PharmaSUG2011 - Paper CD05

Methods of Building Traceability for ADaM Data

Songhui Zhu, K & L Consultant Services, Fort Washington, PA
Lin Yan, Celgene Corp, Basking Ridge, NJ

ABSTRACT

Traceability is the property that enables the understanding where the analysis data come from. It facilitates transparency and is an essential component for building confidence in analysis results. It is built by recording the path from a record to its predecessor(s) [1]. For simple cases, one may use variables SRCDOM, SRCVAR, and SRCSEQ to build traceability [1]. However, establishing clear traceability for ADaM datasets is not always as simple as this. In fact, sometime, it is challenging. This paper presents methods of building traceability in ADaM datasets by examples of building traceability in ADQS by adding variables RLCRIT and RLFACT when the relation between records in ADaM datasets and their predecessors are complex such as the cases where there are multiple source records, the cases where the multiple source records are not kept in ADaM dataset, and the cases where the sources are nested conditional ones.

1. INTRODUCTION

Traceability in ADaM facilitates transparency and is essential for building confidence in analysis results. It can be built by recording the path from a record to its predecessor(s). It is worth pointing out that keeping traceability is also an essential part in both the CDISC SDTM [2] and Clinical Data Acquisition Standards Harmonization (CDASH) standards [3]. The harmonization of ADaM, SDTM, and CDASH gives the full path from ADaM data to their predecessors, SDTM data, and then from SDTM data to their predecessors, the data collection instruments [1]. The establishment of the full path from ADaM data to data collection relies on the traceability in CDASH, the traceability in SDTM, and the traceability in ADaM. This can be achieved only by the collaboration among the whole study team. For example, it relies on well-designed CRF that keeps enough information to build traceability; it also depends on the efforts of data managers who create SDTM data and the efforts of statistical programmers who create ADaM data. But this paper will focus on traceability in ADaM data only.

Firstly, this paper will use a simple example of questionnaire to show how to build traceability in ADaM data using variables SRCDOM, SRCVAR, and SRCSEQ, where SRCDOM is the variable to store the source SDTM domain, SRCVAR is the variable to store source SDTM variable, and SRCSEQ is the variable to store the source SDTM sequence number. However, establishing clear traceability for ADaM datasets is not always as simple as this. In fact, sometime, it is challenging.

Secondly, this paper will introduce a new method of building traceability in ADaM data through an example of ADQS. In this example, using variables SRCDOM, SRCVAR, and SRCSEQ cannot build clear traceability. This new method does not use the three variables SRCDOM, SRCVAR, and SRCSEQ. Instead, a pair of new variables RLCRIT and RLFACT are added, where RLCRIT is the variable to store the relation between ADaM data and its source SDTM/ADaM data and RLFACT is the variable to store the facts of source SDTM/ADaM data. However, adding variables RLCRIT and RLFACT cannot solve all the problems.

Thirdly, this paper will generalize the method to add *n* pairs of variables RLCRIT1 and RLFACT1, RLCRIT2 and RLFACT2, …, RLCRIT*n* and RLFACT*n* to build traceability in ADaM data, where *n* is an integer whose values depends on the complexity of the relation between ADaM data and their predecessors.

Fourthly, this paper will extend the method to the cases when the source SDTM/ADaM data are not kept in ADaM data.

It is worth to point out that all examples used to demonstrate the new method and the generalization of the new methods are all questionnaires due to the variety of complexity of questionnaires. But this does not mean the new method cannot be used to other ADaM data sets.

2. METHODOLOGY

2.1 Method 1 - Using SRCDOM, SRCVAR, and SRCSEQ

According to ADaM Implementation Guide 1.0, whenever practical, variables to support data point traceability should be included in ADaM datasets. The SDTM variables that can be used to support data point traceability in ADaM are the SDTM domain variable value, the name of SDTM source variable, and the relevant sequence value [1]. Correspondingly, in ADaM datasets, variables SRCDOM, SRCVAR, and SRCSEQ can be used to create the path from an ADaM record to its predecessor, where SRCDOM is the two-character identifier of the SDTM domain that relates to AVAL or AVALC, SRCVAR is the name of the column in the SDTM domain that relates to AVAL or AVALC, and SRCSEQ is the sequence number of the row in the SDTM domain that relates to AVAL or AVALC.

