

New Features in Define-XML V2.0 and Its Impact on SDTM/ADaM Specifications

Hang Pang, Vertex Pharmaceuticals Incorporated, Boston, MA

ABSTRACT

Define-XML is required for NDA/BLA submission (FDA Study Data Technical Conformance Guide, V2.3, Oct. 2015). CDISC Define-XML Specification V2.0 (2013) has some significant changes compared with CRT-DDS (Case Report Tabulation Data Definition Specification (define.xml), V1.0, 2005). This paper will discuss the new features in Define-XML V2.0 (e.g. integrates with industry standard CDISC NCI controlled terminologies, and supports more complicated value-level metadata (VLM) definitions...), and its impact on SDTM/ADaM data specifications for submission readiness. The SDTM/ADaM data specifications will be used as a metadata for Define-XML (V2.0) generation, and simplified the NDA/BLA e-submission preparation process. An example of a SAS macro and ADaM specifications for Define-XML (V2.0) generation will be presented.

INTRODUCTION

Define-XML is required for new drug submission by agencies [3][4]. FDA “**STUDY DATA TECHNICAL CONFORMANCE GUIDE**” (Version 2.3, Oct. 2015) stated that “The data definition file describes the metadata of the submitted electronic datasets, and is considered arguably the most important part of the electronic dataset submission for regulatory review. This data definition specification for submitted datasets defines the metadata structures that should be used to describe the datasets and variables.”[3]

Define-XML V2.0 was published in March 2013 [1], and now is accepted by FDA as the standards for study data submissions [5]. It is the first update since CRT-DDS V1.0 was published in 2005 [2]. Define-XML V1.0 was designed with SDTM in mind. V2.0 is intended for use with both SDTM and ADaM. Some things that used to be simple but inflexible are now flexible but complicated. Define-XML V2.0 enables more precise description of the data, but demands more details to achieve it.

The new standards in Define-XML V2.0 have quite a few new features and requirements. For example, the new standards allow the omission of decode if the code is actually an enumerated item of controlled terminology, and support CDISC NCI controlled terminology terms with the C-codes. For the value level metadata (VLM), a Where clause has been introduced, any variable can now have associated value-level metadata, not just --TESTCD or PARAMCD. The new standards support clear links to external files such as aCRF, SAP, reviewer’s guide, SAS programs or other supplementary documents. It also supports the analysis result metadata in ADaM Define-XML. Most of these changes in define file requirements/components imply the changes in SDTM/ADaM specifications.

The purpose of this paper is to compare the Define-XML V1.0 with V2.0, discuss the new features in Define-XML V2.0, and its impact on SDTM/ADaM data specifications for submission readiness.

CONSIDERATION AND DISCUSSION

1. COMPARE CDISC DEFINE-XML V1.0 VS V2.0

Define-XML includes following key components:

- Dataset Definitions
- Dataset Variable Definitions
- Value Level Metadata Definitions
- Controlled Terminology Definitions
- Computational Method Definitions
- Comment Definitions
- Supplemental Data Definition Documents (e.g. Annotated CRF (acrf.pdf), Study Data Reviewer’s Guide (SDRG.pdf), Analysis Data Reviewer’s Guide (ADRG.pdf), and SAS programs, etc.)
- Key Analysis Results Definitions in ADaM Define-XML (optional)

1.1 Dataset Definitions

Dataset definition identifies each dataset that is submitted, brief description and its location within the submission. However, some datasets need more explanation than you can put in a dataset label with more substantive description. Sometimes they need lots of explanation. For example, you may want to be able to hyperlink to the relevant pages of an external documents, e.g. Study (or Analysis) Data Reviewer's Guide.

1.1.1 Dataset Definitions (V1.0)

In Define-XML V1.0, it includes Dataset, Description, Class, Structure, Purpose, Keys, and Location.

