the actual age used for the grouping;
- Creating BRTHDT to calculate AAGE;
- Keeping the predecessor of DM.BRTHDTC to show how it was imputed to BRTHDT.

Table 1: Example ADSL Variable Metadata

Variable Name	Variable Metadata
STUDYID	Predecessor: DM.STUDYID
USUBJID	Predecessor: DM.USUBJID
SUBJID	Predecessor: DM.SUBJID
SITEID	Predecessor: DM.SITEID
SEX	Predecessor: DM.SEX
RACE	Predecessor: DM.RACE
AGE	Predecessor: DM.AGE
AGEU	Predecessor: DM.AGEU
BRTHDTC	Predecessor: DM.BRTHDTC
ARM	Predecessor: DM.ARM
ARMCD	Predecessor: DM.ARMCD
BRTHDT	Derived: Numeric version of DM.BRTHDTC. If only month and year are collected, impute day to 15; else if only year is collected, impute month to 07 and day to 01; else if missing, do not impute.

Variable Name	Variable Metadata
BRTHDTF	Derived: If only day is imputed, set to 'D'; else if both day and month are imputed, set to 'M'. Missing when no imputation is done.
RANDDT	Derived: Numeric version of DS.DSSTDTC when DS.DSTERM = 'RANDOMIZED'. If any part of the date is missing, do not impute.
AAGE	Derived: YRDIF(BRTHDT, RANDDT, 'AGE'). Missing if either date is missing.
AAGEGR1	Derived: If AAGE is missing then AAGEGR1 is missing; else if AAGE <41 then set to "< 41"; else if AAGE < 61 then set to "41-60"; else set to "61 or older".
TRTSEQP	Derived: Set to ADSL.TRT01P || " - " || ADSL.TRT02P. If ADSL.TRT02P is null, set to ADSL.TRT01P.
TRT01P	Derived: Set to the first component of DM.ARM before "-". Leave as null if DM.ARMCD = "SCRNFAIL" or "NOTASSGN".
TRT02P	Derived: If there are two components in DM.ARM separated by "-", set to the second component of DM.ARM. Otherwise, leave as null.
TRTSDT	Derived: Numeric version of the earliest EX.EXSTDTC. If any part of the date is missing, do not impute.
TRTEDT	Derived: Numeric version of the last EX.EXENDTC. If any part of the date is missing, do not impute.
TR01SDT	Derived: Numeric version of the earliest EX.EXSTDTC when EX.EPOCH= "DOUBLE-BLIND TREATMENT". If any part of the date is missing, do not impute.
TR01EDT	Derived: Numeric version of the last EX.EXENDTC when EX.EPOCH= "DOUBLE-BLIND TREATMENT". If any part of the date is missing, do not impute.
TR02SDT	Derived: Numeric version of the earliest EX.EXSTDTC when EX.EPOCH= "OPEN-LABEL TREATMENT". If any part of the date is missing, do not impute.
TR02EDT	Derived: Numeric version of the last EX.EXENDTC when EX.EPOCH= "OPEN-LABEL TREATMENT". If any part of the date is missing, do not impute.

INPUT AND ANALYSIS DATA

The following tables show examples of the input SDTM domains, including DM, DS, and EX. Only variables needed for illustration are included in the dataset examples.

Table 2: Input Data Example DM

Row	STUDYID	USUBJID	SUBJID	SITEID	SEX	RACE	AGE	AGEU	BRTHDTC	ARM	ARMCD
1	ABC123	ABC12301001	001	01	M	WHITE		YEARS	1958-12	Drug A – Drug B	AB
2	ABC123	ABC12301002	002	01	F	ASIAN	50	YEARS	1975-05-10	Placebo – Drug B	PB
3	ABC123	ABC12302003	003	02	M	WHITE	53	YEARS	1963-09-03	Drug A	A

Table 3: Input Data Example DS

Row	USUBJID	DSTERM	DSSTDTC
1	ABC12301001	RANDOMIZED	2016-05-17
2	ABC12301002	RANDOMIZED	2016-02-07

Row	USUBJID	DSTERM	DSSTDTC
3	ABC12302003	RANDOMIZED	2016-10-25

Table 4: Input Data Example EX

Row	USUBJID	EXSEQ	EXTRT	EXDOSE	EXDOSEU	EXSTDTC	EXENDTC	EPOCH
1	ABC12301001	1	Drug A	50	mg	2016-05-24	2016-07-22	DOUBLE-BLIND TREATMENT
2	ABC12301001	2	Drug B	100	mg	2016-08-01	2017-01-30	OPEN-LABEL TREATMENT
3	ABC12301002	1	Placebo	0	mg	2016-02-15	2016-04-16	DOUBLE-BLIND TREATMENT
4	ABC12301002	2	Drug B	100	mg	2016-04-25	2016-10-28	OPEN-LABEL TREATMENT
5	ABC12302003	1	Drug A	50	mg	2016-11-01	2016-11-29	DOUBLE-BLIND TREATMENT

Table 5: Output Data Example ADSL

Row	STUDYID	USUBJID	SUBJID	SITEID	SEX	RACE	AGE	AGEU	BRTHDTC	ARM	ARMCD
1	ABC123	ABC12301001	001	01	M	WHITE		YEARS	1958-12	Drug A – Drug B	AB
2	ABC123	ABC12301002	002	01	F	ASIAN	40	YEARS	1975-05-10	Placebo – Drug B	PB
3	ABC123	ABC12302003	003	02	M	WHITE	53	YEARS	1963-09-03	Drug A	A

Row	BRTHDT	BRTHDTF	RANDDT	AAGE	AAGEGR1	TRTSEQP	TRT01P	TRT02P
1(cont)	15DEC1958	D	17MAY2016	57	41-60	Drug A – Drug B	Drug A	Drug B
2(cont)	10MAY1975		07FEB2016	50	<41	Placebo – Drug B	Placebo	Drug B
3(cont)	03SEP1963		25OCT2016	53	41-60	Drug A	Drug A	

Row	TRTSDT	TRTEDT	TR01SDT	TR01EDT	TR02SDT	TR02EDT
1(cont)	24MAY2016	30JAN2017	24MAY2016	22JUL2016	01AUG2016	30JAN2017
2(cont)	15FEB2016	28OCT2016	15FEB2016	16APR2016	25APR2016	28OCT2016
3(cont)	01NOV2016	29NOV2016	01NOV2016	29NOV2016		

Variable-level metadata, as shown above, is useful for variables that are copied or derived from easily-referenced data. Because ADSL is one record per subject, there is no opportunity to include variables such as sequence number to provide data point traceability. For many ADSL variables, including those mentioned here, variable-level traceability is sufficient. An additional example is being developed for the ADaM Traceability Examples document to show how an intermediate dataset prior to ADSL can be used to provide additional traceability for more complex derivations; that example is not included in this PharmaSUG paper.

