PharmaSUG 2018 - Paper SS-15

It’s All About Getting the Source and Codelist Implementation Right for ADaM Define.xml v2.0

Supriya Davuluri, PPD, LLC, Morrisville, NC

ABSTRACT

There are some obvious challenges in generating Analysis Data Model define.xml v2.0 with confusion prevailing over populating appropriate source for ADaM variables and implementation of codelists, which often causes lack of structural consistency. The ADaM Metadata team proposed conventions for populating attribute Type of element *source* in a consistent manner to distinguish between Predecessor, Assigned and Derived along with some guidelines for implementation of Control Terminology (CT) i.e. NCI/CDISC or User specified CT for ADaM Variables [3]. This paper summarizes the best practices to be followed while populating the variable source, reflecting its occurrence in the given ADaM datasets along with the CT implementation considerations such as usage of Enumerated Lists vs Codelists, CT naming conventions, dealing with Subsets of CTs and specifying CT for a variable to set up the metadata correctly when generating an ADaM define.xml v2.0.

INTRODUCTION

An appropriately functional define.xml is the metadata describing the content and format of standardized electronic datasets, it is an important component of submission for regulatory review. Study specific information is crucial for reviewers as they are limited to the submitted metadata [1]. Variable’s source, codelist and derivation should be clear and easily accessible from the define.xml and allow the reviewers to interpret submission data and speed up the review.

Major Problems in Define.xml

Lack of Structural Consistency and Traceability:

Common deficiencies in ADaM define.xml v2.0 usually encountered by the reviewers are establishing traceability, lack of consistency for variable source which should often reflect the situation in the given ADaM dataset to understand the construction and codelist implementation. Regulatory review may be compromised if traceability is not well established. There are instances where the existing implementation guides [2] do not provide specific instructions on representation of study data and causing obvious challenges in generating ADaM define.xml v2.0 with confusion prevailing over populating appropriate source for ADaM variables and implementation of Codelists, which often results in lack of structural consistency and the metadata is interpreted differently.

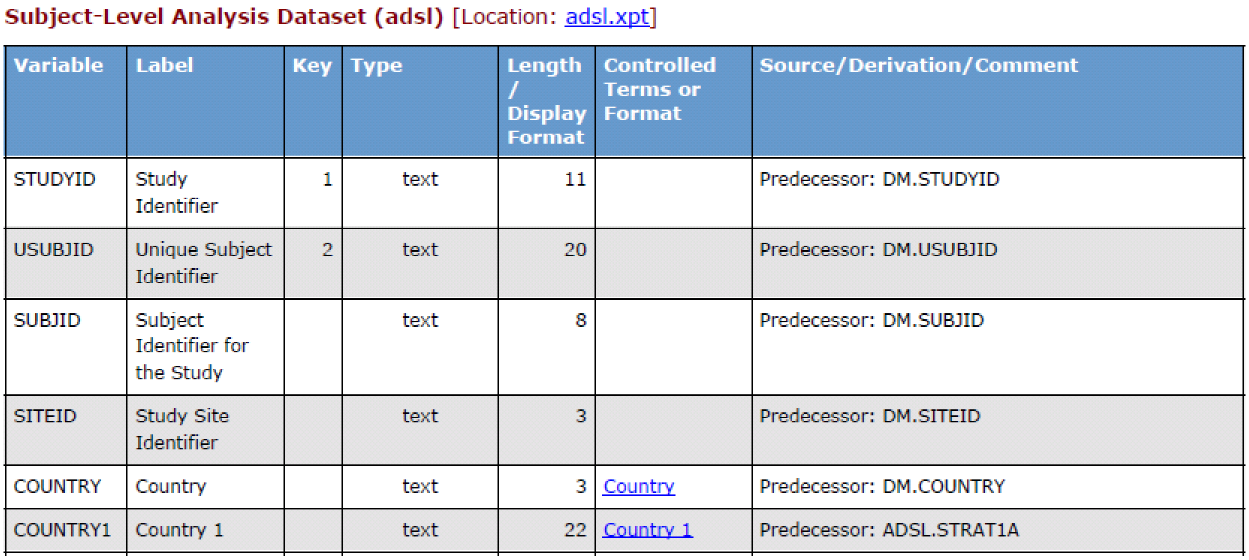
The ADaM metadata team proposed conventions for populating attribute Type of element *source* in a consistent manner [3] to distinguish between Predecessor, Assigned and Derived along with some guidelines for implementation of Control Terminology (CT) [4] i.e. NCI/CDISC or User specified CT for ADaM Variables to generate a well-documented define.xml for significant benefits. Listed below are the best practices to be implemented while populating the variable source, reflecting its occurrence in the given ADaM datasets along with the CT implementation considerations such as usage of Enumerated Lists vs Codelists, CT naming conventions, dealing with subsets of CTs and specifying CT for a Variable to set up the metadata correctly when generating an ADaM define.xml v2.0.

SOURCE FOR VARIABLES

Source of an ADaM variable is an element with type attribute and this could be Predecessor, Assigned or Derived reflecting the situation of a variable from the respective ADaM dataset.

Predecessor: If the ADaM Variable is populated with contents of another variable copied from SDTM, ADaM, Subset of source records or if it is a same variable with a different name as illustrated in the examples below the source of the variable is Predecessor.

Example 1: Variable populated with the contents of another variable which could be a copy from a different variable from SDTM the source for the variables is Predecessor. COUNTRY1 variable is a copy over of ADSL.STRAT1A with in ADSL with a different name as seen in Figure 1 below.

