Paper 205-2012

Examples of Building Traceability in CDISC ADaM Datasets for FDA Submission

Xiangchen (Bob) Cui , Vertex Pharmaceuticals, Inc., Cambridge, MA, USA
Hongyu Liu, Vertex Pharmaceuticals, Inc., Cambridge, MA, USA
Tathabbai Pakalapati, Vertex Pharmaceuticals, Inc., Cambridge, MA, USA

ABSTRACT

Traceability in context of ADaM data sets means providing the method followed to derive an analysis endpoint from source SDTM data. CDISC ADaM IG 1.0 strongly recommends the incorporation of traceability feature in ADaM data sets submitted to FDA. Traceability in derived data sets increases confidence and provides transparency to agency reviewers which might help in expediting the review and approval process. This paper provides examples in applying the inherent traceability features available in ADaM Basic Data Structures (BDS), adding SRCDOM, SRCVAR, and SRCSEQ variables and with examples about adding Relation Criteria and Relation Factor variables in ADaM data sets [2]. This paper tries to provide insight on tradeoffs and limitations of traceability. The examples in this paper were from FDA submissions.

INTRODUCTION

To assist review, analysis datasets and metadata must clearly communicate how the analysis datasets were created. A CDISC-compliant submission includes both SDTM and ADaM datasets; therefore, the relationship between SDTM and ADaM must be clear. This paper highlights the importance of traceability between the input data (SDTM) and the analyzed data (ADaM) [1]. There are two levels of traceability:

Metadata Traceability: Metadata means the information about data i.e. origin of variable, algorithm used to derive the variable etc. It establishes traceability by describing the algorithm used to derive or populate an analysis value from its predecessor.

Data Point Traceability: enables users (agency reviewers, QC programmers, Biostatisticians etc.) to go directly to the specific predecessor record(s) used to derive an analysis value. This level of traceability is very useful when a user is trying to trace a complex data manipulation path. It can be established by providing clear links in the data to the specific data values used as an input from predecessor to derive an analysis value.

Goals that can be achieved by incorporating traceability feature in ADaM datasets are:

- Facilitate transparency in submitted data
- Build confidence in analysis results
- Effective programming validation
- Speed up the overall review process by FDA reviewers
- Build good relationship with FDA reviewers

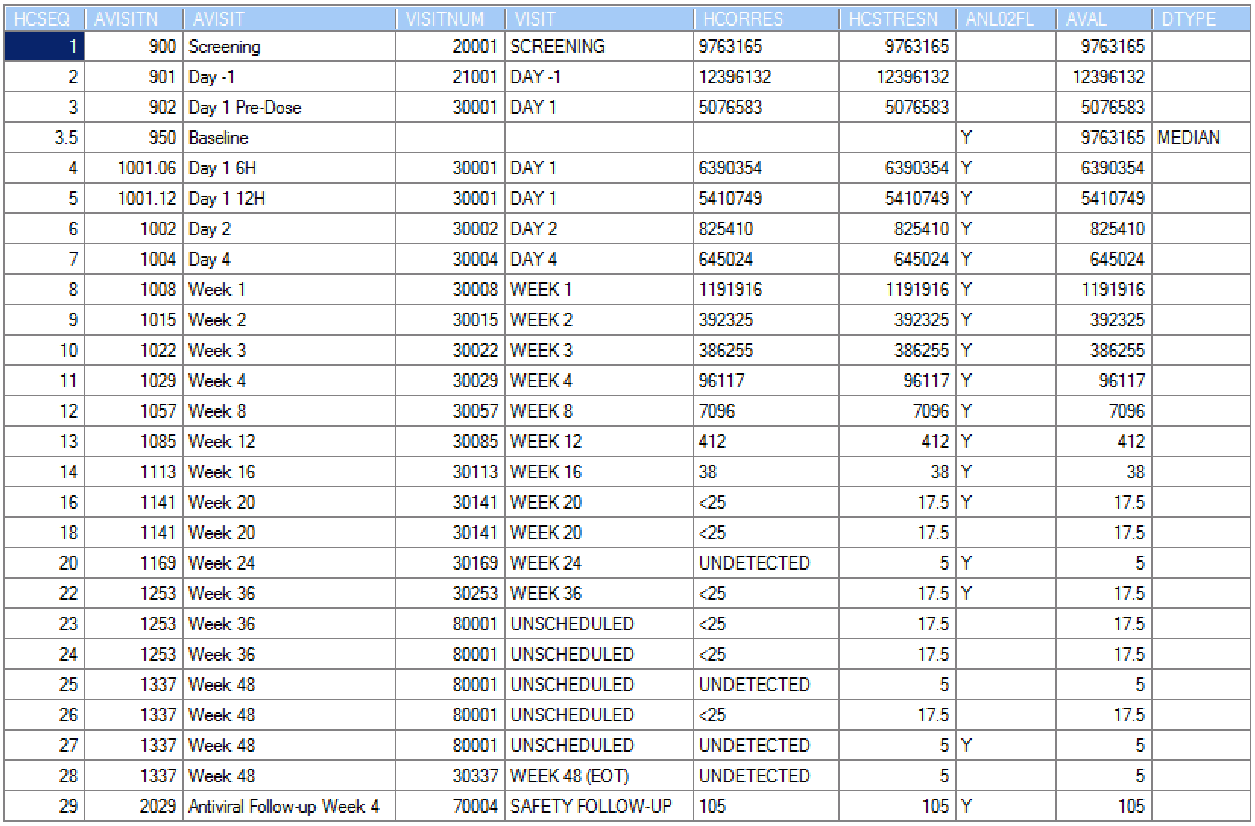
Firstly, this paper tries to present the inherent traceability features available in ADaM Basic Data Structures (BDS) and establishing metadata traceability with examples. Secondly, this paper will explore in detail on establishing Data Point Traceability with examples from FDA submissions and SAS sample codes. This section discusses three methods of establishing data point traceability, using SRCDOM, SRCVAR, and SRCSEQ triplet, using RLCRIT and RLFACT pair, and establishing traceability for Character Data Values Derived from Multiple Source Domains. Thirdly, this paper tries to explain the tradeoff of having traceability feature in ADaM datasets and limitations in incorporating traceability.

Hepatitis C Virus analysis data set and Cystic Fibrosis Clinical Event analysis data set will be used as examples to illustrate various traceability features in this paper.

ADAM BASIC DATA STRUCTURES AND METADATA TRACEABILITY

ADAM BASIC DATA STRUCTURE (BDS)

The concept of BDS does not limit number of analysis datasets that one can have in a study or number of variables/records an analysis dataset can have. So ADaM datasets can retain all those variables from SDTM datasets or add additional variables/records that help in establishing traceability. Typical examples of variables from SDTM domains that help in establishing traceability in ADaM are Sequence Variables (__SEQ), Sponsor Defined Identifiers (__SPID), Group Identifiers (__GRPID), Timing Variables (VISIT, VISITNUM, EPOCH, __DTC, __DY) etc. Examples of additional variables that can be added in ADaM to achieve some level of traceability are Analysis Flag variables (ANLzzFL) - to indicate the records that were chosen for analysis among the multiple visits that fall within the same analysis time point windows, Criterion variables CRITy - text description defining the conditions necessary to satisfy the presence of the criterion and CRITyFL - character indicator of whether the criterion described in CRITy was met. If additional records were added to analysis datasets for analyses purposes, to establish traceability, BDS allows the usage of variable DTYPE (Derivation Type) which precisely populates the derivation algorithm used to derive an analysis value.

