PharmaSUG2010 - Paper CD09

CREATING SV AND SE FIRST

Henry B. Winsor, WinsorWorks, Limited, San Mateo, CA
Mario Widel, Genentech, Inc., South San Francisco, CA

ABSTRACT

One current concern of the FDA is that many sponsors are not submitting both the SV and SE data sets when submitting studies for review. The Authors address reasons why the data sets are useful other than as a piece of a final submission, even to the point of encouraging the sponsor to create these data sets as soon as clinical data is available. One method for easy creation is demonstrated, along with a verification method.

INTRODUCTION

Amongst the many complaints about current SDTM practices that have been shared by the FDA, the Agency complains that few sponsors are populating the SV and SE data sets appropriately. Indeed, if a sponsor even populates the data sets, this is often done at the last moment before data submission, so the sponsor has no benefit of the use of these data sets during the process of data cleaning and report preparation. We believe this to be a mistake, in that these data sets can be vary useful and should be the first data sets that are populated when clinical data is made available.

It should be noted that both authors are strong believers in the standardization of data structures within a company. While SDTM is really designed as a submission data source, it is also quite useful during the data cleaning stages. We are convinced that the rewards of designing and keeping to standardized data structures greatly outweigh the additional costs and time involved in remapping the raw data to fit within the data structures. While SDTM is not an ideal data structure for cleaning and reviewing data, it certainly beats not having a single standard within a company.

WHY BOTHER?

So why populate the SV and SE data sets first? A major reason is to avoid circularity in your programs, i.e., where program A creates data for program B to use, which creates the data for Program A to use in creating the data for Program B. We would think that this danger is obvious, but reports from the field indicate that this hazardous technique is still being inflicted upon companies by programmers who cannot otherwise generate data.

DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	EPOCH
TA	CR	Controlled Release	0	PRE	Pre-Treatment		Pre-Study
TA	CR	Controlled Release	1	TITUP	Titration Up	Randomized to CR	Up
TA	CR	Controlled Release	2	TRT	Treatment of Interest		Controlled Release
TA	CR	Controlled Release	3	TITDN	Titration Down		Down
TA	CR	Controlled Release	99	POST	Post-Treatment		Post-Study
TA	IR	Immediate Release	0	PRE	Pre-Treatment		Pre-Study
TA	IR	Immediate Release	1	TITUP	Titration Up	Randomized to IR	Up
TA	IR	Immediate Release	2	TRT	Treatment of Interest		Immediate Release
TA	IR	Immediate Release	3	TITDN	Titration Down		Down
TA	IR	Immediate Release	99	POST	Post-Treatment		Post-Study

We assume that you have already populated the TV, TA and TE data sets, which can and should be populated before you receive any clinical data. We'll need the VISITNUM, and VISIT columns from TV, and then some columns from TA.

STUDYID	DOMAIN	ETCD	ELEMENT	TESTRL	TEENRL
EX	TE	PRE	Pre-Treatment	**	**
EX	TE	TITUP	Titration Up	**	**
EX	TE	TRT	Treatment of Concern	**	**
EX	TE	TITDN	Titration Down	**	**
EX	TE	POST	Post-Treatment	**	**

STUDYID	DOMAIN	VISITNUM	VISIT	VISITDY	TVSTRL	TVENRL
EX	TV	0	Screening	-20	**	**
EX	TV	1	Visit 1	1	**	**
EX	TV	2	Visit 2	8	**	**
EX	TV	3	Visit 3	15	**	**
EX	TV	4	Visit 4	22	**	**

** Values omitted for brevity.

Populating first SV and then SE from SV allows us to use these data sets to populate the VISITNUM, VISIT and VISITDY and EPOCH variables in the other SDTM data sets from one data source without worrying about circularity. Assigning the VISIT and ELEMENT variables in this fashion allows us to globally modify such things as visit names and epoch names without relying upon the raw clinical data for anything more than a date. This can be a powerful tool, especially when you have to combine data from multiple trials into a single data source.