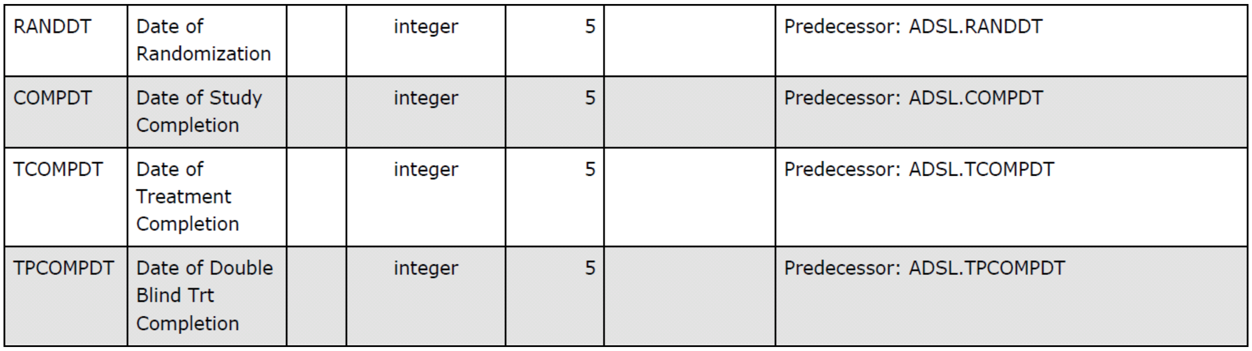


Subject-Level Analysis Dataset (adsl) [Location: adsl.xpt]

Variable	Label	Key	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	1	text	11		Predecessor: DM.STUDYID
USUBJID	Unique Subject Identifier	2	text	20		Predecessor: DM.USUBJID
SUBJID	Subject Identifier for the Study		text	8		Predecessor: DM.SUBJID
SITEID	Study Site Identifier		text	3		Predecessor: DM.SITEID
COUNTRY	Country		text	3	Country	Predecessor: DM.COUNTRY
COUNTRY1	Country 1		text	22	Country 1	Predecessor: ADSL.STRAT1A

Figure 1. Variable populated with the contents of another variable from SDTM or different name with in ADaM dataset.

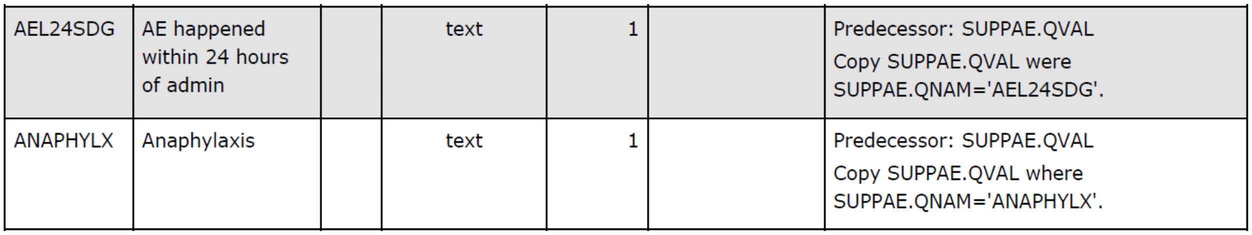
Example 2**:** ADSL Variables may be copied to basic structure datasets to support traceability or enable analysis across other ADaM datasets. Source for such variables is Predecessor as seen in the Figure 2 below.



RANDDT	Date of Randomization		integer	5		Predecessor: ADSL.RANDDT
COMPDT	Date of Study Completion		integer	5		Predecessor: ADSL.COMPDT
TCOMPDT	Date of Treatment Completion		integer	5		Predecessor: ADSL.TCOMPDT
TPCOMPDT	Date of Double Blind Trt Completion		integer	5		Predecessor: ADSL.TPCOMPDT

Figure 2. Variable populated with the contents of another variable from ADaM.

Example 3: If an ADaM variable is an exact copy of the contents of a specific source variable from a *subset* of source records the source for the variable is Predecessor as seen in the Figure 3.

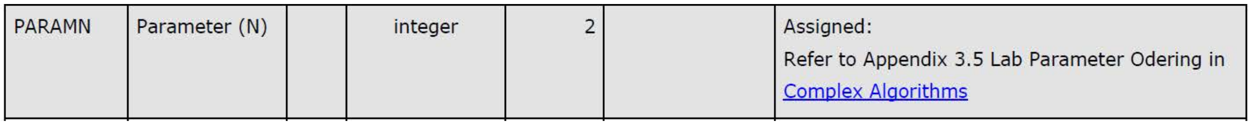


AEL24SDG	AE happened within 24 hours of admin		text	1		Predecessor: SUPPAE.QVAL Copy SUPPAE.QVAL were SUPPAE.QNAM='AEL24SDG'.
ANAPHYLX	Anaphylaxis		text	1		Predecessor: SUPPAE.QVAL Copy SUPPAE.QVAL where SUPPAE.QNAM='ANAPHYLX'.

Figure 3. ADaM Variable is an exact copy of a specific source from a subset of records.

Assigned: Numeric counterparts for primary analysis variables in character format, Character counterparts for numeric primary analysis variables, BDS parameter variables PARAM, PARAMCD, PARAMN as a general convention, even if values may be built from contents in respective --TEST, --STRESU, --TESTCD variables from SDTM, variables populated with contents from medical dictionaries etc. are recommended to have the source value as Assigned as illustrated in the examples below.

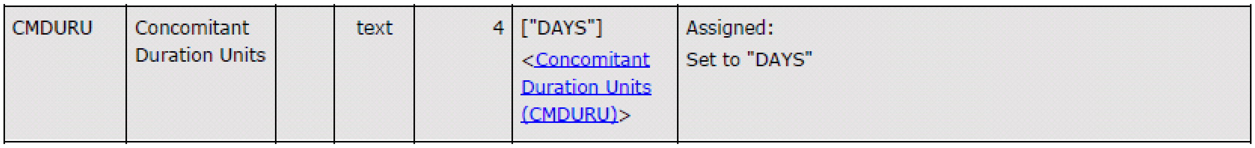
Example 1: Numeric counterparts for primary analysis variables in character format



PARAMN	Parameter (N)		integer	2		Assigned: Refer to Appendix 3.5 Lab Parameter Odering in Complex Algorithms

Figure 4. Numeric counterparts for primary analysis variables in character format

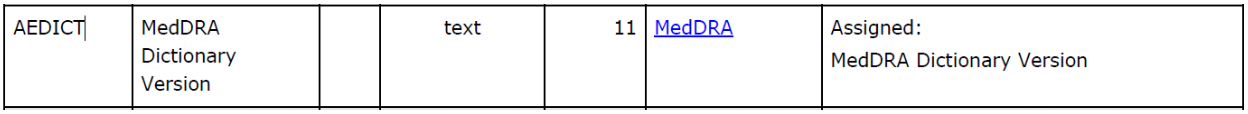
Example 2: CMDURU is assigned a value based on the units of duration collected.



CMDURU	Concomitant Duration Units		text	4	["DAYS"] <Concomitant Duration Units (CMDURU)>	Assigned: Set to "DAYS"

Figure 5. BDS parameter variable CMDURU

Example 3: AEDICT is populated from the MedDRA Medical dictionary.