EXAMPLE 2: TRACEABILITY WITH PARAMETERS FROM MULTIPLE INPUT DATASETS

The ADaMIG section 4.4 includes an example of a time-to-event analysis dataset with input data from multiple SDTM domains. Content of this example is being expanded in the ADaM Traceability Examples document to illustrate how to maintain traceability when there are multiple input datasets. The example in section 4.4 of the ADaMIG describes that the event is based on the earliest study day of:

- hospitalization due to high blood pressure, or

- systolic blood pressure that exceeds 140, or
- diastolic blood pressure that exceeds 90.

If the subject did not have the event, then the subject is censored based on the final disposition.

DATA FLOW

Data for this example time-to-event analysis dataset (ADHYP) is from at least SDTM hospitalization (HO), vital signs (VS), and disposition (DS). Figure 2 demonstrates that data flow. It is possible for data to come from additional input sources, but for this illustration we are only looking at input data that is used to derive these particular parameter(s) of interest. Also, not shown in the diagram is any data coming from ADSL.

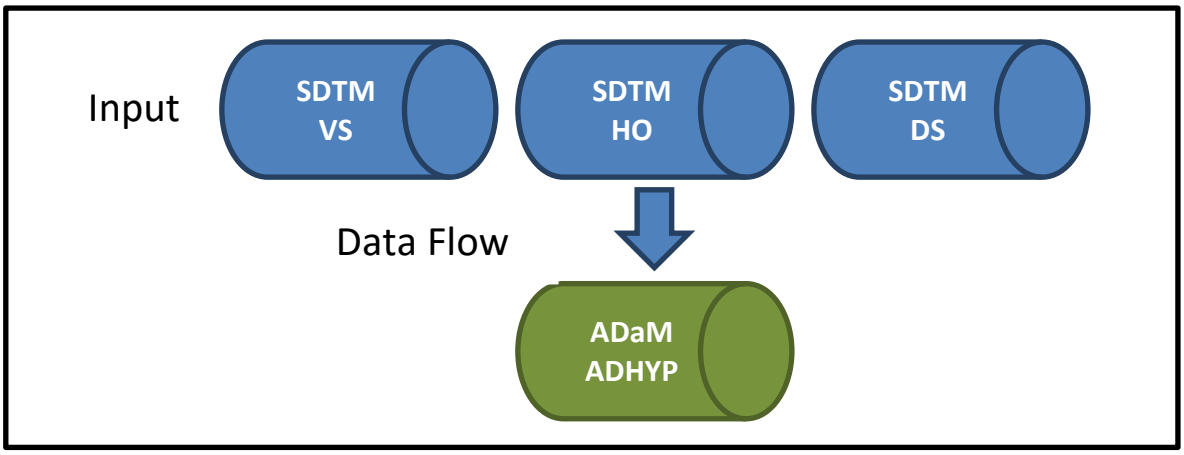

Figure 2: Example Time-to-Event Data Flow

TRACEABILITY NEEDS

Keeping sub-events (i.e., hospitalization, diastolic blood pressure > 90 and systolic blood pressure > 140) provides the data to support the hypertension event and is part of the traceability of this analysis dataset. Table 6 shows example variable metadata for specific ADaM dataset variables, and Table 7 shows Parameter-Value-Level metadata for selected variables. Note that only variables used to illustrate the concept of traceability are shown in these tables .

Table 6: Variable Metadata for ADHYP

Variable Name	Variable Label	Type	Codelist/ Controlled Terms	Variable Metadata
USUBJID	Unique Subject Identifier	Char		VS.USUBJID if subject had either systolic blood pressure > 140 and/or diastolic blood pressure > 90 HO. USUBJID if subject is hospitalized due to high blood pressure DS. USUBJID if subject did not have hospitalization or hypertension
PARAM	Parameter	Char	Time to First Hospital Admission (day); Time to First DBP > 90 (day); Time to First SBP > 140 (day); Time to Hypertension Event (day)	Time to Hypertension Event parameter is used for the analysis of time to hypertension. The other 'time to' parameters are sub-events and are included.
PARAMCD	Parameter Code	Char	HOSPADM; DBP; SBP; HYPEREVT	Create one record for each PARAMCD even if the subject did have the event.
AVAL	Analysis Value	Num		*See parameter-level metadata below*

Variable Name	Variable Label	Type	Codelist/ Controlled Terms	Variable Metadata
CNSR	Censored	Num	1; 0	*See parameter-level metadata below*
EVNTDESC	Event or Censoring Description	Char		*See parameter-level metadata below*
SRCDOM	Source Data	Char	HO; VS; DS	*See parameter-level metadata below*
SRCVAR	Source Variable	Char	HOSTDY; VSDY; DSSTDY	*See parameter-level metadata below*
SRCSEQ	Source Sequence Number	Num		The sequence number --SEQ of the row in the input dataset identified in the SRCDOM that relates to the analysis value being derived.

