PharmaSUG 2020 - Paper DS-062

Untangling the Subject Elements Domain

Christine McNichol, Covance Inc.

ABSTRACT

The Subject Elements (SE) domain is unique and challenging in its sources and mapping. Without much encouragement, a map of the sources, derivations and interaction between records in SE can start to look more like a tangled mass of spaghetti than common linear SDTM mapping. Why is this? SE can have multiple sources for the data points needed to derive each element's start and end. Compared to other domains, SE also has a good deal more direct correlation to values in other domains. For successful implementation, it is critical to understand SE's purpose and requirements, the unique mapping path from source to SE and how this differs from other common domains, the steps needed to successfully derive SE, and programming methods that can be used. This paper will explore the inner workings of SE and explain how to successfully create SE one manageable bite at a time.

INTRODUCTION

Subject Elements (SE) provides a high-level view of which parts of a study a subject has participated in and when. It is extremely helpful to use in the creation of some other Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) data set variables. However, creating the SE domain is a bit more complicated than most other SDTM domains. It is not a domain that contains data that is directly collected on the Case Report Form (CRF) in one or two consolidated places so the flow of the mapping from raw to SDTM is unique. Additionally, there are derivations needed to create SE where most other SDTM domains capture data directly from the CRF. More thought and planning are needed. But an understanding of SE and its intricacies can help in understanding how to go about successfully creating it.

SE BASICS

SE is a special purpose SDTM domain containing one record per subject per study element. Its purpose is to describe the study elements that a subject participated in during the study and the dates associated with those elements. While Trial Arms (TA) and Trial Elements (TE) contain all the planned study elements, SE contains only those that a subject participates in, but also may contain unplanned study elements if one occurs for a subject.

For each subject, there are key SE specific variables derived. Element Code (ETCD) and Description of Element (ELEMENT) are present on each record to describe each study element in which a subject participated. Along with this, the Planned order of Element within Arm (TAETORD) may be present to provide the intended ordering of those elements. For each element present, the Start Date/Time of Element (SESTDTC) and End Date/Time of Element (SEENDTC) of the subject's involvement in that element are captured. The Epoch (EPOCH) that corresponds to the Element may also be defined. If applicable to the subject and study, a Description of Unplanned Element (SEUPDES) may be present in the case where the subject had an element that was not planned and not present in TA and TE.

SE PROGRAMMATIC USES

Once SE is created, it can be used programmatically for several purposes such as the creation of values in other SDTM data sets and input into analysis data sets. SE might be used by other domains to assign EPOCH values to records. It may be used in consistency checking. SE may also feed into a study's Analysis Data Subject Level (ADSL) data set to define the analysis period start and stop dates. Because of these important uses for SE, it is critical that careful thought is put into its creation and that it is created in accordance with any study specific definitions of the study periods.

COVANCE

DATA SOURCES AND MAPPING FLOWS

When creating SDTM domains for the interventions, events, findings, and some special purpose domain classes, there are some common general methods of creation.

THE STACK

When creating events and interventions domains, a stacking of records may be called for. One example may be creating a Medical History (MH) domain by selecting records from a medical history data set and stacking with adverse events that started prior to dosing. Another example could be creating the Concomitant Medications (CM) domain by stacking raw prior medication data with on study medication data. This could also be seen in cases where Exposure (EX) is created by setting raw data sets from multiple administration types, or with Disposition (DS) setting raw subject disposition with informed consent records. In each of these examples, the source data has similar variables and record structure to the target SDTM data set, and the records are being set or stacked on top of each other to construct the final SDTM.

Figure 1 is an example of stacking raw data to create the CM domain.

