Evaluating SDTM SUPP Domain For AdaM - Trash Can Or Buried Treasure

Xiaopeng Li, Yi Liu and Chun Feng
Celerion, Lincoln, NE USA

celerion
Applied Translational Medicine

ABSTRACT

Study Data Tabulation Model (SDTM) is commonly expected as industry standard for clinical study electronic submissions to the US Food and Drug Administration (FDA). SDTM has its own standard regulations on data structure for capturing and categorizing variables across all the SDTM parent domains. Analysis Data Model (ADaM), as the FDA recommended analysis submission data model, is generated based on SDTM. Due to the distinctive structure of SDTM data, programmers who generate ADaM compliant data sets frequently encounter difficulties locating or deriving ADaM-oriented variables from SDTM parent domains. This is often difficult because customized or sponsor-specific analysis information needed in ADaM data may not be captured or allowed in the SDTM parent domains. Therefore, maintaining all needed information in SDTM supplemental domains becomes an efficient solution under the SDTM data structure. The paper discusses the importance of supplemental domains in terms of traceability of data and supports for analysis, and illustrates how beneficial SDTM supplemental domains are to ADaM programmers using real-life clinical data examples.

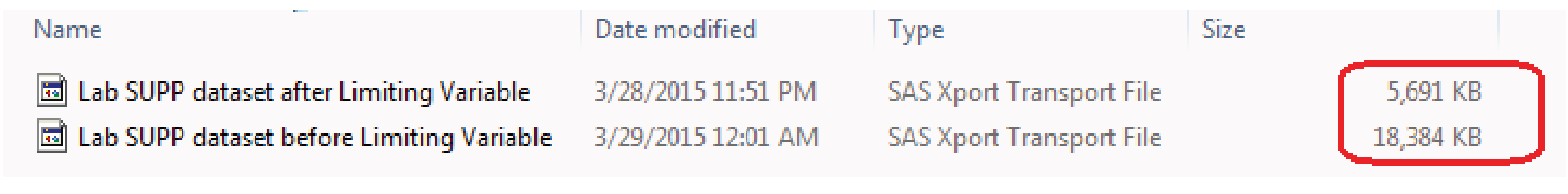
INTRODUCTION

Study Data Tabulation Model (SDTM) has become part of the FDA intent-to-require submission package which demonstrates an interchange standard with specifications of the structure and metadata for clinic data. Besides, SDTM is the source for ADaM or analysis data. All clinic-collected data in the ADaM data must be in the SDTM domains. According to the SDTM implementation guide (SDTMIG), there are three types of pre-defined core variables in SDTM parent domains: required, expected, and permissible variables. Other non-standard variables can be included in supplemental (SUPP) domains, if there are any. SUPP domains are intended to capture additional sponsor-specific variables or customized analysis variables for ADaM data which do not fit within the SDTM parent domain. The records in SUPP domains and parent domains are linked by the same set of keys which is --GRPID in parent domains and identifying variable value (IDVARVAL) in SUPP domains. SUPP domain records include the identifying variable (IDVAR) which identifies the related record(s) to the parents domain such as sequence number (--SEQ), --GRPID, etc, IDVARVAL, the qualifier variable label (QNAM), the qualifier variable label (QLABEL), data value (QVAL), the origin of the value (QORIG), and the evaluator (QEVAL). Therefore, due to the SDTM standardized data structure and restricted SDTM Control Terminology (CT), SDTM data provides consistent source information for ADaM data. It creates consistent ADaM data across studies which could be beneficial for the downstream tables, figures and listings (TFLs).