Table 7: Parameter Level Metadata for ADHYP

Variable	Where	Controlled Terms/ Formats	Source/Derivation/Comment
AVAL	PARAMCD = 'HOSPADM'		Predecessor: If subject was admitted to the hospital, then AVAL = HO.HOSTDY. Otherwise set to DS.DSSTDY.
AVAL	PARAMCD = 'DBP'		Predecessor: If subject had diastolic blood pressure > 90, then AVAL = VS.VSDY. Otherwise set to DS.DSSTDY.
AVAL	PARAMCD = 'SBP'		Predecessor: If subject had systolic blood pressure > 140, then AVAL = VS.VSDY. Otherwise set to DS.DSSTDY.
AVAL	PARAMCD = 'HYPEREVT'		DERIVED: If subject had a sub-event then set to earliest of the date of the sub-event.
CNSR	PARAMCD = 'HOSPADM'	1; 0	DERIVED: If subject was not admitted to the hospital, then CNSR = 1. Otherwise, CNSR = 0.
CNSR	PARAMCD = 'DBP'	1; 0	DERIVED: If subject never had diastolic blood pressure > 90, then CNSR = 1. Otherwise, CNSR = 0.
CNSR	PARAMCD = 'SBP'	1; 0	DERIVED: If subject never had systolic blood pressure > 140, then CNSR = 1. Otherwise, CNSR = 0.
CNSR	PARAMCD = 'HYPEREVT'	1; 0	DERIVED: If all of the sub-events (HOSPADM, DBP, SBP) had CNSR = 1, then CNSR = 1. Otherwise, CNSR = 0.
EVNTDESC	PARAMCD = 'HOSPADM'		Assigned: If subject was admitted to the hospital, then EVNTDESC = 'FIRST HOSPITAL ADMISSION'. Otherwise if DS.DSDECOD = 'COMPLETED' then EVNTDESC = 'COMPLETED THE STUDY'. Otherwise EVNTDESC = DS.DSDECOD.
EVNTDESC	PARAMCD = 'DBP'		Assigned: If subject had diastolic blood pressure > 90, then EVNTDESC = 'FIRST DBP > 90'. Otherwise if DS.DSDECOD = 'COMPLETED' then EVNTDESC = 'COMPLETED THE STUDY'. Otherwise EVNTDESC = DS.DSDECOD.
EVNTDESC	PARAMCD = 'SBP'		Assigned: If subject had systolic blood pressure > 140, then EVNTDESC = 'FIRST SBP > 140'. Otherwise if DS.DSDECOD = 'COMPLETED' then EVNTDESC = 'COMPLETED THE STUDY'. Otherwise EVNTDESC = DS.DSDECOD.
EVNTDESC	PARAMCD = 'HYPEREVT'		DERIVED: If at least one of the sub-events (HOSPADM, DBP, SBP) had EVNTDESC that indicated 'FIRST …', then EVNTDESC = 'HYPERTEN. EVENT'. Otherwise if DS.DSDECOD = 'COMPLETED' then EVNTDESC = 'COMPLETED THE STUDY'. Otherwise EVNTDESC = DS.DSDECOD.

Variable	Where	Controlled Terms/ Formats	Source/Derivation/Comment
SRCDOM	PARAMCD = 'HOSPADM'	HO; DS	Assigned: If subject was admitted to the hospital, then SRCDOM = 'HO'. Otherwise SRCDOM = 'DS'.
SRCDOM	PARAMCD = 'DBP'	VS; DS	Assigned: If subject had diastolic blood pressure > 90, then SRCDOM = 'VS'. Otherwise SRCDOM = 'DS'.
SRCDOM	PARAMCD = 'SBP'	VS; DS	Assigned: If subject had systolic blood pressure > 140, then SRCDOM = 'VS'. Otherwise SRCDOM = 'DS'.
SRCDOM	PARAMCD = 'HYPEREVT'	HO; VS; DS	DERIVED: Using sub-events determine the earliest event time and set SRCDOM accordingly. Otherwise SRCDOM = 'DS'.
SRCVAR	PARAMCD = 'HOSPADM'	HOSTDY; DSSTDY	Assigned: If subject was admitted to the hospital, then SRCVAR = 'HOSTDY'. Otherwise, SRCVAR = 'DSSTDY'.
SRCVAR	PARAMCD = 'DBP'	VSDY; DSSTDY	Assigned: If subject had diastolic blood pressure > 90, then SRCVAR = 'VSDY'. Otherwise, SRCVAR = 'DSSTDY'.
SRCVAR	PARAMCD = 'SBP'	VSDY; DSSTDY	Assigned: If subject had systolic blood pressure > 140, then SRCVAR = 'VSDY'. Otherwise, SRCVAR = 'DSSTDY'.
SRCVAR	PARAMCD = 'HYPEREVT'	HOSTDY; VSDY; DSSTDY	DERIVED: Using sub-events determine the earliest event time and set SRCVAR accordingly. Otherwise, SRCVAR = 'DSSTDY'.

INPUT AND ANALYSIS DATA

As indicated previously HO, VS, and DS datasets are three inputs that are needed for the creation of the analysis dataset (ADHYP).

Tables 8, 9, and 10 contain example input data illustrating how ADHYP is created based on the variable metadata in Table 6 and parameter-level metadata in Table 7. Only variables pertinent to the example are included here.

Table 8: Input Data Example VS

Row	USUBJID	VISITNUM	VSSEQ	VSDTC	VSDY	VSTESTCD	VSSTRESN
1	2010	1	22	2004-08-05	1	SYSBP	115
2	2010	1	23	2004-08-05	1	DIABP	75
3	2010	2	101	2004-08-12	8	SYSBP	120
4	2010	2	102	2004-08-12	8	DIABP	90
5	2010	3	207	2004-08-19	15	SYSBP	135
6	2010	3	208	2004-08-19	15	DIABP	92
7	2010	4	238	2004-08-25	21	SYSBP	138
8	2010	4	239	2004-08-25	21	DIABP	95
9	3082	1	27	2004-09-08	1	SYSBP	120
10	3082	1	28	2004-09-08	1	DIABP	80
11	3082	2	119	2004-09-15	8	SYSBP	125
12	3082	2	120	2004-09-15	8	DIABP	84

