PharmaSUG2016 - Paper PO22

Automatic Consistency Checking of Controlled Terminology among SDTM Datasets, Define.xml, and NCI/CDISC Controlled Terminology for FDA Submission

Min Chen, Alkermes Inc., Waltham, MA

Xiangchen (Bob) Cui, Alkermes Inc., Waltham, MA

ABSTRACT

In FDA electronic submission, the most current version of NCI/CDISC controlled terminology for SDTM variables was expected to be submitted in define-xml. With the large amount of controlled terminology and the occasional update of NCI/CDISC controlled terminology, the controlled terminology applied in the SDTM programming was often out-of-date, and subject to the OPENCDISC report. It is desirable to ensure the consistency among SDTM datasets, define.xml, and NCI/CDISC controlled terminology to achieve the technical accuracy.

In this paper we created a library of controlled terminology in SDTM specifications spreadsheet which contains both standard NCI/CDISC controlled terminology and sponsor-defined controlled terminology. A SAS® macro tool was applied to automate consistency-checking of controlled terminology between SDTM datasets and define.xml. The macro tool also checked the update of the NCI/CDISC controlled terminology, and if needed, automatically updated the library of controlled terminology accordingly. This Macro-based comprehensive approach can ensure consistency between SDTM datasets and define.xml, as well as between controlled terminology in define.xml and NCI/NDISC controlled terminology, for final FDA submission. The high quality of the submissions can be achieved in a cost-effective and efficient way.

INTRODUCTION

CDISC Controlled Terminology (CT) is a finite set of allowable value lists that are used in a clinical trial for data collection, analysis and submission.

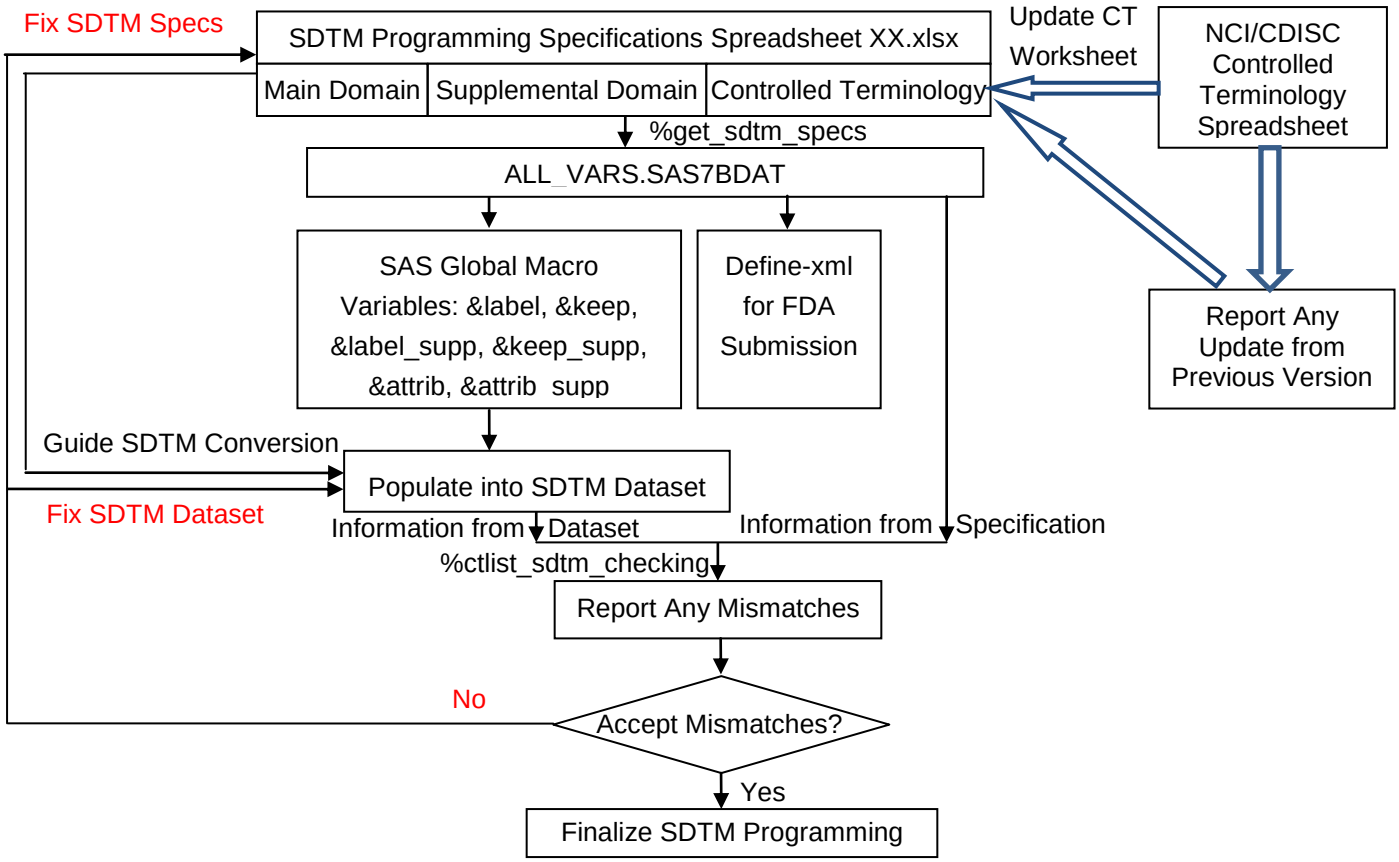
The FDA Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER) considered it important that all controlled terms submitted in datasets should be consistent with NCI/CDISC controlled terminology or external dictionary such as MedDRA, which builds study data standardization and facilitates the use and the development of automated review and analysis tools [1]. A consistent and single version of controlled terminology is specifically crucial to pooled analysis, such as Integrated Summaries of Safety (ISS) analysis, across the multiple studies as it 'ensures a consistent and coherent comparison of clinical and scientific concepts' across the studies [1]. Sponsor may also propose their own controlled terminology for variables with no standard controlled terminology or insufficient terminology defined in NCI/CDISC controlled terminology. However, any sponsor-defined controlled terminology or an extension of a standard dictionary should be well documented in define files: define-xml and SDTM Study Data Reviewer's Guide (SDRG).