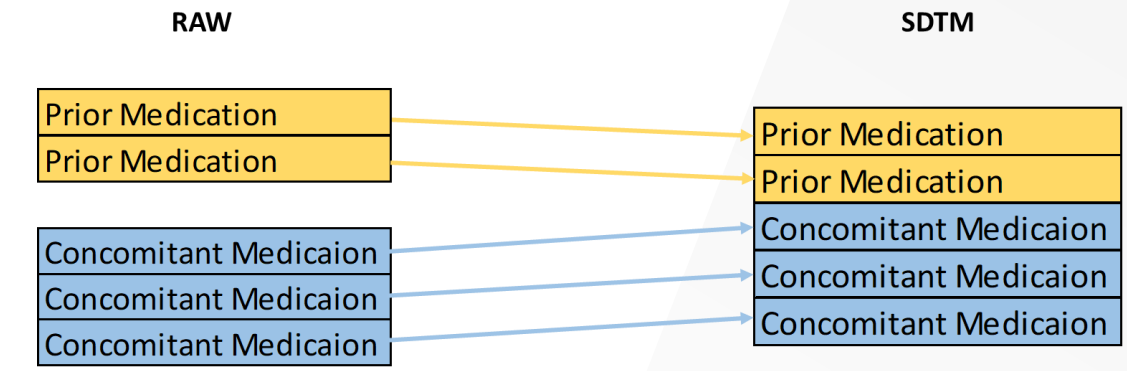

Figure 1. Stacking raw data to create CM.

THE FLIP

Findings domains are in a normalized structure. However, the source data for these are not always in a normalized structure. The source data for the Vital Signs (VS) domain may be in a wide structure instead of normalized having one record per time point with separate variables for the Systolic Blood Pressure, Diastolic Blood Pressure and Weight parameters. To create VS, this data will need to be transposed or flipped such that there is one record per parameter created from this existing record. For this one record with three results in the source data, there will be three records with one result each in the VS domain created. This process may be needed for a variety of domains including VS, LB, IE for example.

Figure 2 is an example of flipping raw data to create the VS domain.

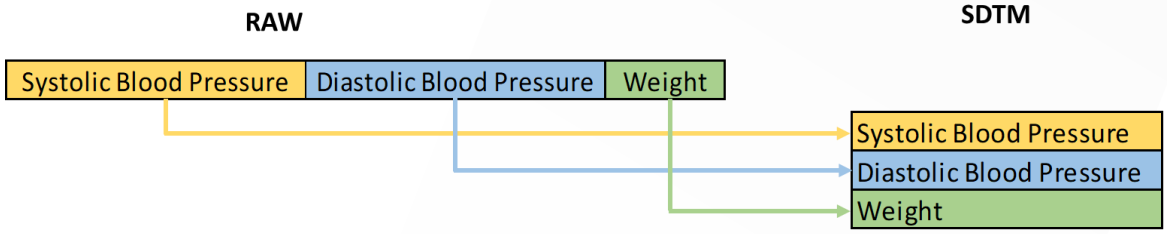

Figure 2. Flipping raw data to create VS.

PARALLEL

Even with the creation of some of the special purpose domains, there is an existing raw data set with a structure that is somewhat parallel to the intended SDTM source structure. For Demographics (DM), a one record per subject raw demographic data set is likely present which can be used as a starting point for the DM structure, and then other values from other source data sets, such as first dose date (RFXSTDTC) from exposure may be added as variables onto that basic existing one record per subject structure. Even the Subject Visits (SV) domain may have a similar raw data structure equivalent. There may be a raw visits data set. Though it may not be exactly the final SV structure with unscheduled visits labeled in a particular way or it may contain only CRF visits and not dates of extra visits from external lab data for example, it has most of the structure that would be present in the final SV domain. Similarly, if CM source data already contains all of the applicable records or if VS data was already in a normalized structure, that would follow a similar parallel creation approach where raw data records would essentially slide into SDTM and the same basic structure would be maintained.

Figure 3 is an example of demographics data with a parallel structure to the DM domain.

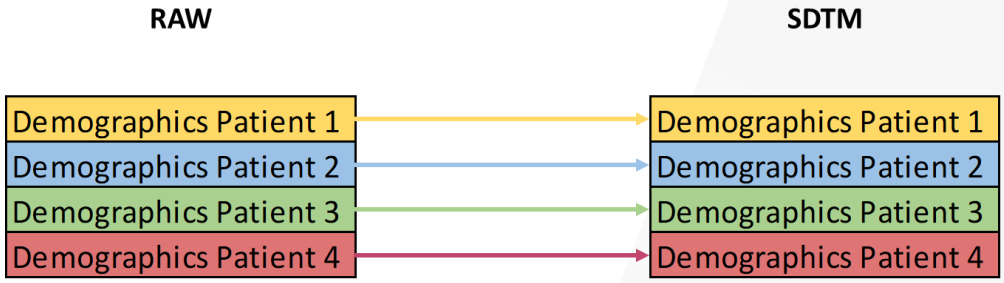

Figure 3. Raw demographics data with a parallel structure to DM.