If a record in ADaM dataset is created from SDTM data, the traceability from ADaM to SDTM data can be easily achieved. This is illustrated by the first example.

Example 1 SF-36[4] is a questionnaire consisting of 36 questions. The SDTM dataset QS for questionnaire SF-36 may look like as shown in Table 1 if the creation of SDTM data set QS follows the CDISC SDTM Implementation Guide 1.0.

Table 1 Illustration of QS for SF-36 Questionnaires

USUBJID	VISITNUM	QSTESTCD	QSTEST	QSORRES	QSSTRESN	QSSEQ
1001	3	GH1	Health	Very Good	4.4	0021
1001	3	GH11A	Sick a little easier	Mostly False	4	0022
1001	3	GH11B	Healthy as anybody	Mostly True	4	0023

According to ADaM Implementation Guide 1.0, variables SRCDOM, SRCVAR, and SRCSEQ can be used to create the path from an ADaM record to its predecessor. The corresponding records in ADaM data set ADQS look like as those shown in Table 2.

Table 2 Illustration of keeping traceability in ADQS for SF-36 Questionnaires

SUBJID	AVISITN	PARAMCD	PARAM	AVAL	SRCDOM	SRCVAR	SRCSEQ
1001	3	SFGH1	Health	4.4	QS	QSSTRESN	0021
1001	3	SFGH11A	Sick a little easier	4	QS	QSSTRESN	0022
1002	3	SFGH11B	Healthy as anybody	4	QS	QSSTRESN	0023

2.2 Method 2 - Adding variable pair RLCRIT and RLFACT

Please note that in the first example, each ADQS record is created from one and only one source record in QS. However, for questionnaires, it is very common that an analysis parameter in ADaM is derived from multiple SDTM records. For example, a total score of SF-36 is derived from 36 SDTM records. Sometimes, a composite score is even a conditional combination of values of multiple records. In these cases, a record in ADaM data set comes from multiple source records in SDTM data set. As a result, using SRCDOM, and SRCVAR, and SRCSEQ is not enough to keep good traceability.

Example 2 Modified American College of Rheumatology Response (ACR) Assessment [5]. It may be designed as follows:

Table 3 Modified American College of Rheumatology Response Assessment

78 Joint Score

N = No pain/tenderness N = No swelling

Y = Pain/tenderness Y = Swelling

ND = Not Done ND = Not Done

Joints	Right Side	
	Pain/tenderness	Swelling
	Y N ND	Y N ND
Temporomandibular		
Sternoclavicular		
Acromioclavicular		
Shoulder		
…..		
Joints	Left Side	
	Pain/tenderness	Swelling
	Y N ND	Y N ND
Temporomandibular		
Sternoclavicular		
Acromioclavicular		
Shoulder		
…..		

In this questionnaire, the disease activity score for ACR Rheumatoid Arthritis (ACR-N) is the minimum of percent change in the number of pain joints, percent change in the number of swelling joints, and median of 5 scores of Patient's global assessment from VAS questionnaire, Physician's global assessment from VAS questionnaire, Pain score from VAS questionnaire, Physical function from HAQ-DI questionnaire, and Acute-phase reactants (CPR) test value from lab test. The algorithm for the computation of the disease activity score can be clearly shown in Table 4.

In the definition of Disease activity score for ACR Rheumatoid Arthritis , the percent change in the number of swollen joints and the percent change in the number of tender joints can be found in field PCHG of total number of swollen/pain joints which is derived from multiple records in QS. The median involved in the calculation of the disease activity score is also a derived value -

median of five assessments from other SDTM domains. As the median is not needed directly for analysis, there is no need to create a separate record for it.