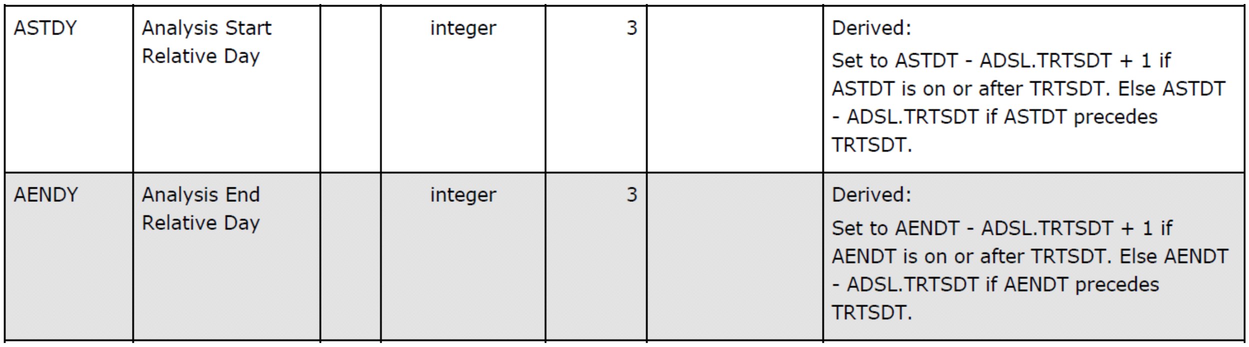


AEDICT	MedDRA Dictionary Version		text	11	MedDRA	Assigned: MedDRA Dictionary Version

Figure 6. Variable populated with contents from medical dictionaries

Derived: If a variable’s content is the result of a derivation, involving one or more other variables: such as calculation of a numeric duration, ISO8601 datetime string content to numeric date (/time/datetime) conversion or derivation of population flags based on an algorithm involving one or more other variables etc. then the recommended source for the variable is Derived as illustrated in the examples below.

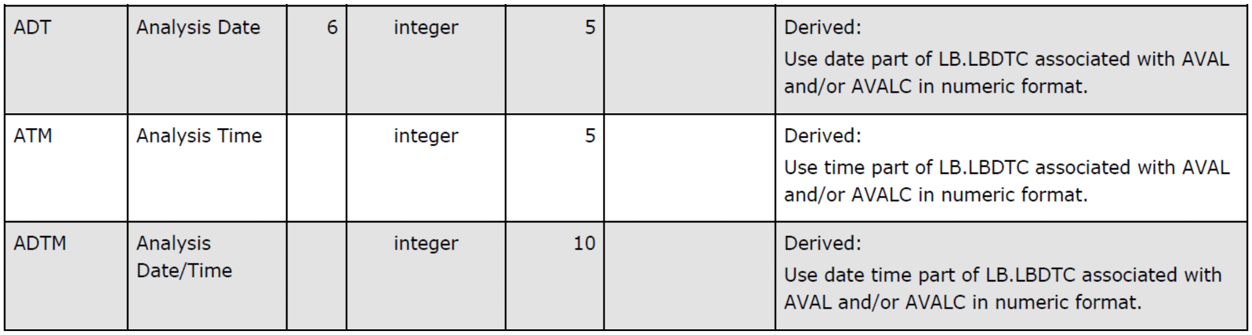
Example 1: Calculation of a Numeric Duration.



ASTDY	Analysis Start Relative Day		integer	3		Derived: Set to ASTDT - ADSL.TRTSDT + 1 if ASTDT is on or after TRTSDT. Else ASTDT - ADSL.TRTSDT if ASTDT precedes TRTSDT.
AENDY	Analysis End Relative Day		integer	3		Derived: Set to AENDT - ADSL.TRTSDT + 1 if AENDT is on or after TRTSDT. Else AENDT - ADSL.TRTSDT if AENDT precedes TRTSDT.

Figure 7. Derived Variables calculating the duration.

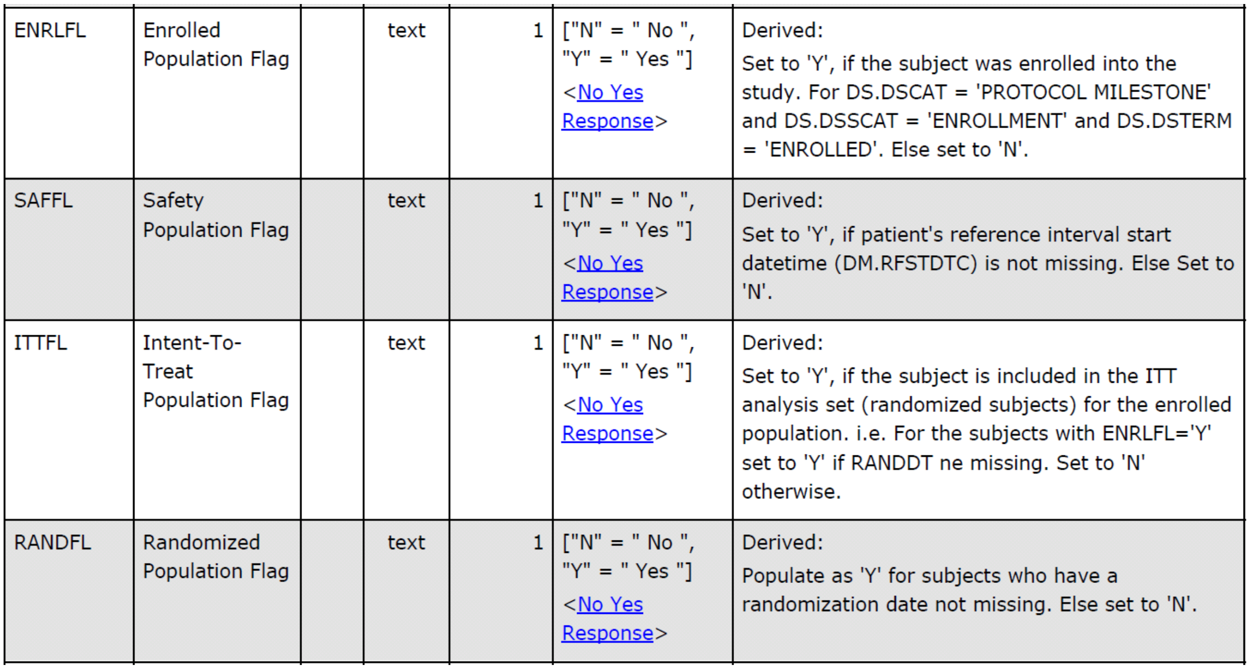
Example 2: Date/Time associated with AVAL and/or AVALC in numeric format.