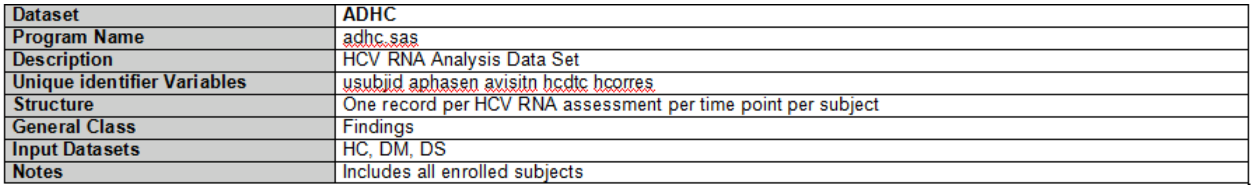


HCSEQ	AVISITN	AVISIT	VISITNUM	VISIT	HCORRES	HCSTRESN	ANL02FL	AVAL	DTYPE
1	900	Screening	20001	SCREENING	9763165	9763165		9763165	
2	901	Day -1	21001	DAY -1	12396132	12396132		12396132	
3	902	Day 1 Pre-Dose	30001	DAY 1	5076583	5076583		5076583	
3.5	950	Baseline					Y	9763165	MEDIAN
4	1001.06	Day 1 6H	30001	DAY 1	6390354	6390354	Y	6390354	
5	1001.12	Day 1 12H	30001	DAY 1	5410749	5410749	Y	5410749	
6	1002	Day 2	30002	DAY 2	825410	825410	Y	825410	
7	1004	Day 4	30004	DAY 4	645024	645024	Y	645024	
8	1008	Week 1	30008	WEEK 1	1191916	1191916	Y	1191916	
9	1015	Week 2	30015	WEEK 2	392325	392325	Y	392325	
10	1022	Week 3	30022	WEEK 3	386255	386255	Y	386255	
11	1029	Week 4	30029	WEEK 4	96117	96117	Y	96117	
12	1057	Week 8	30057	WEEK 8	7096	7096	Y	7096	
13	1085	Week 12	30085	WEEK 12	412	412	Y	412	
14	1113	Week 16	30113	WEEK 16	38	38	Y	38	
16	1141	Week 20	30141	WEEK 20	<25	17.5	Y	17.5	
18	1141	Week 20	30141	WEEK 20	<25	17.5		17.5	
20	1169	Week 24	30169	WEEK 24	UNDETECTED	5	Y	5	
22	1253	Week 36	30253	WEEK 36	<25	17.5	Y	17.5	
23	1253	Week 36	80001	UNSCHEDULED	<25	17.5		17.5	
24	1253	Week 36	80001	UNSCHEDULED	<25	17.5		17.5	
25	1337	Week 48	80001	UNSCHEDULED	UNDETECTED	5		5	
26	1337	Week 48	80001	UNSCHEDULED	<25	17.5		17.5	
27	1337	Week 48	80001	UNSCHEDULED	UNDETECTED	5	Y	5	
28	1337	Week 48	30337	WEEK 48 (EOT)	UNDETECTED	5		5	
29	2029	Antiviral Follow-up Week 4	70004	SAFETY FOLLOW-UP	105	105	Y	105	

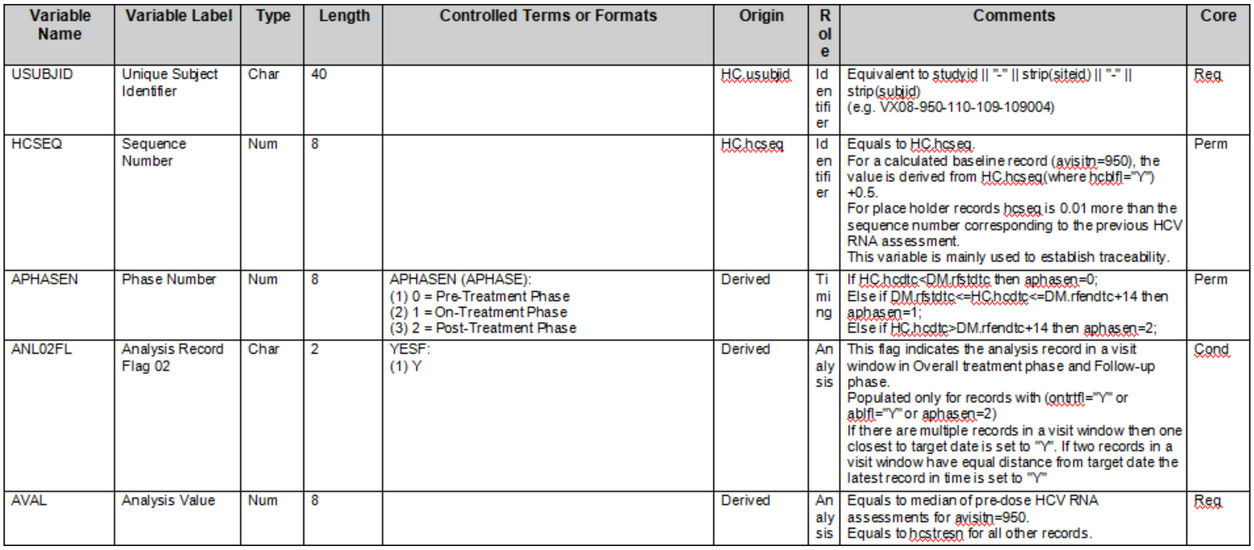
Display 1. Illustration of Usage of ANLzzFL and DTYPE Variables in ADaM Datasets

METADATA TRACEABILITY

Metadata traceability establishes traceability by describing the algorithm used to derive or populate an analysis value from its predecessor via metadata. Well defined and detailed programming specification document (define.pdf) and Define.xml is the only means of building Metadata Traceability. Display 2 shows an example of a programming specification document that enables the user to understand the relationship of an analysis variable to its source dataset(s) and variable(s).



Dataset	**ADHC**
Program Name	adhc.sas
Description	HCV RNA Analysis Data Set
Unique identifier Variables	usubjid aphasen avisitn hcdtc hcorres
Structure	One record per HCV RNA assessment per time point per subject
General Class	Findings
Input Datasets	HC, DM, DS
Notes	Includes all enrolled subjects



Variable Name	Variable Label	Type	Length	Controlled Terms or Formats	Origin	Role	Comments	Core
USUBJID	Unique Subject Identifier	Char	40		HC.usubjid	Identifier	Equivalent to studyid || "-" || strip(siteid) || "-" || strip(subjid) (e.g. VX08-950-110-109-109004)	Req
HCSEQ	Sequence Number	Num	8		HC.hcseq	Identifier	Equals to HC.hcseq. For a calculated baseline record (avisitn=950), the value is derived from HC.hcseq(where hcblfl="Y") +0.5. For place holder records hcseq is 0.01 more than the sequence number corresponding to the previous HCV RNA assessment. This variable is mainly used to establish traceability.	Perm

APHASEN	Phase Number	Num	8	APHASEN (APHASE):
				(1) 0 = Pre-Treatment Phase
				(2) 1 = On-Treatment Phase
				(3) 2 = Post-Treatment Phase
					Derived	Timing	If HC.hcdtc<DM.rfstdtc then aphasen=0;
							Else if DM.rfstdtc<=HC.hcdtc<=DM.rfendtc+14 then aphasen=1;
							Else if HC.hcdtc>DM.rfendtc+14 then aphasen=2;
								Perm

ANL02FL	Analysis Record Flag 02	Char	2	YESF: (1) Y	Derived	Analysis	This flag indicates the analysis record in a visit window in Overall treatment phase and Follow-up phase. Populated only for records with (ontrtfl="Y" or ablfl="Y" or aphasen=2) If there are multiple records in a visit window then one closest to target date is set to "Y". If two records in a visit window have equal distance from target date the latest record in time is set to "Y"	Cond
AVAL	Analysis Value	Num	8		Derived	Analysis	Equals to median of pre-dose HCV RNA assessments for avisitn=950. Equals to hcstresn for all other records.	Req

Display 2. Illustration of an ADaM Programming Specification Document

DATA POINT TRACEABILITY

This section presents the methods that can be implemented in ADaM datasets to establish Data Point Traceability of numeric data, namely, using SRCDOM, SRCVAR and SRCSEQ triplet and using RLCRIT and RLFACT variable pair [2]. It also discusses usage of SRCDOM, SRCVAR and other variables to establish Data Point Traceability of character values originating from multiple source domains.

SRCDOM, SRCVAR AND SRCSEQ TRIPLET

SDTM DOMAIN variable value, the name of the SDTM source variable, and the relevant SDTM domain --SEQ value serves as primary candidates for data point traceability [1]. ADaM implementation guide V1.0 recommends using SRCDOM, SRCVAR and SRCSEQ triplet along with derived analysis variable so that one can link back to the source SDTM records used to derive the analysis value.

