Paper 212-2010

SDTM Attribute Checking Tool

Ellen Xiao, Merck & Co., Inc., Rahway, NJ

ABSTRACT

Converting clinical data into CDISC SDTM format is a high priority of many pharmaceutical/biotech companies. Most of these companies are ready to invest in obtaining a SDTM conversion tool that can do the job specifically designed for their company or to hire a CRO to convert the data for them. CDISC STDM conversion may sound like an easy job given the various available tools. What is important not to overlook is the validation of the resulting CDISC SDTM data. It can be quite challenging and time consuming validate SDTM submissions manually. Three targets or perspectives of validation can be categorized as structure, content and attributes. Here, I'd like to focus specifically on a SDTM Attribute Checking Tool used to check between SDTM data and the define.xml.

INTRODUCTION

The 1999 FDA "Electronic Submission (eSub) Guidance" and the "Electronic Common Technical Document (eCTD) Specification" specify that a document describing the content and structure of the included data should be provided as part of the submission. This document is known as the data definition document (e.g., "define.pdf" in the 1999 guidance). The Data Definition Document provides a list of the datasets included in the submission along with a detailed description of the contents of each dataset. To increase the level of automation and improve the efficiency of the regulatory review process, the define.xml as published by the CDISC define.xml team is the preferred type of data definition file which is more suitable for providing the different types of metadata required to adequately describe data in the SDTM format. An additional benefit of define.xml is its machine-readability.
Inconsistency between the document and real datasets is unavoidable. Common inconsistencies can be a difference in labels (truncation, typo in the label, extra blanks, special non-displayable characters etc.), different lengths of variables, extra/fewer variables etc. Any discrepancies may lead the regulatory agency reviewer question the overall quality of the deliverables which may cause unexpected delays of drug approval. Thus, capturing and correcting these discrepancies is as important as creating the documents and CDISC data themselves. To capture all these inconsistencies manually is time consuming and error-prone. Developing a tool to do the consistency check is a more practical way of solving this kind of issue. And this kind of tool can actually be used across platforms and companies.

SDTM ATTRIBUTE CHECKING TOOL

A series of SAS macros have been developed that can be used to check each SDTM domain against the define.xml to maintain consistency between the datasets and the define.xml. The SAS macro is called `INF1-3CheckInFormWithDefine` which includes a calling program to `%compare0define` and macros (%parse, `%exdde, %exdde, %parse2, %charnum, %defaut, %read` etc.).

1. %PARSE2 MACRO:

This macro reads in a token list separated by a delimiter defined in the macro parameter and assigned each token to different global macro variables from *&macronam.1* to *&macronam&j (&j* is the total number of tokens listed) and assign total number of tokens to *&macronam.0*.

```
%LET sub_lst=%scan(&lst,&i,&dlm);
      %do %while("&sub_lst" ne "");
          %do;
             %let j=%eval(&j+1);
             %global &macronam&j;
             %let &macronam&j=&sub_lst;
          %end;
             %let i=%eval(&i+1);
             %LET sub_lst=%scan(&lst,&i,&dlm);
      %end;
 %let &macronam.0=&j;
```

2. %EXDDE MACRO:

This macro is mainly used for importing data from Excel (define.xml) into SAS which provides many flexible features in the data importing process. Such as, reading multiple sub-sheets nested within one Excel file with a single macro call which reduces the CPU time for data transfer.

2.1 Set up default values for some macro parameters.
This feature allows the user to minimize the effort and allows use of the default values for the parameters and to define others.