OBJECTS OF SUPP DOMAINS

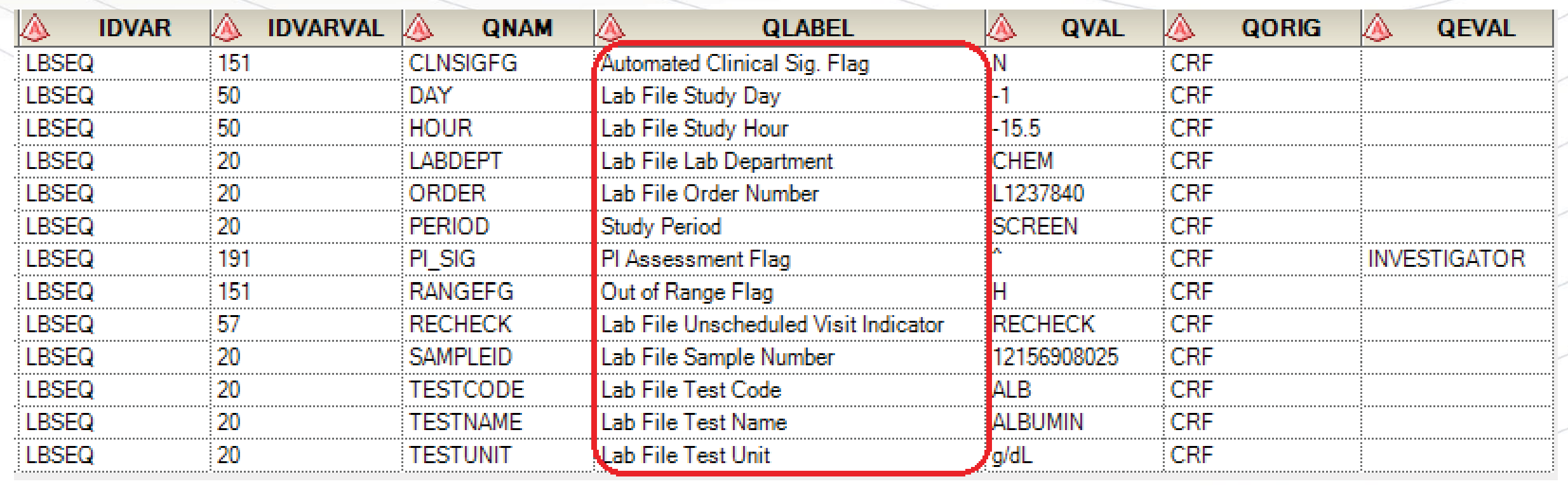
Traceability and analysis support are two primary expectations for capturing source information from the Case Report Form (CRF) and clinical database in SUPP domains. Traceability shows the heritage relationship between the source data sets (SDTM) and analysis data sets (ADaM). Because of this, traceability provides programmers and reviewers the transparency to understand where and how the information is collected and retained. Analysis support of SUPP domains allow analysis-related variables from source data to be kept in SUPP domains. These analyses-related variables support downstream analysis, especially ADaM and TFLs programming. In SUPP domains, a programmer can keep as many source data variables as possible theoretically, but it is not appropriate to keep all the clinic data in SUPP domains practically. It is important to retain necessary variables in the SDTM SUPP domains for analysis support and traceability purposes, because large SAS transport file size can cause potential storage space issues for the FDA. Additionally, it may be time consuming for FDA reviewers analyzing sizable SUPP SDTM data sets.

One simple approach to help address file size limitations and meet the FDA submission size requirements is to efficiently reduce the SUPP domain variables and only include those needed for downstream analysis use. The result of limiting variables in the SUPP domain yields a significant decline in data file size. Displays 1-3 show an example of comparison of file sizes before and after applying limit variables on laboratory (LB) SUPP domain. The size of the LB SUPP data is 18384 KB including all the information from Clinical Data Acquisition Standards Harmonization (CDASH). The size drops significantly to 5691 KB after keeping specified analysis-driven variables. Through managing variable numbers in SUPP domain, it can optimize the data set size and remove unnecessary and unused variables in the data set.



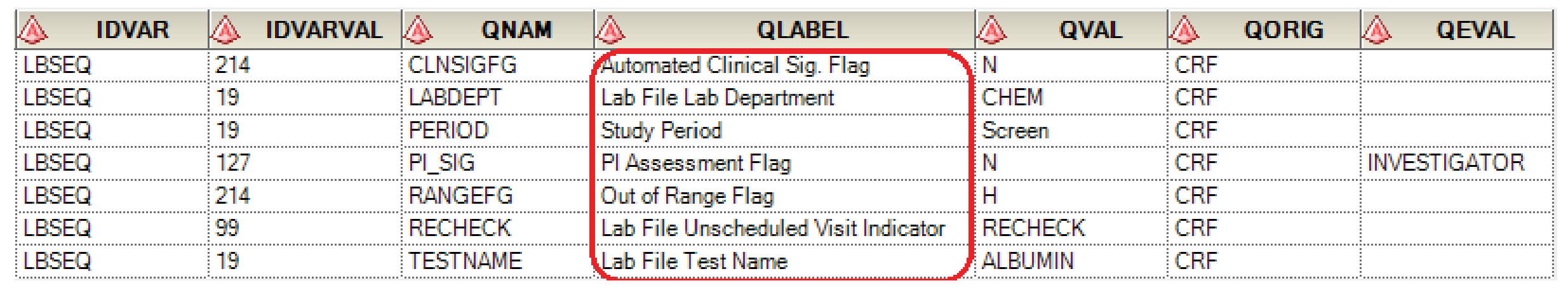
Name	Date modified	Type	Size
Lab SUPP dataset after Limiting Variable	3/28/2015 11:51 PM	SAS Xport Transport File	5,691 KB
Lab SUPP dataset before Limiting Variable	3/29/2015 12:01 AM	SAS Xport Transport File	18,384 KB

Display 1. Lab Data Sets File Size before and after Limiting Variable



IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
LBSEQ	151	CLNSIGFG	Automated Clinical Sig. Flag	N	CRF	
LBSEQ	50	DAY	Lab File Study Day	-1	CRF	
LBSEQ	50	HOUR	Lab File Study Hour	-15.5	CRF	
LBSEQ	20	LABDEPT	Lab File Lab Department	CHEM	CRF	
LBSEQ	20	ORDER	Lab File Order Number	L1237840	CRF	
LBSEQ	20	PERIOD	Study Period	SCREEN	CRF	
LBSEQ	191	PI_SIG	PI Assessment Flag	^	CRF	INVESTIGATOR
LBSEQ	151	RANGEFG	Out of Range Flag	H	CRF	
LBSEQ	57	RECHECK	Lab File Unscheduled Visit Indicator	RECHECK	CRF	
LBSEQ	20	SAMPLEID	Lab File Sample Number	12156908025	CRF	
LBSEQ	20	TESTCODE	Lab File Test Code	ALB	CRF	
LBSEQ	20	TESTNAME	Lab File Test Name	ALBUMIN	CRF	
LBSEQ	20	TESTUNIT	Lab File Test Unit	g/dL	CRF	