SE MAPPING FLOW DIFFERENCES

The SE domain rarely has an equivalent raw data set of similar structure. There is not usually a parallel raw data set that can be used as a base for the creation of the SE structure. Most times, SE structure will need to be completely constructed one data point at a time. For this construction of SE, the 'Stack', the 'Flip' and the 'Parallel' style of creation will not work. Instead, SE mapping might be described more accurately as the 'Tangle'. This flow is complicated by the multiple data sources, the need to build the SE structure with references to the Trial Design Module (TDM) data sets Trial Arms (TA) and Trial Elements (TE), derivation of element dates, and potential algorithm complexities.

MULTIPLE DATA SOURCES

SE may not just have one or two raw data sources as is the case with many other domains. For example, it may be necessary to have available some or all of DM or raw demographics data, informed consent date, randomization, first dose, last dose, change of dose dates, dates of key visits beginning a particular study period, treatment/study discontinuation, subject planned visit dates, or dates for common visit specific recordings (lab/vitals).

BUILDING SE STRUCTURE WITH REFERENCES TO TDM

Since there is not an equivalent structure of SE elsewhere, the rows of SE must be identified and constructed. TA will contain the planned elements specific to each planned arm in the study. It must be used to identify the values of the ELEMENT variable and which records we may have for a given subject. TE must then be referenced for a definition of the start and stop of those elements.

DERIVATION OF ELEMENT DATES

Even once the source dates are all identified and the row structure of the ELEMENT values in SE is created, there is still work to be done to get to the final SE. These individual dates may not carry directly into one element's start or stop date. Though a rarity in SDTM mapping, derivations will need to be

applied. Multiple dates will likely go into a derivation for each start and end date and a particular input date may go into several of the start/end date derivations. SESTDTC may look at the minimum of a certain set of dates. SEENDTC may look at the maximum of certain dates with conditions or presence of some dates impacting as well.

Also keep in mind that each of the Elements in SE covers a portion of the subject's trial path. The combination of elements would cover the entire trial experience. So SESTDTC for one element is equivalent to the SEENDTC of the prior element without a gap between elements. So SESTDTC and SEENDTC within the same element and from one element to the next must be cross checked to ensure that they are equivalent and there are neither gaps nor overlaps between elements.

ALGORITHM COMPLEXITIES

In many cases the SESTDTC/SEENDTC values may be defined by a simple derivation looking for a minimum or maximum of raw dates. However, that may not always be the case. Other considerations may need to be made for use or imputation of partial dates, time presence on some elements, multiple raw data sources for similar values or derivation intricacies where a very small change in wording can significantly change the resulting value.

THE TANGLE

Taking these confounding factors into account, the flow of mapping from raw data to SE begins to look more like a spaghetti diagram than the usual SDTM mapping.

Figure 4 is an example of mapping multiple sources to the SE domain.