Additionally, the SV and SE data sets are the only data sources within SDTM that put all the visit related study dates into one data set, allowing you to spot date issues early in the process. Even if your Data Management group does not populate every Case Report Form page with a different date field, you still have enough different collections going on that you need to check these dates for synchronicity. The earlier you do it, the earlier you can get them fixed and not have to deal with conflicting dates later on in the reporting process.

One might argue that it is easier to create the SV and SE data sets last, after all the clinical data has been entered and verified. While this is true when taking into consideration only the SV and SE data sets, this also requires that you create the VISIT and ELEMENT variables in a number of programs, which doesn't sound at all easier. You also have the task of checking your visit dates for synchronicity without the aid of a program that allows you to do that work early and easily as a byproduct. It should be obvious that creating the SV and SE data sets early in the process of completing your SDTM work can be a real time and effort saver.

CREATING SV

Here's the task at hand. We are going to create the SV data set first, use it to create the SE data set, then use those two data sets to populate the VISIT and ELEMENT variables in the rest of the data sets, including the reference start date and end date in DM. All we need are the TV, TE and TA data sets and the date values found in your raw data.

The first step to creating SV is you need to identify all the potential visit date information in your raw data. You exclude dates like the subject birth date (although some companies consider the birth date to be the first date in the screening period, to each his own), event dates unrelated to visits such as Adverse Event dates and Concomitant Medication dates. The rule of thumb is if isn't scheduled to be done at a planned visit, we don't want the date. The easiest way to start this is to take a set of Case Report Forms, preferably annotated, and identify all of the date fields of potential interest and then trace them to the data source. The annotation values are important, because they identify what the Case Report Form designer thinks are the visits and allow us to remap the data as we choose.

Each collection data base will have different names for the variables that contain the annotation values, so you'll need to be familiar with these in your own system. For example, DLB/ERT databases store two values, one called EVENTID and one called PAGE, the combination of the two allows the user to know exactly on which Case Report Form page the data record of interest appeared.

Other databases will use similar variables and these variables allow us to easily remap data to whatever destination

we want. Suppose your Case Report Forms use several EVENTID values to indicate measurements taken during what is really the first visit, we can easily remap all of those different values to one value, Visit One. This ability to remap keeps you from being locked into whatever system Data Management needs to have for their purposes. You have the flexibility of using other visit labeling rules and names, already stored in the TV data set.

After you have identified all of your dates and done the necessary remapping so that you can still identify the data source yet fit the dates within your TV visit structure, you need to set all of the dates so you can check for coherency and synchronicity. Do all of a subject's Visit One dates roughly coincide with each other? If a visit is supposed to take place on one day, do all the dates agree? Are any different by a year or a month, indicating an entry problem? You need to build a report at this time, and you are going to have to review it manually for the most part, but any date problems will stand out like an elephants on a putting green.

The report should look something like this:

SUBJID	EVENTID	Date	Sources
1026	Visit One	05JUL2009	AD VS PE BL BG MH LB
	Visit Two	25JUL2009	VS PE BL LB
		26JUL2009	LB
	Visit Three	02AUG2009	VS BL LB
	Visit Four	09AUG2009	VS BL LB
	Termination	16AUG2009	VS PE LB
1027	Visit One	10JUL2009	AD VS PE BL BG MH LB
	Visit Two	25JUL2009	VS PE BL LB
	Termination	01AUG2009	VS PE LB

…

Note that the second record for subject 1026 at Visit Two is not necessarily an error. Subjects do not always behave as desired and it is entirely likely that the subject had some of the lab draw done on the 26th of July while the rest and the other measurements were taken on July 25th. In this case the subjects' Visit Two is of two days duration, which is why the SV data set has both a Visit Start date and End date as two separate fields. There is also no requirement that all of the EVENTID values be the same for a particular visit. For instance, suppose the BL dates have EVENTID values Baseline, PreRand and EndRand. Then the data will look like this:

SUBJID	EVENTID	Date	Sources
1026	Visit One	05JUL2009	AD VS PE BL BG MH LB
	Baseline	25JUL2009	BL
	Visit Two		VS PE
		26JUL2009	LB
	PreRand	02AUG2009	BL
	Visit Three		VS LB
	PostRand	09AUG2009	BL
	Visit Four		VS LB
	Termination	16AUG2009	VS PE LB
1027	Visit One	10JUL2009	AD VS PE BL BG MH LB
	Baseline	25JUL2009	BL
	Visit Two		VS PE LB
	Termination	01AUG2009	VS PE LB

