PharmaSUG2011 – Paper CD22

Truncation, Variable Association, Controlled Terminology, and Some Other Pitfalls in the SDTM Mapping Process

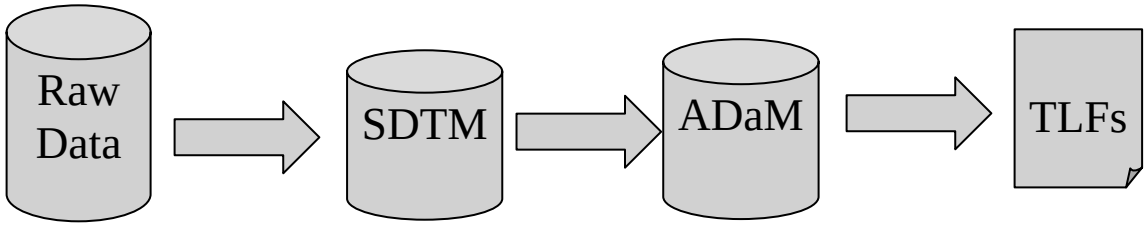
Na Li, XenoPort, Inc., Santa Clara, CA
Gary de Jesus, Infovision, Inc., Richardson, TX
Daniel Bonzo, XenoPort, Inc., Santa Clara, CA

ABSTRACT

This paper discusses a class of likely pitfalls during the SDTM mapping process. Problems in truncation, variable association, controlled terminology, and mapping SUPPQUAL usually occur when SDTM mapping proceeds from raw to SDTM and then using SDTM to generate ADaM. In this pathway, understanding data standards and data capture and reporting instruments from CRF to TLFs (Tables, Listings and Figures) is critical to mitigating potential errors that are embedded in the mapping process. Some collective experience in identifying and preventing these pitfalls will be shared.

INTRODUCTION

Since acceptability of using Study Data Tabulation Model (SDTM) format has been established for electronic submissions, pharmaceutical companies are moving forward with the implementation of SDTM and Analysis Data Model (ADaM). There are several pathways to implement the mapping process from raw clinical data to generating TLFs (Tables, Listings, and Figures). One particular pathway is to use the mapping Raw Data→SDTM→ADaM→TLFs (as shown below). In this approach, the raw data (SAS® datasets) come from Oracle Clinical (OC) extract using in-house database standards which are SDTM-like. The data are mapped from OC to SDTM using CDISC-SDTM 3.1.2 IG (implementation guide). ADaM data sets are then generated using CDISC-ADaM 2.1 IG.



Note that both SDTM and ADaM data are required to have a vertical structure, while raw data normally come in horizontal structure. The preservation of data integrity from the horizontal structure into the vertical structure can be problematic at times. Furthermore, the creation of ADaM data sets using SDTM data sets as source can also be problematic not to mention the generation of the resulting TLFs. This paper will focus on mapping process and will not attempt to cover resulting problems in the creation of define.xml.

TRUNCATION OF LONG TEXT FIELDS

The SDTM 3.1.2 IG instruction for long text field with more than 200 characters is detailed in SDTM section 4.1.5.3.2. The instruction states to keep the first of the 200 characters in the standard domain and keep the rest 200 characters of text in Supplemental Qualifiers (SUPP--) domain. Long text fields should be given special attention during the mapping process. If segmentation is not done correctly the information from the original data can be lost. Furthermore, in order to generate TLFs using such information, the standard SDTM domain needs to be merged with SUPP-- domains. Correctly merging all of the segments together is essential in maintaining the data integrity of the TLFs. The potential long text information may also be mapped in the following manner: CO for comment related fields; TI for trial inclusion/exclusion criteria; DV for protocol deviation detail; and LB/EG for abnormality interpretation.

Long text fields in CO and TS (Trial Summary) are allowed in its standard domain. The first 200 characters of text can be put under --VAL and the remaining text can be put into --VAL1 to --VALn for each 200 character segments for the rest of the text.

Long texts in TI can be handled in the metadata. If the length of the text criterion is <= 200 characters, putting it in IETEST should be sufficient. If the length of the text is >200 characters, a meaningful text should be put in IETEST and the full text can be put in the metadata.

In other domains, the first 200 characters can be mapped into the standard domain variable and data set and the remaining text can be mapped into the SUPP-- domains.

Consider the Excel file below as a source data shown below. It contains information on protocol deviations. In this case, the information can be mapped into the DV domain. Note that the Deviation Description column's contents are more than 200 characters long.