Table 9: Input Data Example HO

Row	USUBJID	HOSEQ	HOTERM	HODECOD	HOSTDTC	HOENDTC	HOSTDY	HOENDY
1	2010	99	HOSPITAL	HOSPITAL	2004-08-13	2004-08-15	9	11
2	2010	199	HOSPITAL	HOSPITAL	2004-08-20	2004-08-22	16	18

Table 10: Input Data Example DS

Row	USUBJID	DSSEQ	DSSTDTC	DSSTDY	DSDECOD	DSTERM
1	2010	25	2004-08-05	1	RANDOMIZED	Subject Randomized
2	2010	301	2004-08-26	22	COMPLETED	Subject Completed
3	3082	20	2004-09-08	1	RANDOMIZED	Subject Randomized
4	3082	130	2004-09-17	10	COMPLETED	Subject Completed

Utilizing the SRCDOM, SRCVAR, and SRCSEQ variables in ADHYP allows us to trace each record back to the source dataset. For example, Table 11 shows the resulting analysis dataset ADHYP.

Table 11: Output Data Example ADHYP

Row	USUBJID	PARAM	PARAMCD	AVAL	CNSR	EVNTDESC	SRCDOM	SRCVAR	SRCSEQ
1	2010	Time to First Hospital Admission (day)	HOSPADM	9	0	FIRST HOSPITAL ADMISSION	HO	HOSTDY	99
2	2010	Time to First DBP>90 (day)	DBP	15	0	FIRST DBP>90	VS	VSDY	208
3	2010	Time to First SBP>140 (day)	SBP	22	1	COMPLETED THE STUDY	DS	DSSTDY	301
4	2010	Time to Hypertension Event (day)	HYPEREVT	9	0	HYPERTEN. EVENT	HO	HOSTDY	99
5	3082	Time to First Hospital Admission (day)	HOSPADM	10	1	COMPLETED THE STUDY	DS	DSSTDY	130
6	3082	Time to First DBP>90 (day)	DBP	10	1	COMPLETED THE STUDY	DS	DSSTDY	130
7	3082	Time to First SBP>140 (day)	SBP	10	1	COMPLETED THE STUDY	DS	DSSTDY	130
8	3082	Time to Hypertension Event (day)	HYPEREVT	10	1	COMPLETED THE STUDY	DS	DSSTDY	130

Notice that we can trace the value of Row 1 back to HO data with HOSEQ = 99 (Row 1 in HO) and trace the value of Row 2 back to VS data with VSSEQ = 208 (Row 6 in VS).

EXAMPLE 3: TRACEABILITY WHEN CREATING ROWS IN BDS

It is not uncommon to have an analysis need whereby one needs to derive an analysis value from multiple rows from a preceding SDTM dataset. The ADaM Basic Data Structure (BDS) variable DTYPE is used to indicate when a new derived row has been added to a parameter, and to briefly describe how the analysis value was derived. This example demonstrates the case where electrocardiogram values were measured in triplicate at each time point in a study, and the average of these triplicate values is what is needed for the analysis.

ANALYSIS NEED

In this example, the analysis requirement is to summarize the average of the triplicate ECG interval values (AVAL) as well as change from baseline (CHG), where baseline (BASE) is defined as the average of the triplicate ECG intervals collected on the visit prior to the first administration of study drug. This summary will be performed by analysis visit (AVISIT).

Dataset=ADEG

Table x.x.x
Actual Values and Change from Baseline in ECG Parameters by Time point
(Safety Analysis Set) SAFFL

PARAM
AVISIT

Parameter / Time point	Statistic	TRTA: Placebo (N=xx) Actual Value (AVAL)	Placebo (N=xx) Change from Baseline (CHG)	Active xx mg (N=xx) Actual Value	Active xx mg (N=xx) Change from Baseline	Active yy mg (N=xx) Actual Value	Active yy mg (N=xx) Change from Baseline
QTcF Interval (msec)							
Baseline	n	xx		xx		xx	
	Mean (SD)	xx.x (xx.xx)		xx.x (xx.xx)		xx.x (xx.xx)	
	Median	xx.x		xx.x		xx.x	
	Min, Max	xx, xx		xx, xx		xx, xx	
Visit 1							
	N	xx	xx	xx	xx	xx	xx
	Mean (SD)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)
	Median	xx.x	xx.x	xx.x	xx.x	xx.x	xx.x
	Min, Max	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx
Visit 2	n	xx	xx	xx	xx	xx	xx
	Mean (SD)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)
	Median	xx.x	xx.x	xx.x	xx.x	xx.x	xx.x
	Min, Max	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx
. . .							

Where DTYPE='AVERAGE'

The average of triplicate or available ECG measurements collected at each nominal time point are used for analysis.
Cross-reference: Listing 8.3.1

Program: xxxxxxx.sas, Output: xxxxxxx.rtf, Generated on: DDMONYYYY xx:xx Page x of y
Programming Notes:
Note [1]: Repeat for all ECG parameters and time points including Day 28, but NOT Day 28/EDV.
Note [2]: Present ECG parameters ordered as follows: Heart rate, PR interval, QRS duration, QTcB, QTcF.

Figure 3: Example Table Showing Analysis Need for Averaging Values Across Each Visit

DATA FLOW

Data for ADEG in this example is coming directly from SDTM domain EG plus ADSL.