Display 2. Lab Data Sets Full Variables before Limiting Variables



IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
LBSEQ	214	CLNSIGFG	Automated Clinical Sig. Flag	N	CRF	
LBSEQ	19	LABDEPT	Lab File Lab Department	CHEM	CRF	
LBSEQ	19	PERIOD	Study Period	Screen	CRF	
LBSEQ	127	PI_SIG	PI Assessment Flag	N	CRF	INVESTIGATOR
LBSEQ	214	RANGEFG	Out of Range Flag	H	CRF	
LBSEQ	99	RECHECK	Lab File Unscheduled Visit Indicator	RECHECK	CRF	
LBSEQ	19	TESTNAME	Lab File Test Name	ALBUMIN	CRF	

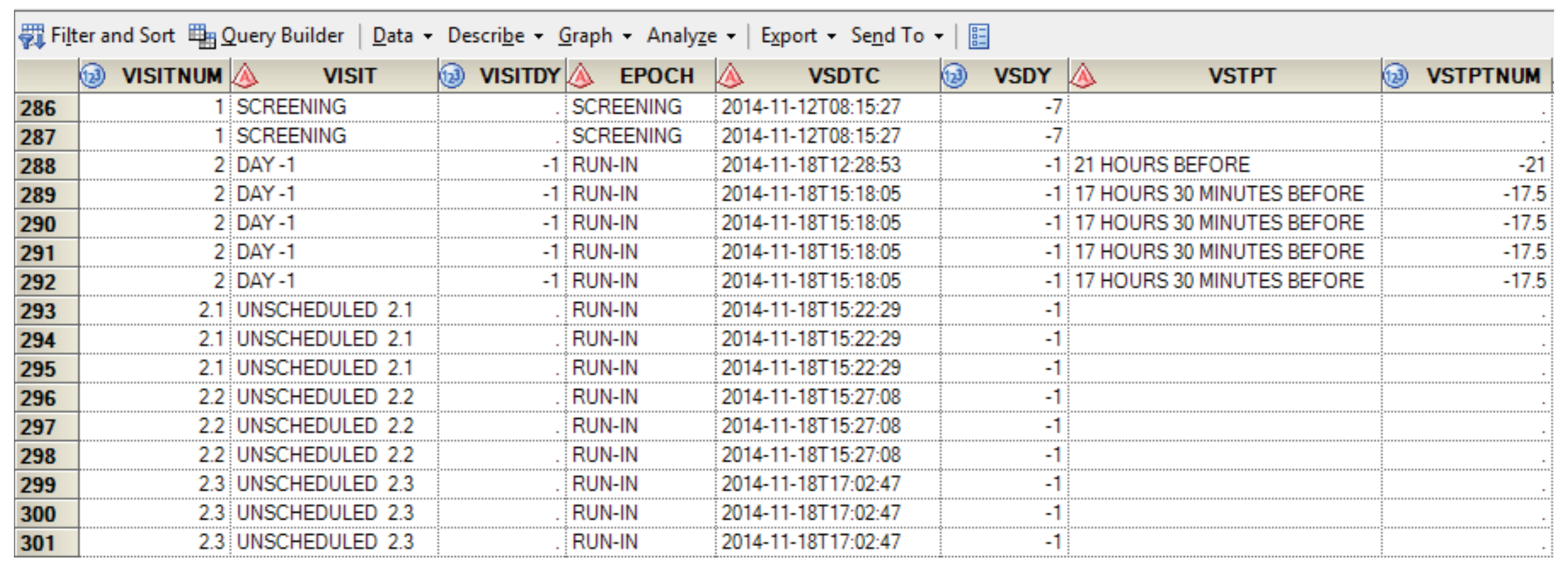
Display 3. Lab Data Sets Appropriate Variables after Limiting Variables

In addition to preserving valuable analysis variables and excluding non-analysis variables in SDTM SUPP domains, there is an increase in programming efficiency and reduction in complexity of programming. Programmer scan quickly seize the useful analysis information in a decreased number of variables in SUPP domains as opposed to seeking the information within massive data sets containing unusable variables in SUPP domains. To prevent unnecessary informationfrom being contained in SUPP domain, the authors suggest users follow Statistical Analysis Plan (SAP) and discuss the indispensable information within the SUPP domain with the analysis team.

SCENARIOS

UNSCHEDULED OR RECHECK

Dealing with data collected as rechecks, unscheduled, or early termination time points can be challenging for programmers generating TFLs. Early termination, recheck, and unscheduled time points are mapped to “UNSCHEDULED XX.X” under VISIT variable following the current SDTM structure, making them difficult to assign back to the appropriate visit if necessary to use in analysis.