SOURCE	LEVEL	Deviation Table number	Deviation item number	Unique Subject Identifier	SCOPE	Deviation Description
CRA	MINOR	10	2	XXX	DATA	According to the protocol, PK samples must be stored at -70°C±10°C until shipment. All whole blood and plasma PK samples for the Period 4 time points, Hour -24.08 through Hour 5.00 were out of range on 25-September-2009, reaching a high of -58°C for approximately 3.5 hours.
CRA	MINOR	10	2	XXX	DATA	According to the protocol, whole blood and plasma PK samples must be stored at -70±10°C within 30 minutes of quenching/collection, respectively. The samples listed below were late to freezer in error: Period 1, Hour -12.08, Whole blood 6 min late; Period 3, Hour 0.50, Plasma 2 min late.

Following the suggested mapping process, the information in DV domain should look like the table below.

DOMAIN	USUBJID	DVSEQ	DVSPID	DVTERM	DVCAT
DV	XXX	3	10.2	According to the protocol, PK samples must be stored at -70°C±10°C until shipment. All whole blood and plasma PK samples for the Period 4 time points, Hour -24.08 through Hour 5.00 were out of range	MINOR
DV	XXX	4	10.2	According to the protocol, whole blood and plasma PK samples must be stored at -70±10°C within 30 minutes of quenching/collection, respectively. The samples listed below were late to freezer in	MINOR

As shown above, only the first segment with about 200 characters of text was kept in DVTERM field in DV. The remaining text information was put under SUPPDV and should look like the table below.

RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG
SUPPDV	XXX	DVSEQ	3	DVTERM1	Deviation Text	on 25-September-2009, reaching a high of -58°C for approximately 3.5 hours.	DERIVED
SUPPDV	XXX	DVSEQ	4	DVTERM1	Deviation Text	error: Period 1, Hour -12.08, Whole blood 6 min late; Period 3, Hour 0.50, Plasma 2 min late.	DERIVED

Now, in order to correctly report protocol deviations in the required TLFs, SUPPDV domain needs to be merged with the DV domain using the IDVARVAL and DVSEQ as the keys. This can be done using the SAS codes below:

```
proc sort data=sdtm.suppdv out=suppdv;
by usubjid idvarval qnam;
run;

proc transpose data=suppdv out=tdv(drop=_name_ _label_);
var qval;
id qnam;
idlabel qlabel;
by usubjid idvarval;
run;

data tdv;
 length dvseq 8;
 set tdv;

  dvseq=input(idvarval,8.0);

  drop idvarval;
run;

data dv;
 merge sdtm.dv tdv;
 by usubjid dvseq;
 length fulltext $ 2000;

      ** depends on the # of text string segments;
 fulltext=strip(dvterm)||' '||strip(dvterm1)||' '||strip(dvterm2);
run;
```

The resulting listing report should come out as follows:

Subject No.	Type of Deviation	Deviation Identifier	Description of Deviation/Violation
xxx	MINOR	10.2	According to the protocol, PK samples must be stored at -70°C±10°C until shipment. All whole blood and plasma PK samples for the Period 4 time points, Hour -24.08 through Hour 5.00 were out of range on 25-September-2009, reaching a high of -58°C for approximately 3.5 hours.

Subject No.	Type of Deviation	Deviation Identifier	Description of Deviation/Violation
	MINOR	10.2	According to the protocol, whole blood and plasma PK samples must be stored at -70±10°C within 30 minutes of quenching/collection, respectively. The samples listed below were late to freezer in error: Period 1, Hour -12.08, Whole blood 6min late; Period 3, Hour 0.50, Plasma 2min late.

VARIABLE ASSOCIATION ERRORS IN A GROUP OF RELATED RECORDS

SDTM IG 3.1.2 section 8 focuses on representing relationships and data. Section 8.1 describes the relationship among a group of records for a given subject within the same dataset.The use of Group Identifier (--GPRID) to link related records for a subject is recommended. Also in IG section 6.4, the detailed instruction on "Findings About Events or Interventions" describes how to group the associated information in the FA domain. The variable FAOBJ is designated for such a purpose. This section shows an example of a pitfall in generating the FA domain. The same principles apply to other SDTM standard domains using --GRPID.

In OC-based electronic data capture systems, eCRFs are designed such that a group of variables/questions are related to a specific record. Such information is collected as columns in a horizontal structure. As columns need to be transposed to rows for SDTM domains, the relationship among these columns might be lost or mapped incorrectly. Furthermore, attention is needed when using such mapped information for analysis and reporting to prevent information loss.