Datasets for Study 1234					
Dataset	Description	Structure	Purpose	Keys	Location
DM	Demographics	Special Purpose - One record per event per subject	Tabulation	STUDYID, USUBJID	crt/datasets/1234/dm.xpt
TE	Trial Elements	Trial Design - One Record Per Element	Tabulation	STUDYID, ELEMENT	crt/datasets/1234/te.xpt
TA	Trial Arms	Trial Design - One Record per Element for each Arm	Tabulation	STUDYID, ARM	crt/datasets/1234/ta.xpt
TV	Trial Visits	Trial Design - One Record per Visit per Arm	Tabulation	STUDYID, VISIT	crt/datasets/1234/tv.xpt

1.1.2 Dataset Definitions (V2.0)

Define-XML V2.0 adds a column 'Documentation', which supports to explicit links to external reference documents such that derivation rules, SAP, data reviewer's guide, and dataset generation programs. The links can precisely link to the PDF page numbers, page ranges, or named destinations.

It also adds a column 'Class', usually the SDTM data are listed in order of class TRIAL DESIGN, SPECIAL PURPOSE, INTERVENTIONS, EVENTS, FINDINGS, RELATIONSHIP (e.g. RELREC, SUPPQUAL domains), the ADaM datasets are listed in order of class ADSL, OCCDS (Occurrence Data Structure e.g. ADAE, ADCM), and BDS (Basic Data Structure), then in alphabetical order within the class.

SDTM:

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
TA	Trial Arms	TRIAL DESIGN	One record per planned Element per Arm	Tabulation	STUDYID, ARMCD, TAETORD	ta.xpt	
TE	Trial Elements	TRIAL DESIGN	One record per planned Element	Tabulation	STUDYID, ETCD	te.xpt	
TV	Trial Visits	TRIAL DESIGN	One record per planned Visit per Arm	Tabulation	STUDYID, VISITNUM, ARMCD	tv.xpt	
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID, USUBJID	dm.xpt	See Reviewer's Guide, Section 2.1 Demographics Reviewers Guide

ADaM:

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
ADSL	Subject-Level Analysis	SUBJECT LEVEL ANALYSIS DATASET	one record per subject	Analysis	USUBJID	adsl.xpt	Screen Failures are excluded since they are not needed for this study analysis
ADQSADAS	ADAS-Cog Analysis	BASIC DATA STRUCTURE	One record per subject per parameter per analysis visit per analysis date	Analysis	USUBJID, PARAMCD, AVISIT, ADT	adqsadas.xpt	See referenced dataset creation program and ADRG, Section 2.1 adqsadas.sas Analysis Data Reviewer's Guide

1.2 Dataset Variable Definitions

1.2.1 Dataset Variable Definitions (V1.0)

Demographics Dataset (DM)						
Variable	Label	Type	Controlled Terms or Format	Origin	Role	Comment
STUDYID	STUDY IDENTIFIER	text		CRF Page	Identifier	Demographics CRF Page 4
DOMAIN	DOMAIN ABBREVIATION	text		CRF Page	Identifier	DOMAIN ABBREVIATION
USUBJID	UNIQUE SUBJECT IDENTIFIER	text		CRF Page	Identifier	Demographics CRF Page 4
SUBJID	SUBJECT IDENTIFIER	text		CRF Page	Topic	Demographics CRF Page 4

1.2.2 Dataset Variable Definitions (V2.0):

The 'Origin' indicates the variable is given a value directly or by mapping from another variable. 'Origin' allowed values for SDTM are Protocol, CRF page(s), eDT, Derived, and Assigned. It supports annotated CRF page references for hyperlinks, even to multiple aCRFs. 'Derived' should be used for all other derivations. If the 'Origin' is Derived, a Method is required, which may also include a document/page reference.

Column 'Controlled Terms or Format' should be populated if the variable is governed by a codelist or format mapping. The possible valid values are 1) a CT name from the Codelist tab; 2) a CT name from the NCI tab.

Links may be placed in column 'Source/Derivation/Comment', the syntax is described in section 2.2.7.