It is very important to ensure that the define files are consistent with the corresponding SDTM datasets for FDA submissions since "the regulatory review team identified lack of consistency between the Define file and the data as a problem in many submissions" [2]. Many pharmaceuticals companies retrieve the controlled terminology directly from SDTM data into define-xml to ensure the consistency between define-xml and SDTM data, and this "data-driven" method for populating controlled terminology will lose the information for permissible values not presented in SDTM data, while CDISC SDTM Implementation Guide V3.2 [3] requires that "all values in the permissible value set for the study should be included, whether they are represented in the submitted data or not". In order to address the requirement, we will introduce a "metadata-driven" method used in our clinical trials, by which we created a controlled terminology library within SDTM programming specifications. A SAS Macro SAS Tool was developed for automatic checking of the consistency of controlled terminology between SDTM datasets and the specifications. Since the SDTM programming specifications are the unique sources to manage metadata and automatically generate define files, the consistency of controlled terminology between SDTM datasets and the programming specifications ensures consistency between SDTM datasets and define files.

Another common issue in FDA submission is the lack of consistency between submitted SDTM data and NCI/CDISC controlled terminology [1]. Usually NCI/CDISC controlled terminology will be released every 3 months. In each release, new controlled terminology may be added and sometimes old controlled terms could be updated as well. As the results, the customized controlled terminology library in the SDTM programming specifications could be out-of-date after the new release of NCI/CDISC controlled terminology, and trigger the warnings/errors in OPENCDISC report. Another SAS Macro Tool was also presented in this paper to automatically check the update of the NCI/CDISC controlled terminology, and if needed, automatically update the library of controlled terminology accordingly. The automation Tool can serve as a solution to this regulatory concern.

Since the automation of consistency checking is conducted from beginning of SDTM programming to end for FDA submission, the high quality of the submissions can be achieved in a cost-effective and efficient way.

Display 1 shows the process flow.



Display 1. Overview of Process Flow

AN INTRODUCTION OF NCI/CDISC CONTROLLED TERMINOLOGY

National Cancer Institute (NCI) **Enterprise Vocabulary Services** (EVS) maintains and distributes SDTM controlled terminology as part of NCI Thesaurus in Excel®, text, odm.xml, pdf, html, and OWL/RDF formats.

An example of NCI/CDISC Controlled Terminology Spreadsheet can be seen as followed, and CDISC Questionnaires, Ratings, and Scales (QRS) Terminology has been merged into SDTM terminology since 2015-12-18 release.

Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	CDISC Definition	NCI Preferred Term
C66742		No	No Yes Response	NY	No Yes Response	A term that is used to indicate a question with permissible values of yes/no/unknown/not applicable.	CDISC SDTM Yes No Unknown or Not Applicable Response Terminology
C49487	C66742		No Yes Response	N	No	The non-affirmative response to a question. (NCI)	No
C48660	C66742		No Yes Response	NA	NA; Not Applicable	Determination of a value is not relevant in the current context. (NCI)	Not Applicable
C17998	C66742		No Yes Response	U	U; Unknown	Not known, not observed, not recorded, or refused. (NCI)	Unknown
C49488	C66742		No Yes Response	Y	Yes	The affirmative response to a question. (NCI)	Yes
C101819		No	Hamilton Anxiety Rating Scale Clinical Classification Test Code	HAM-A TESTCD	Hamilton Anxiety Rating Scale Clinical Classification Test Code	Hamilton Anxiety Rating Scale test code.	CDISC Clinical Classification HAMA Test Code Terminology
C102073	C101819		Hamilton Anxiety Rating Scale Clinical Classification Test Code	HAMA101	HAMA1-Anxious Mood	Hamilton Anxiety Rating Scale - Anxious mood: worries, anticipation of the worst, fearful anticipation, irritability.	HAMA - Anxious Mood
C102074	C101819		Hamilton Anxiety Rating Scale Clinical Classification Test Code	HAMA102	HAMA1-Tension	Hamilton Anxiety Rating Scale - Tension: feelings of tension, fatigability, startle response, moved to tears easily, trembling, feelings of restlessness, inability to relax.	HAMA - Tension

We use this spreadsheet as the source of standard controlled terminology in our library, based on which a SAS data was generated as shown in Display 2. The SAS data contains the information exactly the same as the source spreadsheet, but for Questionnaire, Variable '**CODELIST**' is changed to QSTESTCD or QSTEST to follow the codelist names assigned in SDTM IG.

CODELIST_CODE	CODELIST	CTORDER	CODE	CODEVAL	DECODEVAL	Codelist_Name	Codelist_Extensible	CDISC_Submission_Value	CDISC_Synonym_s_	NCI_PREFERRED_TERM
C66742	NY	1	C49487	N		No Yes Response	No	N	No	No
C66742	NY	2	C48660	NA		No Yes Response	No	NA	NA; Not Applicable	Not Applicable
C66742	NY	3	C17998	U		No Yes Response	No	U	U; Unknown	Unknown
C66742	NY	4	C49488	Y		No Yes Response	No	Y	Yes	Yes
C101819	QSTESTCD	1	C102073	HAMA101	HAMA1-Anxious Mood	Hamilton Anxiety Rating Scale Clinical Classification Test Code	No	HAMA101	HAMA1-Anxious Mood	HAMA - Anxious Mood
C101819	QSTESTCD	2	C102074	HAMA102	HAMA1-Tension	Hamilton Anxiety Rating Scale Clinical Classification Test Code	No	HAMA102	HAMA1-Tension	HAMA - Tension

Display 2. A SAS Data for NCI/CDISC Controlled Terminology

AN INTRODUCTION OF CUSTOMIZED SDTM DOMAIN SPREADSHEET TEMPLATES

In our 'metadata-driven' method, the customized CDISC compliant SDTM specifications for individual domains were created as standard templates for our department to provide all the metadata-related information needed for SDTM programming and for Define-XML generating. The standard SDTM specifications consist of three worksheets: Main Domain Worksheet XX, Supplemental Domain Worksheet SUPPXX, and Controlled Terminology Worksheet XXCT, where XX represents the domain name.

The Main Domain Worksheet, as shown in Display 3 (a), was modified based on CDISC SDTM IG Metadata Excel Workbook sdtmv1.3_sdtmigv3.1.3_metadata.xls, and provided the dataset information and dataset variable information for both Main Domain Dataset and Supplemental Domain Dataset if any. Codelist name for each variable subject to the controlled terminology was assigned in the Column '**Controlled Terminology**'. An example is that codelist name **AESEV** is assigned in Column '**Controlled Terminology**' for variable **AESEV**, and codelist name **MedDRA** for variable **AEDECOD** in Display 3 (a).

The Supplemental Domain Worksheet provided the information of Value Level Metadata for QVAL per QNAM in supplemental domains, as shown in Display 3 (b). The value of QVAL per QNAM may also be subject to the controlled terminology, a codelist name will be written in the Column '**Controlled Terminology**' for such a QVAL. An example in Display 3 (b) shows that the value of QVAL is subject to codelist NY when QNAM=AEDESCYN.