Consider as an example a dataset containing information for subjects who experienced heart burn symptoms on a questionnaire CRF. When the symptom occurs the subject is asked if the symptom awakened him and, in addition, a question in coughing is inquired. The symptoms' severities are collected in a scale from 1 to 4. In this case, severity is entered only when a symptom exists. Suppose the analysis focuses on number of heart burn symptoms, number of coughing symptoms, and number of awakening due to heart burn symptoms by week. The severity associated with each symptom may also be analyzed by week.

An example of how the raw data looks like is given in the table below:

PATID	SYM_DT	SYM_TM	HEARTBRN (hear burn symptom)	HBSEV (heart burn severity)	AWAKEN	COUGH	COUGHSEV (cough severity)
1	22-Oct-09	21:20	No	.		No	.
2	27-Oct-09	22:00	Yes	3	No	No	.
3	28-Oct-09	3:05	Yes	3	Yes	Yes	4

Using the recommended approach, the resulting FA domain should be as follows:

DOMAIN	USUBJID	FASEQ	FATESTCD	FATEST	FAOBJ	FAORRES	FADTC
FA	1	1	OCCUR	Occurrence	**HEARTBRN**	N	2009-10-22T21:20
FA	1	2	OCCUR	Occurrence	COUGH	N	2009-10-22T21:20
FA	2	3	OCCUR	Occurrence	**HEARTBRN**	Y	2009-10-27T22:00
FA	2	4	SEV	Severity	**HEARTBRN**	3	2009-10-27T22:00
FA	2	5	AWAKEN	Awaken	**HEARTBRN**	N	2009-10-27T22:00
FA	2	6	OCCUR	Occurrence	COUGH	N	2009-10-27T22:00
FA	3	7	OCCUR	Occurrence	**HEARTBRN**	Y	2009-10-28T03:05
FA	3	8	SEV	Severity	**HEARTBRN**	3	2009-10-28T03:05
FA	3	9	AWAKEN	Awaken	**HEARTBRN**	Y	2009-10-28T03:05
FA	3	10	OCCUR	Occurrence	**COUGH**	Y	2009-10-28T03:05
FA	3	11	SEV	Severity	**COUGH**	4	2009-10-28T03:05

In the above approach, the first 2 rows are derived from the first row of the raw data. Rows 3 to 6 are derived from the second row of the raw data. Rows 7 to 11 are derived from the third row of the raw data. When the raw data contains

a value of No in the HEARTBRN/COUGH field, N is set as a value in FAORRES and no severity and awaken information are kept in FA. When the raw data contains a value of Yes in the HEARTBRN/COUGH field, Y is set as a value in FAORRES in a row with FATEST as Occurrence. Severity that is associated with the symptom is then put in another row with FATEST as Severity. Awaken (whether such symptom awakens the subject during sleep) is put in another row with FATEST as Awaken. The variable FAOBJ with value HEARTBRN groups all the information of symptom severity awakening together (rows 3 to 5 and rows 7 to 9). When COUGH is present, both Occurrence and Severity (rows 10 and 11) are grouped together with FAOBJ value as COUGH.

From the above example, when mapping SDTM variables with association attention needs to be focused on the grouping variable in order to preserve the relationship among these variables. In the FA domain, the FAOBJ variable helps to serve this purpose. Group Identifier (--GRPID) in other SDTM domains should follow the same principle.

Not only is the use of the correct grouping identification critical, it is also important to know the purpose of data collection. In the above example, the Awaken question is associated with heartburn symptom but not coughing. If such knowledge is missed, the Awaken information might be mistakenly mapped under the cough-related FAOBJ or dropped completely.

In addition to understanding variable relationships, special attention should be focused on generating the ADaM datasets that utilize the SDTM domains. Since the data structure changed from horizontal structure in raw data to vertical structure in SDTM, the correlation among the SDTM columns is not necessarily intuitive. When mapping the ADaM datasets, understanding the data structure and data collection is essential. For example, in ADaM we need to acquire the severity for the awakening heart burn symptom per week per subject. In the raw data for subject 3, the information is obvious with the severity as 3. However, after the mapping process, the information can be obtained using a two-step process. First, we need to acquire the date under FADTC when FAORRES=Y, FATEST=Awaken and FAOBJ=HEARTBRN. Then, we need to merge with the same FADTC and FAOBJ for the same subject to get the severity information in the row with FATEST=Severity (This is a value of 3, where FASEQ=8).