SDTM:

Comparing to Define-XML V1.0, V2.0 added columns 'Key' and 'Length', removed the column 'Role', and Column 'Comment' was changed to 'Derivation/Comment'. Column 'Key' is for the key variable sequence which indicates that this variable is a key for the enclosing group. It also provides an ordering for the keys.

Demographics (DM) [Location: [dm.xpt](#)]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier	1	text	7		Protocol	
DOMAIN	Domain Abbreviation		text	2	["DM" = "Demographics"] < Domain Abb. (DM) >	Assigned	
USUBJID	Unique Subject Identifier	2	text	14		Derived	Concatenation of STUDYID and SUBJID
SUBJID	Subject Identifier for the Study		text	6		CRF Page 3	

ADaM:

Comparing to V1.0, Define-XML V2.0 added a column 'Length / Display Format', removed column 'Role', and combined two columns 'Source' and 'Comment' into column 'Source/Derivation/Comment'

'Source' value could be Predecessor (SDTM domain name, ADaM dataset name), Assigned, and Derived.

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

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	text	12		Predecessor: DM.STUDYID
USUBJID	Unique Subject Identifier	text	11		Predecessor: DM.USUBJID
SUBJID	Subject Identifier for the Study	text	4		Predecessor: DM.SUBJID
SITEID	Study Site Identifier	text	3		Predecessor: DM.SITEID
SITEGR1	Pooled Site Group 1	text	3		Derived: refer to SAP, Section 7.1 - if not pooled then SITEGR1=SITEID. If pooled, SITEGR1 will be 900

1.3 Value Level Metadata (VLM) Definitions:

Value level metadata provides information about the value of single variable in relationship to the value of another variable(s), e.g. AVAL (standard results) related to PARAMCD.

1.3.1 Value Level Metadata Definitions (V1.0):

Value level metadata describes a variable's values that are associated with a specific value of --TESTCD or PARAMCD in Define-XML V1.0. Only one variable is described, its identity is assumed but not specified. In the Value Level Metadata in V1.0, each value of --TESTCD or PARAMCD should be enumerated even some of them share the same metadata.

Value Level Metadata							
Source Variable	Value	Label	Type	Controlled Terms or Format	Origin	Role	Comment
SCTESTCD	ALLERGY	Allergy Status	integer	YESNOUNK	CRF Page		Subject Characteristics CRF Page 4
SCTESTCD	EDLEVLN	EDUCATIONAL LEVEL-DVN	float		CRF Page		Subject Characteristics CRF Page 4
SCTESTCD	EXCLSN	EXERCISE CLASSIFICATION-DVN	float		CRF Page		Subject Characteristics CRF Page 4
SCTESTCD	RACEOTH	Text for other race	text		CRF Page		Subject Characteristics CRF Page 4

1.3.2 Value Level Metadata Definitions (V2.0):

Instead of Define-XML V1.0 with columns 'Source Variable' (--TESTCD or PARAMCD), 'Value', 'Label', V2.0 replaced them with columns 'Variable' (could be any variable) and 'Where' clause. V2.0 also added column 'Length/Display Format', removed the column 'Role', and column 'Comment' was changed to 'Derivation/Comment'.

Define-XML V2.0 identifies record types with a WHERE clause. Every item with the same WHERE clause belongs to the same record type. With this method you can provide a value-level description of any combination of variables in the dataset, not just --TESTCD or PARAMCD. The Where clause can even handle more complicated situation where more than one variable is allowed,

SDTM:

Value Level Metadata - DA [DAORRES]						
Variable	Where	Type	Length / Display Format	Controlled Terms or Format	Origin	Derivation/Comment
DAORRES	DATESTCD EQ DISPAMT (Dispensed Amount)	integer	2		CRF Page 19	
DAORRES	DATESTCD EQ RETAMT (Returned Amount)	integer	2		CRF Page 19	

ADaM:

Parameter Value List - ADQSADAS [DTYPE]						
Variable	Where	Type	Length / Display Format	Controlled Terms or Format	Origin	Derivation/Comment
DTYPE	PARAMCD NOTIN ("ACTOT" (Adas-Cog(11) Subscore))	text	10	["LOCF" = "Last Observation Carried Forward"] <Derivation Type>	Assigned	Value: null