Table 4 Definition of Disease activity score for ACR Rheumatoid Arthritis

Minimum of	1. percent change in the number of swollen joints	
	2. percent change in the number of tender joints	
	3. Median of	a) Patient’s global assessment from VAS questionnaire
		b) Physician’s global assessment from VAS questionnaire
		c) Pain score from VAS questionnaire
		d) Physical function from HAQ-DI questionnaire
		e) Acute-phase reactants test value from lab test

When deriving ADaM data set for this questionnaire, keeping traceability is not straightforward since a final score involves so many parameters. But it is not impossible to keep good traceability by a good design as follows:

a. use metadata to describe the derivation rules of the derived parameters.

b. set up the data point traceability in ADQS by adding variables to connect derived records and their source records. For example, one may add RLCRIT (relation criteria) and RLFACT (relation factors) to keep traceability, where variable RLCRIT stores data source (ADaM or SDTM data sets) with source variables in order; variable RLFACT stores the values of those source variables in the same order. For a record derived from multiple source records, RLCRIT stores paths that can be used to trace back to multiple source records, some of which may be in ADaM datasets, while some others are in SDTM datasets. Please see Table 5 for details.

Table 5 Illustration of RLCRIT and RLFACT in Metadata

Variable Name	**Variable Label**	**Type**	**Derivation**
RLCRIT	Parameter Relation Criteria	Char	Median of (Four values of AVAL from ADQS when PARAMCD in (“EGAD”, “PGAD”, “PAIN”, “HAQDITOT”) and LBSTRESN from LB when LBTESTCD = “CRP” and date part of LB.LBDTC = ADQS.ADT)
RLFACT	Parameter Relation Factor	Char	Values of RLCRIT in order separated by delimiter ‘,’

2.3 Method 3 - Adding multiple pairs of variables RLCRIT and RLFACT

Please note that this pair of variables only establishes a part of the traceability of the disease activity score – traceability to the source of median of the five variables. More variables are needed to establish the traceability to the two percent changes. For example, one may add another pair of RLCRIT and RLFACT. Considering all these together, the metadata can be as follows:

Table 6 Illustration of metadata for complete traceability

Variable Name	**Variable Label**	**Type**	**Derivation**
RLCRIT1	Parameter Relation Criteria 1	Char	Median of (Four values of AVAL from ADQS when PARAMCD in (“EGAD”, “PGAD”, “PAIN”, “HAQDITOT”) and value of LBSTRESN from record with sequence number = LBSEQ in LB when LBTESTCD = “CRP” and date part of LB.LBDTC = ADQS.ADT)
RLFACT1	Parameter Relation Factor 1	Char	Values of AVAL from ADQS when PARAMCD in (“EGAD”, “PGAD”, “PAIN”, “HAQDITOT”) and value of LBSTRESN from record with sequence number = LBSEQ in LB when LBTESTCD = “CRP” and date part of LB.LBDTC = ADQS.ADT separated by delimiter ‘,’
RLCRIT2	Parameter Relation Criteria 2	Char	ADQS. JSTOTAL. ADT.PCHG ADQS. JPTOTAL. ADT. PCHG
RLFACT2	Parameter Relation Factor 2	Char	Values, separated by delimiter “,”, of PCHG when PARAMCD in (“JSTOTAL” “JPTOTAL”) when date = ADT

In ADQS, the data can be like as follow:

Table 7 Illustration of ADQS with complete traceability

PARAM	**RLCRIT1**	**RLFACT1**	**RLCRIT2**	**RLFACT2**
ACR-N	Median of(ADQS.EGAD.2009-11-10, ADQS.PGAD. 2009-11-10, ADQS.PAIN. 2009-11-10, ADQS.HAQDITOT. 2009-11-10, LB.CRP.0001)	7.1, 8,2, 5.0, 2.9, 10.5	ADQS.JSTOTAL. 2009-11-10 ADQS. JPTOTAL. 2009-11-10	40, 38
ACR-N	Median of(ADQS.EGAD. 2009-12-10, ADQS.PGAD. 2009-12-10, ADQS.PAIN. 2009-12-10, ADQS.HAQDITOT. 2009-12-10, LB.CRP.0021)	6.9, 7.8, 4.8, 2.0, 8.5	ADQS.JSTOTAL. 2009-12-10 ADQS. JPTOTAL. 2009-12-10	20, 18