2.1.1 Convert parameters STARTC and ENDC to numbers if they are letters by calling macro *%charnum*. Since the head of the Excel file column was displayed as character(s), the user may define macro variables `&StartC` (start column to read) or `&EndC` (end column to read) as character(s). This macro convert `&StartC` and/or `&EndC` from character(s) to number(s).

```
%charnum(StartC); %charnum(EndC);
```

2.1.2 Set up default values for macro parameters not specified such as, start row, start column, end row and end column of the excel spread sheets to read in etc..

```
%if %length(&&&dvar) eq 0 %then %do;
    %let &dvar = &dvarv;
    %if &dvar eq fmt %then %do;
        %if &var ^= %str(col&StartC.-col&EndC.) %then %do;
               %parse2(lst=&var,macronam=varlst,dlm=%str( ));
                    %let &dvar = sysfunc(translate(%sysfunc(repeat
                                ($30.c,%eval(&varlst0- 1))),," ","c"));
         %end;
    %end;
%end;
```

2.2 If logic checks fail, error messages will be printed in the log file/window and the macro will end execution.

The logic checking was established at the beginning of the macro. Instead of running the whole program, the logic checking makes it possible to stop the program execution when logic checking failed which provide great efficiency for the user to identify the error(s) in the very early stage.

The following logic checks are included in the program:
2.2.1 Check if the Excel file exist.
2.2.2 Check if the required macro parameters exist.
2.2.3 Check if "start row number", "start column number", "end row number" and "end column number" values defined from macro parameters larger than 0.
2.2.4 Check if "start row", "end row" have specified with numeric number instead of character.
2.2.5 Check if "end row" is larger than or equal to "start row" and "end column" is larger than or equal to "start column".

Below is an example of failed logic checking and the related error message that is generated at the SAS log file/window.

```
@@@@@@@@@@@@@@@@@@ Message from Macro EXDDE @@@@@@@@@@@@@@@@
@ Logic_Err = 2.                                                  @
@ Parameters RAWDIR, FILE, and OUTDATA are required,              @
@ and MUST BE SPECIFIED.                                          @
@ Please try again.                                               @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@;
```

The error message clearly indicates which part of the program needs to be modified.

2.3 Read in data from each worksheet, and combine them into one dataset.

2.3.1 Get the domain list from metadata and create a macro variable `&domainlist` which can be used to define a macro parameter to read the define.xml sheet by sheet.

```
select distinct scan(memname,1,"$") into: domainlist SEPARATED by  '|'
from sashelp.vtable where libname='DEFINE' and
length(scan(memname,1,"$"))=2 or upcase(memname)='RELREC$';
```

2.3.2 Call *%parse2* to separate the Excel file (define.xml) sheet by sheet into different macro variables and save the total number of sheets read in into a macro variable &sh0.

2.3.3 Macro *%read*: use dynamic data exchange to read the Excel file (define.xml) which provide the feature to handle format inconsistency within a column. It drops columns not needed and flag the sub-sheet origin of the data. It checks if the Excel file is open, and the related sheet name may need to be verified .

```
filename one dde
"Excel|&rawdir\[&file..xls]&&sh&i!r&startR.c&startC:r&endR.c&endC"
```

2.3.4 Open Excel file.
The DDE approach require the Excel file (define.xml) to be opened during the data Processing. This macro integrates the function of opening the Excel file with an X command. If the macro parameter &Excel has been defined as "ON", the user need to manually open the Excel file before submit the SAS program, otherwise the program will open the Excel file. By defining the macro parameter `&sleep`, the user can specify the number of seconds that a SAS data step is suspended from execution while opening up the Excel file.

2.3.5 Close Excel window if not comparing the data between Excel & SAS. By define the macro parameter &debug, the user can control if the Excel file need to leave as open for validation purpose.

2.3.6 Combine data sets from different sub-sheets. In this step multiple sheets from one Excel file are combined into one SAS data set with the variable FLAG identifying the sheet origin of each record.

2.3.7 Delete blank records

CONCLUSION

SDTM Attribute Checking Tool is a handy tool to perform user acceptance check by end users. For both internal and external CRO SDTM conversion, we can always use the same tool to make sure the consistency of the SDTM data. It is a cost effective way to provide greater flexibility to users by providing options to generate customized checks and reports specific to user requirements, both for SDTM domains and for user-defined datasets. Detail implementation of this solution is also available in the macro user menu. Together with SDTM Checker, the whole SDTM validation process can be complete and thorough. It also keeps users away from tedious work and keeps users focus on other important tasks.