DTYPE	PARAMCD EQ ACTOT (Adas- Cog(11) Subscore)	text	10	["LOCF" = "Last Observation Carried Forward"] < Derivation Type >	Assigned	Value: LOCF denotes that the LOCF imputation method was used to impute the value for the given parameter and analysis visit.
-------	---	------	----	--	----------	--

1.4 Controlled Terminology Definitions:

1.4.1 Controlled Terminology Definitions (V1.0):

In Define-XML V1.0, the code lists are always presented in pair, Code and Decode. In most of cases of enumerated terms (e.g. AEACN), the identical values of codes and decodes are redundant. In addition, CDISC/NCI defined more and more controlled terms. You can not differentiate the CDISC terminology (C-codes), sponsor terminology, and sponsor extensions to CDISC terminology.

Controlled Terminology (Code Lists) Section	
Code Value	Code Text
YESNO, Reference Name (YESNO)	
Y	YES
N	NO
U	NOT APPLICABLE

1.4.2 Controlled Terminology Definitions (V2.0):

The enhancements in Define-XML V2.0 simplify the study controlled terminology metadata by 1) Differentiates "Enumerated" codelists with Permitted Value (Code) (e.g. Mild, Moderate, ...) from Code/Decode codelists (Permitted Value/Display Value) (e.g. 1 = Mild, 2 = Moderate, ...), allowing the controlled terms displayed without the redundant "Decode" part. 2) Identifies codelist values as CDISC or other standard terminology (including sponsor-defined), or as sponsor extensions to standard terminology. The CDISC NCI C-codes are able to be displayed, if applicable. 3) Supports greater control over ordering of codelist elements.

Enumerated Codelist:

Action Taken with Study Treatment [CL.AC.N, C66767]
Permitted Value (Code)
DOSE NOT CHANGED [C49504]
DOSE REDUCED [C49505]
DRUG INTERRUPTED [C49501]
DRUG WITHDRAWN [C49502]

Code/Decode Codelist:

No Yes Response Subset [CL.NY, C66742]	
Permitted Value (Code)	Display Value (Decode)
N [C49487]	No
Y [C49488]	Yes

1.5 Computational Method and Comment Definitions

1.5.1 Computational Method Definitions (V1.0):

The computational Method definition in Define-XML V1.0 is used for the case of complex derivations or derivations used more than once.

Computational Algorithms Section	
Reference Name	Computation Method
COMPMETHOD.MEANBP	Sum all mean blood pressure readings for visits 0, 1, and 2 and divide by the total number of readings

1.5.2 Computational Method and Comment Definitions (V2.0):

SDTM:

The “**Computational Algorithm**” Definition in SDTM Define-XML V2.0 automatically polled all the derivation algorithms from the Variable metadata column ‘Derivation/Common’ (non-missing text) when the ‘Origin’=**Derived**.

The “**Comments**” Definition in Define-XML V2.0 automatically polled all the comments from the Dataset metadata column ‘Documentation’ (non-missing text), and Variable metadata column ‘Derivation/Common’ (non-missing text) when the ‘Origin’=**Assigned**.

Computational Algorithms		
Method	Type	Description
Algorithm to derive AEENDY	Computation	AEENDY = AEENDTC - RFSTDTC+1 if AEENDTC is on or after RFSTDTC. AEENDTC - RFSTDTC if AEENDTC precedes RFSTDTC
Algorithm to derive AESTDY	Computation	AESTDY = AESTDTC - RFSTDTC+1 if AESTDTC is on or after RFSTDTC. AESTDTC - RFSTDTC if AESTDTC precedes RFSTDTC
Algorithm to derive the AETRTEM flag	Computation	AETRTEM = "Y" if Adverse Event was not present prior to the RFSTDTC, or it was present prior to the RFSTDTC but increased in severity during the treatment period. Null otherwise.
Algorithm to derive AGE	Computation	Age at Screening Date (Screening Date - Birth date). For the complete algorithm see the referenced external document. Complex Algorithms (complexalgorithms.pdf)

ADaM:

The “**Analysis Derivations**” Definition in ADaM Define-XML V2.0 automatically polled all the analysis derivation algorithm from the Variable metadata column ‘Source/Derivation/Common’ (non-missing text) when the ‘Source’=**Derived**.