Variable Name	Variable Label	Type	CDISC Notes
SRCDOM	Source Domain	Char	The 2-character identifier of the SDTM domain that relates to the derived analysis value
SRCVAR	Source Variable	Char	The name of the column (in the SDTM domain identified by SRCDOM) that relates to the derived analysis value
SRCSEQ	Source Sequence Number	Num	The sequence number SEQ of the row (in the SDTM domain identified by SRCDOM) that relates to the derived analysis value

Table 1. Definitions for SRCDOM, SRCVAR and SRCSEQ Triplet

Example of usage of SRCDOM, SRCVAR and SRCSEQ Triplet

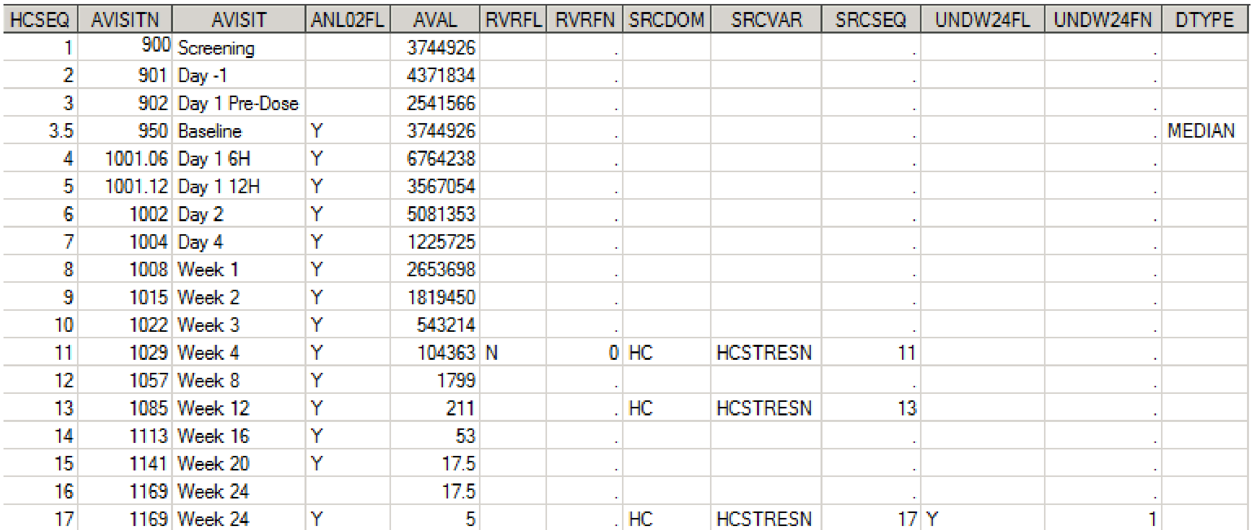
Endpoints Rapid Viral Response (RVR) defined as undetectable HCV RNA at week 4 and undetectable HCV RNA at week 24 in HCV RNA lab analysis data will be used to demonstrate the usage of SRCDOM, SRCVAR and SRCSEQ triplet in ADaM dataset. Table 2 shows specification (metadata) for endpoints RVRFL, RVRFN, UNDW24FN, UNDW24FL and for SRCDOM, SRCVAR and SRCSEQ triplet variables building data point traceability followed by a sample SAS code that populates these variables. Display 3 shows the snapshot of these variables in an analysis dataset.

Variable Name	Variable Label	Type	Length	Controlled Terms or Formats	Comments
RVRFL	Rapid Viral Response Flag	Char	2		Equals to "Y" if a subject has undetectable HCV RNA at Week 4 i.e.

					HCORRES="UNDETECTED" at avisitn=1029 and anl02fl="Y". Else equals to "N".
RVRFN	Rapid Viral Response Flag (N)	Num	8	YESNOFN (RVRFL): (1) 1 = Y (2) 0 = N	Equals to 1 if rvrfl="Y". Equals to 0 if rvrfl="N".
UNDW24FL	Undetectable HCV RNA at Week 24	Char	2		Equals to "Y" if a subject has undetectable HCV RNA at Week 24 i.e HCORRES="UNDETECTED" at avisitn=1169 and anl02fl ="Y". Else equals to "N".
UNDW24FN	Undetectable HCV RNA at Week 24 (N)	Num	8	YESNOFN (UNDW24FL): (1) 1 = Y (2) 0 = N	Equals to 1 if undw24fl="Y" Equals to 0 if undw24fl="N"
SRCDOM	Source Domain	Char	4		= "HC" for avisitn=1029 and anl02fl ="Y" = "HC" for avisitn=1169 and anl02fl ="Y"
SRCVAR	Source Variable	Char	8		Equal to “HCSTRESN” for avisitn=1029 and anl02fl="Y" Equal to “HCSTRESN” for avisitn=1169 and anl02fl="Y"
SRCSEQ	Source Sequence Number	Num	8		= HCSEQ for avisitn=1029 and anl02fl ="Y" = HCSEQ for avisitn=1169 and anl02fl ="Y"

Table 2. Specification of Analysis Endpoints and Triplet Variables

```
data adhc;
     set hc_1;
     by usubjid hcseq;
 /**aphasen=1 means on-treatment phase and avisitn=1029 means analysis Week 4**/
 /******anl02fl=”Y” selects the analysis record at a Visit*******/
   if aphasen=1 and avisitn=1029 and (anl02fl="Y" or dtype="PLACE HOLDER") then do;
                 if aval=5 then do;rvrfn=1;rvrfl="Y";end;
                 else do;rvrfn=0;rvrfl="N";end;
                  srcdom="HC";
                  srcvar="HCSTRESN";
                  srcseq=hcseq;
   end;
   if aphasen=1 and avisitn=1169 and (anl02fl="Y" or dtype="PLACE HOLDER") then do;
                 if aval=5 then do;undw24fn=1;undw24fl="Y";end;
                  else do;undw24fn=0;undw24fl="N";end;
                 srcdom="HC";
                 srcvar="HCSTRESN";
                 srcseq=hcseq;
   end;
run;
```



HCSEQ	AVISITN	AVISIT	ANL02FL	AVAL	RVRFL	RVRFN	SRCDOM	SRCVAR	SRCSEQ	UNDW24FL	UNDW24FN	DTYPE
1	900	Screening		3744926		.			.		.	
2	901	Day -1		4371834		.			.		.	
3	902	Day 1 Pre-Dose		2541566		.			.		.	
3.5	950	Baseline	Y	3744926		.			.		.	MEDIAN
4	1001.06	Day 1 6H	Y	6764238		.			.		.	
5	1001.12	Day 1 12H	Y	3567054		.			.		.	
6	1002	Day 2	Y	5081353		.			.		.	
7	1004	Day 4	Y	1225725		.			.		.	
8	1008	Week 1	Y	2653698		.			.		.	
9	1015	Week 2	Y	1819450		.			.		.	
10	1022	Week 3	Y	543214		.			.		.	
11	1029	Week 4	Y	104363	N	0	HC	HCSTRESN	11		.	
12	1057	Week 8	Y	1799		.			.		.	
13	1085	Week 12	Y	211		.	HC	HCSTRESN	13		.	
14	1113	Week 16	Y	53		.			.		.	
15	1141	Week 20	Y	17.5		.			.		.	
16	1169	Week 24		17.5		.			.		.	
17	1169	Week 24	Y	5		.	HC	HCSTRESN	17	Y	1	

Display 3. Illustration of SRCDOM, SRCVAR and SRCSEQ Triplet Establishing Data Point Traceability for Rapid Viral Response and Undetectable HCV RNA at Week 24 Endpoints in an Analysis Dataset

The above example shows that SRCDOM, SRCVAR and SRCSEQ triplet builds a clear path from an ADaM record to its predecessor in source SDTM. Same SRCDOM, SRCVAR and SRCSEQ triplet can be applied for multiple derived variables at different time points. It also demonstrates the importance of traceability when multiple HCV RNA records are present at visit window week 24.

Limitations of SRC Triplet are:

- Can be applied only if the predecessor record used to derive an analysis variable comes from single SDTM domain
- Can be applied only if the derived analysis variable depends only on a single record and single variable from the source SDTM

RLCRIT AND RLFACT PAIR

Both SRC triplet method or CRITy and CRITyFL method mentioned in ADaM implementation guide V1.0 cannot be used to build Data Point Traceability when an analysis value depends on multiple records corresponding to different time points from a source SDTM domain. This limitation can be overcome by using RLCRIT and RLFACT variable pair. RLCRIT – Relation Criteria variable stores data source (ADaM or SDTM) along with source variables used in the derivation of analysis value in the derivation rule. RLFACT – Relation Factor variable stores values of those variables mentioned in RLCRIT in the same order.

