Proceedings

# First SUGCDSC Conference

## August 31, 2018

Silver Spring Civic Building at Veterans Plaza
8525 Fenton Street, Silver Spring, MD 20910
Washington DC Area, USA



**SAS User Group for Clinical Data Standards and Codes** (SUGCDSC)
PO Box 1032, Rockville, MD 20849, USA
https://sugcdsc.org

# Improving Traceability for Complex Algorithms in ADaM Datasets

Priya Saradha, Levstat® LLC

## ABSTRACT

Currently, ADaM datasets implement traceability using the variable triplet SRCSEQ, SRCDOM, and SRCVAR to the maximum extent possible. These variables enable the user to identify source records from SDTM or other ADaM datasets. The usage of these variables has been largely limited to identifying a single record from the source dataset. Traceability becomes a challenge when multiple records from a single dataset or multiple datasets contribute to the derivation of a single record in the output dataset. To address this challenge, we adapted the usage of SRCSEQ by creating it as a character variable, SRCSEQS, to include all the sequence numbers and their respective sources contributing to a specific record. This paper intends to discuss this simple, yet effective, approach in detail providing implementation examples.

## INTRODUCTION

Traceability in ADaM establishes various invisible threads that connect statistical results to raw data collected in the CRF. Traceability has greatly enhanced data review experience and has improved transparency. Based on the ADaM Implementation Guide (ADaMIG), the variable triplet SRCSEQ, SRCDOM and SRCVAR is extensively used to implement traceability in ADaM datasets. However, programmers still tend to keep records and variables from SDTM solely for traceability purposes as the usage of these variables is largely limited to identifying a single record from the source dataset. In most cases, SRCSEQ reflects the contents of xxSEQ or ASEQ variable taken from the respective source dataset.

## CURRENT CHALLENGE

Due to the increasing complexity of data analysis in clinical trials, there is a huge demand to bring information from multiple SDTM and/or ADaM datasets which poses a great challenge to implement traceability. There are two common situations where traceability is a challenge –

1. Multiple records from any single data source contributing to any single analysis result/value
2. Multiple datasets contributing to the derivation of any single analysis result/value

Some of the examples include –

1. Cumulative dose computation in Exposure Analysis Dataset
2. Imputation of the analysis value in Questionnaire Analysis Dataset
3. Creation of total score based on a specific algorithm in Questionnaire Analysis Dataset
4. Calculation of time-to-event variable in Time-to-event Analysis Dataset

## SOLUTION

Our team has had great success in implementing traceability for such complex scenarios. We identified that defining SRCSEQ as a numeric variable is the key reason for this challenge. The numeric variable limits the programmer to store only one value. To overcome this limitation we decided to define SRCSEQ as a character variable, SRCSEQS, enabling it to accept more than one value, while SRCDOM and SRCVAL variables continue to remain as character variables.

To make this more efficient and address traceability, we decided to establish a pattern while storing all sequence numbers contributing to each analysis value as a character-based SRCSEQS. So we decided

to define SRCSEQS as a combination of individual blocks separated by a comma (,). Each block will have two sections, separated by a hyphen (-). The first section is the source dataset name, either SDTM domain abbreviation or ADaM dataset name (ADxx). The second section comprises of the sequence numbers, either as a single value or as a range separated by a hyphen (-). The syntax of this pattern is as shown below –

*<source>-<sequence number/range> [, <source>-<sequence number/range>]*

where the sequence number is an individual value and the range is bound by a lower and upper value separated by '-'.

The above syntax can be visualized better with the below examples –

a. CM-29, CM-40

   This example indicates that the records from CM identified by CMSEQ values 29 and 40 contribute to the analysis result/value of the current record in the analysis dataset.

b. TR-11-13, RS-32

   This example indicates that the analysis result/value of the current record in the analysis dataset is derived using TR, RS and SE domains. The range 11-13 from TR indicates that records with TRSEQ as 11, 12, and 13 are used. Similarly, the record from RS with RSSEQ=32 is used to derive the current analysis value along with the records from TR domain.

## IMPLEMENTATION

Below are some of the implementation examples where the aforementioned solution has been implemented to demonstrate the effectiveness of the solution.