The “**Comments**” Definition in ADaM Define-XML V2.0 automatically polled all the comments from the Dataset metadata column ‘Documentation’ (non-missing text), and Variable metadata column ‘Source/Derivation/Common’ (non-missing text) when the ‘Source’=**Assigned**.

Analysis Derivations		
Method	Type	Description
CM.ADQSADAS.AVISIT	Computation	Derived based on windowing algorithm described in SAP, Section 8.2
CM.ADQSADAS.ADY	Computation	ADY = ADT - TRTSDT + 1, if ADT >= TRTSDT. ADY = ADT - TRTSDT, if ADT < TRTSDT.
CM.ADQSADAS.ADT	Computation	SAS date from QS.QSDTC
CM.ADQSADAS.AVAL	Computation	Derivations are described per parameter in the parameter value level metadata

1.6 Supplemental Data Definition Documents:

Under Define-XML V2.0, enhanced links in Documentation allow sponsors to provide information in external files such as annotated CRFs, a SAS program, a data reviewer’s guide (SDRG or ADRG), or SAP. According to the FDA requirements, the SAS programs for ADaM datasets, tables and figures of primary and secondary efficacy endpoints can be provided as ASCII text file or PDF file, which stored in the folder “programs” as the eCTD structure. The data reviewer’s guide or other supplement documents are in the same folder of Define.xml [2].

The Study Data Reviewer’s Guide (SDRG.pdf) for SDTM and Analysis Data Reviewer’s Guide (ADRG.pdf) for ADaM are recommended as an integral part of a standards-compliant study data submission by FDA [3]. The SDRG/ADRG should describe any special considerations or directions that may facilitate a FDA reviewer’s use of the submitted data:

- 1) Study protocol title, number, and version
- 2) Study design
- 3) Standards, formats, and terminologies and their versions
- 4) Description of study datasets or analysis data
- 5) Data standards validation rules, versions, and issues
- 6) Description of all sponsor decisions related to data standard implementations

Suggest to use PhUSE Industry Standard SDRG (V1.2) & ADRG (V1.1) Templates (2015-01-16) for the SDRG and ADRG.

2. SDTM/ADAM SPECIFICATIONS AS METADATA FOR DEFINE-XML V2.0

To increase the efficiency of developing the Define-XML and reducing the redundant metadata preparation process, examples will be given to illustrate what changes in SDTM/ADaM specifications are made in order to be served as the metadata and embrace the new Define-XML V2.0 features and requirements.

2.1 SDTM/ADaM Data Be Re-Sized

FDA requested that “The allotted length for each column containing character (text) data should be set to the maximum length of the variable used across all datasets in the study. This will significantly reduce file sizes. For example, if USUBJID has a maximum length of 18, the USUBJID’s column size should be set to 18, not 200.” [3]

To prepare the submitted data, call the resize macro (resize.sas) to use the actual variable longest length. Please note SDTM and ADaM datasets must be re-sized together, because some SDTM variables used to derive ADaM variables are also kept in the ADaM data for the traceability purpose.

2.2 SDTM/ADaM Specification (Excel Sheet):

Here just taking ADaM specification as an example, SDTM specification has a similar structure.

2.2.1 Study Tab

The Study tab contains study level information. The first 4 columns are required and the last two columns for dictionary version should be filled if study has AE and/or CM.

