PharmaSUG 2014 - Paper PO04

CDISC Mapping and Supplemental Qualifiers

Arun Raj Vidhyadharan, inVentiv Health, Somerset, NJ
Sunil Mohan Jairath, inVentiv Health, Somerset, NJ

ABSTRACT

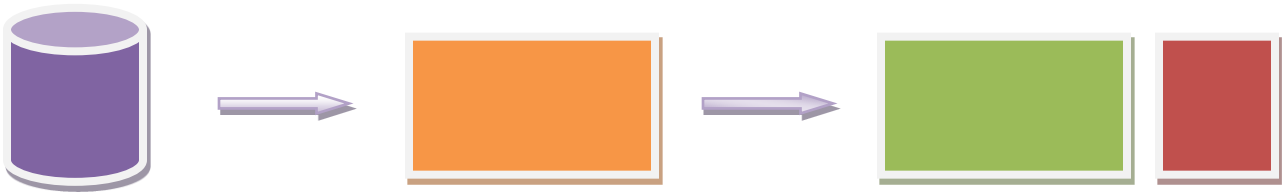
Mapping of datasets from sponsor defined data structure, otherwise known as Clinical Data Management (CDM) data structure, to CDISC SDTM structure can be one of most trickiest and complex programming situations. There are several methods devised by companies for this purpose. Some use SAS Mapping tools, some use tools based on SAS, VB and Excel while some use just SAS programs. Irrespective of the technique used, the basic fundamentals of mapping remain the same. This paper covers the various factors to be considered while mapping, certain unique scenarios one can encounter and possible solutions to them. This paper also covers the creation of SUPPLEMENTAL QUALIFIERS and custom domains and talks about a powerful SAS procedure to validate your SDTM datasets.

INTRODUCTION

Before initiating the data mapping and conversion process it is crucial to have a basic understanding of the SDTM specifications. CDISC provides implementation guides for all of the CDISC data standards on their Website (www.cdisc.org). The SDTM Implementation Guide (SDTMIG) is an essential tool for anyone involved with the metadata mapping or programming associated with the creation of SDTM data sets. The SDTM Implementation Guide contains the specifications and metadata for all of the SDTM data domains and guidance for producing SDTM domain files.

Identifying and practicing a mapping process depends on the extent to which the CDM data structure is compliant to the CDISC data structure, the level of compliance with SDTM standards you wish to achieve (SDTM, SDTM-plus or SDTM-minus) and the number of datasets you plan to map.

DEFINING THE PROCESS



CDM Data Structure

SDTM-plus/minus

SDTM + Supplemental Qualifiers

Figure 1.

Some companies practice a two stage process of mapping. This 2 stage mapping generates the SDTM-plus/minus datasets and then generates the SDTM datasets. Firstly, we discuss the various scenarios one encounters while creating the SDTM-plus/minus datasets from CDM data. The creation of SDTM datasets and Supplemental Qualifiers from SDTM-plus/minus datasets is described later in the paper.

The various steps involved in a typical mapping process are as follows:

1) Identify the CDM datasets that you are planning to map.
2) Identify the SDTM datasets that correspond to the CDM datasets you identified in the previous step (For example: CDM dataset DEMO corresponds to SDTM dataset DM).
3) Have the metadata of those identified CDM datasets and their corresponding SDTM metadata handy.
4) For CDM datasets that have a corresponding SDTM dataset, continue to step 5. Other datasets would become custom domains.
5) Map variables in CDM datasets to SDTM domain variables.

This is the actual mapping process and typically involves the following scenarios:

- *Direct carry forward:*

Variables that conform 100% to SDTM standards are directly carried forwarded to the SDTM dataset without any modification. This happens when the CDM datasets has some level of compliance with corresponding SDTM dataset.

- *Renaming variables:*

Some variables have to be renamed to map to the corresponding SDTM standard variable. For example, variable GENDER in CDM dataset DEMO should be renamed to variable SEX in SDTM dataset DM.

- *Changing variable attributes like label, type, length and format:*

Another important thing apart from mapping the variable names is to map the variable attributes. Attributes like label, type, length and format has to conform to the SDTM variable attributes.

- *Reformat:*

The actual value being represented does not change, only the format in which is stored changes, such as converting a SAS date to an ISO8601 format character string.

- *Combining:*

In some cases, multiple variables need to be combined to form a single SDTM variable.

- *Splitting:*

There could also be cases where a variable in CDM needs to be split into 2 or more SDTM variables.

- *Derivation:*

Some SDTM variables are derived using variables in CDM dataset. For example, derivation of AE Duration (AEDUR) variable using the start date and end date of AE in CDM dataset.