Controlled Terminology Worksheet was a library designed for each individual domain for all the variables or variable values in SUPPXX.QVAL with controlled terminologies, as shown in Display 3 (c). Specially, for controlled terminology from variable values, we populated Column **Variable** as 'QVAL.' concatenating with the value of QNAM. For example, SUPPAE.QVL was subject to external codelist NY when QNAM=AEDESCYN. In Controlled Terminology Worksheet, we populated **Variable** Column as **QVAL.AEDESCYN**.

These worksheets were well-designed for the automation purpose.

Selected	Class	Domain	Variable	Label	Key	Type	Length	Controlled Terminology	Origin	Core	Comment	Form.Item
Y	Events	AE	STUDYID	Study Identifier	1	Char	20		CRF Page 1	Req		TrialNo
D	Events	AE	DOMAIN	Domain Abbreviation		Char	2	DOMAIN	Assigned	Req	AE	Constant: 'AE'
D	Events	AE	USUBJID	Unique Subject Identifier	2	Char	40		Derived	Req	STUDYID+ '-' + SUBJID	TrialNo PatientNo
D	Events	AE	AESEQ	Sequence Number		Num	8		Derived	Req	Sort by STUDYID, USUBJID, AETERM, AESTDTC then assign value. Start at 1 for each subject. No duplicates allowed within a subject within a domain.	Sequence
			…									
D	Events	AE	AEDECOD	Dictionary-Derived Term	3	Char	100	MedDRA	CRF Page 9	Req	MedDRA dictionary assigned.	AE.AEDECOD
Y	Events	AE	AESEV	Severity/Intensity		Char	8	**AESEV**	CRF Page 8	Perm		AE.AESEV

(a) Design of Main Domain Worksheet – Controlled Terminology for Dataset Variables

Selected	Domain	QNAM	QLABEL	Controlled Terminology	Origin	Comment	Form.Item
Y	SUPPAE	VMEDDRA	MedDRA Version Number		Assigned	MedDRA dictionary assigned.	
Y	SUPPAE	AEDESC1	Subject Description of Adverse Event		CRF Page 4		AE.AEDESC11
Y	SUPPAE	AEDESCYN	Any comments-Invest. Or Medical review?	**NY**	CRF Page 4		AE.AEDESCYN

(b) Design of Supplemental Domain Worksheet – Controlled Terminology for Value Level Metadata

Controlled Terminology Selection

Controlled Terminology Section – Enumerated Items

Dataset Structure, Not CT

Controlled Terminology Section – Codelist

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST
Y	DOMAIN	DOMAIN	1	One record per adverse event per subject	AE	Adverse Events
Y	RDOMAIN	DOMAIN	1		AE	Adverse Events
N	AECAT	AECAT	2	ADVERSE EVENTS		
N	AECAT	AECAT	1	BASELINE SIGNS AND SYMPTOMS		
Y	AESEV	AESEV	1	MILD		
Y	AESEV	AESEV	2	MODERATE		
Y	AESEV	AESEV	3	SEVERE		
	…					
N	QVAL.AEDESCYN	NY	4	A		
Y	QVAL.AEDESCYN	NY	1	N		
N	QVAL.AEDESCYN	NY	3	U		
Y	QVAL.AEDESCYN	NY	2	Y		

(c) Design of Controlled Terminology Worksheet - Codelist

Display 3. SDTM Specification Designed for Define-XML V2.0

Design of Controlled Terminology Worksheet by Types of Controlled Terminology

In CDISC SDTM Implementation Guide (IG) V3.2, the controlled terminology was represented in the Column 'Controlled Terms, Codelist or Format' by three categories as shown in Display 4 [3]:

1. An external codelist from CDISC/NCI codelist or MedDRA/WHODRUG dictionary, indicated by codelist name with a pair of parentheses,

2. Sponsor-defined controlled terminology or controlled terminology with no specific terms available at the current time, indicated by a single asterisk (*), and

3. A list of controlled terms directly defined in the Column 'Controlled Terms' instead of defined externally.

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
LBTEST	Lab Test or Examination Name	Char	(LBTEST)	Synonym Qualifier	Verbatim name of the test or examination used to obtain the measurement or finding. Note any test normally performed by a clinical laboratory is considered a lab test. The value in LBTEST cannot be longer than 40 characters. Examples: Alanine Aminotransferase, Lactate Dehydrogenase.	Req
LBCAT	Category for Lab Test	Char	*	Grouping Qualifier	Used to define a category of related records across subjects. Examples: such as HEMATOLOGY, URINALYSIS, CHEMISTRY.	Exp

MHENRTPT	End Relative to Reference Time Point	Char	BEFORE, AFTER, COINCIDENT, ONGOING, U	Timing	Identifies the end of the event as being before or after the reference time point defined by variable MHENTPT.	Perm

Display 4. Controlled Terminology Classification in CDISC SDTM IG V3.2

This kind of classification provides the source of codelist populated in our Controlled Terminology Worksheet. We used NCI/CDISC CT spreadsheet shown in Display 4 as a centralized standard template. For external codelists from NCI CT, the set of allowable value list values were retrieved for each variable subject to controlled terminology, the allowable values for internal codelists were built from SDTM IG, and the sponsor-defined value list values were created manually across multiple projects per project need as a template, which are subject to extension for new values collected in new studies.

In NCI/CDISC controlled terminology spreadsheet on 2015-12-18, there are totally 16876 standard controlled terms for 208 "codelists", many of which were never used in the projects as NCI CT is a library for all clinical trials. Together with the sponsor-defined controlled terminology, the number of the whole standard controlled terminology for the department will be very large. It will be of a great pain to check all the controlled terms in a study in one centralized customized controlled terminology spreadsheet, especially when SDTM datasets were done by multiple programmers. That is the reason that we created controlled terminology worksheet for individual domains, which greatly simplifies the selection of the controlled terms and avoids the unexpected selection of controlled terms simultaneously by multiple programmers.

For an external codelist from CDISC/NCI codelist spreadsheet, the codelist name assigned in Column '**Controlled Terminology**' at Main Domain worksheet must be consistent with that in CDISC/NCI codelist spreadsheet. For MedDRA/WHODRUG dictionary, the codelist name is the assigned as 'MedDRA' and 'WHODD', respectively. For sponsor-defined CT and internally defined CT, we assign a sponsor-defined codelist name which is consistent across multiple studies.

In define-xml v2.0, there are three types of controlled terminology in format.

1. Codelist for Code-Decode Variables

A typical example is a pair of variables –TESTCD and –TEST in SDTM findings domain. Both variables are subject to controlled terminology and have 1:1 mapping of test code and test name as shown in Display 5 (a).