…

All that is important is that you can identify the date sources so that they map uniquely into visit intervals. For instance, dates with EVENTID values Baseline and Visit 2 go into the same interval. Also, problem dates will clearly stand out as anomalous entries, so they can be queried. Suppose SUBJID 1026 had a Baseline date of 25JUL2010. The Baseline EVENTID would sort to be the last entry for the patient, so it would be obvious that there is a date problem. In that case, these type of questionable dates should be removed from the data until it is clear that the date is correct and the patient really did have a visit that took 366 days to complete.

We have said nothing about unscheduled visits up to this point, and that is for a good reason. We are only concerned with scheduled event dates at this time, we will return to unscheduled events later. Do not try to include unscheduled visit days into the process yet, you'll only be making things hard for yourself.

An observer familiar with the variables in SV will note that the above report has pretty much everything we need to start remapping data into the SV data set. If you have already gone over the data, marked questionable values and reported them to Data Management and excluded them from the data set, then we are ready to map our data into SV. SUBJID becomes USUBJID, EVENTID is mapped into VISITNUM and VISIT, and the Date is mapped into SVSTDTC and SVENDTC.

STUDYID	DOMAIN	USUBJID	VISITNUM	VISIT	VISITDY	SVSTDTC	SVENDTC
EX	SV	1026	0	Screening	-20	05JUL2009	05JUL2009
EX	SV	1026	1	Visit 1	1	25JUL2009	26JUL2009
EX	SV	1026	2	Visit 2	8	02AUG2009	02AUG2009
EX	SV	1026	3	Visit 3	15	09AUG2009	09AUG2009
EX	SV	1026	4	Visit 4	22	16AUG2009	16AUG2009
EX	SV	1027	0	Screening	-20	10JUL2009	10JUL2009
EX	SV	1027	1	Visit 1	1	25JUL2009	25JUL2009
EX	SV	1027	2	Visit 2	8	01AUG2009	01AUG2009

A couple of things should be mentioned here. Note that the text in Visit no longer matches what was in EVENTID, nor do the VISITNUM values correspond to the visit numbers as before. This is deliberate and shows what you can do with remapping if you do it in the SV data set. You no longer have to put re-labeling code in every data set program, you'll be able to put it in one program and use the data in a consistent fashion in the other programs. Also, while the variables SVSTDY and SVENDY do not appear here, that is only for space reasons. Relative days are so useful when reviewing data for context issues that they should always be populated. Just remember to use SVSTDTC where VISITNUM = 1 as the reference date in your calculations, as you don't have a DM data set, nor should you have one. The DM data set will use SV as input only.

A word about SVSTDTC and SVENDTC is appropriate here. Note that we are only using dates, not full datetime variables. This is not accidental, and you should avoid populating these variables with datetime values whenever you can. One, nobody starts tracking a subject when they enter the clinic to start a visit, so any time collected for start and/or end is a bit of a stretch. More importantly, the job of assigning Visits and Epochs to individual data sets later in the process will be greatly simplified. There will be times that you cannot avoid a datetime value, such as when a day is split into several visits for whatever reason by your clinical study designers, but these cases are thankfully quite rare and we are certain that you'll be able to account for those intraday visits.

UNSCHEDULED VISITS

Now it is time to address the unscheduled visits that may have been done by your subjects. If there are none, so much the better, but it's a rare trial that doesn't have at least one. Subjects get ill, have elevated lab tests that need to be replicated; the list of potential reasons goes on and on.

For our purposes, the most complicated thing about unscheduled visits is how to map them between the scheduled trial visits. The Case Report Form s are seldom much help, as most DM groups will have a set of generic unscheduled visit Case Report Form s for use with all unscheduled visits, so you can't use the annotation variables to properly order the visits. But, since you have already done the scheduled visits, you have a structure that you can use to in mapping the unscheduled visits.