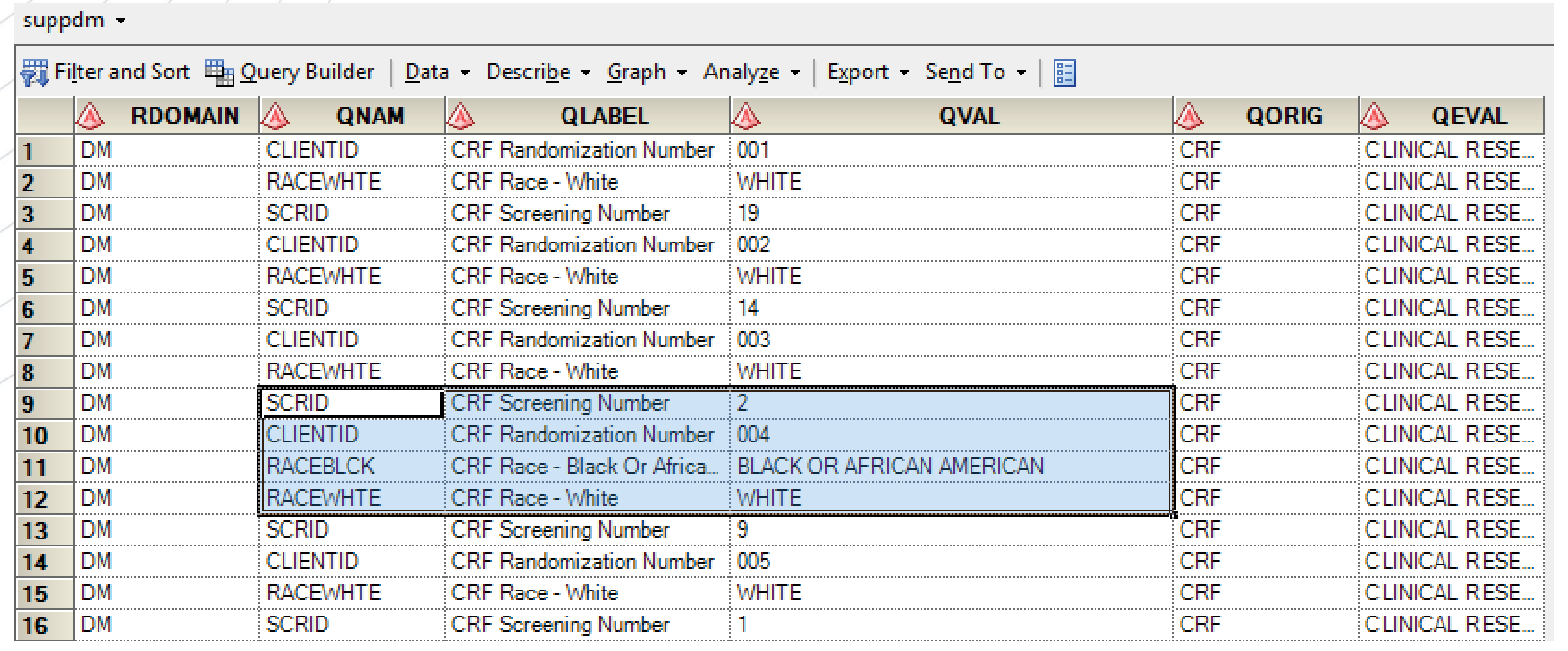
	VISITNUM	VISIT	VISITDY	EPOCH	VSDTC	VSDY	VSTPT	VSTPTNUM
286	1	SCREENING	.	SCREENING	2014-11-12T08:15:27	-7		.
287	1	SCREENING	.	SCREENING	2014-11-12T08:15:27	-7		.
288	2	DAY -1	-1	RUN-IN	2014-11-18T12:28:53	-1	21 HOURS BEFORE	-21
289	2	DAY -1	-1	RUN-IN	2014-11-18T15:18:05	-1	17 HOURS 30 MINUTES BEFORE	-17.5
290	2	DAY -1	-1	RUN-IN	2014-11-18T15:18:05	-1	17 HOURS 30 MINUTES BEFORE	-17.5
291	2	DAY -1	-1	RUN-IN	2014-11-18T15:18:05	-1	17 HOURS 30 MINUTES BEFORE	-17.5
292	2	DAY -1	-1	RUN-IN	2014-11-18T15:18:05	-1	17 HOURS 30 MINUTES BEFORE	-17.5
293	2.1	UNSCHEDULED 2.1	.	RUN-IN	2014-11-18T15:22:29	-1		.
294	2.1	UNSCHEDULED 2.1	.	RUN-IN	2014-11-18T15:22:29	-1		.
295	2.1	UNSCHEDULED 2.1	.	RUN-IN	2014-11-18T15:22:29	-1		.
296	2.2	UNSCHEDULED 2.2	.	RUN-IN	2014-11-18T15:27:08	-1		.
297	2.2	UNSCHEDULED 2.2	.	RUN-IN	2014-11-18T15:27:08	-1		.
298	2.2	UNSCHEDULED 2.2	.	RUN-IN	2014-11-18T15:27:08	-1		.
299	2.3	UNSCHEDULED 2.3	.	RUN-IN	2014-11-18T17:02:47	-1		.
300	2.3	UNSCHEDULED 2.3	.	RUN-IN	2014-11-18T17:02:47	-1		.
301	2.3	UNSCHEDULED 2.3	.	RUN-IN	2014-11-18T17:02:47	-1		.