For this kind of controlled terminology, the controlled terminology worksheet will be designed to list all possible code values in Column '**TESTCD**' and the corresponding decode values in Column '**TEST**'. Column '**Controlled Term**' will be left missing.

VSTEST [CL.VSTEST]

Permitted Value (Code)
Weight [*C25208*]
Height [*C25347*]
Heart Rate [*C49677*]
Systolic Blood Pressure [*C25298*]
Diastolic Blood Pressure [*C25299*]
Respiratory Rate [*C49678*]
Temperature [*C25206*]
Body Mass Index [*C16358*]

VSTESTCD [CL.VSTESTCD]

Permitted Value (Code)	Display Value (Decode)
WEIGHT [*C25208*]	Weight
HEIGHT [*C25347*]	Height
HR [*C49677*]	Heart Rate
SYSBP [*C25298*]	Systolic Blood Pressure
DIABP [*C25299*]	Diastolic Blood Pressure
RESP [*C49678*]	Respiratory Rate
TEMP [*C25206*]	Temperature
BMI [*C16358*]	Body Mass Index

(a) define-xml for controlled terminology of VSTESTCD-VSTEST with 1:1 mapping

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST
Y	VSTESTCD	VSTEST	1		WEIGHT	Weight
Y	VSTESTCD	VSTEST	2		HEIGHT	Height
Y	VSTESTCD	VSTEST	3		HR	Heart Rate
N	VSTESTCD	VSTEST	4		PULSE	Pulse Rate
Y	VSTESTCD	VSTEST	5		SYSBP	Systolic Blood Pressure
Y	VSTESTCD	VSTEST	6		DIABP	Diastolic Blood Pressure
Y	VSTESTCD	VSTEST	7		RESP	Respiratory Rate
Y	VSTESTCD	VSTEST	8		TEMP	Temperature
Y	VSTESTCD	VSTEST	9		BMI	Body Mass Index

(b) An Example of Codelist for Controlled Terminology of VSTESTCD-VSTEST Pair in SDTM.VS

Display 5. An Example of Codelist for Code-Decode Variables

2. Enumerated Codelist

Enumerated Codelist defines all the possible code values for a character variable. **Codelist LBCAT** is a typical 'enumerated item' which provided all the possible value of **Variable LBCAT** as shown in Display 6 (a)**.** For this kind of controlled terminology, each code value is populated in Column '**Controlled Terminology'** of CT Worksheet as shown in Display 6 (b)**. Columns 'TESTCD'** and **'TEST'** will be blank. **Please note:** Column **Controlled Terminology** for **Codelist DOMAIN** of Variable **DOMAIN** is specially designed for dataset structure, and will not be handled as enumerated item**.**

LBCAT [CL.LBCAT]

Permitted Value (Code)
HEMATOLOGY
CHEMISTRY
URINALYSIS
ALCOHOL BREATH TEST

(a) define-xml for controlled terminology of Enumerated Codelist LBCAT

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST
Y	LBCAT	LBCAT	1	HEMATOLOGY		
Y	LBCAT	LBCAT	2	CHEMISTRY		
Y	LBCAT	LBCAT	3	URINALYSIS		
N	LBCAT	LBCAT	5	SEROLOGY		

(b) An Example of Enumerated Codelist LBCAT for variable LBCAT in SDTM.LB

Display 6. An Example of Enumerated Codelist for LBCAT

3. External Codelist: MedDRA and WHODD

The sponsor is expected to provide a subsection for external code list references in define-xml, like dictionary name and version, to be used to map the terms. The MedDRA or WHO Drug dictionaries name will be provided as Global Macro Variables &MedDRA and &WHODRUG in programming setup. In Main Domain Worksheet, if any variable with Controlled Terminology MedDRA or WHO Drug is selected, the external published source, MedDRA dictionaries, will be shown in define-xml as shown in Display 7. We do not need to fill in our CT Worksheet in SDTM specifications for dictionary version.

External Dictionaries

Reference Name	External Dictionary	Dictionary Version
Adverse Event Dictionary (CL.MedDRA)	MedDRA	18.0

(a) define-xml for External Codelist MedDRA

Selected	Domain	Variable	Label	Key	Type	Length	Controlled Terminology
D	AE	AEDECOD	Dictionary-Derived Term		Char	100	MedDRA

(b) Assign Codelist Name in Main Domain Worksheet Only

Display 7. An Example of External CodeList, MedDRA Dictionary

With the design of our SDTM programming specification template, for a specific study, programmers only need to select/unselect the study-specific variables in Main Domain Worksheet, select/unselect QNAM in Supplemental Domain Worksheet, and select/unselect the controlled terminology defined in eCRF for an individual domain.

A MACRO TOOL FOR AUTOMATIC CONSISTENCY CHECKING OF CONTROLLED TERMINOLOGY BETWEEN SDTM DATASETS AND PROGRAMMING SPECIFICATION

Once study-specific SDTM programming specifications is finalized based on the standard SDTM specification templates, a macro **%get_sdtm_specs** is developed to read the individual SDTM programming specification, and automatically retrieves dataset information, dataset variable information, value level metadata information, and controlled terminology information from the specs. The controlled terms in the SDTM CT Worksheet will merge with that in NCI/CDISC CT spreadsheet by **CODELIST**, and get the standard NCI CT information in variables **Codelist_Code, Codelist_Name, ,** and **Code.** Variable **Codelist_Code** provides the unique code for each codelist name, **Codelist_Name** the description of the codelist name, **Codelist_Extensible** the extensibility of the codelist, and **Code** the unique code for each value list. All information was stored into SAS datasets, with variable information in data **XX_VARS**, QNAM-QVAL information in data **SUPPXX_QNAM** if supplemental domain exists, and controlled terminology information in **XX_CTS**, as shown in Display 7. SAS datasets, named as **ALL_VARS, SUPPQUAL,** and **ALL_CTS**, will also be generated cumulatively each time to incorporate all the variable information from all existed domains when individual SDTM specification programs were run. The SAS data **XX_CTS,** or **ALL_CTS** was used for consistency checking purposes.