ADT	Analysis Date	6	integer	5		Derived: Use date part of LB.LBDTC associated with AVAL and/or AVALC in numeric format.
ATM	Analysis Time		integer	5		Derived: Use time part of LB.LBDTC associated with AVAL and/or AVALC in numeric format.
ADTM	Analysis Date/Time		integer	10		Derived: Use date time part of LB.LBDTC associated with AVAL and/or AVALC in numeric format.

Figure 8. ISO8601 datetime string content to numeric date (time/datetime) conversion.

Example 3: Subject level indicator variables such as analysis population flags required for every population that is defined in the statistical analysis plan are created based on the algorithms involving more than one variable as seen in Figure 9 below.



ENRLFL	Enrolled Population Flag		text	1	["N" = " No ", "Y" = " Yes "] <No Yes Response>	Derived: Set to 'Y', if the subject was enrolled into the study. For DS.DSCAT = 'PROTOCOL MILESTONE' and DS.DSSCAT = 'ENROLLMENT' and DS.DSTERM = 'ENROLLED'. Else set to 'N'.
SAFFL	Safety Population Flag		text	1	["N" = " No ", "Y" = " Yes "] <No Yes Response>	Derived: Set to 'Y', if patient's reference interval start datetime (DM.RFSTDTC) is not missing. Else Set to 'N'.
ITTFL	Intent-To-Treat Population Flag		text	1	["N" = " No ", "Y" = " Yes "] <No Yes Response>	Derived: Set to 'Y', if the subject is included in the ITT analysis set (randomized subjects) for the enrolled population. i.e. For the subjects with ENRLFL='Y' set to 'Y' if RANDDT ne missing. Set to 'N' otherwise.
RANDFL	Randomized Population Flag		text	1	["N" = " No ", "Y" = " Yes "] <No Yes Response>	Derived: Populate as 'Y' for subjects who have a randomization date not missing. Else set to 'N'.

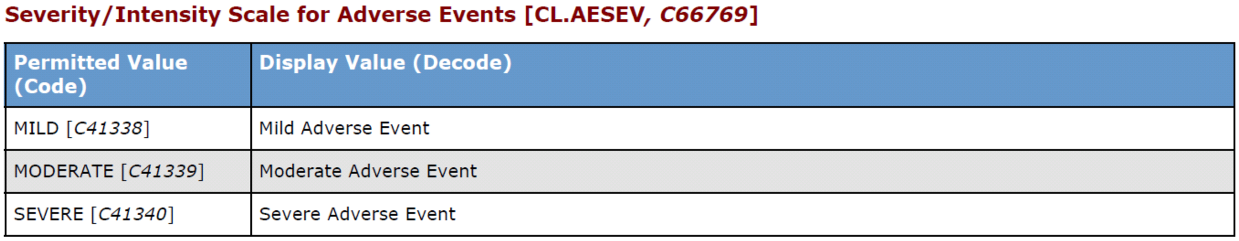
Figure 9. Subject – Level population flags derived based on an algorithm involving comparison of one or more variables.

Working with Control Terminology (CT) in ADaM define.xml:

ADaM metadata team recommends the following considerations while building the CT in ADaM Define.xml.

CT should be provided for every variable with a finite set of valid values. Even though a variable has only one valid value, it may be advantageous to specify a codelist for validation purposes. Multiple variables can reference the same codelist. The "Codelist / Controlled Terms" in the ADaM IG are generally a good reference when deciding on the need for CT for a specific variable. All values in the permissible value set for the study should be included, whether they are represented in the submitted data or not. It is always recommended to ensure the CT for SDTM variables carried into ADaM are consistent with SDTM define.xml for traceability.

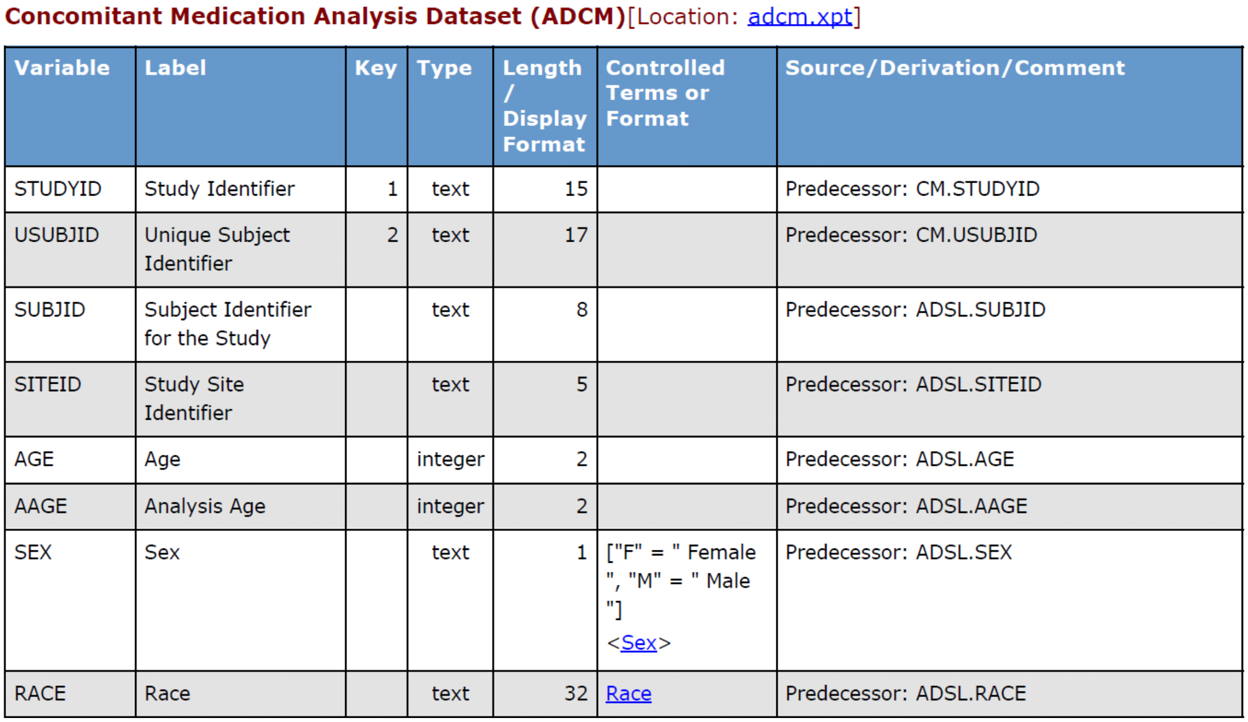
For example, if there was the possibility to classify severity as "MILD", "MODERATE" or "SEVERE", but only events of mild severity were reported, the full list of possible values, i.e. "MILD", "MODERATE", "SEVERE" should be included on the define.xml file per the SDTM IG v3.2, Section 4.1.3.3 Controlled Terminology Values [6].The same rule is applicable to ADaM define.xml.