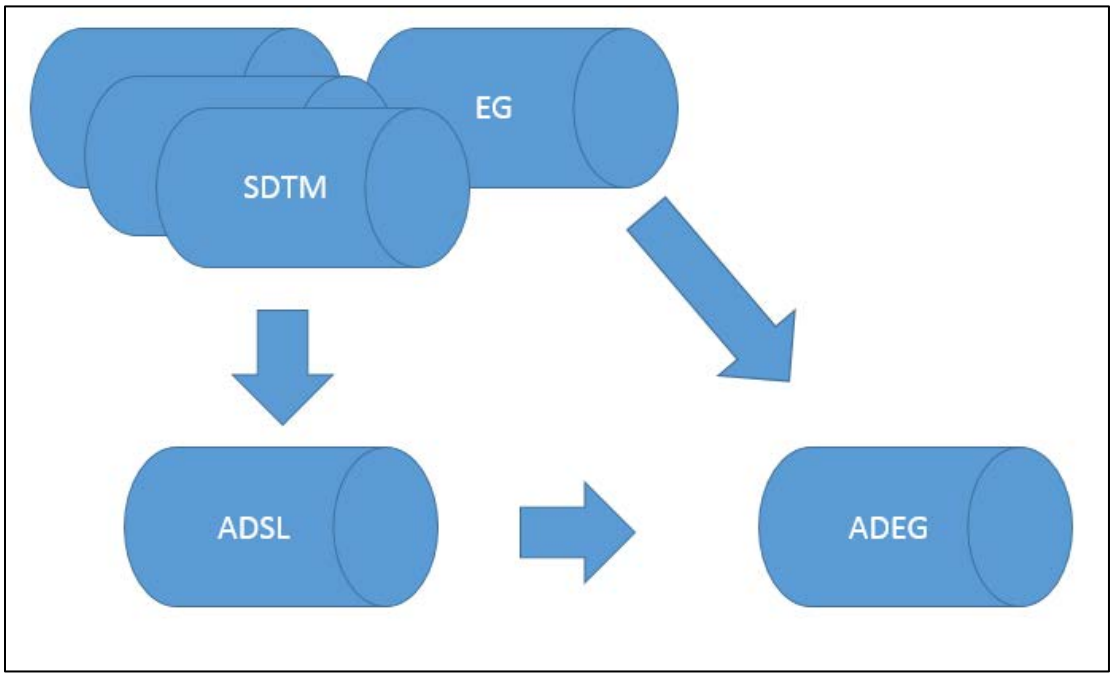

Figure 4: Example ADEG Data Flow

TRACEABILITY METADATA

Dataset-level, variable-level, and parameter-value-level metadata, all useful for traceability, are shown below. Only content pertinent to the example are included here.

Table 12: Dataset Metadata Example ADEG

Dataset	Dataset Description	Data Structure	Class of Dataset
ADEG	Electrocardiogram Analysis Dataset	One record per subject, parameter, analysis visit, reference ID, derivation type	BASIC DATA STRUCTURE

Table 13: Variable Metadata Example ADEG

Variable Name	Variable Label	Code List/Controlled Terminology	Source/Derivation/Comment
STUDYID	Study Identifier		Predecessor: ADSL.STUDYID
USUBJID	Unique Subject Identifier		Predecessor: ADSL.USUBJID
EGSEQ	Sequence Number		Predecessor: EG.EGSEQ
EGREFID	ECG Reference ID	"1st Measure" "2ndMeasure" "3rd Measure"	Predecessor: EG.EGREFID
PARAM	Parameter		Derived: Derived from EGTESTCD and EGSTRESU as follows: **EGTESTCD** / **EGSTRESU** / **PARAM** HR / bpm / Heart Rate (beats/min) PR / msec / PR interval (msec) QRS / msec / QRS Interval (msec) QTCB / msec / QTc Interval Bazett (msec) QTCF / msec / QTc Interval Frederica (msec)
VISIT	Visit Name		Predecessor: EG.VISIT
AVISIT	Analysis Visit		Derived: Derivation is explained in analysis data reviewer's guide, Section 7.5.2 Analysis Data Reviewer's Guide
EGDTC	Date/Time of ECG		Predecessor: EG.EGDTC
BASE	Baseline Value		Derived: For post-baseline records it is value of AVAL for each subject and parameter where ABLFL in Y
AVAL	Analysis Value		*See parameter-level metadata below*
CHG	Change from Baseline		Derived: AVAL – BASE. It is populated for all post-baseline records.
DTYPE	Derivation Type	["AVERAGE" = "Average"]	Derived: Value is AVERAGE for created records added for each visit as the average of the triplicate values collected for each parameter.
ABLFL	Baseline Record Flag	"Y"="Yes"	Derived: For each subject and parameter the baseline flag is set to Y for the last record prior to treatment start where DTYPE="AVERAGE"

Variable Name	Variable Label	Code List/Controlled Terminology	Source/Derivation/Comment
TRTA	Actual Treatment	Placebo Active 20mg Active 40mg	Assigned: Value of ADSL.TRT01A for a particular subject.
SAFFL	Safety Population Flag	["N"="No", "Y"="Yes"]	Predecessor: ADSL.SAFFL

Table 14: Example Parameter-Level Metadata for ADEG Variable AVAL

Variable	Where	Source/Derivation/Comment
AVAL	DTYPE='AVERAGE'	DERIVED: Average of the triplicate values collected at each visit for the parameter.
AVAL	DTYPE Not Equal 'AVERAGE'	Predecessor: EG.EGSTRESN

INPUT AND ANALYSIS DATA

Below are example ECG data from SDTM EG. Only content pertinent to the example are included here.

Table 15: Input Data Example EG

USUBJID	EGSEQ	EGREFID	EGTESTCD	EGSTRESN	EGSTRESU	EGBLFL	VISIT	EGDTC
XYZ-1001	1	1st MEASURE	QTCFAG	385	msec		SCREENING	2016-02-24T07:50:16
XYZ-1001	2	2nd MEASURE	QTCFAG	399	msec		SCREENING	2016-02-24T07:52:59
XYZ-1001	3	3rd MEASURE	QTCFAG	396	msec	Y	SCREENING	2016-02-24T07:56:07
XYZ-1001	4	1st MEASURE	QTCFAG	384	msec		VISIT 2	2016-03-08T09:45:11
XYZ-1001	5	2nd MEASURE	QTCFAG	393	msec		VISIT 2	2016-03-08T09:48:07
XYZ-1001	6	3rd MEASURE	QTCFAG	388	msec		VISIT 2	2016-03-08T09:51:04
XYZ-1001	7	1st MEASURE	QTCFAG	385	msec		VISIT 3	2016-03-22T10:45:03
XYZ-1001	8	2nd MEASURE	QTCFAG	394	msec		VISIT 3	2016-03-22T10:48:07
XYZ-1001	9	3rd MEASURE	QTCFAG	402	msec		VISIT 3	2016-03-22T10:51:05
XYZ-1002	10	1st MEASURE	QTCFAG	399	msec		SCREENING	2016-02-22T07:55:02
XYZ-1002	11	2nd MEASURE	QTCFAG	410	msec		SCREENING	2016-02-22T07:58:05
XYZ-1002	12	3rd MEASURE	QTCFAG	392	msec	Y	SCREENING	2016-02-22T08:01:06
XYZ-1002	13	1st MEASURE	QTCFAG	401	msec		VISIT 2	2016-03-06T09:50:04
XYZ-1002	14	2nd MEASURE	QTCFAG	407	msec		VISIT 2	2016-03-06T09:53:51
XYZ-1002	15	3rd MEASURE	QTCFAG	400	msec		VISIT 2	2016-03-06T09:56:21
XYZ-1002	16	1st MEASURE	QTCFAG	412	msec		VISIT 3	2016-03-24T10:50:07
XYZ-1002	17	2nd MEASURE	QTCFAG	414	msec		VISIT 3	2016-03-24T10:53:08
XYZ-1002	18	3rd MEASURE	QTCFAG	402	msec		VISIT 3	2016-03-24T10:56:05