DOMAIN	VARSEQ	VARIABLE	LABEL	TYPE	LENGTH	DISPLAYFORMAT	CODELIST	CORE	COMMENT
AE	1	STUDYID	Study Identifier	Char	20			Req	
AE	2	DOMAIN	Domain Abbreviation	Char	2		DOMAIN	Req	AE
AE	3	USUBJID	Unique Subject Identifier	Char	40			Req	TRNO+ '-' + PatientNo
AE	4	AESEQ	Sequence Number	Num	8			Req	Sort by STUDYID, USUBJID, AESTDTC, AETERM then assign value. Start at 1 for each subject. No duplicates allowed within a subject within a domain.
AE	9	AEDECOD	Dictionary-Derived Term	Char	100		MedDRA	Req	MedDRA dictionary assigned.
AE	19	AESEV	Severity/Intensity	Char	8		AESEV	Perm	
AE	20	AESER	Serious Event	Char	1		NY	Exp	
AE	21	AEACN	Action Taken with Study Treatment	Char	40		ACN	Exp	

(a) A SAS Data for Variable Information – AE_VARS, or Part of ALL_VARS

DOMAIN	VARIABLE	QNAM	QLABEL	CODELIST	ORIGIN	COMMENT
SUPPAE	QVAL	AEDESC1	Subject Description of Adverse Event		CRF Page 4	
SUPPAE	QVAL	AEDESCYN	Any comments-Invest. Or Medical review?	NY	CRF Page 4	

(b) A SAS Data for QVAL-QNAM Information – SUPPAE_QNAM, or Part of SUPPQUAL

DOMAIN	VARIABLE	VARIABLE VALUE	ORDER	DATA TYPE	CODELIST_CODE	CODE LIST	CODELIST_NAME	CODELIST_EXTENSIBLE	CODE	CODEVAL	DECODEVAL
AE	AEACN		2	text	C66767	ACN	Action Taken with Study Treatment	No	C49504	DOSE NOT CHANGED	
AE	AEACN		4	text	C66767	ACN	Action Taken with Study Treatment	No	C49501	DRUG INTERRUPTED	

AE	AEACN		5	text	C66767	ACN	Action Taken with Study Treatment	No	C49502	DRUG WITHDRAWN	
AE	AEACN		6	text	C66767	ACN	Action Taken with Study Treatment	No	C48660	NOT APPLICABLE	
AE	AESER		1	text	C66742	NY	No Yes Response	No	C49487	N	No
AE	AESER		2	text	C66742	NY	No Yes Response	No	C49488	Y	Yes
SUPPAE	QVAL	AEDESCYN	1	text	C66742	NY	No Yes Response	No	C49487	N	No
SUPPAE	QVAL	AEDESCYN	2	text	C66742	NY	No Yes Response	No	C49488	Y	Yes

(c) A SAS Data for Controlled Terminology Information – AE_CTS, or Part of ALL_CTS

Display 7. An Example of SAS datasets to store the information in individual domains

A SAS macro tool **%ctlist_checking _sdtm** was developed to check the proper use of Controlled Terminology to ensure the submission quality. It can be performed at any stage of the programming cycle in order to facilitate finalizing SDTM programming specifications earlier. The macro compares the controlled terminology and QNAM-QLABEL pair assigned in the SDTM Programming Specifications with ones in the SDTM datasets, detects any mismatches, and generates inconsistency report in RTF format if any exists.

Consistency checking for controlled terminology can be performed during the development of individual SDTM dataset by assigning an SDTM dataset name to domain name **XX**. In the final run all SDTM domains will be checked for consistency of the controlled terminology by selecting macro variable &domain as **_ALL_**.

Please note:

1. The comparison of controlled terminology will ONLY be performed for the variables or variable values assigned with codelist name. Therefore, the macro reads SAS datasets, **XX_VARS (or ALL_VARS)** and **SUPPXX (or SUPPQUAL)**, to select the target domain, target variables, and target variable values for consistency checking.

2. Retrieve codelists from SDTM datasets ONLY if they are assigned CT in the programming specifications.

3. The macro assumes that the controlled terminology is correctly assigned in SDTM specifications templates. The consistency of between the CDISC/NCI controlled terminology and that in SDTM specifications Controlled Terminology Worksheet will be guaranteed by another macro tool which will be discussed in the next section.

4. Compare the code lists from the SDTM datasets with ones from the programming specifications, detect the mismatches, and output non-consistency reports.

Display 8 shows two intermediate SAS datasets for Enumerated Codelists, one from SDTM specifications and another one from SDTM datasets.

DOMAIN	VARSEQ	VARIABLE	CODELIST	ORDER	CODEVAL	DECODEVAL	VAR_LABEL
VS	11	VSSTRESU	VSRESU	6	beats/min		Standard Units
VS	11	VSSTRESU	VSRESU	7	breaths/min		Standard Units
VS	11	VSSTRESU	VSRESU	8	cm		Standard Units
VS	11	VSSTRESU	VSRESU	1	kg		Standard Units
VS	11	VSSTRESU	VSRESU	11	kg/m2		Standard Units
VS	11	VSSTRESU	VSRESU	3	mmHg		Standard Units

(a) A SAS Data for Enumerated Controlled Terminology Information – Codelist from All SDTM Specifications

DOMAIN	VARIABLE	VAR_LABEL	CODELIST	CODEVAL
VS	VSSTRESU	Standard Units	VSRESU	beats/min
VS	VSSTRESU	Standard Units	VSRESU	breaths/min
VS	VSSTRESU	Standard Units	VSRESU	cm
VS	VSSTRESU	Standard Units	VSRESU	kg
VS	VSSTRESU	Standard Units	VSRESU	kg/m2
VS	VSSTRESU	Standard Units	VSRESU	mmHg

(b) A SAS Data for Enumerated Controlled Terminology Information – Codelist from All SDTM Datasets

Display 8. Intermediate SAS datasets for Enumerated Codelists

Display 9 shows two intermediate SAS datasets for code-decode codelists, one from SDTM specifications and another one from SDTM datasets.

DOMAIN	VARSEQ	VARIABLE	CODELIST	ORDER	VAR_LABEL	CODEVAL	DECODEVAL
LB	6	LBTESTCD	LBTESTCD	231	Lab Test or Examination Short Name	KETONES	Ketones
LB	6	LBTESTCD	LBTESTCD	235	Lab Test or Examination Short Name	LDH	Lactate Dehydrogenase
LB	6	LBTESTCD	LBTESTCD	236	Lab Test or Examination Short Name	LDL	LDL Cholesterol

(a) A SAS Data for Code-Decode CT Information – Codelist from All SDTM Specifications

DOMAIN	VARIABLE	VAR_LABEL	CODEVAL	CODELIST
LB	LBTESTCD	Lab Test or Examination Short Name	LBALL	Labs Data
LB	LBTESTCD	Lab Test or Examination Short Name	LDH	Lactate Dehydrogenase
LB	LBTESTCD	Lab Test or Examination Short Name	LDL	LDL Cholesterol

(b) A SAS Data for Code-Decode Controlled Terminology Information – Codelist from All SDTM Datasets

Display 9. Intermediate SAS datasets for Code-Decode Codelists

Display 10 shows a typical report of non-consistency of enumerated codelists between SDTM datasets and specifications. Display 11 shows a typical report of non-consistency of code-decode codelists between SDTM datasets and specifications. Decision will be made to update either the programming specifications or the SDTM derivation program to handle these mismatches, which will be explained in the next section.

