PhUSE US Connect 2019

Paper PP19

Taking Advantage of CDISC Standards and Planning the Study Specifications

Siddharth Kumar Lokineni, Syneos Health, Cary, USA
Vara Prasad Reddy Sakampally, Syneos Health, Cary, USA

ABSTRACT

Implementing the CDISC® standards is a challenging task and will impact study time lines, budget, and quality if they are not considered before setting up a new study for programming. Addressing the key compliance issues and considering the necessary files needed for submission before stepping into the programming phase will help in a smooth flow of the study and avoid rework of many items. This paper will describe a method to design the specifications template to aid in annotating the CRF, checking mapping specifications against CDISC data standards, in generating the supplemental data definition document files and generating the define.xml that go along with the submission packet. Also, it will provide additional information on performing the gap analysis and additional checks on mapping all the data components captured in a study.

INTRODUCTION

The CDISC data standards facilitate uniformity across the industry to represent data collected during a clinical study. With the fast paced deliverables and challenging timelines to deliver quality analysis to clients, it is very important to get dataset programming correct the first time. It is vital to devise an efficient way of designing the specifications to get accuracy and consistency in programming and corresponding output data for CDISC SDTM mapping and ADaM data derivation. This in turn will help in speeding up programming cycle, avoid budgeting issues and rework. A specification is traditionally a multi-tabbed EXCEL file, displaying one domain per sheet, one row for each variable, and spreadsheet columns mimicking the CDISC implementation guide. But here the idea is to generate a specification file in a single tab instead of multiple tabs. This way we can reuse it for CRF annotation, Define.XML easily. A well-defined and organized specification document minimizes the time needed to acclimatize with the metadata required when working on Define documentation during the later stages, which significantly reduces the overall time for review. Note that the specs are reviewed before the dataset programming is started. This ensures quality when metadata information is reused in a submission for regulatory review.

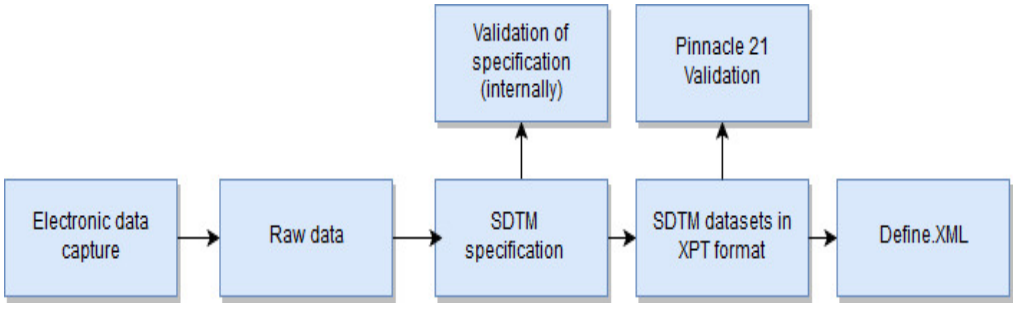

Figure 1. Representation of typical process cycle in SDTM development.

CDISC guidelines will be used to create these specifications. Although the paper begins with a brief description of the specifications structure, some prior knowledge of specification document and excel will be helpful.

SINGLE-TAB SPECIFICATION USES

A single tab spec has the following advantages.

- Variable Metadata update – Metadata has many defining properties about the variable. The information about the variable label, type, format/CT, origin, role, length, and format are these properties. Usually there are certain variables that are repeated across domains. When there is a need to update the variable attributes across domains, a single tab spec will be really useful. The user will be filtering the EXCEL spreadsheet by the variable to update the attributes.

  Example: In the below figure (2), the variable EPOCH is present in multiple domains. If an update is warranted to any of the attributes, having a single tab makes it much easier. In the below example, we can filter the third column to the variable EPOCH and make the change across domains.

  As shown below in the figure 2 – we can see the variable EPOCH is designed to be presented in the multiple datasets. For the EPOCH attributes to be updated in order to make them consistent across the domains, we will filter the third column (Variable name) and change the attributes. Changing would be easier as only the EPOCH variable will be shown as the filter condition is applied like below. This process will make sure that we are not missing on updating any domains where the variable EPOCH is present. In this example, the variable EPOCH is present in the AE, DA, and CM. This is a fool proof method of obtaining a consistent specification. This will be an extremely useful practice when we are working on pooling datasets as there will not be having issues with the differing attributes. One more attribute that can be sorted with using a single tab specification is the length of the labels. Per the CDISC standards, label length cannot be more than 40 characters. This style of specification will make it extremely easy to check the lengths of the labels. This will be useful when this document is used in the creation of a define.xml file.

Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Length	Format
AE	EPOCH	EPOCH	Epoch	Char		Protocol	20	$20.
CM	EPOCH	EPOCH	Epoch	Char		Protocol	20	$20.
DA	EPOCH	EPOCH	Epoch	Char		Protocol	20	$20.

Figure 2. SDTM specification filtered by the variable EPOCH.

- Preparing Define.XML during specification and dataset: It has often been observed that when a metadata discrepancy has been noticed in the define.xml/ define.pdf, the entire cycle of updating and rerunning the datasets process has to be performed. To avoid this, we can use the single tabbed spec file which will be very similar to the specification file that Pinnacle21 uses as input to generate the define-xml/pdf. In other words the proposal here is that the define documentation should be done during the SDTM/ADaM development, particularly while dataset mapping specification are created as shown in the figure 3. Generating define initially will subject it to multiple review cycles and also speeds up the entire process.

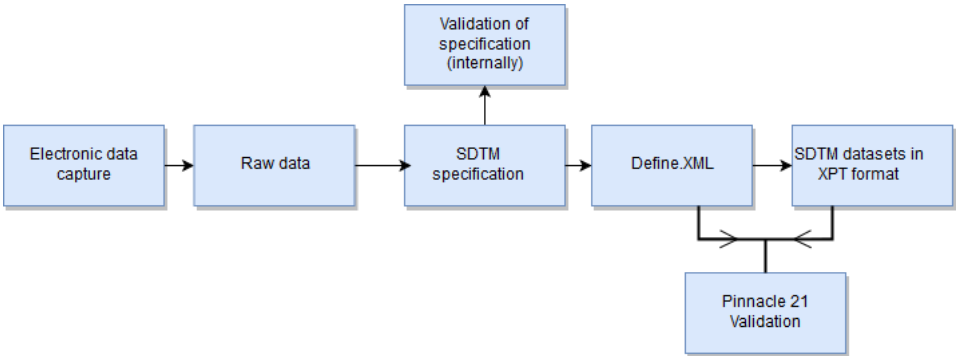

Figure 3. Proposed creation of the define documents at the time of generation of the specs and datasets.

Typically in the industry, define.xml is generated towards the final submission stage and assumes that it has to be done after the SDTM/ADaM development lifecycle. The figure 3 above demonstrates SDTM development and implementation life cycle. Here, the single tabbed specification is made use of and the define.xml is created right after the specifications are worked on unlike the traditional way. Once the SDTM datasets are developed and validated, the datasets will also be passed through the Pinnacle21 to check for the CDISC compliance.

- Ease of review for internal reviewers and clients: Having a single tabbed specification file will make the job of validators and reviewers simpler. It is much convenient to check the logical consistency among similar variables. This also eases the communication with the clients. If clients find any inconsistency – it is easy to point out all the variables and cross check the logic implemented in other datasets.

  Example: The –SDTY variable computational algorithm's update would be simple if the specification is filtered by the variable. Figure 4 depicts the updated mapping specification after the column D (highlighted) is filtered. It can be observed that there are multiple data domains in the column C.

B	C	D	E	F	G	I	J	K	L	M
Observation Class	Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Origin	Role	CDISC Notes (for domains) Description (for General Classes)	Core	Mapping Specifications
Events	AE	STDY	AESTDY	Study Day of Start of Adverse Event	Num	Derived	Timing	Study day of start of adverse event relative to the sponsor-defined RFSTDTC.	Perm	AE start date - reference start date (+ 1 day if AE start date >= reference start date)
Interventions	CM	STDY	CMSTDY	Study Day of Start of Medication	Num	Derived	Timing	Study day of start of medication relative to the sponsor-defined RFSTDTC.	Perm	CMSTDTC - RFSTDTC (+ 1 day if CMSTDTC >= RFSTDTC)
Events	DS	STDY	DSSTDY	Study Day of Start of Disposition Event	Num	Derived	Timing	Study day of start of event relative to the sponsor-defined RFSTDTC.	Perm	DSSTDTC - DM. RFSTDTC (+ 1 day if DSSTDTC>= DM. RFSTDTC)
Events	MH	STDY	MHSTDY	Study Day of Start of Observation	Num	Derived	Timing	Actual study day of start of observation expressed in integer days relative to the sponsor-defined RFSTDTC in	Perm	Medical history start date - reference start date (+ 1 day if medical history start date >= reference start date)
Special-Purpose	SV	STDY	SVSTDY	Study Day of Start of Visit	Num	Derived	Timing	Study day of start of visit relative to the sponsor-defined RFSTDTC.	Perm	SVSTDTC - RFSTDTC (+ 1 day if SVSTDTC >= RFSTDTC)

Figure 4: --STDY variables computational algorithm update is shown.