Example 1 of Usage of RLCRIT and RLFACT Variable Pair

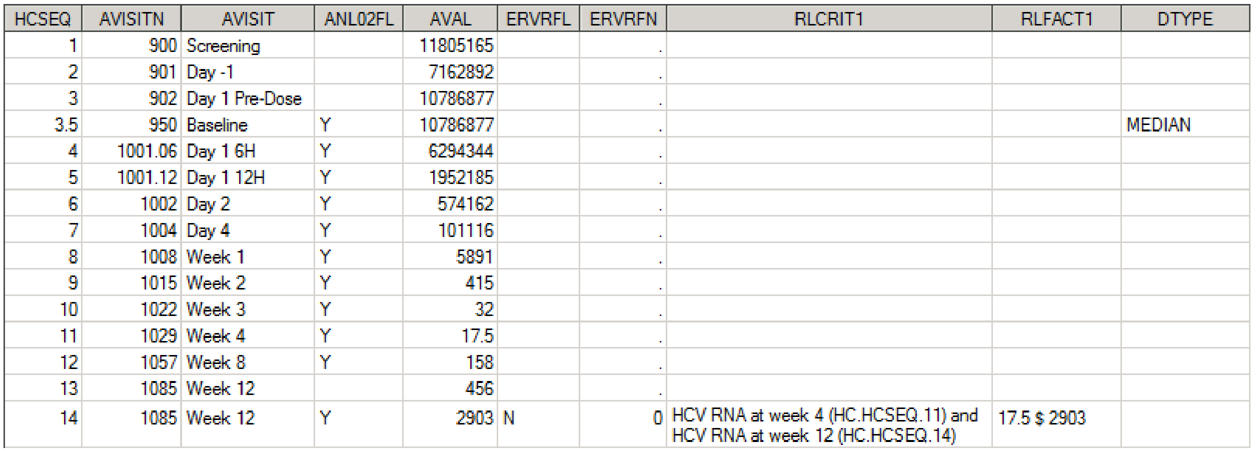
Endpoint Extended Rapid Viral Response (eRVR) in HCV RNA lab analysis data will be used to demonstrate the usage of RLCRIT and RLFACT variable pair in ADaM dataset. Table 3 shows specification (metadata) for the endpoint ERVRFL, ERVRFN and for RLCRIT and RLFACT variable pair building data point traceability followed by a sample SAS code that populates these variables. Display 4 shows the snapshot of these variables in an analysis dataset. eRVR is defined as undetectable HCV RNA (defined as HC.HCSTRESN=5) at week 4 and at week 12. Two records corresponding to week 4 and week 12 and two HCV RNA values (HC.HCSTRESN) are needed to build data point traceability for this endpoint.

Variable Name	Variable Label	Type	Length	Controlled Terms or Formats	Comments
ERVRFL	Extended Rapid Viral Response Flag	Char	2		Equals to "Y" if a subject has undetectable HCV RNA at Week 4 and Week 12 i.e HCORRES="UNDETECTED" at avisitn=1029 and avisitn=1085 and anl02fl="Y". Else equals to "N".
ERVRFN	Extended Rapid	Num	8	YESNOFN (ERVRFL):	Equals to 1 if ervrfl="Y"

<table>
<tr><td></td><td>Viral Response Flag (N)</td><td></td><td></td><td>(1) 1 = Y
(2) 0 = N</td><td>Equals to 0 if ervrfl="N"</td></tr>
<tr><td>RLCRIT1</td><td>Parameter Relation Criteria For ERVR</td><td>Char</td><td>200</td><td></td><td>if aval was non-missing at week 12 then do;
if aval was non-missing at week 4 then do;
RLCRIT1="HCV RNA at week 4(HC.HCSEQ."||
||strip(put(srcseq at week 4,best.)) ||") and HCV
RNA at week 12 at (HC.HCSEQ.”||strip(put(hcseq,
best.))||”)”;
end;
else do;
RLCRIT1="HCV RNA at week 4 was missing! and
HCV RNA at week 12 at (HC.HCSEQ.”||strip(put(
Hcseq,best.))||”)”;
end;
end;
else do;
if aval was non-missing at week 4 then do;
RLCRIT1="HCV RNA at week 4
(HC.HCSEQ."||strip(put(srcseq at week 4,best.))||”)
and HCV RNA at week 12 was missing!";
end;
else do;
RLCRIT1="HCV RNA at week 4 was missing"||"
and HCV RNA at week 12 was missing!";
end;
end;</td></tr>
<tr><td>RLFACT1</td><td>Parameter Relation Factors For ERVR</td><td>Char</td><td>80</td><td></td><td>if aval was non-missing at week 12 then do;
if aval was non-missing at week 4 then do;
RLFACT1=strip(avalc at week4)||" $ “||strip(put(aval,
best.));
end;
else do;
RLFACT1="Missing $ "||strip(put(aval,best.));
end;
end;
else do; if aval was non-missing at week 4 then RLFACT1=strip(avalc at week 4)||" $ Missing";
else RLFACT1="Missing"||" $ Missing";
end;</td></tr>
</table>

Table 3. Specification of Analysis Endpoint eRVR and RLCRIT and RLFACT Variable Pair

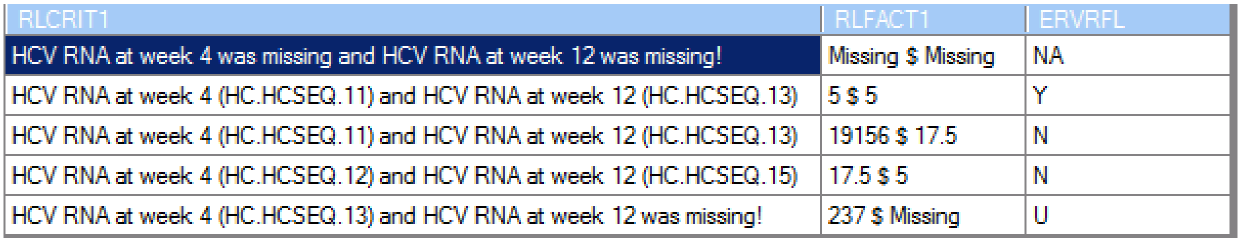
```
data ervr;
   set week4hc week12hc;

   if aphasen=1 and avisitn=1085 and (anl02fl="Y" or dtype="PLACE HOLDER") then do;

              if week4hc=5 and week12hc=5 then do;ervrfn=1;ervrfl="Y";end;
              else do;  ervrfn=0;ervrfl="N";end;

       if week12hc ne . then do;

                 if week4hc ne . then do;rlcrit1="HCV RNA at week 4
                 (HC.HCSEQ."||strip(put(week4seq,best.))||") and HCV RNA at week 12
                 (HC.HCSEQ."||strip(put(week12seq,best.))||")";
                 rlfact1=strip(put(week4hc,best.))||" $ "||strip(put(week12hc,best.));
                end;
                else do;
                 rlcrit1="HCV RNA at week 4 was missing! and HCV RNA at week 12
                 (HC.HCSEQ."||strip(put(week12seq,best.))||")";
                 rlfact1="Missing $ "||strip(put(week12hc,best.));
                 end;
       end;

       else do;
                 if week4hc ne . then do;
                 rlcrit1="HCV RNA at week 4 (HC.HCSEQ."||strip(put(week4seq,best.))||")
                 and HCV RNA at week 12 was missing!";
                 rlfact1=strip(put(week4hc,best.))||" $ Missing";
                 end;
                  else do;
                 rlcrit1="HCV RNA at week 4 was missing and HCV RNA at week 12 was
                 missing!";
                 rlfact1="Missing"||" $ Missing";
                 end;
       end;

   end;
   run;
```



HCSEQ	AVISITN	AVISIT	ANL02FL	AVAL	ERVRFL	ERVRFN	RLCRIT1	RLFACT1	DTYPE
1	900	Screening		11805165		.			
2	901	Day -1		7162892		.			
3	902	Day 1 Pre-Dose		10786877		.			
3.5	950	Baseline	Y	10786877		.			MEDIAN
4	1001.06	Day 1 6H	Y	6294344		.			
5	1001.12	Day 1 12H	Y	1952185		.			
6	1002	Day 2	Y	574162		.			
7	1004	Day 4	Y	101116		.			
8	1008	Week 1	Y	5891		.			
9	1015	Week 2	Y	415		.			
10	1022	Week 3	Y	32		.			
11	1029	Week 4	Y	17.5		.			
12	1057	Week 8	Y	158		.			
13	1085	Week 12		456		.			
14	1085	Week 12	Y	2903	N	0	HCV RNA at week 4 (HC.HCSEQ.11) and HCV RNA at week 12 (HC.HCSEQ.14)	17.5 $ 2903	

Display 4. Illustration of RLCRIT and RLFACT Variable Pair Establishing Data Point Traceability

In above example record with HCSEQ=11 at analysis week 4 and record with HCSEQ=14 at analysis week 12 are used to derive the Extended Rapid Viral Response endpoint for a subject. This information is stored in RLCRIT1 variable. The HCV RNA values at these time points are 17.5 and 2903 are stored in RLFACT1 variable, separated by symbol “$”. RLCRIT and RLFACT variable pair clearly show that ERVRFL=”N” and how it was derived. Display 5 shows various values of RLCRIT and RLFACT for different subjects in a study populated at analysis visit Week 12.