Study Name	Study Description	Protocol	Reviewer's Guide	MedDRA Version	WHO Drug Version
XXX-XXX-001	An Open-Label Study of the Safety, Tolerability, and Pharmacokinetic / Pharmacodynamic Profile of XXX as a Single Agent in Subjects With YYY Disease	XXX-XXX-001	adrg.pdf	19.0	1st March 2016

2.2.2 Domain Definitions Tab

The Domain tab lists all ADaM datasets and provides dataset level information. Note that the order of datasets presented in the Define-XML is the order they are listed in this sheet. Usually, in order of ADSL, OCCDS (Occurrence Data Structure e.g. ADAE, ADCM), and BDS (Basic Data Structure), then in alphabetical order within the class.

Dataset	Description	Class	Structure	Purpose	Key Variables	Documentation
ADSL	Subject Level Analysis Dataset	SUBJECT LEVEL ANALYSIS DATASET	One record per subject	Analysis	STUDYID, USUBJID	Contains Core variables and subject level covariates. See SAS program <link file=./programs/adsl.sas>
ADAE	Adverse Events Analysis Dataset	OCCURRENCE DATA STRUCTURE	One record per event per subject	Analysis	STUDYID, USUBJID, AETERM, ASTDT, AESEQ	See SAS program <link file=./programs/adae.sas>
ADLB	Laboratory Tests Analysis Dataset	BASIC DATA STRUCTURE	One record per lab test per time point per visit per subject	Analysis	STUDYID, USUBJID, PARAMCD, ADTM, AVISITN, ATPTN	Only lab tests specified in the SAP are retained from LB for analysis. See referenced dataset creation program and ADRG, Section 2.1 <link file=./programs/adlb.sas> <link file=reviewerguide.pdf ref=25 display=ADaM Reviewer's Guide p.25>

2.2.3 Variable Definitions Tab

A tab is created for every ADaM dataset to document the source and derivation of each variable, as one dataset one Tab in the specification sheet. The 'Origin' column could be "Predecessor" (source is from SDTM domain or other ADaM dataset), "Assigned", and "Derived".

Dataset	Variable	Label	Type	Length	Controlled Terms or Format	Origin	Source/Derivation/Comment	Core	Value Level Metadata (Where=)
ADSL	STUDYID	Study Identifier	Char	20		Predecessor	DM.STUDYID<link file=/sas/vx12-809-105-IA1-adam-programs.pdf> <link file=/sas/vx12-809-105-IA1-adam-programs.pdf ref=15 display=SAS Programs p.15> <link file=/sas/vx12-809-105-IA1-adam-programs.pdf ref=SAS_advts display=ADVS SAS program>	Req	
ADSL	USUBJID	Unique Subject Identifier	Char	40		Predecessor	DM.USUBJID <link file=/sas/vx12-809-105-IA1-adam-programs.pdf ref=15 display=SAS Programs p.15>	Req	
ADSL	SUBJID	Subject Identifier for the Study	Char	20		Predecessor	DM.SUBJID	Req	
ADSL	SITEID	Study Site Identifier	Char	8		Predecessor	DM.SITEID	Req	
ADSL	AGE	Age	Num	8		Predecessor	DM.AGE	Req	
ADSL	AGEU	Age Units	Char	8	AGEU	Predecessor	DM.AGEU	Req	
ADSL	AGEGR	Pooled Age Group	Char	20	AGEGR	Assigned	Map AGE low-<65=<65 Years' 65-high=>=65 Years' !=Missing'	Perm	

2.2.4 Value Level Metadata (VLM)

For VLM variables, we will have multiple rows, instead of one row. Value-level metadata provides information about the value of single variable in relationship to the value of another variable(s), e.g. AVAL (standard results) related to PARAMCD.