2.4 Method 4 - Establishing data point traceability in ADQS while not keeping source records in ADQS

For questionnaires, it is very common that analyses will be performed on the composite score or total scores of several groups of questions. In those cases, it may not always be necessary to keep all the original questions. When the questionnaire includes many questions and the questionnaire is evaluated at many visits, there will be too many records. If only a final score is to be analyzed, it is quite reasonable to keep only the final score in the ADaM dataset. In this case, how to establish data point traceability? The approach used in previous examples can establish data point traceability without bringing in the source records in ADaM dataset. For example, again, one may add variables RLCRIT (relation criteria) and RLFACT (relation factors) to establish data point traceability.

For example, the total Number of Swelling Joints for DAS28 [6] is the number of the swelling joints among the specified 28 joints. The joint assessment is from a part of ACR joint assessment. The RLCRIT can be defined as the number of records with QSORRES=Yes when QSSPID=DAS28. The data point tracebility of the source can be completed in RLFACT with QSDTC or range of QSDTC's for the group of the records in SDTM data set QS. If the records are measured within a period, the QSDTC can be a range, e.g. QTDTC = [2010-04-10 to 2010-05-10].

Table 8 Illustration of Metadata for ADQS without keeping source record

Variable Name	Variable Label	Type	Derivation
RLCRIT	Parameter Relation Criteria	Char	Number of records with QSORRES=Yes when QSSPID=DAS28 for PARAM = 'Number of Swelling Joints for DAS28'
RLFACT	Parameter Relation Factor	Char	QSDTC

Using this strategy, one can create ADQS that look as follows:

Table 9 Illustration of ADQS without keeping source records

PARAM	RLCRIT	RLFACT
Number of Swelling Joints for DAS28	20	2009-11-10
Number of Swelling Joints for DAS28	18	2009-12-01 to 2009-12-03

3. CONCLUSIONS

This paper presented methods of building traceability in ADaM datasets through examples of questionnaires. A methods of adding variables RLCRIT and RLFACT was presented through an example of creation ADQS for questionnaire for Modified American College of Rheumatology Response (ACR). Then this method was generalized to more complex cases in which there are multi-layer questions so that multiple pairs of variables RLCRIT's RLFACT's were added to build traceability in ADQS. This paper also extended the method to cases in which source records are not kept.

REFERENCES

[1] DISC SDTM Implementation Guide (Version 3.1.2) CDISC Submission Data Standards Team.

[2] CDISC ADaM Implementation Guide Version 1.0 CDISC Analysis Data Model Team.

[3] Clinical Data Acquisition Standards Harmonization (CDASH). Version 1.0.

[4] Short form 36 health survey questionnaire (SF-36). http://www.sf-36.org/tools/sf36.shtml#LIT. Nov. 27, 2010.

[5] A Proposed Revision to the ACR20: The Hybrid Measure of American College of Rheumatology
Response. Arthritis & Rheumatism (Arthritis Care & Research) Vol. 57, No. 2, March 15, 2007, pp 193–202.

[6] Disease Activity Score – 28 Joints (DAS28). http://www.iche.edu/newsletter/DAS28.pdf. Nov. 27, 2010.

ACKNOWLEDGEMENT

Songhui Zhu thanks the K & L Consulting Services and Lin Yan thanks Celgene for the support they received.

CONTACT INFORMATION

Your comments and questions are valuable and appreciated. Authors can be reached at

Songhui Zhu, PhD
K&L Consulting Service
1300 Virginia Dr., #103
Fort Washington, PA 19034 - 3266

Lin Yan
Celgene Corporation,
Basking Ridge, NJ 07920