Severity/Intensity Scale for Adverse Events [CL.AESEV, *C66769*]

Permitted Value (Code)	Display Value (Decode)
MILD [*C41338*]	Mild Adverse Event
MODERATE [*C41339*]	Moderate Adverse Event
SEVERE [*C41340*]	Severe Adverse Event

Figure 10. Codelist for Severity/Intensity scale for Adverse events.

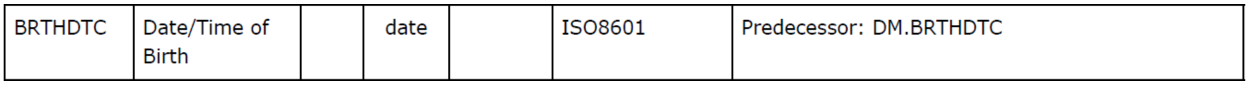


Concomitant Medication Analysis Dataset (ADCM)[Location: adcm.xpt]

Variable	Label	Key	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	1	text	15		Predecessor: CM.STUDYID
USUBJID	Unique Subject Identifier	2	text	17		Predecessor: CM.USUBJID
SUBJID	Subject Identifier for the Study		text	8		Predecessor: ADSL.SUBJID
SITEID	Study Site Identifier		text	5		Predecessor: ADSL.SITEID
AGE	Age		integer	2		Predecessor: ADSL.AGE
AAGE	Analysis Age		integer	2		Predecessor: ADSL.AAGE
SEX	Sex		text	1	["F" = " Female ", "M" = " Male "] <Sex>	Predecessor: ADSL.SEX
RACE	Race		text	32	Race	Predecessor: ADSL.RACE

Figure 11. ADaM Variables with Control Terminology.

ADaM Metadata team suggests ISO 8601 is considered a presentation format rather than CT; therefore, no CT reference is specified when compiling the metadata for an ISO 8601 formatted datetime or duration. The stylesheet provided with the Define-XML v2.0 package automatically populates the Controlled Terms/Codelist column with "ISO 8601" wherever applicable (as seen in Figure 12 below).



BRTHDTC	Date/Time of Birth		date		ISO8601	Predecessor: DM.BRTHDTC

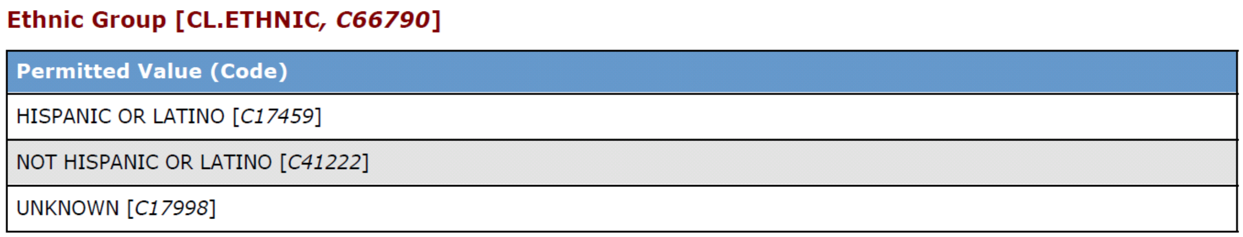
Figure 12. ISO8601 format displayed for BRTHDTC.

The enhancements in Define-XML v2.0 simplify the study controlled terminology metadata by 1) differentiating "Enumerated" Items list from Code/Decode Lists, allowing the controlled terms to be displayed without the redundant "Decode" part. 2) Identifying codelist values as CDISC or other standard terminology (including sponsor-defined), or as sponsor extensions to standard terminology. The CDISC NCI C-codes can be displayed, if applicable. 3) Supports greater control over ordering of codelist elements.

Types of CT definitions

Define-XML provides the possibility of including two types of CT definitions:

- **Enumerated Item Lists** - these include a simple list of valid values, the set of values from the list itself is sufficient for data interpretation. See example CT for ETHNIC in Figure 13.