CONTROLLED TERMINOLOGY NON-COMPLIANCE

Controlled Terminology is required by CDISC. Detailed information can be found at http://www.cancer.gov/cancertopics/cancerlibrary/new-terminologyresources/cdisc. Following this standard requires familiarity with the source data collection instruments, especially involving test units and the case sensitivity of these values (e.g. DA, CM, EX, LB, VS, and ECG).

Ensuring consistency among values in SDTM should be a primary focus. In the example below, the CM domain is generated by combining two raw datasets (concomitant medication and rescue medication). The variable CMDOSU (dose unit) depends on the medication type and dosing level. When combining the two raw datasets, the units should be made consistent (e.g. TABLET) on the CM domain. An example of an inconsistent CMDOSU is shown below, as the dose unit is Tab in rescue medication raw data but the dose unit is TABLET in concomitant medication.

SUBJID	CMDOSE	CMDOSU	CMCAT
1	20	Mg	RESMED
1	1	Tab	RESMED
1	1	Tab	RESMED
1	2	TABLET	CONMED
1	4	TABLET	CONMED
1	4	TABLET	CONMED

In order to prevent such inconsistency, running frequency on raw data is suggested. It is also good practice to periodically compare the raw data with the list of the controlled terminology.

Case sensitivity is another pitfall area. For example, the unit for height is either “cm” in lower case or “IN” in upper case, and the unit for weight is either “kg” in lower case or “LB” in upper case. In the LB domain, the lab test units often require more mapping effort in matching with the SDTM controlled terminology. For example, the unit for “A measure of an antigen potency defined as a number of antigen units per one milliliter of product is “AgU/mL”. Unit for square inch is “in2” with all lower case but unit inch is “IN” in upper case. In order to comply with the nitty-gritty needs for the controlled terminology, a procedure such as running PROC FREQ on the raw and SDTM domains needs to be established in the mapping process to ensure the consistency of values.

SUPPQUAL ISSUES

SUPPQUAL domains are essential for accommodating non-standard variables into the SDTM model. However, it is very easy to overlook the importance of IDVARVAL in relating a SUPP domain to its parent domain. Usually, the --SEQ variable is used as IDVARVAL and extra care must be taken on how this value is generated within the parent domain.

In order to populate the --SEQ variable, the sort order has to be properly specified such that it will result in a unique key for each record within each subject. An incomplete specification of the sort order can make it very challenging to validate its corresponding SUPP domain. Domains that are historical or cumulative in nature can potentially contain records that are very similar except for some variables. CM, MH or AE domains can be especially prone to this type of problem.

Take for example the following raw CM domain with the following records:

ROW	PATID	CMTRT	CMDECOD	CMDOSE	CMDOSU	CMDOSFRQ	CMSTDTC
1	X00001	Advil	IBUPROFEN	10	mg	PRN	2010-11-29
2	X00001	Midol Liquigels	IBUPROFEN	10	mg	PRN	2010-11-29
3	X00001	Versed	MIDAZOLAM HYDROCHLORIDE	20	mg	PRN	2010-11-30
4	X00001	Versed	MIDAZOLAM HYDROCHLORIDE	20	mg	PRN	2010-11-30

ROW	ROUTE	ROUOTH	INDICATION
1	ORAL		Headache
2	ORAL		Menstrual cramps
3	OTHER	SUBCUTANEOUS	EGD Sedation
4	IV		EGD Sedation

The first 2 records show 2 differently named medications resolving to the same code taken on the same day while the last 2 records show the same medication being take via different routes on the same day. The proper sort key for this domain could be: PATID, CMTERM, CMDECOD, CMSTDTC and ROUTE resulting with CMSEQ populated as follows:

ROW	DOMAIN	USUBJID	CMSEQ	CMTRT	CMDECOD	CMDOSE	CMDOSU	CMDOSFRQ
1	CM	ABC-X00001	1	Advil	IBUPROFEN	10	mg	PRN
2	CM	ABC-X00001	2	Midol Liquigels	IBUPROFEN	10	mg	PRN
3	CM	ABC-X00001	3	Versed	MIDAZOLAM HYDROCHLORIDE	20	mg	PRN
4	CM	ABC-X00001	4	Versed	MIDAZOLAM HYDROCHLORIDE	20	mg	PRN