DESIGN OF THE SPECIFICATION

The spec will contain the following sheets, each serving a purpose.

1. Change log sheet: Usually it is observed that specifications are not tracked for the changes like the way the SAS programs are. The change log sheet will help us to document all the changes that are made to the specification during the course of the study. Below screenshot of the change log sheet has the following columns: Version, Date, Changes, Changed by, Reviewed by and Review Date. "Version" column documents the version of the spread sheet. "Date", "Changes" and "Changed by" columns signify the date on which the changes were made, what were the changes and who made the changes respectively. The last two columns "Reviewed by" and "Review date" indicate who was the reviewer and when was the review done.

A	B	C	D	E	F
Version:	Date:	Changes:	Changed by:	Reviewed by:	Review Date:
1.1	5-Dec-2018	Globally changes STUDYID to "XXXXXXXXXX" Modified AE.AESER, AE.AECONTRT, AE.AEACNOTH	Sid Lokineni	Prasad Sakampally	12/8/2018
	5-Dec-2018	QS: Updated VISITNUM, VISIT, QSBLFL, LB: LBBLFL DM: RFENDTC. Globally changes USUBJID. VS: VSSSTAT AE: AESCONG, AESDISAB, AESDTH, AESHOSP, AESLIFE	Sid Lokineni	Prasad Sakampally	12/8/2018
	9-Dec-2018	DM: updated RFXSTDTC and RFXENDTC. SUPPDM: added RANDNUM SUPPQS: Spec added. DS: EPOCH need to be added for SUBDISP records XA: Updated XASTRESC to decode the value per XAORRES	Sid Lokineni	Prasad Sakampally	12/10/2018

Figure 5: Change log documentation.

2. Datasets sheet: This sheet tells the datasets that are needed for the study. It has the following details presented– 78Dataset, Dataset label, Dataset class per CDISC guidelines, Dataset structure per CDISC guidelines and Key variables used in final sort. CRF pages will give information of the pertinent CRF page for the related information. Dependencies column will help to sort the dependent and independent datasets which can be worked on in appropriate hierarchy.

Dataset	Dataset Label	Dataset Class	Structure	Keys	Repeating	Reference Data?	CRF Pages	External Data	Dependencies
TS	Trial Summary Domain	Trial Design	One record per trial summary parameter val	TSPARMCD, TSSEQ, TSVAL	No	Yes			
TI	Trial Inclusion and Exclusion Criteria Domain	Trial Design	One record per I/E criterion	IETESTCD	No	Yes			
TE	Trial Elements Domain	Trial Design	One record per planned Element	ETCD	No	Yes			
TA	Trial Arms Domain	Trial Design	One record per planned Element per Arm	ARMCD, TAETORD	No	Yes			
TV	Trial Visits Domain	Trial Design	One record per planned Visit per Arm	VISITNUM, ARMCD	No	Yes			
DM	Demographics Domain	Special-Purpose	One record per subject	USUBJID	No	No	Demography, Randomization		
SUPPDM	Supplemental Qualifiers for DM	Relationship	One record per subject per IDVAR, IDVARVAL, and QNAM	USUBJID, IDVAR, IDVARVAL, QNAM	Yes	No			
SE	Subject Element Domain	Special-Purpose	One record per subject per actual Element	STUDYID, USUBJID, TAETORD, ETCD, SESTDTC	Yes	No			
SV	Subject Visits Domain	Special-Purpose	One record per subject per actual visit	USUBJID, SVSTDTC, VISITNUM	Yes	No	Date of Visit, Unscheduled		
SUPPSV	Supplemental Qualifiers for SV	Relationship	One record per subject per IDVAR, IDVARVAL, and QNAM	USUBJID, IDVAR, IDVARVAL, QNAM	Yes	No			
AE	Adverse Event Domain	Events	One record per subject per adverse event	USUBJID, AETERM, AESTDTC, AEENDTC	Yes	No	Adverse Events		
CM	Concomitant Medication Domain	Interventions	One record per subject per recorded intervention occurrence or constant-dosing interval	USUBJID, CMTRT, CMSTDTC, CMENDTC, CMINDC, CMDOSTOT, CMSPID	Yes	No	CM		

Figure 6: Dataset information.