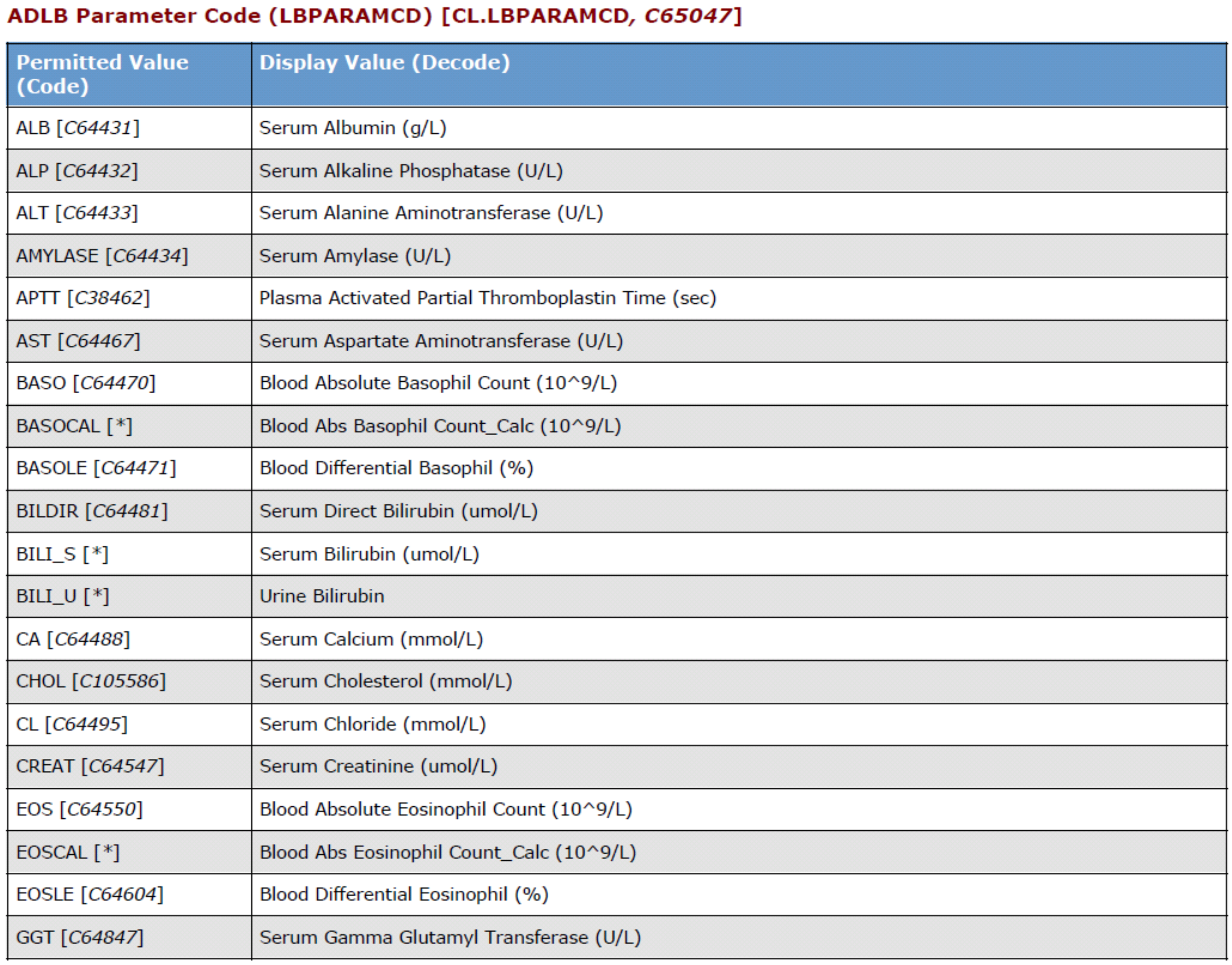


Ethnic Group [CL.ETHNIC, *C66790*]

Permitted Value (Code)
HISPANIC OR LATINO [*C17459*]
NOT HISPANIC OR LATINO [*C41222*]
UNKNOWN [*C17998*]

Figure 13. Enumerated Items List for Ethnic group.

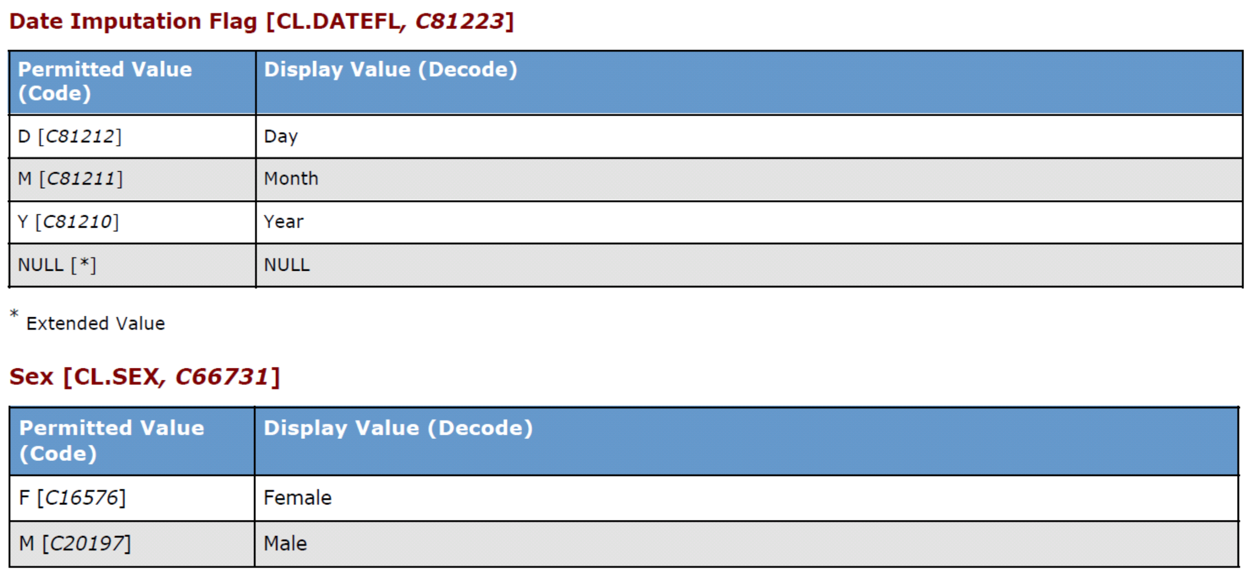
- **Code/Decode Lists** - Codelists are mainly used when decodes for the valid values facilitate data interpretation (e.g. CTs for variables like PARAMCD, DATEFL, SEX, ADaM numeric variables with suffix N like AGEGRP1N or RACEN for their primary character counterparts).



ADLB Parameter Code (LBPARAMCD) [CL.LBPARAMCD, *C65047*]

Permitted Value (Code)	Display Value (Decode)
ALB [*C64431*]	Serum Albumin (g/L)
ALP [*C64432*]	Serum Alkaline Phosphatase (U/L)
ALT [*C64433*]	Serum Alanine Aminotransferase (U/L)
AMYLASE [*C64434*]	Serum Amylase (U/L)
APTT [*C38462*]	Plasma Activated Partial Thromboplastin Time (sec)
AST [*C64467*]	Serum Aspartate Aminotransferase (U/L)
BASO [*C64470*]	Blood Absolute Basophil Count (10^9/L)
BASOCAL [*]	Blood Abs Basophil Count_Calc (10^9/L)
BASOLE [*C64471*]	Blood Differential Basophil (%)
BILDIR [*C64481*]	Serum Direct Bilirubin (umol/L)
BILI_S [*]	Serum Bilirubin (umol/L)
BILI_U [*]	Urine Bilirubin
CA [*C64488*]	Serum Calcium (mmol/L)
CHOL [*C105586*]	Serum Cholesterol (mmol/L)
CL [*C64495*]	Serum Chloride (mmol/L)
CREAT [*C64547*]	Serum Creatinine (umol/L)
EOS [*C64550*]	Blood Absolute Eosinophil Count (10^9/L)
EOSCAL [*]	Blood Abs Eosinophil Count_Calc (10^9/L)
EOSLE [*C64604*]	Blood Differential Eosinophil (%)
GGT [*C64847*]	Serum Gamma Glutamyl Transferase (U/L)

Figure 14.1 PARAMCD Codelist with Code/Decode and applicable NCI C - codes.