Display 4 Unscheduled and Rechecked Information on SDTM

Display 4 shows an example of VISIT information including unscheduled records in SDTM Vital Sign domain. As illustrated in this example, unscheduled 2.1 and unscheduled 2.2 are rechecks and unscheduled 2.3 is an unscheduled record. The decision to either include or not include unscheduled, recheck, or early termination records should be specified in the SAP. The SAP of the example in Display 4 required the inclusion of recheck records and omission of unscheduled records in the analysis. In this case, the recheck information must be captured in the SDTM SUPP domain for analysis use (ADaM and TFLs).

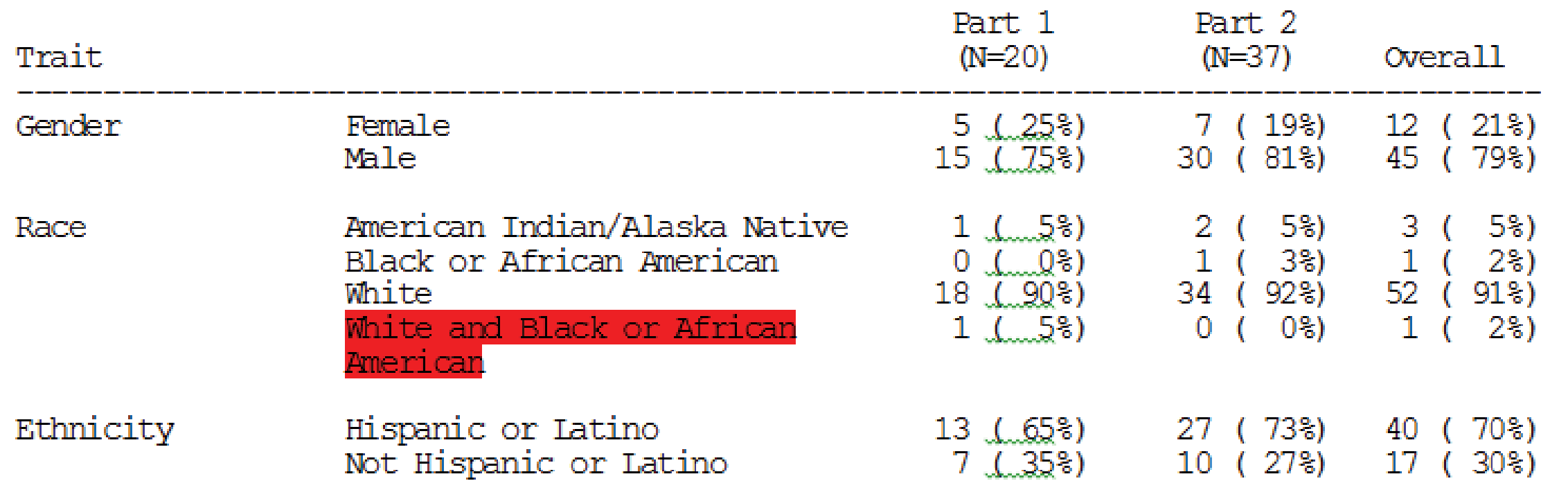
CRF RACE INFORMATION

Race is required to be presented in demographic listings and tables. Specified categories for races in SDTM terminology include: WHITE, BLACK OR AFRICAN AMERICAN, AMERICAN INDIAN OR ALASKA NATIVE, NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER, and ASIAN. However, these race codes do not meet all the analysis needs especially if there are multiple races selected for a subject. For example, when a multiracial subject is recorded as WHITE and BLACK OR AFRICAN AMERICAN in the CRF, the subject’s race could be mapped to the values of OTHER, or MULTIPLE in the SDTM demographic (DM) domain without matching SDTM CT. If the race information (WHITE and BLACK OR AFRICAN AMERICAN) collected from the CRF or source data is not kept in the SUPPDM domain (Display 5), the race information for the subject would not be available for analysis. Therefore, the DM summary table would lose valuable race information in the category (Display 6). Additionally, keeping the original race information, not useable in the controlled terminology, allows full traceability for SDTM.