APPENDIX

CDISC SUBMISSION DATASET DEFINITION METADATA EXAMPLE (STUDY DATA TABULATION MODEL IMPLEMENTATION GUIDE: HUMAN CLINICAL TRAILS. WWW.CDISC.ORG)

Dataset	Description	Location	Structure	Class	Purpose	Keys
DM	Demographics	dm.xpt	One record per subject	Special Purpose Domains	Tabulation	STUDYID, USUBJID
CO	Comments	co.xpt	One record per comment per subject	Special Purpose Domains	Tabulation	STUDYID, USUBJID, COSEQ
SE	Subject Elements	se.xpt	One record per actual Element per subject	Special Purpose Domains	Tabulation	STUDYID, USUBJID, ETCD, SESTDTC
SV	Subject Visits	sv.xpt	One record per actual visit per subject	Special Purpose Domains	Tabulation	STUDYID, USUBJID, VISITNUM
CM	Concomitant Medications	cm.xpt	One record per recorded medication occurrence per subject	Interventions	Tabulation	STUDYID, USUBJID, CMTRT, CMSTDTC
EX	Exposure	ex.xpt	One record per constant dosing interval per subject	Interventions	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC
SU	Substance Use	su.xpt	One record per substance type per reported occurrence per subject	Interventions	Tabulation	STUDYID, USUBJID, SUTRT, SUSTDTC
AE	Adverse Events	ae.xpt	One record per adverse event per subject	Events	Tabulation	STUDYID, USUBJID, AEDECOD, AESTDTC
DS	Disposition	ds.xpt	One record per disposition status or protocol milestone per subject	Events	Tabulation	STUDYID, USUBJID, DSSTDTC, DSDECOD
MH	Medical History	mh.xpt	One record per medical history event per subject	Events	Tabulation	STUDYID, USUBJID, MHDECOD
DV	Protocol Deviations	dv.xpt	One record per protocol deviation per subject	Events	Tabulation	STUDYID, USUBJID, DVTERM, DVSTDTC

EXAMPLE OF DEMOGRAPHICS – DM (COPY FROM CDISC.ORG)