ROW	CMSTDTC	CMROUTE
1	2010-11-29	ORAL
2	2010-11-29	ORAL
3	2010-11-30	SUBCUTANEOUS
4	2010-11-30	INTRAVENOUS

The corresponding SUPPCM should look something like:

RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QVALUE
CM	ABC-X00001	CMSEQ	1	INDICATION	HEADACHE
CM	ABC-X00001	CMSEQ	2	INDICATION	MENSTRUAL CRAMPS
CM	ABC-X00001	CMSEQ	3	ROUTE	OTHER
CM	ABC-X00001	CMSEQ	3	ROUOTH	SUBCUTANEOUS
CM	ABC-X00001	CMSEQ	3	INDICATION	EGD
CM	ABC-X00001	CMSEQ	4	ROUTE	IV
CM	ABC-X00001	CMSEQ	4	INDICATION	EGD

If either CMTERM or ROUTE is omitted from the sort specification, the key is no longer unique per subject and will result in different values for CMSEQ and therefore potentially making independent validation of these domains more complex. The programmer should be aware of the peculiarities that the data may have so that the sort specification can be properly enumerated.

Another thing that can be easily overlooked is the fact that --SEQ in the parent domain is a numeric variable while IDVARVAL is a character variable and truncation can occur if the conversion is not properly done. The number of SUPP items per USUBJID can easily be exceeded if the format used is too narrow to fit all the digits truncating the IDVARVAL value.

Consider the following AE domain snippet from a study with more than a hundred AEs per subject:

STUDYID	DOMAIN	USUBJID	AESEQ
ABC	AE	ABC-X00001	99
ABC	AE	ABC-X00001	100
ABC	AE	ABC-X00001	101
ABC	AE	ABC-X00001	102

In this example, the programmer developed a canned macro that automatically created SUPP domains and had no problems using this macro from previous studies. However, embedded within the macro is the unfortunate line: IDVARVAL = left(put(AESEQ,2.0)); resulting in the following SUPPAE snippet:

STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL
ABC	AE	ABC-X00001	AESEQ	99
ABC	AE	ABC-X00001	AESEQ	0
ABC	AE	ABC-X00001	AESEQ	1
ABC	AE	ABC-X00001	AESEQ	2

In addition to the above examples, there are numerous areas that could lead to mapping errors. Regarding the EX domain, special effort should be made in understanding the dosing schedule and data collection methods. A paper from David C. Izard at the 2010 PharmaSUG 2010 details the complexities of the EX domain mapping process. It might be tricky to preserve the contiguity of records when working on the DA domain.

CONCLUSION

Whether mapping is done in-house, contracted to vendors, or facilitated through the use of a third party software, one needs to pay attention to mapping pitfalls. The mapping process might be correct following the SDTM guidelines, pass the open CDISC checking tool and successfully loaded to JANUS, but it could still be an incorrect reflection of the source data. Not only should the SDTM domains preserve the information from the source data correctly, but the

ADAM datasets should also faithfully preserve the information from the SDTM domains. If the understanding on both SDTM and ADaM structures in the context of a clinical study design is not done right, reporting tools such as TLFs will not represent the source data correctly.

In order to preserve the original data collected and accurately present the information in the TLFs, a good understanding of the source data structure and study design is critical. Paying attention on long texts and potential comment-like fields can help to mitigate long text truncation error. Understanding the data collection tool and the relationship among variables can be helpful in preventing mapping error on relationship-related data. Using pre-process frequency check can help in preventing errors when working on controlled terminologies. For creating SUPPQUAL domains, having a very good understanding of data peculiarities can help clearly define the relationship between SUPP and its corresponding parent domain.

REFERENCES

Study Data Tabulation Model, Version 1.2 Final. Published by CDISC November 12, 2008.

Study Data Tabulation Model Implementation Guide: Human Clinical Trials, Version 3.1.2. Published by CDISC November 12, 2008.

David C. Izard (Octagon Research Solutions). “Help! The EX Domain is now an Analysis Dataset! What Do I Do?!?” PharmaSUG 2010 Paper CD05.

AKNOWLEDGMENTS

We would like to thank our colleagues, Gie B Hubilla and Tosh Kameda for their review of this paper.

CONTACT INFORMATION

For comments and questions regarding this paper, please contact:

Na Li
XenoPort Inc.
3410 Central Expressway
Santa Clara, CA 95051
Office: (408) 616-7146
E-mail: nli@xenoport.com