Domain	Variable	Variable Label	Codelist Name	Controlled codelist in Dataset	Controlled codelist in Specs.	codelists In Specs. NOT in Dataset	codelists In Dataset NOT In Specs.
DM	DTHFL	Subject Death Flag	NY		Y	Yes	
VS	VSORRESU	Original Units	VSRESU	beats/min			Yes
			VSRESU		BEATS/MIN	Yes	

Display 10. Non-Consistency Report of Enumerated Codelists between SDTM Datasets and Specifications

Domain	Variable	Variable Label	Code Value	Decode Value Label in Dataset	Decode Value Label in Specs.	Codelists In Specs. NOT in Dataset	Codelists In Dataset NOT In Specs.	Different Controlled Terms
EG	EGTESTCD	ECG Test or Examination Short Name	EGHRMN	ECG Mean Heart Rate	Summary (Mean) Heart Rate			Yes
LB	LBTESTCD	Lab Test or Examination Short Name	LBALL	Labs Data			Yes	
			BNZDZLO	Lorazepam				
			BNZDZLO		Lorazepan	Yes		

Display 11. Non-Consistency Report of Code-Decode Codelists between SDTM Datasets and Specifications

5. Report any variables with defined code list name in the Main Domain Worksheet, but with no value list selected in the Controlled Terminology Worksheet. The report can identify the omissions of specification for Controlled Terminology. The programmers should review the report and check these variables to make sure the possible value lists are selected in the Controlled Terminology Worksheet.

Display 12 shows a typical report for variables subject to controlled terminology but without any value list selected. The value of Controlled terminology for TI.IECAT should be selected in the Controlled Terminology Worksheet.

Domain	Order in Data	Variable	Variable Label	Codelist Name
TI	5	IECAT	Inclusion/Exclusion Category	IECAT

Display 12. Report of Variables subject to controlled terminology but without value list selected

6. Report any variables not selected in Main Domain Worksheet, but the value list is selected in the Controlled Terminology Worksheet. This report can identify the discrepancy between controlled terminology information in the Main Domain Worksheet and Controlled Terminology Worksheet. The programmers should review the report and check these variables to make sure whether they should be selected in the Main Domain Worksheet or the value of controlled terminology should not be selected in the Controlled Terminology Worksheet.

Display 13 shows a typical report when Codelist for variable AEHOSP was mistakenly selected in CT Worksheet, while variable AEHOSP was not collected.

Domain	Variable	Codelist Name	Code	Decode
AE	AEHOSP	NY	N	No
			Y	Yes

Display 13. Report of Variables not selected in the Main Domain Worksheet, but the value list is selected in the Controlled Terminology Worksheet

7. Report any variables subject to controlled terminology, but empty in SDTM datasets. The programmers should review the report and check these variables to make sure whether the data is correctly populated or it is data-driven warnings. Display 14 shows a typical report for this condition. CM.EPOCH should be populated per derivation rule defined in Main Domain Worksheet.

Domain	Variable	Variable Label	Codelist Name
CM	EPOCH	Epoch	EPOCH

Display 14. Report of An Empty Variable which are Subject to Controlled Terminology

DECISION MAKING ON THE MISMATCHES BETWEEN SDTM SPECS AND DATASETS

There are 5 scenarios of mismatches between SDTM datasets and specifications.

1. The Controlled Terms are not in the Datasets but in the Specifications.

SDTM IG required that the controlled terminology be included in the define.xml with all possible value set for the study, no matter whether they are in the submitted SDTM data or not. Therefore, those code lists correctly defined in the programming specifications but not shown in the SDTM datasets are acceptable, and no further action is needed for them. This scenario usually occurs when the information about these controlled terms are defined in the eCRF, but not collected in the raw data. An example can be seen in Display 10 for codelist **NY** for Variable **DM.DTHFL**.

2. The Controlled Terms are in the Datasets but not in the Specifications.

The specification does not list all the possible values for the controlled terms or value lists. This kind of mismatches is not acceptable, and selecting the missing controlled terms or value lists in the controlled terminology worksheet is the solution. This scenario usually occurs when the value set of controlled terminology is not checked in the controlled terminology worksheet. An example can be found in Display 11 for codelist **LBTESTCD**. Code value 'LBALL' is in SDTM.LB for variable LBTESTCD, but is not selected in CT Worksheet.

3. The Code Lists are Differently Defined in the Datasets from that in the Specifications.

This kind of mismatches is not acceptable, and the revision should be done either in the datasets or in the specifications to make them consistent. An example can be found in Display 10 for codelist **VSRESU**. The lower case was populated in VS data, while upper case in the specifications. In this case, we need to resort to NCI/CDISC controlled terminology spreadsheet as the industry standard.

4. The Decode Values for the Same Code Value are Differently Defined in the Datasets from that in the Specifications.

This kind of mismatches is not acceptable, and the revision should be done either in the datasets or in the specifications to make them consistent. An example can be found in Display 11 with different decode values when EG.EGTESTCD=EGHRMN. In this case, we need to resort to NCI/CDISC controlled terminology spreadsheet as the industry standard.

5. Typo Occurrence in SDTM Derivation Programs

Correct the typos. In Display 11, LBTEST is mistakenly spelled as 'Lorazepan' for LBTESTCD=BNZDZLO in the LB Controlled Terminology Worksheet.

A summary of these 5 scenarios is shown in Table 1.

#	Scenario	Condition	Action Taken
1	Controlled Terms are not in the Datasets but in the Specifications	Code lists are correctly defined in specifications	No Action Needed
2	Controlled Terms are in the Datasets but not in the Specifications	Specification does not select all the possible values for the controlled terms	Select Missing Controlled Terms in CT Worksheet
3	The Code Lists are Differently Defined in the Datasets from that in the Specifications	Code Values in datasets is not consistent with CT Worksheet	Check and Follow NCI/CDISC CT spreadsheet or standard sponsor-defined codelist
4	The Decode Value Lists are Differently Defined in the Datasets from that in the Specifications for Same Code Value	Decode Values in datasets is not consistent with CT Worksheet	Check and Follow NCI/CDISC CT spreadsheet or standard sponsor-defined codelist
5	Typo Occurrence in SDTM Derivation Programs		Revise SDTM Datasets or CT Worksheet to correct the typo

Table 1. Summary of 5 Scenarios of Mismatches between SDTM Datasets and Specifications

In our metadata-driven SDTM programming process, define-xml was wholly generated by and completely consistent with SDTM programming specifications [4]. Our macro-based SAS tool ensures the consistency of controlled terminology between SDTM datasets and SDTM programming specifications, therefore further ensures consistency between SDTM datasets and define files.