Below are example ECG data from ADaM ADEG. Only content pertinent to the example are included here.

Table 16: Output Data Example ADEG

USUBJID	EGSEQ	EGREFID	PARAM	VISIT	AVISIT	EGDTC	BASE	AVAL	CHG	DTYPE	ABLFL
XYZ-1001	1	1st MEASURE	QTcF Interval (msec)	SCREENING	Baseline	2016-02-24T07:50:16		385			
XYZ-1001	2	2nd MEASURE	QTcF Interval (msec)	SCREENING	Baseline	2016-02-24T07:52:59		399			
XYZ-1001	3	3rd MEASURE	QTcF Interval (msec)	SCREENING	Baseline	2016-02-24T07:56:07		396			
XYZ-1001			QTcF Interval (msec)	SCREENING	Baseline			393.3		AVERAGE	Y
XYZ-1001	4	1st MEASURE	QTcF Interval (msec)	VISIT 2	Visit 2	2016-03-08T09:45:11	393.3	384	-9.3		
XYZ-1001	5	2nd MEASURE	QTcF Interval (msec)	VISIT 2	Visit 2	2016-03-08T09:48:07	393.3	393	-0.3		
XYZ-1001	6	3rd MEASURE	QTcF Interval (msec)	VISIT 2	Visit 2	2016-03-08T09:51:04	393.3	388	-5.3		
XYZ-1001			QTcF Interval (msec)	VISIT 2	Visit 2		393.3	388.3	-5.0	AVERAGE	
XYZ-1001	7	1st MEASURE	QTcF Interval (msec)	VISIT 3	Visit 3	2016-03-22T10:45:03	393.3	385	-8.3		
XYZ-1001	8	2nd MEASURE	QTcF Interval (msec)	VISIT 3	Visit 3	2016-03-22T10:48:07	393.3	394	0.7		
XYZ-1001	9	3rd MEASURE	QTcF Interval (msec)	VISIT 3	Visit 3	2016-03-22T10:51:05	393.3	402	8.7		
XYZ-1001			QTcF Interval (msec)	VISIT 3	Visit 3		393.3	393.7	0.3	AVERAGE	
XYZ-1002	11	2nd MEASURE	QTcF Interval (msec)	SCREENING	Baseline	2016-02-22T07:58:05		410			
XYZ-1002	12	3rd MEASURE	QTcF Interval (msec)	SCREENING	Baseline	2016-02-22T08:01:06		392			
XYZ-1002			QTcF Interval (msec)	SCREENING	Baseline			400.3		AVERAGE	Y
XYZ-1002	13	1st MEASURE	QTcF Interval (msec)	VISIT 2	Visit 2	2016-03-06T09:50:04	400.3	401	0.7		
XYZ-1002	14	2nd MEASURE	QTcF Interval (msec)	VISIT 2	Visit 2	2016-03-06T09:53:51	400.3	407	6.7		
XYZ-1002	15	3rd MEASURE	QTcF Interval (msec)	VISIT 2	Visit 2	2016-03-06T09:56:21	400.3	400	-0.3		
XYZ-1002			QTcF Interval (msec)	VISIT	Visit 2		400.3	402.7	2.3	AVERAGE	
XYZ-1002	16	1st MEASURE	QTcF Interval (msec)	VISIT 3	Visit 3	2016-03-24T10:50:07	400.3	412	11.7		
XYZ-1002	17	2nd MEASURE	QTcF Interval (msec)	VISIT 3	Visit 3	2016-03-24T10:53:08	400.3	414	13.7		
XYZ-1002	18	3rd MEASURE	QTcF Interval (msec)	VISIT 3	Visit 3	2016-03-24T10:56:05	400.3	402	1.7		
XYZ-1002			QTcF Interval (msec)	VISIT 3	Visit 3		400.3	409.3	9.0	AVERAGE	

OTHER USES

This example demonstrated how to maintain traceability when you create new records in a BDS ADaM dataset. The traceability is both metadata driven (i.e., variable-level and parameter-level metadata defining how the new record is derived) and data-driven (e.g., maintaining the source records and variables from SDTM in the ADEG dataset, such as using EGSEQ). There are numerous use cases similar to this example, such as:

- An endpoint analysis where new records are created for an analysis visit of “Endpoint”, using a derivation type (DTYPE) of LOCF or WOCF.
- Creating a composite endpoint.
- Interpolating missing values (when not using Mixed modeling methods).
- Creating group scores (e.g., total symptom scores as sum of individual symptom scores from an allergy diary).
- Creating additional collapsed record rows in an adverse event analysis dataset, when a single adverse event was collected across multiple records.

EXAMPLE 4: TRACEABILITY WHEN MULTIPLE ANALYSIS VARIABLES ARE NEEDED ON THE SAME ROW

In cases of statistical modelling that features multiple dependent and/or independent variables, statistical software many require all analysis variables to be in the same record for processing. The ADaM BDS supports only one analysis variable per row in variable AVAL and/or AVALC. This example shows a way to support multiple analysis variables on one row within ADaM and still maintain the ADaM principle of traceability.