	RDOMAIN	QNAM	QLABEL	QVAL	QORIG	QEVAL
1	DM	CLIENTID	CRF Randomization Number	001	CRF	CLINICAL RESE...
2	DM	RACEWHTE	CRF Race - White	WHITE	CRF	CLINICAL RESE...
3	DM	SCRID	CRF Screening Number	19	CRF	CLINICAL RESE...
4	DM	CLIENTID	CRF Randomization Number	002	CRF	CLINICAL RESE...
5	DM	RACEWHTE	CRF Race - White	WHITE	CRF	CLINICAL RESE...
6	DM	SCRID	CRF Screening Number	14	CRF	CLINICAL RESE...
7	DM	CLIENTID	CRF Randomization Number	003	CRF	CLINICAL RESE...
8	DM	RACEWHTE	CRF Race - White	WHITE	CRF	CLINICAL RESE...
9	DM	SCRID	CRF Screening Number	2	CRF	CLINICAL RESE...
10	DM	CLIENTID	CRF Randomization Number	004	CRF	CLINICAL RESE...
11	DM	RACEBLCK	CRF Race - Black Or Africa...	BLACK OR AFRICAN AMERICAN	CRF	CLINICAL RESE...
12	DM	RACEWHTE	CRF Race - White	WHITE	CRF	CLINICAL RESE...
13	DM	SCRID	CRF Screening Number	9	CRF	CLINICAL RESE...
14	DM	CLIENTID	CRF Randomization Number	005	CRF	CLINICAL RESE...
15	DM	RACEWHTE	CRF Race - White	WHITE	CRF	CLINICAL RESE...
16	DM	SCRID	CRF Screening Number	1	CRF	CLINICAL RESE...

Display 5. CRF Race Information on SDTM

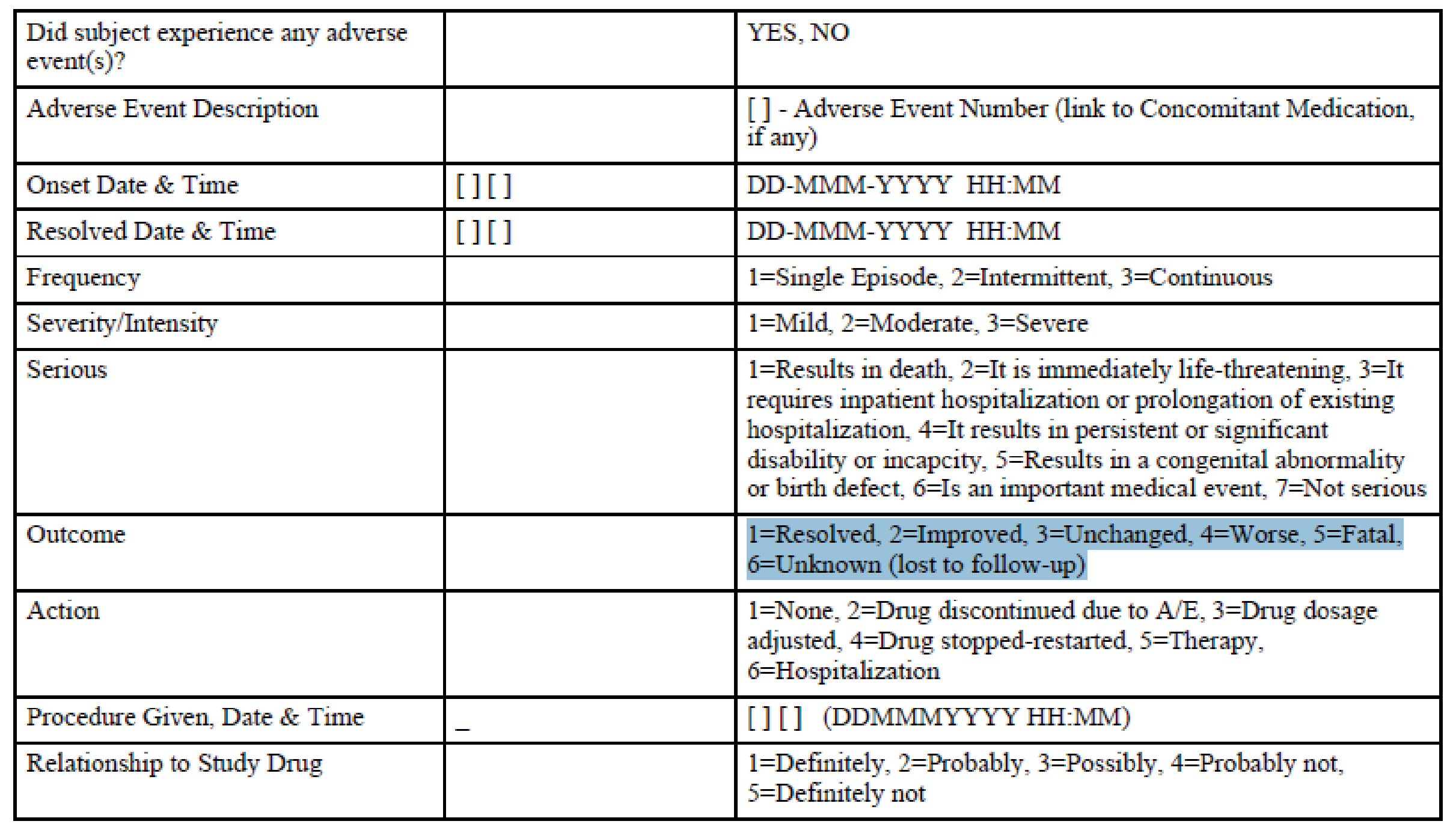


```
                                                         Part 1       Part 2
Trait                                                    (N=20)       (N=37)      Overall
-------------------------------------------------------------------------------------------
Gender              Female                                5 ( 25%)     7 ( 19%)   12 ( 21%)
                    Male                                 15 ( 75%)    30 ( 81%)   45 ( 79%)

Race                American Indian/Alaska Native         1 (  5%)     2 (  5%)    3 (  5%)
                    Black or African American             0 (  0%)     1 (  3%)    1 (  2%)
                    White                                18 ( 90%)    34 ( 92%)   52 ( 91%)
                    White and Black or African            1 (  5%)     0 (  0%)    1 (  2%)
                    American

Ethnicity           Hispanic or Latino                   13 ( 65%)    27 ( 73%)   40 ( 70%)
                    Not Hispanic or Latino                7 ( 35%)    10 ( 27%)   17 ( 30%)
```