We always use integers for scheduled VISITNUM values. This allows us to use decimals to properly insert the unscheduled visits between scheduled visits. First, count the number of unscheduled visits per subject and find the largest number that fall between two given scheduled visits. If it's less than 10, then you can sequence your unscheduled visits by taking the integer value of the preceding scheduled visit and adding .1 or .2, etc., to the VISITNUM value, so that unscheduled visits fit between or after scheduled visits. The text for the corresponding VISIT variable is simply Unscheduled Visit 0.1, Unscheduled Visit 2.2, etc. Combine all of the unscheduled visit records with the previously created scheduled visit records and you have a complete SV data set.

STUDYID	DOMAIN	USUBJID	VISITNUM	VISIT	VISITDY	SVSTDTC	SVENDTC
EX	SV	1026	0	Screening	-20	05JUL2009	05JUL2009
EX	SV	1026	1	Visit 1	1	25JUL2009	26JUL2009

EX	SV	1026	2	Visit 2	8	02AUG2009	02AUG2009
EX	SV	1026	3	Visit 3	15	09AUG2009	09AUG2009
EX	SV	1026	4	Visit 4	22	16AUG2009	16AUG2009
EX	SV	1027	0	Screening	-20	10JUL2009	10JUL2009
EX	SV	1027	1	Visit 1	1	25JUL2009	25JUL2009
EX	SV	1027	1.1	Unsched 1.1		27JUL2009	27JUL2009
EX	SV	1027	1.2	Unsched 1.2		29JUL2009	29JUL2009
EX	SV	1027	2	Visit 2	8	01AUG2009	01AUG2009

VISITDY is not populated for unscheduled visits, and the SVUPDES column is omitted. In the case of USUBJID = 1027, we will note that he had extremely elevated liver function tests on his baseline draw. The decision was made to terminate study drug on the 26th for safety reasons, yet he returned for two additional visits to check on the values before finally coming in on August 1st for his termination visit assessments. For this patient, you would enter something "Follow-up Safety Lab" in SVUPDES for the unscheduled visits only.

CREATING SE

With a completed SV data set, it's time to work on creating SE. The secret to creating an SE data set is this; you should be determining almost all of your start and stop points for individual SE records from your visits. If you think about it, visits are usually the boundaries between the study periods. This should be a no-brainer if you have picked your study visit day values properly. For instance, when does the screening period end? It ends the day before the first treatment visit. To populate SE, you should be mostly using SV dates.

The one exception is the end of treatment. Most people will assume that treatment end is marked by the date of the termination visit, but this is not necessarily true. You need to think of the date of termination as a separate entity that floats around and does not necessarily anchor a study period. The end of any treatment period is the last day study drug was taken, so map that value from your raw exposure data, and let the termination date fall where it does. Some people terminate on the last day of treatment, some terminate the next day, some terminate a week later. It's isn't a problem unless you want too make it into one by insisting that termination day must be a boundary date.

DOMAIN	USUBJID	SESEQ	ETCD	ELEMENT	SESTDTC	SEENDTC	EPOCH
SE	1026	0	PRE	Pre-Treatment	05JUL2009	24JUL2009	Pre-Study
SE	1026	1	TITUP	Titration Up	25JUL2009	01AUG2009	Up
SE	1026	2	TRT	Treatment of Concern	02AUG2009	08AUG2009	Controlled Release
SE	1026	3	TITDN	Titration Down	09AUG2009	15AUG2009	Down
SE	1026	4	POST	Post-Treatment	16AUG2009	16AUG2009	Post-Study
SE	1027	0	PRE	Pre-Treatment	10JUL2009	24JUL2009	Pre-Study
SE	1027	1	TITUP	Titration Up	25JUL2009	26JUL2009	Up
SE	1027	2	POST	Post-Treatment	27JUL2009	01AUG2009	Post-Study

Here is what our SE data looks for the two sample patients. Note the STUDYID column was dropped for space reasons, and the TAETORD column does not appear for the same reason. This is not a real problem as there is always a 1-1 map between TAETORD and EPOCH. As mentioned before, the SESTDTC values mostly come from the SVSTDTC column, the exception is the start of the Post-Treatment Epoch. The start of this Epoch is always defined as the day after last day of dosing. In the case of USUBJID 1026, that date coincides with the date of the termination visit, but in the case of USUBJID 1027, it clearly does not.