In this example, the scores of a Motor Function Questionnaire is to be analyzed. There are multiple scores collected at baseline and after one month of treatment. The program to generate a statistical generalized linear model requires these scores to be in the same row for processing, therefore it was decided to produce a horizontal ADaM dataset.

The approach demonstrated is to first create a BDS dataset named ADQS, making use of traceability built into the BDS standard to explain the origin, derivation, imputation, and any other complexity behind each analysis value. The values of PARAMCD and PARAM are chosen with the intention of using them as the variable name and label in a subsequent wide format dataset. The BDS dataset is finally transposed into a wide format (structure of ADaM Other) dataset named ADQST to support statistical analysis and review.

This concept of creating a BDS and transposing is not new, and it was described in the ADaM Examples in Commonly Used Statistical Analysis Methods, example 6 (Multivariate Analysis of Variance). Text in that section describes how a BDS dataset would need to be transposed in order to be truly analysis-ready.

DATA FLOW

In this example, ADaM dataset ADQS has data from SDTM QS plus ADSL, and ADaM dataset ADQST is derived directly from and using only ADQS.

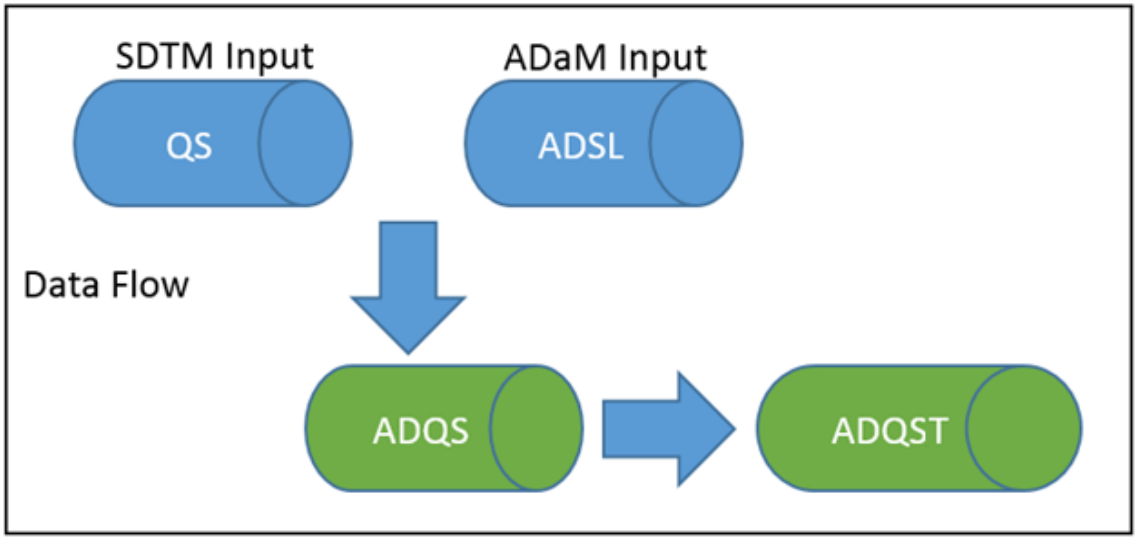

Figure 5: Example Data Flow When Transposing BDS

TRACEABILITY METADATA

Dataset-level, variable-level, and parameter-value-level metadata are shown below. Only content pertinent to the example are included here.

The variable metadata for ADQS provides traceability to the source SDTM data variables and describes the process of deriving new DTYPE='SUM' records:

Table 17: Variable Metadata Example ADQS

Variable Name	Where Condition	Variable Metadata
STUDYID		QS.STUDYID
USUBJID		QS.USUBJID
TRTP		ADSL.TRT01P
VISIT		QS.VISIT
PARAMCD		Keep QS records where QS.QSCAT='MOTOR FUNCTION QUESTIONNAIRE' and VISIT in ('BASELINE' 'MONTH 1'), set PARAMCD=QS.QSTESTCD and PARAM=QS.QSTEST. Create 4 derived records per subject, with PARAMCD, PARAM values as: UPPER=Upper Body Motor Score LOWER=Lower Body Motor Score For timepoints BASELINE and MONTH 1
PARAM		See PARAMCD
DTYPE		Set to "SUM" where PARAMCD="UPPER" or "LOWER"
AVAL	DTYPE NE 'SUM'	QS.QSSTRESN

Variable Name	Where Condition	Variable Metadata
	DTYPE EQ 'SUM'	Where PARAMCD='UPPER', the sum of the scores for questions 1-3 Where PARAMCD='LOWER', the sum of the scores for questions 4-6 If any scores are missing, do not impute sum
ABLFL		Set to Y where VISIT=BASELINE
BASE		AVAL value from the record where ABLFL=Y, populate for post-baseline records only
CHG		AVAL-BASE
QSSEQ		QS.QSSEQ
QSCAT		QS.QSCAT

Table 18: Dataset Metadata Example ADQST

Dataset	Description	Class	Structure	Description
ADQST	Transposed ADQS	ADAM OTHER	One record per subject	This dataset is derived from ADQS by transposing CHG by USUBJID, using the values of PARAMCD as new variable names and the values of PARAM as new variable labels

The variable metadata for ADQST is relatively simple, describing the transpose process and providing the predecessor origins for each variable:

Table 19: Variable Metadata Example ADQST

Variable Name	Variable Label	Variable Metadata
STUDYID	Study Identifier	ADQS.STUDYID
USUBJID	Unique Subject Identifier	ADQS.USUBJID
TRTP	Planned Treatment	ADQS.TRTP
VISIT	Visit Name	ADQS.VISIT Only keep records for Month 1
S01	Score 1	ADQS.CHG where PARAMCD='S01'
S02	Score 2	ADQS.CHG where PARAMCD='S02'
S03	Score 3	ADQS.CHG where PARAMCD='S03'
S04	Score 4	ADQS.CHG where PARAMCD='S04'
S05	Score 5	ADQS.CHG where PARAMCD='S05'
S06	Score 6	ADQS.CHG where PARAMCD='S06'
UPPER	Upper Body Score	ADQS.CHG where PARAMCD='UPPER'
LOWER	Lower Body Score	ADQS.CHG where PARAMCD='LOWER'

INPUT AND ANALYSIS DATA

Below are example ADQS data. Only content pertinent to the example are included here.