Condition can be based on values of multiple variables with more complicated conditions (not just EQUAL). The available logic operators in VLM are EQ, NE, IN, NOTIN, LT, LE, GT, GE. When more than one variable is involved, however the only permitted conjunction is "&". Here are some examples:

- VSTESTCD=SYSBP & VSPOS=STANDING
- PARAMCD IN (ACITM01, ACITM012, ACITM02, ACITM04)
- PARAMCD NE ACTTOT

	A	B	C	D	E	F	G	H	I	J	K
1	Variable Selected	Dataset	Variable	Label	Type	Length	Controlled Terms or Format	Origin	Source/Derivation/Comment	Core	Value Level Metadata (Where=)
28	Y	ADLB	ANL02FL	Analysis Record Flag 02 (Max Value)	Char	1			Derivations are described in the parameter value level metadata	Cond	
29							NYNULL	Derived	Equals to Y for the earliest maximum on-treatment R2A1HI for PARAMCD where APHASEN=1		PARAMCD in ('ALT', 'AST')
30							ANL02FL	Derived	Equals to Y for the earliest maximum on-treatment R2A1HI for PARAMCD wher APHASEN=1		PARAMCD in ("BILI", "GLUC")
31								Assigned	NULL		OTHER

2.2.5 Controlled Terminology Tab (Codelist)

The Codelist tab is used to define the sponsor's controlled terminologies. Column 'Codelist Name' defines the name of the codelist.

	A	B	C
1	Codelist Name	Permitted Value (Code)	Display Value (Decode)
85	SEVn	1	MILD
86	SEVn	2	MODERATE
87	SEVn	3	SEVERE
88	SEX	M	MALE
89	SEX	F	FEMALE
90	SEXn	1	MALE
91	SEXn	2	FEMALE
92	SEXn	3	UNKNOWN

2.2.6 CDISC NCI Controlled Terminology Tab

The CDISC **NCI** controlled terminology tab displays the CTs published by the National Cancer Institute. The published spreadsheet may be downloaded from <http://www.cancer.gov/research/resources/terminology/cdisc> and the first 6 columns (A-F) can be copied into this tab. The version of NCI CT used should be the same for ADaM and SDTM within a study.

When a variable is governed by a NCI CT, variable metadata should display the CT name in the 'Controlled Terms or Format' column. The NCI CT name is located in column "CDISC Submission Value". When column "Codelist Extensible (Yes/No)" equals "No", this is a non-extensible codelist meaning values not listed by NCI are not permitted in data.

When specifying a NCI codelist, only values that appear in data will be shown in the Define-XML. If it is necessary to display possible values which do not appear in data, create a codelist with the same name in the 'Codelist' tab listing all possible values. The macro will display the associated codes from the NCI CT if available.

	A	B	C	D	E	F	G
	Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	Version: 2015-03-27
1							
2	C66767		No	Action Taken with Study Treatment	ACN	Action Taken with Study Treatment	
3	C49503	C66767		Action Taken with Study Treatment	DOSE INCREASED		
4	C49504	C66767		Action Taken with Study Treatment	DOSE NOT CHANGED		
5	C49505	C66767		Action Taken with Study Treatment	DOSE REDUCED		
6	C49501	C66767		Action Taken with Study Treatment	DRUG INTERRUPTED		
7	C49502	C66767		Action Taken with Study Treatment	DRUG WITHDRAWN		
8	C48660	C66767		Action Taken with Study Treatment	NOT APPLICABLE	NA; Not Applicable	
9	C17998	C66767		Action Taken with Study Treatment	UNKNOWN	U; Unknown	
10	C101865		No	Acute Coronary Syndrome Presentation	ACSPCAT	Acute Coronary Syndrome Presentation	
11	C80383	C101865		Acute Coronary Syndrome Presentation	NON-ST ELEVATION	NON-STEMI; NSTEMI	
12	C17649	C101865		Acute Coronary Syndrome Presentation	OTHER	Other	
13	C101888	C101865		Acute Coronary Syndrome Presentation	ST ELEVATION	STEMI	
14	C66914	C101865		Acute Coronary Syndrome Presentation	STABLE ANGINA		
15	C66911	C101865		Acute Coronary Syndrome Presentation	UNSTABLE ANGINA		

2.2.7 External Files Link

Links can be created in the dataset, variable, and VLM metadata to any external file using the following syntax: <link file= **relative-path-to-file** >, such as: <link file=./programs/adsl.txt>. The relative path should take into consideration the relative location of the Define-XML and the linked document in the final esub folder structure.