dm.xpt, Demographics — Version 3.1.2, February 21 2007. One record per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core	References
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req	SDTMIG 2.4.4
DOMAIN	Domain Abbreviation	Char	DM	Identifier	Two-character abbreviation for the domain.	Req	SDTMIG 2.4.4, SDTMIG 4.1.2.2 SDTMIG Appendix C2
USUBJID	Unique Subject Identifier	Char		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	SDTMIG 2.4.4, SDTMIG 4.1.2.3
SUBJID	Subject Identifier for the Study	Char		Topic	Subject identifier, which must be unique within the study. Often the ID of the subject as recorded on a CRF.	Req	SDTMIG 2.4.4
RFSTDTC	Subject Reference Start Date/Time	Char	ISO 8601	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 subjects.	Exp	SDTMIG 4.1.4.1
RFENDTC	Subject Reference End Date/Time	Char	ISO 8601	Record Qualifier	Reference End Date/time for the subject in ISO 8601 character format. Usually equivalent to the date/time when subject was determined to have ended the trial, and often equivalent to date/time of last exposure to study treatment. Required for all randomized subjects; null for screen failures or unassigned subjects.	Exp	SDTMIG 4.1.4.1
SITEID	Study Site Identifier	Char		Record Qualifier	Unique identifier for a study site.	Req	
INVID	Investigator Identifier	Char		Record Qualifier	An identifier to describe the Investigator for the study. May be used in addition to SITEID. Not needed if SITEID is equivalent to INVID.	Perm	
INVNAM	Investigator Name	Char		Synonym Qualifier	Name of the investigator for a site.	Perm	
BRTHDTC	Date/Time of Birth	Char	ISO 8601	Record Qualifier	Date/time of birth of the subject.	Perm	SDTMIG 4.1.4.1
AGE	Age	Num		Record Qualifier	Age expressed in AGEU. Usually derived from RFSTDTC and BRTHDTC, but BRTHDTC may not be available in all cases (due to subject privacy concerns).	Exp	
AGEU	Age Units	Char	(AGEU)	Variable Qualifier	Units associated with AGE.	Exp	SDTMIG Appendix C7
SEX	Sex	Char	(SEX)	Record Qualifier	Sex of the subject.	Req	
RACE	Race	Char	*	Record Qualifier	Race of the subject. Sponsors should refer to "Collection of Race and Ethnicity Data in Clinical Trials" (FDA, September 2005) for guidance regarding the collection of race (http://www.fda.gov/cder/guidance/5656fnl.htm). If multiple race responses are collected and one is designated as the primary race, RACE should hold the primary race. If the primary race was collected via an "Other, Specify" field and the sponsor chooses not to map the value to one of the 5 designated values, then the value of RACE should be 'OTHER'. If multiple races are collected and none is designated as primary, then the value of RACE should be 'MULTIPLE', and the additional information will be included in the Supplemental Qualifiers dataset. If a subject does not provide race information, the value of RACE should be 'UNKNOWN'.	Exp	
ETHNIC	Ethnicity	Char	(ETHNIC)	Record Qualifier	The ethnicity of the subject. Sponsors should refer to "Collection of Race and Ethnicity Data in Clinical Trials" (FDA, September 2005) for guidance regarding the collection of ethnicity (http://www.fda.gov/cder/guidance/5656fnl.htm).	Perm	
ARMCD	Planned Arm Code	Char	*	Record Qualifier	Short name for ARM (may be up to eight characters).	Req	SDTMIG 4.1.2.1
ARM	Description of Planned Arm	Char	*	Synonym Qualifier	Name of the Arm to which the subject was assigned.	Req	SDTMIG 4.1.2.1, SDTMIG 4.1.2.4
COUNTRY	Country	Char	(COUNTRY)	Record Qualifier	Country of the investigational site in which the subject participated in the trial.	Req	
DMDTC	Date/Time of Collection	Char	ISO 8601	Timing	Date/time of demographic data collection.	Perm	SDTMIG 2.4.5, SDTMIG 4.1.4.1
DMDY	Study Day of Collection	Num		Timing	Study day of collection measured as integer days.	Perm	SDTMIG 2.4.5, SDTMIG 4.1.4.1

** Indicates variable may be subject to controlled terminology, (Parenthesis indicates CDISC/NCI code-list code value)*

EXAMPLE OF COMPARISION RESULT

Discrepancy of extra variable definition

Study: XXXXXX-XXX 11:51 Sunday, May 17, 2009
SDTM Data Vs. Mapping Specification:
List of Variables in SDTMPLUS, but not in Define.xml

Obs	Memname	Name	Label	Vartype
53	SUPPQUAL	IDVAR	IDVAR	Char

54	SUPPQUAL	IDVARVAL	IDVARVAL	Char
55	SUPPQUAL	QEVAL	OEVAL	Char
56	SUPPQUAL	QLABEL	OLABEL	Char

Discrepancy of label

Study: XXXXXX-XXX 19:51 Sunday, May 24, 2009
SDTMPLUS Data Vs. Mapping Specification:
List of Variables with Non-matched Labels

Obs	Memname	Name	Exllabel	Label
203	CF	CFTEST	Test	CFTEST
204	CF	CFTESTCD	Test Short Name	Test Short Name
205	CF	CFTPT	Planned Time Point Name	CFTPT
206	CF	CFTPTNUM	Planned Time Point Number	CFTPTNUM

Discrepancy of variable length

Study: XXXXXX-XXX 19:51 Sunday, May 24, 2009
SDTMPLUS Data Vs. Mapping Specification:
List of Variables with Non-matched Variable Length

Obs	Memname	Name	Exltype	Vartype	Exlength	Length
749	QS	QSBLFL	Char	Char	20	1
750	QS	QSDRVFL	Char	Char	20	1
751	QS	QSDTC	Char	Char	19	16
753	QS	QSDTI	Char	Char	19	4000

REFERENCES

[1] Define.xml
http://www.cdisc.org/content1057

CONTACT INFORMATION

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

Ellen Xiao
Merck & Co., Inc.,
Rahway, NJ 07065
E-mail: hong_xiao2@merck.com