- *Mapping variable values to conform to SDTM formats and applying new code list:*

Some variable values need to be recoded or mapped to match with the values of corresponding variable in SDTM dataset. This mapping should be done very carefully and is recommended for variables with a code list attached that has non-extensible controlled terminology. It is also advised to map for all values in the controlled terminology than just for the values present in the dataset. This would cover for values that are not in the dataset currently but may come in during future dataset updates. A very simple example of this mapping is demonstrated below:

CDM dataset ADVERSE				
Variable Name	Type	Codelist Attached	Value	Formatted Value
AENY	Num	YesNo	1	No
AENY	Num	YesNo	2	Yes

Figure 2.

SDTM dataset AE				
Variable Name	Type	Codelist Attached	Value	Formatted Value
AEOCCUR	Char	NY	N	N
AEOCCUR	Char	NY	Y	Y

Figure 3.

The above table shows how the variable AENY (Adverse Event Occurrence) in CDM dataset ADVERSE is mapped to the corresponding SDTM variable AEOCCUR in AE dataset. The original CDM variable AENY has numeric values 1 – that represents 'No' and 2 – that represents 'Yes' with codelist YesNo attached. However, as per the SDTM standards, the corresponding variable i.e. AEOCCUR has character values 'N' and 'Y' with codelist NY attached. So, to conform to SDTM standards, the values in CDM variable AENY has to be re-coded/mapped to the values in SDTM variable AEOCCUR. Once the values are mapped, the new codelist NY can be applied which resolves the values to 'N' and 'Y'. Note that apart from renaming, mapping and applying new codelist, other attributes like label, length, type etc. should also be changed to conform to the SDTM variable attributes.

- *Transposing data from horizontal structure to vertical structure:*

Sometimes the structure of the CDM dataset will be completely different from its corresponding SDTM dataset structure. In such cases, we might even have to transform the CDM dataset to a structure that is compliant with SDTM requirements. A classic example of this is the Vital Signs dataset. Some sponsors collect vital signs data in their dataset in wide form. This means that every test and their recorded value are stored in separate variables as shown below:

Subject Identifier	Visit	Visit Date	Heart Rate	Heart Rate Unit	Systolic BP	Systolic BP Unit	Diastolic BP	Diastolic BP Unit	Weight	Weight Unit
100	Week 1	01JAN21013	71	/MIN	140	MMHG	75	MMHG	54	KG
100	Week 2	08JAN21013	80	/MIN	139	MMHG	63	MMHG	49	KG
100	Week 3	18JAN21013	75	/MIN	122	MMHG	70	MMHG	49	KG

Figure 4.

As this structure is non-compliant to the SDTM requirements, where data is stored in lean form, the dataset is transposed to have all the tests, values and unit under 3 variables. The number of observations in the resultant dataset would increase exponentially and the number of variables would reduce reciprocatively, depending on the number of vital signs tests.

Subject Identifier	Visit	Visit Date	Vital Signs Test	Result or Finding in Original Unit	Original Unit
100	Week 1	01JAN21013	Heart Rate	71	/MIN
100	Week 1	01JAN21013	Systolic Blood Pressure	140	MMHG
100	Week 1	01JAN21013	Diastolic Blood Pressure	75	MMHG
100	Week 1	01JAN21013	Weight	54	KG
100	Week 2	08JAN21013	Heart Rate	80	/MIN
100	Week 2	08JAN21013	Systolic Blood Pressure	139	MMHG
100	Week 2	08JAN21013	Diastolic Blood Pressure	63	MMHG

100	Week 2	08JAN21013	Weight	49	KG
100	Week 3	18JAN21013	Heart Rate	75	/MIN
100	Week 3	18JAN21013	Systolic Blood Pressure	122	MMHG
100	Week 3	18JAN21013	Diastolic Blood Pressure	70	MMHG
100	Week 3	18JAN21013	Weight	49	KG

Figure 5.

However, there could still be variables in the CDM dataset that cannot be mapped to a variable as per SDTM standard. Such variables would go into supplemental qualifiers, which is the topic of the next section.

SUPPLEMENTAL QUALIFIERS AND CUSTOM DOMAINS

The SDTM standard includes a rule that no new variables can be added to any data domain. If a user has additional data for a domain, which cannot be fit into the domain using the standard variables, they must use a Supplemental Qualifiers special-purpose dataset for this purpose. This is a separate dataset from the "parent" domain in question, and it has a vertical structure that allows the user to add any supplemental data in a "variable name – variable value" structure. Every SUPPQUAL contains ten variables because variables in a SUPPQUAL domain are either "Required" or "Expected": five Key variables that reference a specific record in its Parent domain and five Q-variables that contain the actual supplemental data. Figure 6 and figure 7 show the structure and attributes of a typical Supplemental Qualifier, respectively.

SUPPQUAL									
STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL

Figure 6.