AUTOMATICALLY CHECK THE UPDATE OF NCI/CDISC CONTROLLED TERMINOLOGY

NCI frequently updated NCI/CDISC controlled terminology per users' need. Usually we got the release at least every 3 months. Although our macro-based SAS tool can ensure the consistent of Controlled Terminology in SDTM datasets and define-xml, we cannot guarantee our customized controlled terminology worksheet reflected the update of NCI/CDISC CT in timely manner. For example, compared with 2015-09-25 release of NCI/CDISC SDTM CT, there were 427 new codelists added in 2015-12-18 release, 13 old codelists deleted, and 120 codelists updated. Manually update of the all individual customized Controlled Terminology Worksheet for multiple studies not only a labor-intensive effort, but also an error-prone process. It is highly desirable to handle this process by automation. A SAS macro tool was developed for this need to automatically check the update of NCI/CDISC Controlled Terminology, and if needed automatically update the standard customized CT Worksheet in each study.

A macro **%chk_nci** was call to automatically check the update from the earlier version each time when we got a new release of NCI/CDISC CT spreadsheet. We consider a controlled terminology a newly added one if Variable **CODELIST_CODE** and **CODE** were new in the new version. If both Variable **CODELIST_CODE** and **CODE** were the same as in the old version, the update of either CODELIST Name, code values, or decode values will be considered as 'Update' of a controlled terminology. Three datasets will be output as the summary of the CT update as shown in Display 15. In Display 15 (c), the flags **UPDATE_CODELIST_NAME, UPDATE_CODE_VAL, and UPDATE_DECODE_VAL** show the part updated in the new version.

CODELIST_CODE	CODE	CODELIST_NEW	CODEVAL_NEW	DECODEVAL_NEW
C71620	C124463	UNIT	uIU/dL	
C71620	C124464	UNIT	uIU/L	
C103483	C124782	QSTESTCD	PHQ0216	PHQ02-Total Score
C118971	C102118	**CCCAT**	HAM-A	
C118971	C102119	CCCAT	HAMD 21	
C118971	C117997	CCCAT	ESRSA	

(a) A SAS Data, CODE_IN_NEW_ONLY, for Newly Added NCI/CDISC Controlled Terminology in New Version

CODELIST_CODE	CODE	CODELIST_OLD	CODEVAL_OLD	DECODEVAL_OLD
C100129	C102118	**QSCAT**	HAM-A	
C100129	C102119	QSCAT	HAMD 21	
C100129	C117997	QSCAT	ESRSA	
C117738	C117747	HEFATS	West Haven Hepatic Encephalopathy Grade	
C117739	C117747	HEFATSCD	WHHEGR	West Haven Hepatic Encephalopathy Grade
C71620	C103452	UNIT	/mL	

(b) A SAS Data, CODE_IN_OLD_ONLY, for Deleted NCI/CDISC Controlled Terminology from Old Version

CODELIST_CODE	CODE	CODELIST_NEW	CODELIST_OLD	CODEVAL_NEW	CODEVAL_OLD	DECODEVAL_NEW	DECODEVAL_OLD	UPDATE_CODELIST_NAME	UPDATE_CODE_VAL	UPDATE_DECODE_VAL
C96781	C103420	ONCRTS	RSTEST	New Lesion Progression	New Lesion Progression			Y		
C96781	C123619	ONCRTS	RSTEST	Clinical Response	Clinical Response			Y		
C96781	C123620	ONCRTS	RSTEST	Cytogenetic Response	Cytogenetic Response			Y		
C96781	C123622	ONCRTS	RSTEST	Hematologic Response	Hematologic Response			Y		
C100129	C102120	QSCAT	QSCAT	SF36 V1.0 ACUTE	SF36 v1.0 ACUTE				Y	
C100129	C102121	QSCAT	QSCAT	SF36 V2.0 ACUTE	SF36 v2.0 ACUTE				Y	
C100129	C103525	QSCAT	QSCAT	ODI V2.1A	ODI v2.1A				Y	
C65047	C100425	LBTESTCD	LBTESTCD	HDLCLDLC	HDLCLDLC	HDL Cholesterol/LDL Cholesterol	HDL Cholesterol/LDL Cholesterol Ratio			Y
C65047	C92271	LBTESTCD	LBTESTCD	HAIGMAB	HAABIGM	Hepatitis A Virus IgM Antibody	Hepatitis A Virus Antibody IgM		Y	Y

(c) A SAS Data, CODE_UPDATE, for Updated NCI/CDISC Controlled Terminology

Display 15. Report of NCI/CDISC Controlled Terminology Update

These datasets specify the update of each NCI/CDISC Controlled Terminology, and will be output for automation in next step. The code can be shown as below, where &sdtmdt1 is the date for new release:

```
data out.code_in_new_only_&sdtmdt1.(keep=codelist_code code codelist_new codeval_new decodeval_new)
     out.code_in_old_only_&sdtmdt1.(keep=codelist_code code codelist_old codeval_old decodeval_old)
     out.code_update_&sdtmdt1.;
     retain CODELIST_CODE CODE CODELIST_NEW CODELIST_OLD CODEVAL_NEW CODEVAL_OLD DECODEVAL_NEW
            DECODEVAL_OLD Update_Codelist_Name Update_Code_Val Update_Decode_Val;
     merge codelist_new(in=a keep=codelist_code code codelist codeval decodeval
                        rename=(codelist=codelist_new codeval=codeval_new decodeval=decodeval_new))
           codelist_old(in=b keep=codelist_code code codelist codeval decodeval
                        rename=(codelist=codelist_old codeval=codeval_old decodeval=decodeval_old));
     by codelist_code code;
     if codelist_new ne codelist_old and not missing(codelist_new) and not missing(codelist_old)
     then update_codelist_name = 'Y';
     else update_codelist_name = '';
     if codeval_new ne codeval_old and not missing(codeval_new) and not missing(codeval_old)
     then update_code_val = 'Y';
     else update_code_val = '';
     if decodeval_new ne decodeval_old and not missing(decodeval_new) and not missing(decodeval_old)
     then update_decode_val = 'Y';
     else update_decode_val = '';

     if a and not b then output out.code_in_new_only_&sdtmdt1.;
     if b and not a then output out.code_in_old_only_&sdtmdt1.;
     if a and b and (update_codelist_name = 'Y' or update_code_val = 'Y' or update_decode_val = 'Y')
     then output out.code_update_&sdtmdt1.;
  run;
```

AUTOMATICALLY UPDATE THE CUSTOMIZED CONTROLLED TERMINOLOGY TEMPLATES

A SAS macro tool **%update_sdtm_cts** was developed to automatically update the customized Controlled Terminology Worksheet of Individual SDTM specifications. Here is the logic to perform the update:

1. The update of controlled terminology will ONLY be done for newly added codelists, deleted codelists, and the updated codelists in the new release. Therefore, the macro reads SAS datasets, **CODE_IN_NEW_ONLY, CODE_IN_OLD_ONLY** and **CODE_UPDATE**, to select the target codelist name, code values, and decode values.