By default, the name of the file will be used as the hyperlink text, this may be changed by adding the **display=** option: <link file=./programs/adsl.txt display=SAS Programs>

Furthermore, for PDF files, it is possible to specify the page number of the file by using the **ref=** option. Example: <link file=./ADRG.pdf ref=25 display=Analysis Data Review Guide p.25>

2.3 Macro for Define-XML V2.0

A SAS macro was developed for Define-XML V2.0, here is an example of SAS macro call program.

```

** Get the define.xml V2 macro **;
filename pgm "..\macrolib\define";
%inc pgm(define);

** Call macro define.sas for define.xml V2, define the paths **;
%define(
datapath= ..\esub\data\adam,
specpath= ..\esub\program\analysis\ADaM_Specifications.xls,
outpath= ..\esub\analysis\adam\datasets);

```

2.4 Consistency Checks

We need to do the Define-XML adherence checks: 1) availability of datasets and variables; 2) order of variables in dataset; 3) labels and data type; 4) variables with controlled terminologies. 5) check the intra document hyperlinks, e.g. annotated CRF page; 6) additional manual checks are required.

Some validation tools are available for Define-XML V2.0, e.g. Pinnacle 21 (OpenCDISC) Community V2.1 (12/7/2015), free download from web site: <http://www.opencdisc.org/>

CONCLUSION

The inadequacies of Define-XML V1.0, especially for ADaM, and the generality and flexibility of Define-XML V2.0 solutions provide compelling reasons to upgrade. The new features of Define-XML V2.0 can present the submitted data more efficiently. It supports CDISC NCI controlled terminology terms with the C-codes, the Where clause in value level metadata (VLM) leads any variable can now have associated value-level metadata, and the new standards support clear links to external files such as aCRF, reviewer's guide, SAS programs or other supplementary documents. It also supports the analysis result metadata in ADaM Define-XML.

Most of these changes in define file requirements/components imply the changes in SDTM/ADaM specifications. The SDTM/ADaM data specifications illustrated above serve for both submitted data generation and a metadata for Define-XML (V2.0) generation, which could simplify the NDA/BLA e-submission preparation process, and increase the programming efficiency.

REFERENCES

- [1] CDISC Define-XML Specification, Version 2.0. March 5, 2013. Available at <http://www.cdisc.org/define-xml>.
- [2] CDISC Case Report Tabulation Data Definition Specification (define.xml), Version 1.0, February 9, 2005. Available at <http://www.cdisc.org/define-xml>.
- [3] FDA Technical Specifications Document: Study Data Technical Conformance Guide. Version 2.3, October, 2015. Available at <http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm>
- [4] FDA Guidance for Industry: Providing Regulatory Submissions in Electronic Format — Standardized Study Data. December 2014. Available at <http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm>
- [5] FDA Data Standards Catalog. Version 4.4, August 17, 2015. Available at <http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm>
- [6] Lex Jansen (2013). Define-XML v2 – What's New. Proceedings of the 9th Pharmaceutical Users Software Exchange (PhUSE 2013, Brussels, Belgium). <http://www.lexjansen.com/phuse/2013/cd/CD05.pdf>
- [7] Lin Yan (2014). Developing ADaM Specifications to Embrace Define-XML 2.0.0 Requirements. Proceedings of the Pharmaceutical Industry SAS® Users Group (Paper DS10, PharmaSUG 2014, San Diego, CA). <http://www.pharmasug.org/2014-proceedings.html>

ACKNOWLEDGMENTS

The author would like to thank Wayne Zeng and Minying Royston for collaborative effort to develop the ADaM specification template and define.xml V2.0 macro, and Todd Case for review the paper.

CONTACT INFORMATION

The author can be reached at:

Hang Pang
Vertex Pharmaceuticals Incorporated
50 Northern Ave.
Boston, MA 02210
hang_pang@vrtx.com or
hang_pang@yahoo.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.