Variable name	Type	Length	Label
STUDYID	Character	10	Study Identifier
RDOMAIN	Character	2	Related Domain Abbreviation
USUBJID	Character	20	Unique Subject Identifier
IDVAR	Character	8	Identifying Variable
IDVARVAL	Character	100	Identifying Variable Value
QNAM	Character	8	Qualifier Variable Name
QLABEL	Character	40	Qualifier Variable Label
QVAL	Character	100	Data Value
QORIG	Character	50	Origin
QEVAL	Character	50	Evaluator

Figure 7.

Supplemental Qualifiers are created for each domain that has non-SDTM standard variable(s). Hence, if there are 'n' CDM datasets that have non-SDTM standard variable(s), we would create 'n' Supplemental Qualifier datasets. Supplemental Qualifier datasets also follow a naming convention. It will always start with "SUPP" followed by 2 characters that represents the SDTM domain for which they are created. So, if we are creating a Supplemental Qualifier dataset for the demography dataset 'DM', it would be named 'SUPPDM'. An example of supplemental qualifier dataset created for DM SDTM plus dataset is demonstrated below. In this example, the SDTM plus dataset DM contains population flags Intent to Treat (ITTFL) and Per Protocol (PPROTFL). As these are non-standard SDTM variables, we create supplemental qualifier dataset SUPPDM and move these variables there. The SDTM plus dataset DM is depicted in figure 8 below.

DM

STUDYID	DOMAIN	USUBJID	AGE	SEX	RACE	ITTFL	PPROTFL
12345	DM	100	45	M	ASIAN	N	N
12345	DM	200	67	F	WHITE	Y	N
12345	DM	300	34	M	BLACK OR AFRICAN AMERICAN	Y	Y

Figure 8.

The corresponding supplemental qualifier dataset SUPPDM created to contain the population flags would be as follows:

SUPPDM

STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
12345	DM	100			ITTFL	Intent to Treat Population Flag	N	Derived	Sponsor
12345	DM	100			PPROTFL	Per Protocol Population Flag	N	Derived	Sponsor
12345	DM	200			ITTFL	Intent to Treat Population Flag	Y	Derived	Sponsor
12345	DM	200			PPROTFL	Per Protocol Population Flag	N	Derived	Sponsor
12345	DM	300			ITTFL	Intent to Treat Population Flag	Y	Derived	Sponsor
12345	DM	300			PPROTFL	Per Protocol Population Flag	Y	Derived	Sponsor

Figure 9.

Note that the variables IDVAR and IDVARVAL are empty for this supplemental qualifier dataset. This is because the SDTM dataset DM contains only one observation per subject and variable USUBJID is sufficient enough to reference the records. If the SDTM dataset contains multiple observations per subject, the IDVAR and IDVARVAL variables are used to identify specific observations. The sequence variable (e.g. AESEQ) is utilized for this purpose.

There could also be certain CDM datasets that cannot be mapped to any SDTM domains. Such datasets, if necessary, would become custom domains.

VALIDATING YOUR SDTM DATASETS

Finally, once you are done with mapping, you could validate your datasets using the SAS procedure CDISC. The following PROC CDISC code shows the syntax that is required to validate a SAS dataset that conforms to CDISC SDTM version 3.1:

```
proc cdisc model=sdtm;
   sdtm sdtmversion="3.1";
   domaindata data=results.ae domain=ae catagory=events;
run;
```

CONCLUSION

Mapping of sponsor defined CDM datasets to SDTM datasets can be a tedious and laborious task. However, a clear understanding of the metadata of your source and target datasets as well as the basic fundamentals of mapping can help you streamline your process and achieve your results. You could also develop your own tools using SAS and excel based on this understanding for re-use in studies that require SDTM mapping. This paper also explains creation of supplemental qualifiers and discusses various attributes of it. Creation of SUPPQUAL datasets is an absolute requirement if you have SDTM plus datasets and want to make them 100% STDM. Finally, you could utilize the powerful SAS procedure CDISC to validate your SDTM domains.

REFERENCES

- Robert W. Graebner: Practical Methods for Creating CDISC SDTM Domain Data Sets from Existing Data
- Barry R. Cohen: SDTM, Plus or Minus
- John R Gerlach & Glenn O'Brien: Generating SUPPQUAL Domains from SDTM-Plus Domains
- SAS Online Documentation, Version 9.2

CONTACT INFORMATION

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

Name: Arun Raj Vidhyadharan
Enterprise: inVentiv Health Clinical
Address: 500 Atrium Drive
City, State ZIP: Somerset, NJ 08873
Work Phone: 732.652.3490
E-mail: arunraj.vidhyadharan@inventivhealth.com

Name: Sunil Mohan Jairath
Enterprise: inVentiv Health Clinical
Address: 500 Atrium Drive
City, State ZIP: Somerset, NJ 08873
Work Phone: 732.652.3482
E-mail: sunil.jairath@inventivhealth.com