3. Variable Sheet: All the variables in every dataset are present in this sheet and it contains the below columns–

   a. Seq. for Order – Gives the information about the order of the variables to be presented in the dataset.
   b. Observation Class – Per the CDISC SDTM IG the observations class is presented here.
   c. Domain Prefix - Presents the two characater abbreveiation for the domain.
   d. Variable Name(without domain prefix and with domain prefix) – Variable name without the domain prefix will help to filter and sort the similar type of the variables. Variable name with the domain prefix will be domain specific.
   e. Variable Label – Label of the domain.
   f. Type – identifies the type of the variable (Numeric or Character).
   g. Controlled Terms or Format – This will have allowed CT's or the formats for the Age units (Years, Months or days), Domains, Ethinicity, Race, Sex, Country, Date format (ISO 1806) or Y/N format.
   h. Origin – Source of the variable is provided (Protocol, Assigned, Derived or CRF).
   i. Role - Identifies the type of the variable. Per CDISC there are different type of variables (Identifier, Topic, Record Qualifier,Timing etc..).
   j. CDISC Notes (for domains) and Description (for General Classes) – Provides detail information per the CDISC guide lines for the domains and other important information. For example in the below figure, RFSTDTC has information per the SDTM IG and useful information for the programmer.
   k. Core – Tells if the variable is required, expected or permissible. This categorization is based on the CDISC SDTM IG.
   l. Mapping Specification – Gives the data sources or derivation algorithm to get the variable.
   m. Length and Format – Length and Format of the variable can be found here.
   n. Submission Comments- Provision for submission comments that can be used in the define.xml

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Seq. For Orde	Observation Class	Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Role	CDISC Notes (for domains) Description (for General Classes)	Core	Mapping Specifications	Length	Format	Submission Comments
1	Special-Purpose	DM	STUDYID	STUDYID	Study Identifier	Char		Protocol	Identifier	Unique identifier for a study.	Req	Set to "XXXXXXXXXX"	10	$10.	
2	Special-Purpose	DM	DOMAIN	DOMAIN	Domain Abbreviation	Char	(DOMAIN)	Assigned	Identifier	Two-character abbreviation for the domain.	Req	Set to "DM"	2	$2.	
3	Special-Purpose	DM	USUBJID	USUBJID	Unique Subject Identifier	Char		Derived	Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product. This must be a unique number, and could be a compound identifier formed by concatenating STUDYID-SITEID-SUBJID.	Req	STUDYID || "-" || rd_dm.SUBJECTNUMBERSTR	20	$20.	Concatenation of STUDYID-SUBJID
4	Special-Purpose	DM	SUBJID	SUBJID	Subject Identifier for the Study	Char		CRF	Topic	Subject identifier, which must be unique within the study. Often the ID of the subject as recorded on a CRF.	Req	rd_dm.SUBJECTNUMBERSTR	5	$5.	
5	Special-Purpose	DM	RFSTDTC	RFSTDTC	Subject Reference Start Date/Time	Char	ISO 8601	Derived	Record Qualifier	Reference Start Date/time for the subject in ISO 8601 character format. Usually equivalent to date/time when subject was first exposed to study treatment. Required for all randomized subjects; will be null for all subjects who did not meet the milestone the date requires, such as screen failures or unassigned	Exp	rd_DBSMDISP.DBDISPDAT_DTS	20	$20.	Subject's first exposure to study treatment.

Figure 7: Variable information.

4. Value level Spreadsheet: Value level metadata information is presented in this sheet. It is designed to resemble define specifications sheet closely to allow it to copy-paste from this document. This will usually be a live document and will be updated during the course of the study.