RLCRIT1	RLFACT1	ERVRFL
HCV RNA at week 4 was missing and HCV RNA at week 12 was missing!	Missing $ Missing	NA
HCV RNA at week 4 (HC.HCSEQ.11) and HCV RNA at week 12 (HC.HCSEQ.13)	5 $ 5	Y
HCV RNA at week 4 (HC.HCSEQ.11) and HCV RNA at week 12 (HC.HCSEQ.13)	19156 $ 17.5	N
HCV RNA at week 4 (HC.HCSEQ.12) and HCV RNA at week 12 (HC.HCSEQ.15)	17.5 $ 5	N
HCV RNA at week 4 (HC.HCSEQ.13) and HCV RNA at week 12 was missing!	237 $ Missing	U

Display 5. Illustration of Various Possible Values of RLCRIT and RLFACT Variable Pair in a Study for Different Subjects

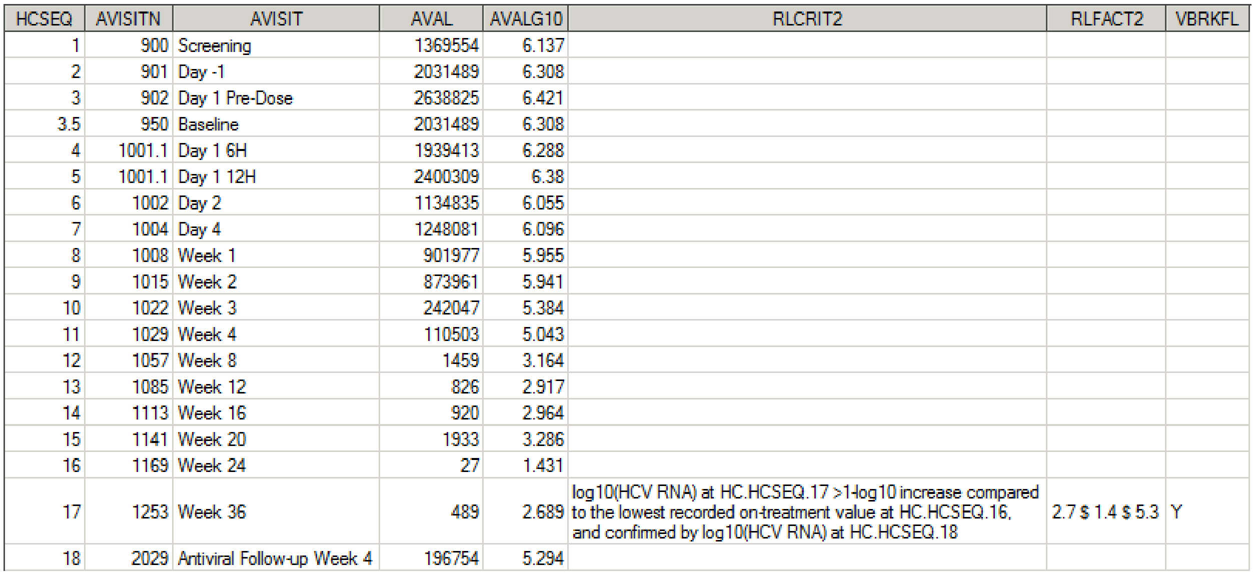
Example 2 of Usage of RLCRIT and RLFACT Variable Pair

Endpoint Viral Breakthrough (VBT) in HCV RNA lab analysis data will be used as a second example to demonstrate the usage of RLCRIT and RLFACT variable pair in ADaM dataset. Table 4 shows specification (metadata) for the endpoint Viral Breakthrough and for RLCRIT and RLFACT variable pair which helps in building data point traceability followed by a sample SAS code that populates these variables. Display 6 and Display 7 shows the snapshot of these variables in an analysis dataset. The definition of VBT as shown in the comments column of VBRKFL requires three HCR RNA records, i.e. the current record, nadir at a prior time point, and a confirmatory record. Hence three “locations” (HCSEQ) and three HCV RNA level (HCSTRESN) are needed for traceability!

Variable Name	Variable Label	Type	Length	Controlled Terms or Formats	Comments
VBRKFL	Viral Breakthrough Flag (Confirmed)	Char	2	YESF: (1) Y	Set to "Y" for the event initiating the confirmed viral breakthrough on-treatment records. Viral Breakthrough is defined as : a) Confirmed >1-log10 IU/mL HCV RNA on-treatment increase from nadir or b) Confirmed >100 IU/mL HCV RNA following an undetectable HCV RNA at a prior time point. Note: New definition from FDA: >LLOQ from >100 IU/mL
RLCRIT2	Parameter Relation Criteria For VBT	Char	200		Equals to "log10(HCV RNA) at HC.HCSEQ."||strip(put(log10incseq,best.))||" >1-log10 increase compared to the lowest recorded on-treatment value at HC.HCSEQ."||strip(put(nadirseq,best.))||", and confirmed by log10(HCV RNA) at HC.HCSEQ."||strip(put(log10incconfirmseq,best.)); or Equals to "HCV RNA at HC.HCSEQ."||strip(put(gt100seq,best.))||" >100 IU/mL, undetectable at HC.HCSEQ."||strip(put(undetectseq,best.))||", and confirmed by HCV RNA at HC.HCSEQ."||strip(put(gt100confirmseq,best.));
RLFACT2	Parameter Relation Factors For VBT	Char	80		=strip(put(round(log10(log10incseqaval),0.1),best.))||" $ "||strip(put(round(log10(nadir),0.1),best.))||" $ "||strip(put(round(log10(log10incconfirmseqaval),0.1),best.)) or = strip(put(gt100seqaval,best.))||" $ 5 $ "||strip(put(gt100confirmseqaval,best.));

Table 4. Specification of Analysis Endpoint Viral Breakthrough and RLCRIT and RLFACT Variable Pair

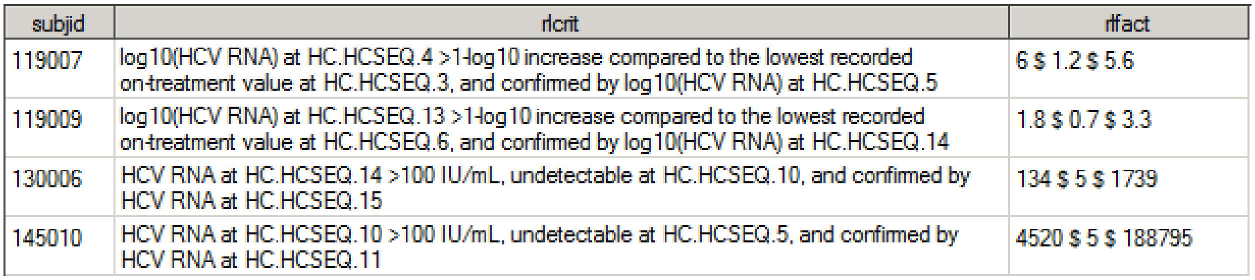
```
proc sort data=hc;by usubjid avisitn hcseq;run;

data hc;
retain nadir vbrkidx1 vbrkidx2 code1 code2 undetect log10incseq log10incconfirmseq
       gt100seq gt100confirmseq undetectseq nadirseq log10incseqaval
       log10incconfirmseqaval gt100seqaval gt100confirmseqaval;
set hc;
by usubjid hcseq;
if first.usubjid then do;/*Initializing State Variables*/
   vbrkidx1=.;/*Identifies Viral Breakthrough Criteria 1*/
   vbrkidx2=.;/*Identifies Viral Breakthrough Criteria 2*/
   code1=0;/*Identifies the >1-log10 increase condition*/
   code2=0;/*Identifies undetectable and HCV RNA>100IU/mL condition*/
   undetect=0;
   nadir=hcstresn;/*Initializing the nadir value equal to first on-treatment value*/

   log10incseq=.;
   log10incconfirmseq=.;
   gt100seq=.;
   gt100confirmseq=.;
   undetectseq=.;
   nadirseq=.;
   log10incseqaval=.;
   log10incconfirmseqaval=.;
   gt100seqaval=.;
   gt100confirmseqaval=.;
   end;

if hcstresn ne . then log10inc=log10(hcstresn)-log10(nadir);/*Log10 inc from Nadir*/
if phasefn in (1,2) then do;
   if log10inc>1 then do;/*Condition 1*/
       if code1=0 then code1=1;
       else if code1=1 then code1=2;/*Confirmation*/
       if code1=1 then do; vbrkidx1=avisitn;log10incseq=hcseq;log10incseqaval=aval;
       end;
       if code1=2 then do;
       vbrkidx1=vbrkidx1;log10incconfirmseq=hcseq;log10incconfirmseqaval=aval;
       end;
   end;

  else if hcstresn ne . then code1=0;/*Reset if not confirmed*/

   if undetect=1 and hcstresn>100 then do;/*Condition 2*/
       if code2=0 then code2=1;
       else if code2=1 then code2=2;/*Confirmation*/
       if code2=1 then do; vbrkidx2=avisitn;gt100seq=hcseq;gt100seqaval=aval;
       end;
       if code2=2 then do;vbrkidx2=vbrkidx2;gt100confirmseq=hcseq;
                          gt100confirmseqaval=aval;
       end;
   end;
   else if hcstresn ne . then code2=0;/*Reset if not confirmed*/
end;

if undetect eq 0 and hcstresn=5 then do;undetect=1;undetectseq=hcseq;end;
       /************Update Nadir if new nadir found***********/
if hcstresn ne . and hcstresn<nadir then do;nadir=hcstresn;nadirseq=hcseq;end;
run;

proc sort data= hc out=vbrk;by usubjid hcseq;
      where (code1=2 and vbrkidx1<2029) or (code2=2 and vbrkidx2<2029);
run;
data vbrk;
   length rlcrit2 $200 rlfact2 $80;
   set vbrk;
   by usubjid hcseq;
   if first.usubjid;
   if nmiss(vbrkidx1,vbrkidx2)=1 then vbrkidx=sum(vbrkidx1,vbrkidx2);
   else vbrkidx=vbrkidx2;
    if vbrkidx1 ne . and vbrkidx2 eq . then do;
       rlcrit2="log10(HCV RNA) at HC.HCSEQ."||strip(put(log10incseq,best.))||" >1-
       log10 increase compared to the lowest recorded on-treatment value at
       HC.HCSEQ."||strip(put(nadirseq,best.))||", and confirmed by log10(HCV RNA) at
       HC.HCSEQ."||strip(put(log10incconfirmseq,best.));

       rlfact2=strip(put(round(log10(log10incseqaval),0.1),best.))||" $
       "||strip(put(round(log10(nadir),0.1),best.))||" $
       "||strip(put(round(log10(log10incconfirmseqaval),0.1),best.));

   end;
   else if vbrkidx2 ne . or nmiss(vbrkidx1,vbrkidx2)=2 then do;
       rlcrit2="HCV RNA at HC.HCSEQ."||strip(put(gt100seq,best.))||" >100 IU/mL,
       undetectable at HC.HCSEQ."||strip(put(undetectseq,best.))||
       ", and confirmed by HCV RNA at HC.HCSEQ."|| strip(put(gt100confirmseq,best.));
       rlfact2=strip(put(gt100seqaval,best.))||" $ 5 $
       "||strip(put(gt100confirmseqaval,best.));
   end;
run;

data hc;
     merge hc(in=a) vbrk(in=b keep=usubjid hcseq vbrkidx rlcrit2 rlfact2);
     by usubjid hcseq;
     if a;
     if b then vbrkfl="Y";
run;
```