### ADEX (EXPOSURE ANALYSIS DATASET)

This example illustrates computation of average dose of study drug within the study for each patient. The below snapshot from Exposure domain, lists the dose administered for each patient/study drug combination.

| STUDYID | USUBJID | EXSEQ | EXTRT | EXDOSE | EXDOSU | EXSTDTC | EXENDTC |
|---|---|---|---|---|---|---|---|
| XYZ | XYZ-01-001 | 10 | Study Drug X | 5 | Mg | 2018-04-17 | 2018-04-28 |
| XYZ | XYZ-01-001 | 11 | Study Drug X | 10 | Mg | 2018-04-29 | 2018-05-03 |
| XYZ | XYZ-01-001 | 14 | Study Drug Y | 5 | Mg | 2018-05-04 | 2018-05-10 |
| XYZ | XYZ-01-001 | 16 | Study Drug Y | 15 | Mg | 2018-05-11 | 2018-05-20 |

The below table shows the average dose computation records where traceability is established by providing either a range of sequence numbers or individual sequence numbers used for calculation.

| STUDYID | USUBJID | PARAMCD | PARAM | AVAL | DTYPE | SRCDOM | SRCSEQS | SRCVAR |
|---|---|---|---|---|---|---|---|---|
| XYZ | XYZ-01-001 | AVGDOSX | Average dose of Study Drug X (mg) | 7.5 | DERIVED | EX | EX-10-11 | EXDOSE |
| XYZ | XYZ-01-001 | AVGDOSY | Average dose of Study Drug Y (mg) | 10 | DERIVED | EX | EX-14, EX-16 | EXDOSE |

### ADQS (QUESTIONNAIRE ANALYSIS DATASET)

This example illustrates how SRCSEQS can be used to reference derived records within the same ADaM dataset. The below records are part of QS domain where 4 Scores are collected for each visit/patient combination. As in all SDTM datasets, QSSEQ differentiates each record. Based on the analysis requirements, the ADaM dataset has to compute the total score after imputation of the missing values based on LOCF method.

| STUDYID | USUBJID | QSSEQ | QSTESTCD | QSTEST | QSSTRESN | QSSTRESC | VISIT |
|---|---|---|---|---|---|---|---|
| XYZ | XYZ-01-001 | 20 | SC01 | Score 1 | 4 | 4 | VISIT 5 |
| XYZ | XYZ-01-001 | 21 | SC02 | Score 2 | 5 | 5 | VISIT 5 |
| XYZ | XYZ-01-001 | 22 | SC03 | Score 3 | . | | VISIT 5 |
| XYZ | XYZ-01-001 | 23 | SC04 | Score 4 | 1 | 1 | VISIT 5 |

In the below table, DTYPE is set to LOCF for the imputed record and PARAMTYP is set to DERIVED for the derived parameter. The SRCSEQS for the imputed record indicates the source sequence number which contributes to the AVAL computation. Similarly, the SRCSEQS for the new parameter includes all records that contribute to the computation of total score.

| STUDYID | USUBJID | ASEQ | PARAMCD | PARAM | AVISIT | AVAL | DTYPE | PARAMTYP | SRCDOM | SRCVAR | SRCSEQS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XYZ | XYZ-01-001 | 32 | SC01 | Score 1 | VISIT 5 | 4 | | | QS | QSSTRESN | QS-20 |
| XYZ | XYZ-01-001 | 33 | SC02 | Score 2 | VISIT 5 | 5 | | | QS | QSSTRESN | QS-21 |
| XYZ | XYZ-01-001 | 34 | SC03 | Score 3 | VISIT 5 | . | | | QS | QSSTRESN | QS-22 |
| XYZ | XYZ-01-001 | 35 | SC03 | Score 3 | VISIT 5 | 5 | LOCF | | QS | QSSTRESN | QS-21 |
| XYZ | XYZ-01-001 | 36 | SC04 | Score 4 | VISIT 5 | 1 | | | QS | QSSTRESN | QS-23 |
| XYZ | XYZ-01-001 | 37 | TSCORE | Total Score | VISIT 5 | 15 | | DERIVED | ADQS | AVAL | **ADQS-32-33, ADQS-34-36** |

The above examples provide traceability for complex derivations that involve more than one source dataset and/or multiple variables. Implementation of this solution will eliminate the need to keep records from SDTM for traceability purposes increasing the analysis and storage efficiencies of ADaM datasets.

## CONCLUSION

This approach provides a clear path traceable from ADaM to its source datasets irrespective of the number of source datasets contributing to its creation. We have been able to establish traceability using this approach for complex data manipulations, including, creation of new parameters, cumulative dose computations, imputation of missing questionnaire results, identifying event dates in time-to-event data etc.

## CONTACT INFORMATION

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

Priya Saradha
Levstat® LLC
priya.saradha@levstat.com
www.levstat.com