Date Imputation Flag [CL.DATEFL, *C81223*]

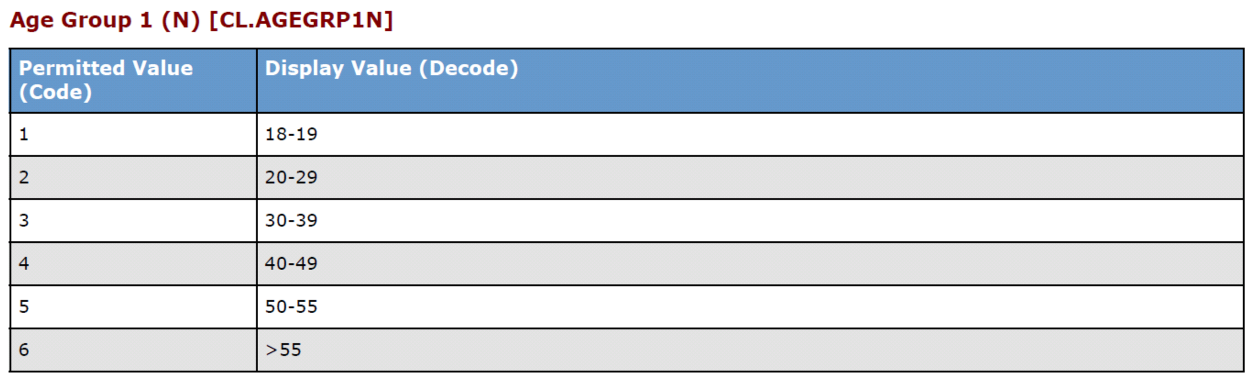
Permitted Value (Code)	Display Value (Decode)
D [*C81212*]	Day
M [*C81211*]	Month
Y [*C81210*]	Year
NULL [*]	NULL

* Extended Value

Sex [CL.SEX, *C66731*]

Permitted Value (Code)	Display Value (Decode)
F [*C16576*]	Female
M [*C20197*]	Male

Figure 14.3 Codelists with Code/Decode and applicable NCI C – codes for ADaM character variables.



Age Group 1 (N) [CL.AGEGRP1N]

Permitted Value (Code)	Display Value (Decode)
1	18-19
2	20-29
3	30-39
4	40-49
5	50-55
6	>55

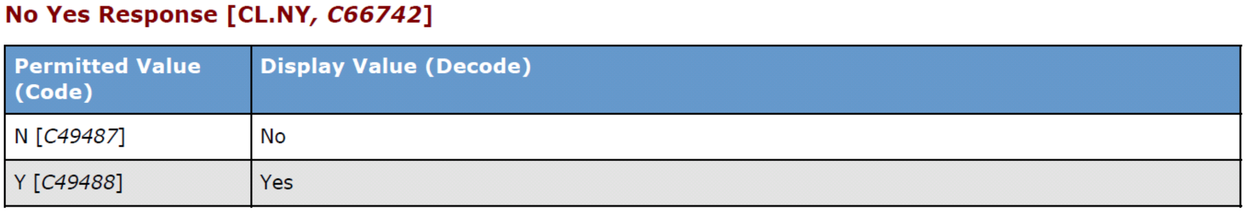
Figure 14.4 Codelist for ADaM Numeric variable AGEGRP1N for character counterpart.

Naming Conventions

For NCI/CDISC CT, ADaM metadata team recommends the CodeList Name attribute must exactly match the CodeList Name from the published Controlled Terminology ODM (see [5] Define-XML v2.0, Section 5.3.12). These codelist names should not be used for sponsor specific CT. If a NCI/CDISC CT is defined as extensible by CDISC, sponsor specific additional values can be added. Before defining an additional value, the suggested value should be checked and it is not a synonym for an available CT value and then the sponsor specific value can be added and marked as an extended value in define.xml

Dealing with subsets of Control Terminology

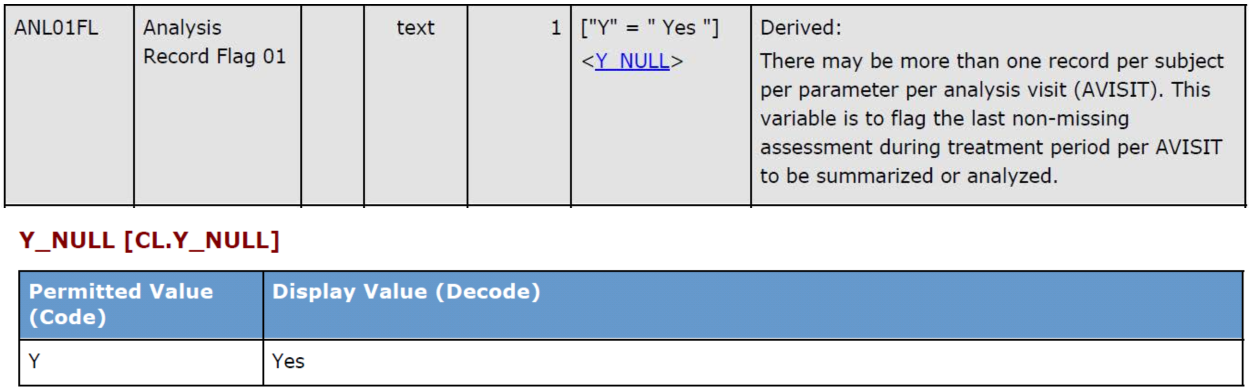
Some variables in ADaM datasets share a common NCI/CDISC CT reference for example the standard "No Yes Response" (NY) CT (Figure 15) is referenced for the variables ABLFL, ITTFL, ITTRFL, etc. However, only a subset of the values defined in the respective NCI/CDISC CT may be applicable for some variables. ITTFL has values "N" and "Y" whereas only "Y" is applicable for ABLFL or ANL01FL (Figure 16) in such cases subsets of codelists are recommended.



No Yes Response [CL.NY, *C66742*]

Permitted Value (Code)	Display Value (Decode)
N [*C49487*]	No
Y [*C49488*]	Yes

Figure 15. "No Yes Response" (NY) Codelist.