HCSEQ	AVISITN	AVISIT	AVAL	AVALG10	RLCRIT2	RLFACT2	VBRKFL
1	900	Screening	1369554	6.137			
2	901	Day -1	2031489	6.308			
3	902	Day 1 Pre-Dose	2638825	6.421			
3.5	950	Baseline	2031489	6.308			
4	1001.1	Day 1 6H	1939413	6.288			
5	1001.1	Day 1 12H	2400309	6.38			
6	1002	Day 2	1134835	6.055			
7	1004	Day 4	1248081	6.096			
8	1008	Week 1	901977	5.955			
9	1015	Week 2	873961	5.941			
10	1022	Week 3	242047	5.384			
11	1029	Week 4	110503	5.043			
12	1057	Week 8	1459	3.164			
13	1085	Week 12	826	2.917			
14	1113	Week 16	920	2.964			
15	1141	Week 20	1933	3.286			
16	1169	Week 24	27	1.431			
17	1253	Week 36	489	2.689	log10(HCV RNA) at HC.HCSEQ.17 >1-log10 increase compared to the lowest recorded on-treatment value at HC.HCSEQ.16, and confirmed by log10(HCV RNA) at HC.HCSEQ.18	2.7 $ 1.4 $ 5.3	Y
18	2029	Antiviral Follow-up Week 4	196754	5.294			

Display 6. Illustration of RLCRIT and RLFACT Variable Pair Establishing Data Point Traceability for Viral Breakthrough Endpoint in an Analysis Dataset

In above example log10 increase in HCV RNA value at analysis Week 36 (HCSEQ=17) is greater by factor 1 compared to the HCV RNA value at Week 24 (HCSEQ=16). This increase by factor 1 is confirmed by HCV RNA value at Follow-up Week 4 (HCSEQ=18). The position of records contributing to Viral Breakthrough is stored in RLCRIT variable in a sequential order. The log10 value (AVALG10) used in the derivation of Viral Breakthrough is stored in RLFACT variable separated by symbol “$” following the same order as in RLCRIT variable.



subjid	rlcrit	rlfact
119007	log10(HCV RNA) at HC.HCSEQ.4 >1-log10 increase compared to the lowest recorded on-treatment value at HC.HCSEQ.3, and confirmed by log10(HCV RNA) at HC.HCSEQ.5	6 $ 1.2 $ 5.6
119009	log10(HCV RNA) at HC.HCSEQ.13 >1-log10 increase compared to the lowest recorded on-treatment value at HC.HCSEQ.6, and confirmed by log10(HCV RNA) at HC.HCSEQ.14	1.8 $ 0.7 $ 3.3
130006	HCV RNA at HC.HCSEQ.14 >100 IU/mL, undetectable at HC.HCSEQ.10, and confirmed by HCV RNA at HC.HCSEQ.15	134 $ 5 $ 1739
145010	HCV RNA at HC.HCSEQ.10 >100 IU/mL, undetectable at HC.HCSEQ.5, and confirmed by HCV RNA at HC.HCSEQ.11	4520 $ 5 $ 188795

Display 7. Illustration of Various Possible Values of RLCRIT and RLFACT Variable Pair for Viral Breakthrough in a Study

Hence using RLCRIT and RLFACT variable pair can establish Data Point Traceability in situations where an analysis endpoint depends on multiple records corresponding to different time points from a single source SDTM domain. The limitation of this method is unable to build traceability when an analysis endpoint/variable depends on multiple records originating from different source SDTM datasets.

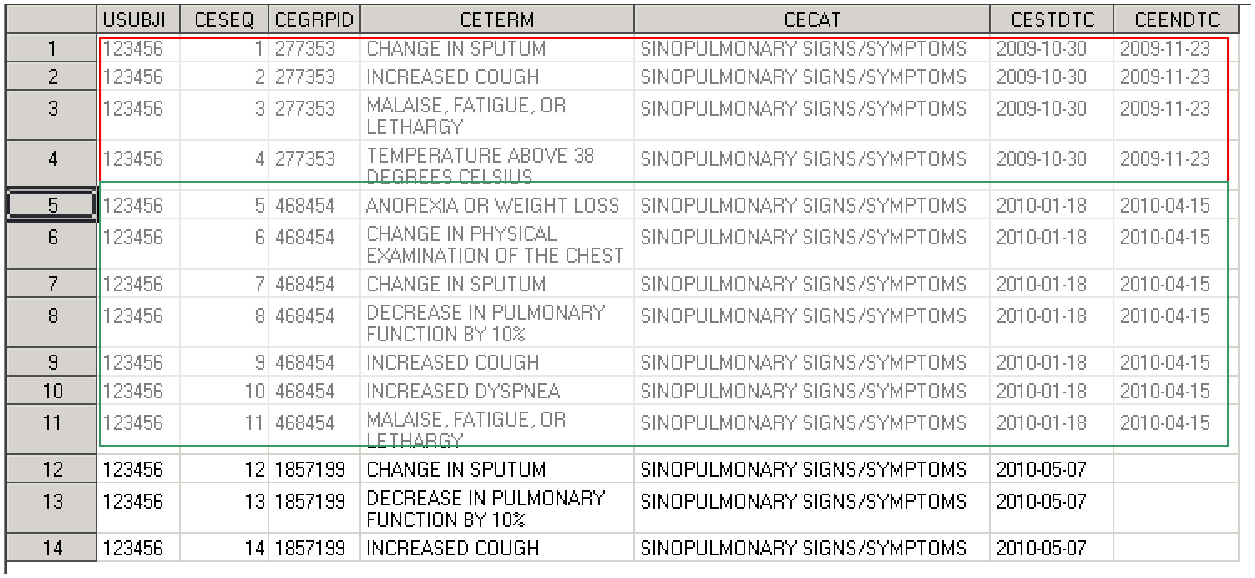
TRACEABILITY FOR CHARACTER DATA VALUES DERIVED FROM MULTIPLE SOURCE DOMAINS

There were 13 predefined clinical events that were analyzed for one of our clinical studies. These events may be derived from character outputs of one or more source SDTM domains. The source domains include SDTM CE, CM, HO, SUPPCM, and SUPPHO. It was more important and complicated to clearly show the traceability. We take the IV antibiotic therapy administrated for pulmonary exacerbation event as an example. A pulmonary exacerbation in the study protocol was defined as ‘An event that a subject has a change in antibiotic therapy (IV, inhaled or oral), due to occurrences of at least four of 12 predefined Sinopulmonary signs/symptoms within an antibiotic therapy course’.