Display 6. Race Summary Information in Table

ADVERSE EVENT INFORMATION

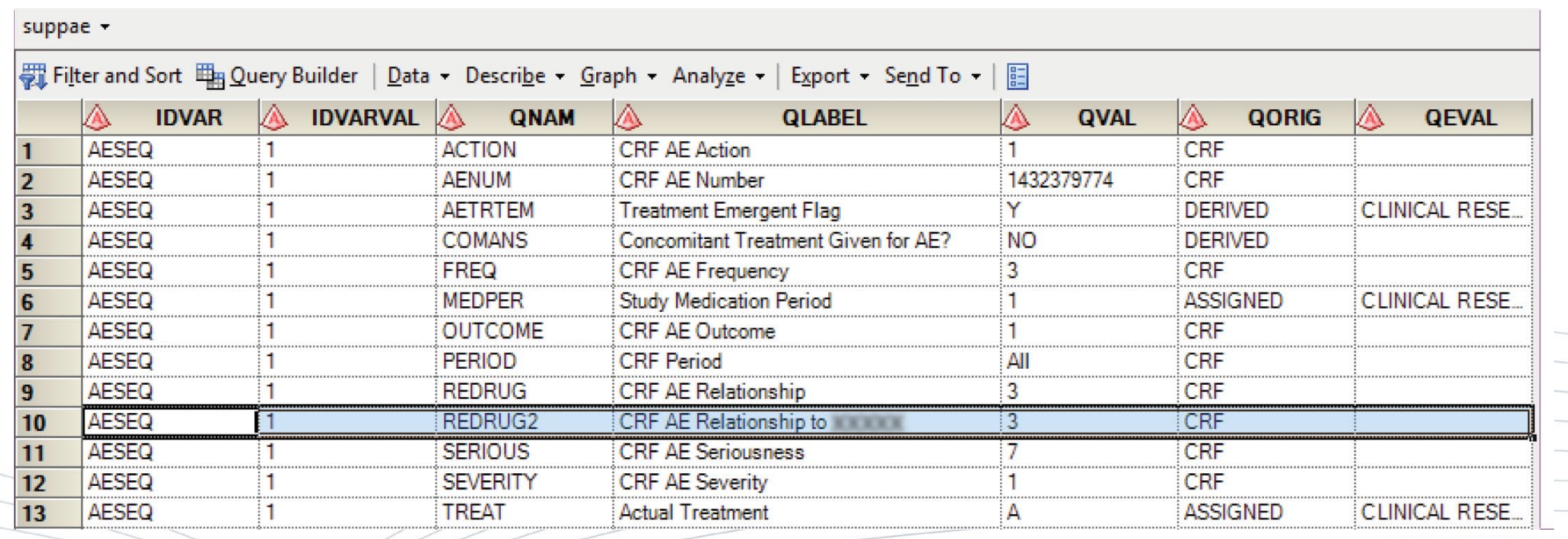
There are two types of codelist in SDTM CT: extensible codelist and non-extensible codelist. The extensible codelist provides the flexibility to make the SDTM variable information match the CRF/source data. The non-extensible codelist CT restricts the options of codes for the variables. For example, a SDTM adverse event (AE) variable AEOUT (Outcome of AE) has a non-extensible codelist which includes: FATAL, NOT RECOVERED/NOTRE SOLVED, RECOVERED/RESOLVED, RECOVERED/RESOLVED WITH SEQUELAE, RECOVERING/ RESOLVING and UNKNOWN. The AE outcome options in CRF (Resolved, Improved, Unchanged, Worse, Fatal, and Unknown [lost to follow-up]) are shown in Display 7. Not all the outcome options can be accurately mapped into SDTM AE. In this case, keeping CRF AE outcome in SUPP AE becomes a practical way to support analysis and maintain traceability.



Did subject experience any adverse event(s)?		YES, NO
Adverse Event Description		[ ] - Adverse Event Number (link to Concomitant Medication, if any)
Onset Date & Time	[ ] [ ]	DD-MMM-YYYY HH:MM
Resolved Date & Time	[ ] [ ]	DD-MMM-YYYY HH:MM
Frequency		1=Single Episode, 2=Intermittent, 3=Continuous
Severity/Intensity		1=Mild, 2=Moderate, 3=Severe
Serious		1=Results in death, 2=It is immediately life-threatening, 3=It requires inpatient hospitalization or prolongation of existing hospitalization, 4=It results in persistent or significant disability or incapcity, 5=Results in a congenital abnormality or birth defect, 6=Is an important medical event, 7=Not serious
Outcome		1=Resolved, 2=Improved, 3=Unchanged, 4=Worse, 5=Fatal, 6=Unknown (lost to follow-up)
Action		1=None, 2=Drug discontinued due to A/E, 3=Drug dosage adjusted, 4=Drug stopped-restarted, 5=Therapy, 6=Hospitalization
Procedure Given, Date & Time	_	[ ] [ ] (DDMMMYYYY HH:MM)
Relationship to Study Drug		1=Definitely, 2=Probably, 3=Possibly, 4=Probably not, 5=Definitely not