Table 20: Intermediate Data Example ADQS

Row	STUDYID	USUBJID	TRTP	VISIT	PARAMCD	PARAM	DTYPE	AVAL
1	XYZ	XYZ-001	DRUG A	BASELINE	S01	Score 1		40
2	XYZ	XYZ-001	DRUG A	MONTH 1	S01	Score 1		55
3	XYZ	XYZ-001	DRUG A	BASELINE	S02	Score 2		30
4	XYZ	XYZ-001	DRUG A	MONTH 1	S02	Score 2		40
5	XYZ	XYZ-001	DRUG A	BASELINE	S03	Score 3		45
6	XYZ	XYZ-001	DRUG A	MONTH 1	S03	Score 3		40
7	XYZ	XYZ-001	DRUG A	BASELINE	S04	Score 4		20
8	XYZ	XYZ-001	DRUG A	MONTH 1	S04	Score 4		30
9	XYZ	XYZ-001	DRUG A	BASELINE	UPPER	Upper Body Score	SUM	115
10	XYZ	XYZ-001	DRUG A	MONTH 1	UPPER	Upper Body Score	SUM	135
11	XYZ	XYZ-001	DRUG A	BASELINE	LOWER	Lower Body Score	SUM	110
12	XYZ	XYZ-001	DRUG A	MONTH 1	LOWER	Lower Body Score	SUM	115

Row	ABLFL	BASE	CHG	QSSEQ	QSCAT
1(cont)	Y			1	MOTOR FUNCTION QUESTIONNAIRE
2(cont)		40	15	2	MOTOR FUNCTION QUESTIONNAIRE
3(cont)	Y			3	MOTOR FUNCTION QUESTIONNAIRE
4(cont)		30	10	4	MOTOR FUNCTION QUESTIONNAIRE
5(cont)	Y			5	MOTOR FUNCTION QUESTIONNAIRE
6(cont)		45	-5	6	MOTOR FUNCTION QUESTIONNAIRE
7(cont)	Y			7	MOTOR FUNCTION QUESTIONNAIRE
8(cont)		20	10	8	MOTOR FUNCTION QUESTIONNAIRE
9(cont)	Y				
10(cont)		115	20		
11(cont)	Y				
12(cont)		110	5		

Note: In the sample data for ADQS shown in Table 20, records that originate from SDTM have a value in QSSEQ and no value in DTYPE, and records which are derived have a value in DTYPE and no value in QSSEQ. Including variable QSSEQ allows us to identify the exact source record in QS that was used for the row in ADQS.

Below are example ADQST data. Only content pertinent to the example are included here.

Table 21: Transposed Data Example ADQST

USUBJID	TRTP	VISIT	S01	S02	S03	S04	S05	S06	UPPER	LOWER
Unique Subject Identifier	Planned Treatment	Visit Name	Score 1	Score 2	Score 3	Score 4	Score 5	Score 6	Upper Body Score	Lower Body Score
XYZ-001	DRUG A	MONTH 1	15	10	-5	10	-5	0	20	5
XYZ-002	DRUG B	MONTH 1	0	5	20	15	5	5	25	25
XYZ-003	DRUG A	MONTH 1	30	10	15	20	25	30	55	75
XYZ-004	DRUG B	MONTH 1	-5	0	-10	0	5	5	-15	10
XYZ-005	DRUG A	MONTH 1	10	0	5	-10	-5	0	15	-15
XYZ-006	DRUG B	MONTH 1	10	5	0	0	5	5	15	10

The sample data for ADQST shown in Table 21 supports the needs of the statistical analysis, and through the dataset, variable, and parameter metadata it is possible to trace each analysis value to a specific record in ADQS, and from there to the source SDTM records. A note of importance is that if ADQS was not produced, and only ADQST provided, the traceability between source and analysis data would be lost.

OTHER USES

This example demonstrates how each data point in a wide multiple analysis variables dataset can be traced back across derivations to its SDTM source using variable metadata and data point traceability provided by the BDS standard. It is not necessary for the final horizontal dataset to be one record per USUBJID. For example, one may create a one record per USUBJID per AVISIT timepoint dataset that arranges all analysis values from that timepoint horizontally.

CONCLUSION

Four examples were shown to demonstrate how data, metadata, and even intermediate datasets, can all provide traceability when creating ADaM datasets. Each of these examples comes from the ADaM Traceability Examples document now in development.

When deciding how to create ADaM datasets, the authors encourage you to ask yourself the following questions:

- Can the end-user determine which data is copied from SDTM and which is derived?
- Can the end-user determine how each variable and row was created in the dataset?
- Can the end-user trace back to the SDTM data that was used to create the value used for analysis?

By considering the perspective of the end-user, traceability can be built in a natural and useful way.

REFERENCES

All CDISC documents referenced in this paper can be downloaded from https://www.cdisc.org/.

RECOMMENDED READING

- *Analysis Data Model Implementation Guide version 1.1*
- *Analysis Data Model (ADaM) Examples in Commonly Used Statistical Analysis Methods*
- *CDISC Define-XML Specification Version 2.0*

CONTACT INFORMATION

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

Sandra Minjoe
PRA Health Sciences
MinjoeSandra@prahs.com

Wayne Zhong
Accretion Softworks
Wayne@asoftworks.com

Quan Jenny Zhou
Eli Lilly and Company
Zhou_quan_jenny@lilly.com

Kent Letourneau
PRA Health Sciences
LetourneauKent@prahs.com

Richann Watson
DataRich Consulting
richann.watson@datarichconsulting.com