A	B	C	D	E	F	G	H	I	J	K
Dataset	Variable Name	TESTCD/ QNAM	TEST/QLABEL	Verbatim (for test description which does not fully fit within --TEST)	Controlled Terms or Format Code List	Origin	Role	Mapping Specifications	Comment	TIVERS
SUPPCM	QVAL	ATCT1C	Level 1 ATC Code			Assigned		rd_cmrpt_cmrept.ATC1C		
SUPPCM	QVAL	ATCT2C	Level 2 ATC Code			Assigned		rd_cmrpt_cmrept.ATC2C		
SUPPCM	QVAL	ATCT3C	Level 3 ATC Code			Assigned		rd_cmrpt_cmrept.ATC3C		
SUPPCM	QVAL	CMRTOTH	Other Route, Specify			CRF		rd_cmrpt_cmrept.CMRTOTH		
SUPPCM	QVAL	CMSPEC	Additional Details			CRF		rd_cmrpt_cmrept.CMSPEC		
VS	VSTESTCD	HEIGHT	Height			CRF		acm_lb.LBTESTCD		
VS	VSTESTCD	WEIGHT	Weight			CRF		acm_lb.LBTESTCD		
VS	VSTESTCD	BMI	Body mass index			CRF		acm_lb.LBTESTCD		
VS	VSTESTCD	TEMP	Temperature			CRF		acm_lb.LBTESTCD		

Figure 8: Value level sheet.

5. Codelist: This sheet lists all possible values as they appear for the data. The data sheet will be populated as new data gets populated. This will help in the final documentation of the SDTM.

A	B	C	D	E	F	G	H
Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	CDISC Definition	NCI Preferred Term
C65047		Yes	Laboratory Test Code	LBTESTCD	Laboratory Test Code	Terminology used for laboratory test codes of the CDISC Study Data Tabulation Model.	CDISC SDTM Laboratory Test Code Terminology
C100429	C65047		Laboratory Test Code	A1AGLP	Alpha-1 Acid Glycoprotein	A measurement of the alpha-1 acid glycoprotein in a biological specimen.	Alpha-1 Acid Glycoprotein
C80167	C65047		Laboratory Test Code	A1ANTRYP	Alpha-1 Antitrypsin; Serum Trypsin Inhibitor	A measurement of the alpha-1 antitrypsin in a biological specimen.	Alpha-1 Antitrypsin Measurement
C100462	C65047		Laboratory Test Code	A1MCREAT	Alpha-1 Microglobulin/Creatinine	A relative measurement (ratio or percentage) of the alpha-1 microglobulin to creatinine in a biological specimen.	Alpha-1 Microglobulin to Creatinine Ratio Measurement

Figure 9: Codelists sheet.

6. SDTM Reference sheet: It is always useful when a SDTM reference document is available handy. It has details and explanation about a variable. This information is documented from the SDTM IG and will come handy to the programmers and CRF annotators.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Seq. For Order	Observation Class	Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Controlled Terms or Format	Orig	Role	CDISC Notes (for domains) Description (for General Classes)	Core	Internal Specifications	Lengt	Form	Submission Commer
1	All Classes		STUDYID	STUDYID	Study Identifier	Char			Identifier	Unique identifier for a					
2	All Classes		DOMAIN	DOMAIN	Domain Abbreviation	Char			Identifier	Two-character abbreviation for the domain most relevant to the observation. The Domain abbreviation is also used as a prefix for					
3	All Classes		USUBJID	USUBJID	Unique Subject Identifier	Char			Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving					
4	All Classes		POOLID	POOLID	Pool Identifier	Char			Identifier	An identifier used to identify a result from a group of subjects that is					

Figure 10: SDTM Reference sheet.

CONCLUSION:

Use of single tabbed specifications will ease the update and review process, which in turn will save time. By standardizing the specification file, proprietary tools can be developed at an organizational level to assist in define.xml creation, CRF annotation, define.pdf creation that saves a lot of time and money for the organizations, thus keeping the projects under budget. When this approach was implemented compared to traditional speculations, it was observed that programming and study are a lot more organized. This user-friendly interface makes it straightforward to apply changes as the study progresses for similar variables across domains.

REFERENCES

[1] SDTMIG v3.3 https://www.cdisc.org/standards/foundational/sdtm.

[2] Yurong Dai, Jiangang Jameson Cai "Conversion of CDISC specifications to CDISC data–specifications driven SAS programming for CDISC data mapping" Proceedings of PharmaSUG 2017 Conference.

CONTACT INFORMATION:

Author Name : Siddharth Kumar Lokineni
Company : Syneos Health
Address : 3201 Beechleaf Ct
City / Postcode : Raleigh, NC 27604
Work Phone : 210 601 6752
Email : Siddharth.Lokineni@syneoshealth.com

Author Name : Vara Prasad Reddy Sakampally
Company : Syneos Health
Address : 3201 Beechleaf Ct
City / Postcode : Raleigh, NC 27604
Work Phone : 270 282 5166
Email : varaprasadreddy.sakampally@syneoshealth.com