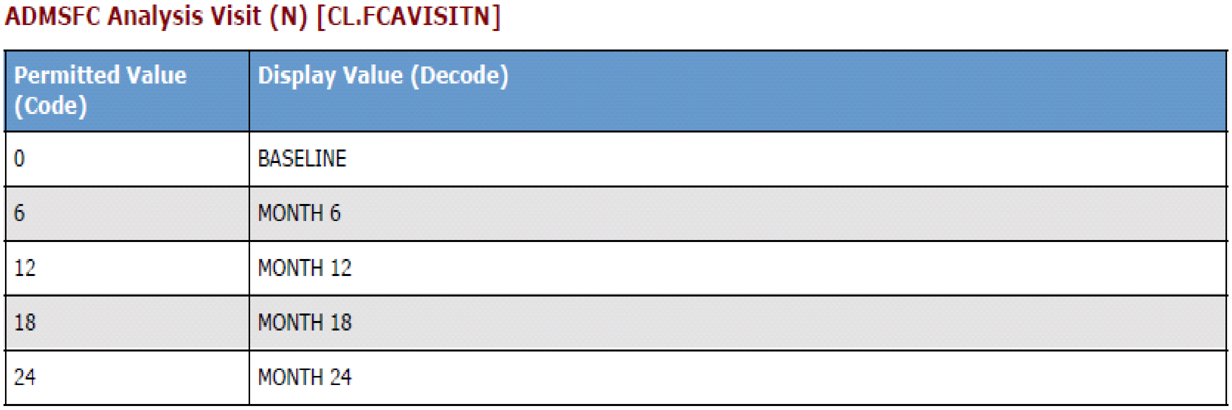
ANL01FL	Analysis Record Flag 01		text	1	["Y" = " Yes "] <Y_NULL>	Derived: There may be more than one record per subject per parameter per analysis visit (AVISIT). This variable is to flag the last non-missing assessment during treatment period per AVISIT to be summarized or analyzed.

Y_NULL [CL.Y_NULL]

Permitted Value (Code)	Display Value (Decode)
Y	Yes

Figure 16. ADaM variable ANL01FL using subset of Codelist.

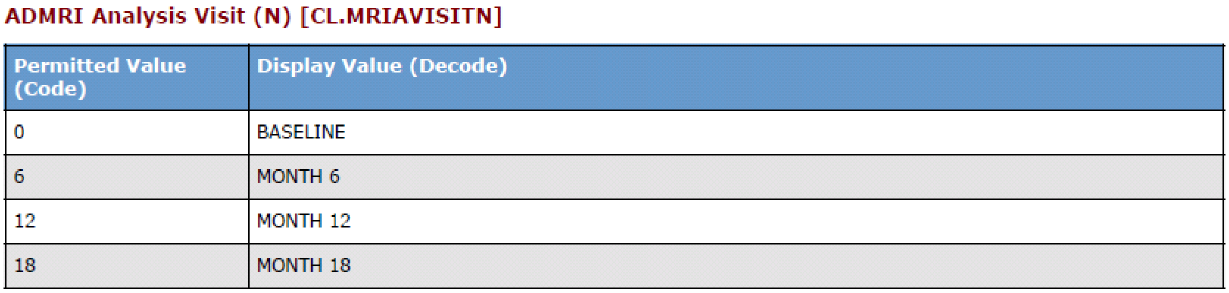
Variables such as AVISIT, AVISITN, --UNITS at times share the same list of discrete values across ADaM datasets. Within a trial, there may be circumstances where a reviewer would benefit from CT lists that are specific for a given domain or data set in order to understand the list of terms applicable to that domain or data set. Therefore, it is recommended to have separate CT for each variable pertaining to the values available within the dataset to avoid confusion for the reviewers (Figure 17a and 17b) in such cases.



ADMSFC Analysis Visit (N) [CL.FCAVISITN]

Permitted Value (Code)	Display Value (Decode)
0	BASELINE
6	MONTH 6
12	MONTH 12
18	MONTH 18
24	MONTH 24

Figure 17a. AVISITN variable using subset of Codelist to display the AVISITN values applicable to the ADMSFC ADaM dataset.



ADMRI Analysis Visit (N) [CL.MRIAVISITN]

Permitted Value (Code)	Display Value (Decode)
0	BASELINE
6	MONTH 6
12	MONTH 12
18	MONTH 18

Figure 17b. AVISITN variable using subset of Codelist to display the AVISITN values applicable to the ADMRI ADaM dataset.

It is always important to be aware of specific set of values that are applicable for a certain variable. This information is also useful for validation. Following convention is suggested by the metadata team when a subset of a CT is required. If the CT is defined by NCI/CDISC

<Codelist name as published in the NCI/CDISC ODM>< unique subset identifier suffix>.

The subset identifier suffix is used to make a distinction between different subsets of CT applicable for different variables.

CONCLUSION

Successful implementation of traceability and structural consistency in ADaM Define.xml v2.0 following the recommended best practices makes it easy to identify derived or imputed information within the ADaM dataset and identify the information coming from SDTM data and helps speed up the review of metadata. There is a need for continuous improvement for achieving higher quality submissions data, perspective users should periodically review the evolving industry standards and recommended best practices while working on the crucial submission aspects to deliver quality outputs.

REFERENCES

[1] Study Data Technical conformance guide

https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

[2] ADaM IG

https://www.cdisc.org/standards/foundational/adam

[3] Origin for ADaM Variables

https://wiki.cdisc.org/display/DEF/Origin+for+ADaM+Variables

[4] Working with Controlled Terminology (CT)

https://wiki.cdisc.org/pages/viewpage.action?pageId=25723999

[5] CDISC Define.XML v2.0 Guidance

https://www.cdisc.org/standards/foundational/define-xml/define-xml-v20

[6] SDTM IG v3.2

https://www.cdisc.org/standards/foundational/sdtmig

ACKNOWLEDGMENTS

I would like to thank Hunter Everton, Theodore Cherico, Ryan Wilkins, LaTonya Murphy, Ken Borowiak, my colleagues, family and friends for their valuable suggestions, support and encouragement throughout.

RECOMMENDED READING

http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm

CONTACT INFORMATION

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

Supriya Davuluri
PPD, LLC.
Morrisville, NC 27560.
Email: Supriya.Davuluri@ppdi.com

DISCLAIMER

The content of this paper are the works of the author and do not necessarily represent the opinions, recommendations, or practices of PPD, LLC.