Within an antibiotic therapy course, there may be more than one antibiotics used. The antibiotic therapy data was collected in SDTM CM domain and the Sinopulmonary signs/symptoms were in SDTM CE domain. Not all signs/symptoms appeared at the same time. Mostly, one or two signs/symptoms trigged antibiotic therapy change, later within the course, additional signs/symptoms occurred. There may be some antibiotic therapy use changes due to 1-3 signs/symptoms within the entire antibiotic course. And those were not qualified as a pulmonary event. In our data we had to collect every signs/symptoms. Therefore, the relationship between signs/symptoms in SDTM CE and CM was not simply 1 to 1. If there was a pulmonary exacerbation occurred for a subject, there must be four or more signs/symptoms in CE domain that related to one or more antibiotic therapies in CM domain. At the same time, antibiotics could be administrated as inhale, oral or intravenous.

In order to build the traceability three variables were used in our analysis dataset: SRCDOM SRCIDVAR and SRCIDVAL.

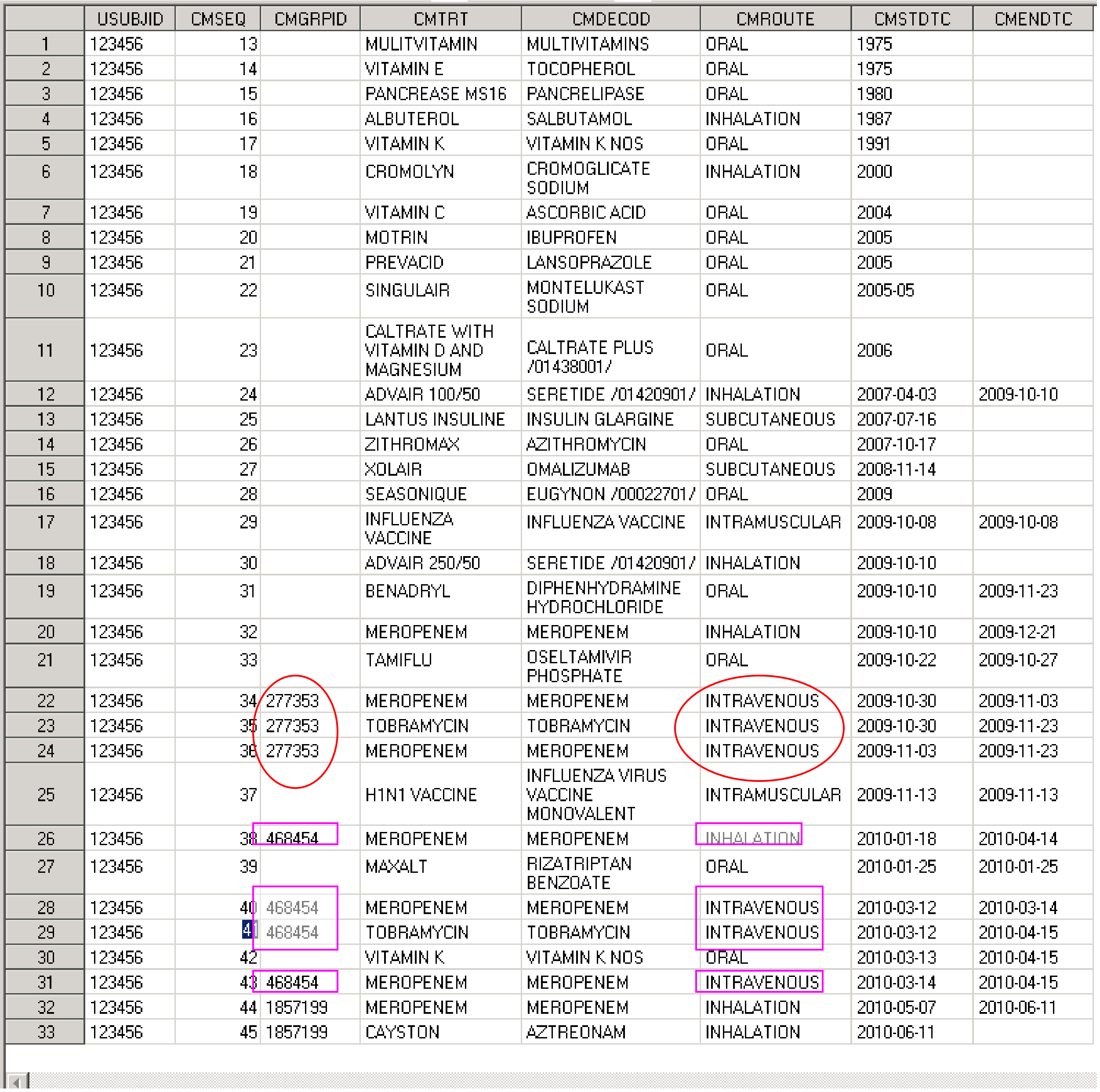
Display 8 shows the signs/symptoms collected in SDTM CE domain for subject “123456”. It shows that this subject has a total of 14 signs/symptoms during the study within three antibiotic therapy courses. The first antibiotic therapy course has four signs/symptoms indicated with a group identifier (CEGRPID) “277353”. The second course has 7 signs/symptoms with a group identifier “468454” and the third one had a group identifier of “1857199” with three signs/symptoms. Therefore, the first two courses are qualified as pulmonary exacerbations while the third is not at the time of the data collection.



	USUBJI	CESEQ	CEGRPID	CETERM	CECAT	CESTDTC	CEENDTC
1	123456	1	277353	CHANGE IN SPUTUM	SINOPULMONARY SIGNS/SYMPTOMS	2009-10-30	2009-11-23
2	123456	2	277353	INCREASED COUGH	SINOPULMONARY SIGNS/SYMPTOMS	2009-10-30	2009-11-23
3	123456	3	277353	MALAISE, FATIGUE, OR LETHARGY	SINOPULMONARY SIGNS/SYMPTOMS	2009-10-30	2009-11-23
4	123456	4	277353	TEMPERATURE ABOVE 38 DEGREES CELSIUS	SINOPULMONARY SIGNS/SYMPTOMS	2009-10-30	2009-11-23
5	123456	5	468454	ANOREXIA OR WEIGHT LOSS	SINOPULMONARY SIGNS/SYMPTOMS	2010-01-18	2010-04-15
6	123456	6	468454	CHANGE IN PHYSICAL EXAMINATION OF THE CHEST	SINOPULMONARY SIGNS/SYMPTOMS	2010-01-18	2010-04-15
7	123456	7	468454	CHANGE IN SPUTUM	SINOPULMONARY SIGNS/SYMPTOMS	2010-01-18	2010-04-15
8	123456	8	468454	DECREASE IN PULMONARY FUNCTION BY 10%	SINOPULMONARY SIGNS/SYMPTOMS	2010-01-18	2010-04-15
9	123456	9	468454	INCREASED COUGH	SINOPULMONARY SIGNS/SYMPTOMS	2010-01-18	2010-04-15
10	123456	10	468454	INCREASED DYSPNEA	SINOPULMONARY SIGNS/SYMPTOMS	2010-01-18	2010-04-15
11	123456	11	468454	MALAISE, FATIGUE, OR LETHARGY	SINOPULMONARY SIGNS/SYMPTOMS	2010-01-18	2010-04-15
12	123456	12	1857199	CHANGE IN SPUTUM	SINOPULMONARY SIGNS/SYMPTOMS	2010-05-07	
13	123456	13	1857199	DECREASE IN PULMONARY FUNCTION BY 10%	SINOPULMONARY SIGNS/SYMPTOMS	2010-05-07	
14	123456	14	1857199	INCREASED COUGH	SINOPULMONARY SIGNS/SYMPTOMS	2010-05-07	

Display 8. Signs/Symptoms Collected in SDTM CE Domain for a Subject

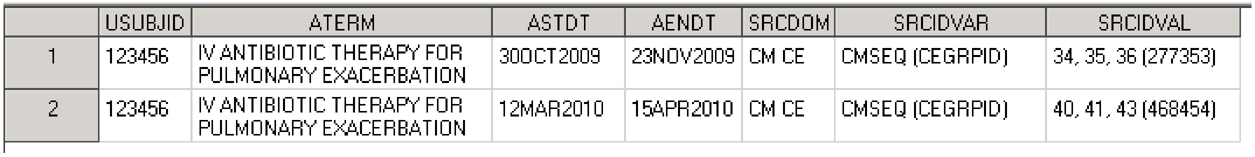
Display 9 below shows the concomitant medication data collected in SDTM CM domain for subject 123456. Corresponding to each antibiotic therapy course group in CE there is at least one antibiotic therapy. The figure shows three antibiotics (all via IV) are used during the first antibiotic course (highlighted in cycle), four (3 IV and 1 inhaled) for the second course and two (all inhaled) for the third course.