Display 7. AE Outcome Information in Blank CRF

In some studies, we are interested in the AE relationship to individual compounds when multiple drugs are co-administered. In the AE parent domain, AEREL is the only variable for the AE relationship to a study drug. So the AE relationship to the other co-administered drugs can only be kept in SUPP domain (Display 8).



	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
1	AESEQ	1	ACTION	CRF AE Action	1	CRF	
2	AESEQ	1	AENUM	CRF AE Number	1432379774	CRF	
3	AESEQ	1	AETRTEM	Treatment Emergent Flag	Y	DERIVED	CLINICAL RESE...
4	AESEQ	1	COMANS	Concomitant Treatment Given for AE?	NO	DERIVED	
5	AESEQ	1	FREQ	CRF AE Frequency	3	CRF	
6	AESEQ	1	MEDPER	Study Medication Period	1	ASSIGNED	CLINICAL RESE...
7	AESEQ	1	OUTCOME	CRF AE Outcome	1	CRF	
8	AESEQ	1	PERIOD	CRF Period	All	CRF	
9	AESEQ	1	REDRUG	CRF AE Relationship	3	CRF	
10	AESEQ	1	REDRUG2	CRF AE Relationship to [redacted]	3	CRF	
11	AESEQ	1	SERIOUS	CRF AE Seriousness	7	CRF	
12	AESEQ	1	SEVERITY	CRF AE Severity	1	CRF	
13	AESEQ	1	TREAT	Actual Treatment	A	ASSIGNED	CLINICAL RESE...

Display 8. AE Relationship to Multiple Co-administered Drugs in SUPPAE

PERIOD

The period variable is used to derive actual treatment from treatment sequence, define a baseline for change from baseline analysis, and assist with summarization as a time point indicator. In ADaM data, APERIOD is a permissible variable which includes period information. Although period is a commonly collected variable in CRF or clinic source data, period is not included in SDTM parent domains according to the current SDTM structure. When period information is not retained in the SDTM SUPP domains, programmers need to derive actual treatments and baselines by merging the results data with the exposure domain (EX). Retaining period in the SDTM data is more efficient for programmers to generate ADaM domains and TFLs. Code 1 and Code 2 depict an example of deriving treatment information in the ADLB domain for a crossover study. Code 1 and Code 2 show two sets of SAS codes, one with and one without retaining the period variable in the SUPP domain. When comparing the two sets of SAS codes below, keeping period in SUPP domain is a more efficient and accurate approach. It is not considered ideal to derive data more than once or remove data that is entered only to derive it later in the process as it can lead to more errors.

Code 1 (with period information in the SUPP domains) :

```
data _null_;
set nodupper;
call symput("period",left(trim(period)));

data adam;
set adam;
if upcase(period) in ( 'SCREEN' , 'SCREENING') then aperiod = .;
else a period = substr(period,1,1) + 0;
%do I = 1 %to & period.;
trt0&i.p = trim(left(substr(armcd,&i.+ 0,1)));
%end;
```

Code 2 (without period information in the SUPP domains) :

```
trt01p = substr(armcd,1,1);
trt01a = substr(armcd,1,1);
trt02p = substr(armcd,2,1);
trt02a = substr(armcd,2,1);
trt03p = substr(armcd,3,1);
trt03a = substr(armcd,3,1);
… …
```

CONCLUSION

SUPP domains are the metaphorical buried treasure. They retain information needed for ADaM data and TFLs while making collected data traceable. Although SUPP domains increase the size of the SDTM transfer package, they can provide important information to support analysis or improve the traceability of SDTM. It is critical to find the balance between controlling the size of SUPP domains and keeping enough analysis variables to support the analysis in data. Based on SDTM IG 3.2, the SUPP domains are critical to the SDTM data. If the future version of SDTM IG can provide more flexibility for the SDTM parent domains, the information for analysis may be included in parent domains.

REFERENCES

1. www.cdisc.org/sdtm
2. www.cdisc.org/adam

CONTACT INFORMATION

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

Name: Xiaopeng Li
Enterprise: Celerion Inc.
Address: 621 Rose Street
City, State ZIP: Lincoln, NE 68502
Work Phone: 402-437-6260
E-mail: xiaopeng.li@celerion.com

Name: Yi Liu
Enterprise: Celerion Inc.
Address: 621 Rose Street
City, State ZIP: Lincoln, NE 68502
Work Phone: 402-437-4778
E-mail: yi.liu@celerion.com

Name: Chun Feng
Enterprise: Celerion Inc.
Address: 621 Rose Street
City, State ZIP: Lincoln, NE 68502
Work Phone: 402-437-4799
E-mail: chun.feng@celerion.com

www.celerion.com