2. Update the Individual Controlled Terminology Worksheet only if the variables in Main Domain Worksheet contain the codelist name in updated NCI/CDISC controlled terminology.

3. Append newly added code list into the Controlled Terminology Worksheet. A flag **Status** will be output to label the status of the codelist update as '**Newly Added**'. Display 16 shows how controlled terminology worksheet was updated for newly added codelist.

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST
N	LBSTRESU	UNIT	167	IU		
Y	LBSTRESU	UNIT	416	mIU/L		

(a) QS Controlled Terminology Worksheet before Update

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST	Status
N	LBSTRESU	UNIT	167	IU			
Y	LBSTRESU	UNIT	416	mIU/L			
	…						
N	LBSTRESU	UNIT	603	uIU/dL			Newly Added
N	LBSTRESU	UNIT	604	uIU/L			Newly Added

(b) LB Controlled Terminology Worksheet after Update

Display 16. An example for Automatic Update of Controlled Terminology Worksheet for Newly Added Codelist

4. Keep old codelists in the Controlled Terminology Worksheet for review for deleted code lists. A flag **Status** will be output as '**Retired**'. Display 17 shows how controlled terminology worksheet was updated for deleted codelist.

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST
Y	LBORRESU	UNIT	20	/mL		

(a) LB Controlled Terminology Worksheet before Update

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST	Status
Y	LBORRESU	UNIT	20	/mL			Retired

(b) LB Controlled Terminology Worksheet after Update

Display 17. An example for Automatic Update of Controlled Terminology Worksheet for Deleted Codelist

5. For updated code lists, keep the original codelist and add the updated codelists in the Controlled Terminology Worksheet, for the codelist before update, a flag **Status** will be output as 'Updated – Old', and for the codelist after update, the flag variable will be labeled as 'Updated – New'. Display 18 shows how controlled terminology worksheet was updated for updated codelist.

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST
Y	LBTESTCD	LBTESTCD	46		HDLCLDLC	HDL Cholesterol/LDL Cholesterol Ratio
N	LBTESTCD	LBTESTCD	49		HAABIGM	Hepatitis A Virus Antibody IgM

(a) LB Controlled Terminology Worksheet before Update

Selected	Variable	Codelist	Order	Controlled Term	TESTCD	TEST	Status
Y	LBTESTCD	LBTESTCD	46		HDLCLDLC	HDL Cholesterol/LDL Cholesterol Ratio	Updated-Old
N	LBTESTCD	LBTESTCD	46		HDLCLDLC	HDL Cholesterol/LDL Cholesterol	Updated-New
N	LBTESTCD	LBTESTCD	49		HAABIGM	Hepatitis A Virus Antibody IgM	Updated-Old
N	LBTESTCD	LBTESTCD	49		HAIGMAB	Hepatitis A Virus IgM Antibody	Updated-New

(b) LB Controlled Terminology Worksheet after Update

Display 18. An example for Automatic Update of Controlled Terminology Worksheet for Updated Codelist

6. Programmer will review the updated Controlled Terminology Worksheet and decide whether to select the newly added controlled terms, unselect the deleted controlled terms, or update the updated controlled terms or not based on the study need. Usually for a completed study, no update will be done, and for a new study, new controlled terminology will be applied. For an ongoing study, if the code list has already been collected in eCRF, the 'deleted' controlled terminology should not be removed from the Controlled Terminology Worksheet for collected variables, otherwise we prefer to update the controlled terminology.

CONCLUSION

In summary, this paper introduced a specially designed SDTM Domain Spreadsheet Templates with Controlled Terminology Worksheet defined for each individual variable, based on which SAS macros were developed for automatic consistency checking of the controlled terminology among SDTM datasets, programming specification, define.xml, and NCI/CDISC controlled terminology. It illustrated five scenarios of mismatches the macro detects, and also provided innovative solutions for mismatches, which was another level of validation for SDTM programming. The automatic update of the controlled terminology worksheet based on NCI/CDISC controlled terminology greatly spared the time and energy to maintain the Controlled Terminology Worksheet of SDTM Programing Specifications Templates.

This macro-based comprehensive approach can ensure consistency among SDTM datasets, define.xml, and most updated NCI/CDISC Controlled Terminology for final FDA submission, and ensure the high quality of the delivery. Since it can be used at any stage of the programming cycle, the high quality of the submissions can be achieved in a cost-effective and efficient way. We hope this approach can assist you in handling SDTM controlled terminology in order to enhance the submission quality.

REFERENCES

1. "CDER Common Data Standards Issues Document". December 2011. http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM254113.pdf

2. "CDISC SDTM/ADaM Pilot Project, Project Report"- http://www.cdisc.org/stuff/contentmgr/files/0/df91a087c6df43275288267c9fe92180/misc/sdtmadampilotprojectreport.pdf

3. CDISC Submission Data Standards Team. "Study Data Tabulation Model Implementation Guide: Human Clinical Trials". November 2013. http://www.cdisc.org/sdtm

4. Min Chen, Xiangchen (Bob) Cui. "A SAS® Macro Tool to Automate Generation of Define.xml V2.0 from SDTM Specification for FDA Submission", PharmaSUG, May 2016.

CONTACT INFORMATION

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

Name: Min Chen**,** Ph.D.
Enterprise: Alkermes, Inc.
Address: 852 Winter Street
City, State ZIP: Waltham, MA 02451
Work Phone: 781-609-6047
Fax: 781-609-5855

E-mail: min.chen@alkermes.com

Name: Xiangchen (Bob) Cui**,** Ph.D.
Enterprise: Alkermes, Inc.
Address: 852 Winter Street,
City, State ZIP: Waltham, MA 02451
Work Phone: 781-609-6038
Fax: 781-609-5855
E-mail:xiangchen.cui@alkermes.com