With the exception of the Post-Study Epoch, all SEENDTC dates are calculated as the day before the first day of the following Epoch. In the Post-Study Epoch, we use the date of last known contact as the ending date. For most subjects, this is the date of termination, but you want to keep that rule flexible to account for any other contacts (post-study adverse events, for instance) that need to be included. A rule of thumb is every date that is reported in the clinical data base should be assignable to one of these intervals except for historical items like Date of Birth, Medical History Occurrence and Prior Medications Start.

VERIFYING THE RESULTS

You are going to need several specialized pieces of code, probably best kept in macro form. The first one converts the SV data set into a one record per subject data set, with VISITNUM, VISIT, VISITDY, SVSTDTC and SVENDTC from all scheduled visits transposed so that they can be processed using arrays. Second, a similar macro for the SE data set, with the SESTDTC, SEENDTC, TAETORD and EPOCH variables transposed. Last is code that merges the transposed SV/SE data onto a data set that contains a date of interest and source, sorted by USUBJID, then checks whether that date lies within the SVSTDTC/SVENDTC interval and/or the SESTDTC/SEENDTC interval. When the date is determined to lie within the SV date interval, the corresponding array values for VISITNUM, VISIT and VISITDY are written to the record. A similar task is performed when the SE date interval is matched, with EPOCH written to the record. At the end, the merged variables are dropped, so you end up with data records that look like the ones below (leaving out TAETORD again):

USUBJID	DATE	SOURCE	VISITNUM	VISIT	VISITDY	EPOCH
1027	14JAN1985	DOB				
	10JUL2009	VS	0	Screening	-20	Pre-Study
…						
	25JUL2009	AE	1	Visit 1	1	Up
	26JUL2009	EX				Up
	01AUG2009	DS	2	Visit 2	8	Post-Study

You should test all known study dates in this fashion, not just the ones we used in mapping the scheduled visits. If you've done everything correctly, each date for a subject should have the correct SV and SE mappings. Note that the subject date of birth maps to nothing, as it shouldn't. The AE date maps to a Visit, but only because it occurs on a visit day, while the last date of study drug dosing maps to an Epoch but not to a Visit.

APPLYING THE RESULTS TO OTHER SDTM DATA SETS

All you need to do is use the three macros that we constructed for verifying the data in your SDTM creation programs. For most data sets, all the macros somewhere towards the bottom of your program, when all other remapping and transformations have been done, and you'll populate the VISITNUM, VISIT, VISITDY, TAETORD and EPOCH variables accurately and consistently. The only exception is the DM data set, you should be looking at specific SESTDTC and SEENDTC values to populate the RFSTDTC and RFENDTC variables.

We'd also like to suggest that you take a look at adding ELEMENT specific visit days to at least a few data sets, especially AE. These visit days are referenced from the start of an EPOCH instead of the start of the study and can be quite useful in answering questions such as how long was the subject on a specific treatment before an event occurred. The CDISC Group should look at adding these variables to the set of timing variables and not push them off to a SUPPQUAL data set.

CONCLUSION

In conclusion, we hope that we've made a good argument for why Sponsors should be taking the creation of the SV and SE data sets more seriously than they currently do. While we'd like all of you to making more use of the SV and SE data sets early on in the process of Clinical Data Review, we hope that you will have at least seen how easy it is to populate these data sets, so that you no longer have a reason not to supply them with your other data in a submission.

REFERENCES

CDISC Study Data Tabulation Model and SDTM IG V3.1.2 at http://www.cdisc.org/sdtm

RECOMMENDED READING

The CDISC message boards at http://www.cdisc.org/discussions/discussions.html.

CONTACT INFORMATION

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

Henry B. Winsor
WinsorWorks, Limited
San Mateo, CA
Email: henry@winsorworks.com

Mario Widel
Genentech Inc.
South San Francisco, CA
Email: mariow@gene.com