	USUBJID	CMSEQ	CMGRPID	CMTRT	CMDECOD	CMROUTE	CMSTDTC	CMENDTC
1	123456	13		MULITVITAMIN	MULTIVITAMINS	ORAL	1975	
2	123456	14		VITAMIN E	TOCOPHEROL	ORAL	1975	
3	123456	15		PANCREASE MS16	PANCRELIPASE	ORAL	1980	
4	123456	16		ALBUTEROL	SALBUTAMOL	INHALATION	1987	
5	123456	17		VITAMIN K	VITAMIN K NOS	ORAL	1991	
6	123456	18		CROMOLYN	CROMOGLICATE SODIUM	INHALATION	2000	
7	123456	19		VITAMIN C	ASCORBIC ACID	ORAL	2004	
8	123456	20		MOTRIN	IBUPROFEN	ORAL	2005	
9	123456	21		PREVACID	LANSOPRAZOLE	ORAL	2005	
10	123456	22		SINGULAIR	MONTELUKAST SODIUM	ORAL	2005-05	
11	123456	23		CALTRATE WITH VITAMIN D AND MAGNESIUM	CALTRATE PLUS /01438001/	ORAL	2006	
12	123456	24		ADVAIR 100/50	SERETIDE /01420901/	INHALATION	2007-04-03	2009-10-10
13	123456	25		LANTUS INSULINE	INSULIN GLARGINE	SUBCUTANEOUS	2007-07-16	
14	123456	26		ZITHROMAX	AZITHROMYCIN	ORAL	2007-10-17	
15	123456	27		XOLAIR	OMALIZUMAB	SUBCUTANEOUS	2008-11-14	
16	123456	28		SEASONIQUE	EUGYNON /00022701/	ORAL	2009	
17	123456	29		INFLUENZA VACCINE	INFLUENZA VACCINE	INTRAMUSCULAR	2009-10-08	2009-10-08
18	123456	30		ADVAIR 250/50	SERETIDE /01420901/	INHALATION	2009-10-10	
19	123456	31		BENADRYL	DIPHENHYDRAMINE HYDROCHLORIDE	ORAL	2009-10-10	2009-11-23
20	123456	32		MEROPENEM	MEROPENEM	INHALATION	2009-10-10	2009-12-21
21	123456	33		TAMIFLU	OSELTAMIVIR PHOSPHATE	ORAL	2009-10-22	2009-10-27
22	123456	34	277353	MEROPENEM	MEROPENEM	INTRAVENOUS	2009-10-30	2009-11-03
23	123456	35	277353	TOBRAMYCIN	TOBRAMYCIN	INTRAVENOUS	2009-10-30	2009-11-23
24	123456	36	277353	MEROPENEM	MEROPENEM	INTRAVENOUS	2009-11-03	2009-11-23
25	123456	37		H1N1 VACCINE	INFLUENZA VIRUS VACCINE MONOVALENT	INTRAMUSCULAR	2009-11-13	2009-11-13
26	123456	38	468454	MEROPENEM	MEROPENEM	INHALATION	2010-01-18	2010-04-14
27	123456	39		MAXALT	RIZATRIPTAN BENZOATE	ORAL	2010-01-25	2010-01-25
28	123456	40	468454	MEROPENEM	MEROPENEM	INTRAVENOUS	2010-03-12	2010-03-14
29	123456	41	468454	TOBRAMYCIN	TOBRAMYCIN	INTRAVENOUS	2010-03-12	2010-04-15
30	123456	42		VITAMIN K	VITAMIN K NOS	ORAL	2010-03-13	2010-04-15
31	123456	43	468454	MEROPENEM	MEROPENEM	INTRAVENOUS	2010-03-14	2010-04-15
32	123456	44	1857199	MEROPENEM	MEROPENEM	INHALATION	2010-05-07	2010-06-11
33	123456	45	1857199	CAYSTON	AZTREONAM	INHALATION	2010-06-11	

Display 9. Concomitant Medication Data Collected in SDTM CM Domain for a Subject

The display below shows the analysis event data in ADCECD domain for subject 123456, showing the IV antibiotic therapy for pulmonary exacerbation events. As we know from previous displays, this subject had IV antibiotic therapies during the first two antibiotic courses. The variable SRCDOM in the figure shows the source domain names (CM and CE) to define the two 'IV antibiotic therapy for pulmonary exacerbation' events. The variable SRCIDVAR indicates the variable names in the source domains that are used to identify the source data. The last variable SRCIDVAL shows the observation identifiers for the source variables in the source domains. Therefore, for the reviewer, it is easy to find the source data to verify or explore more information that defines the two IV antibiotic therapies for pulmonary exacerbation events for the subject 123456 in this study.



	USUBJID	ATERM	ASTDT	AENDT	SRCDOM	SRCIDVAR	SRCIDVAL
1	123456	IV ANTIBIOTIC THERAPY FOR PULMONARY EXACERBATION	30OCT2009	23NOV2009	CM CE	CMSEQ (CEGRPID)	34, 35, 36 (277353)
2	123456	IV ANTIBIOTIC THERAPY FOR PULMONARY EXACERBATION	12MAR2010	15APR2010	CM CE	CMSEQ (CEGRPID)	40, 41, 43 (468454)

TRADEOFF AND LIMITATIONS OF ESTABLISHING TRACEABILITY FEATURES

Establishing traceability in ADaM datasets is not an easy task. It requires lot of effort and overhead like extra SAS code, creation of intermediate datasets and large analysis datasets with additional variables and records. Even though this increases size of analysis datasets and complexity of programs it is strongly recommended by ADaM implementation guide to include as much supporting data as necessary to build traceability. As mentioned earlier traceability will fasten the review process and builds confidence in reviewers and hence these benefits will be a good tradeoff with the extra work required to establish traceability. Also, it is always not feasible to build traceability in ADaM datasets. For example, in Drug Compliance analysis datasets all the exposure records of a subject are used to derive the compliance rate and it is not practical to include all these records in variables supporting traceability. In such cases Metadata traceability will be the best option.

CONCLUSION

This paper provides some examples from FDA submission in applying the inherent traceability features that provides a path between SDTM and ADaM datasets. It explores in detail on establishing Data Point Traceability with examples from FDA submissions and SAS sample codes using SRCDOM, SRCVAR, and SRCSEQ triplet, RLCRIT and RLFACT pair, and a third approach. Lastly, this paper explains the tradeoff of having traceability feature in ADaM datasets and limitations in incorporating traceability.

REFERENCES

[1] CDISC Analysis Data Model Team. “Analysis Data Model (ADaM) Implementation Guide”. December 2009.

http://www.cdisc.org/adam

[2] Zhu, Songhui and Yan, Lin. “Methods of Building Traceability for ADaM Data”. Proceedings of PharmaSUG 2011 Conference.

ACKNOWLEDGMENTS

Appreciation goes to Kelly Blackburn, Stacy Surensky and Lynn Anderson for their valuable support.

CONTACT INFORMATION

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

Name: Xiangchen (Bob) Cui**,** Ph.D.
Enterprise: Vertex Pharmaceuticals, Inc.
Address: 88 Sidney Street
City, State ZIP: Cambridge MA, 02139
Work Phone: 617-444-6069
Fax: 617-460-8060
E-mail: xiangchen_cui@vrtx.com

Name: Hongyu Liu
Enterprise: Vertex Pharmaceuticals, Inc.
Address: 88 Sidney Street
City, State ZIP: Cambridge MA, 02139
Work Phone: 617-444-6918
Fax: 617-460-8060
E-mail: hongyu_liu@vrtx.com

Name: Tathabbai Pakalapati
Enterprise: Vertex Pharmaceuticals, Inc.
Address: 88 Sidney Street
City, State ZIP: Cambridge MA, 02139
Work Phone: 617-444-7404
E-mail: Tathabbai_Pakalapati@vrtx.com