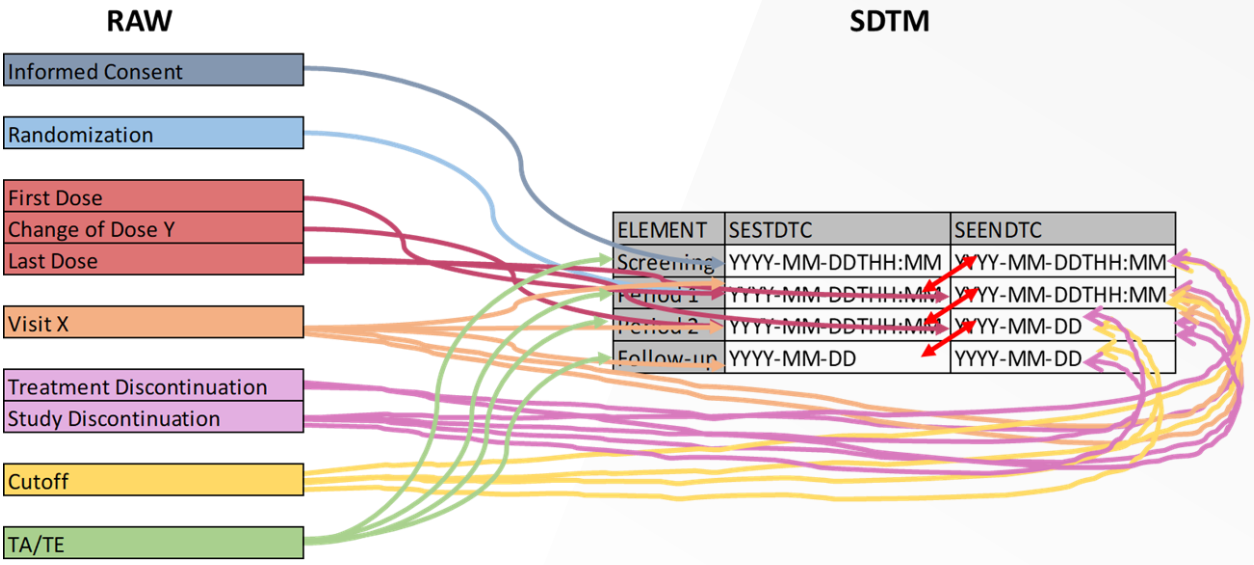

Figure 4. Mapping of SE from multiple sources.

SE CONSTRUCTION

So how do we go about building an SE domain? There are many decisions to make and programming methodologies that can be used but following is one example flow with some decision points and considerations for situations that may arise. At a high level, the two tasks that need to be completed to build SE are the construction of the rows per element per subject and the derivation of the start and end date values.

SE ROWS

Instead of sliding records over from SDTM, commonly SE will need to be constructed. In building SE, the SDTM specified variable structure defines the columns of the data set, but the rows must be built out. DM.ARM can be used to identify the arm in which the subject has been enrolled (ARM).

Figure 5 shows the treatment ARM associated with each subject in the DM domain.

DM

STUDYID	DOMAIN	USUBJID	ARMCD	ARM
ABC	DM	ABC-1	ARMA	TREATMENT ARM A
ABC	DM	ABC-2	ARMB	TREATMENT ARM B

Figure 5. Identification of subject ARM from DM.

Then the records from TA for that ARM value are selected as ELEMENT values that may be present for that subject.

Figure 6 shows each ELEMENT that is associated with the ARM as identified in the TA domain.

TA

STUDYID	DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT
ABC	TA	ARMA	TREATMENT ARM A	1	SCREEN	Screening
ABC	TA	ARMA	TREATMENT ARM A	2	TRT1	Treatment 1
ABC	TA	ARMA	TREATMENT ARM A	3	TRT2	Treatment 2
ABC	TA	ARMA	TREATMENT ARM A	4	FOLLOWUP	Follow-up
ABC	TA	ARMB	TREATMENT ARM B	1	SCREEN	Screening
ABC	TA	ARMB	TREATMENT ARM B	2	TRT3	Treatment 3
ABC	TA	ARMB	TREATMENT ARM B	3	TRT4	Treatment 4
ABC	TA	ARMB	TREATMENT ARM B	4	FOLLOWUP	Follow-up

Figure 6. Identification of each ELEMENT for a given ARM defined in TA.

Rows for each of those ELEMENT values can be created as records for that subject for SE. At this point, all possible planned elements have been created as rows in SE for this subject and there may be more records in SE than are applicable for a subject.

Figure 7 shows the SE structure for a subject built from combining DM and TA.

SE

STUDYID	DOMAIN	USUBJID	SESEQ	ETCD	ELEMENT	SESTDTC	SEENDTC	TAETORD	EPOCH
ABC	SE	ABC-1	1	SCREEN	Screening			1	SCREENING
ABC	SE	ABC-1	2	TRT1	Treatment 1			2	OPEN LABEL TREATMENT
ABC	SE	ABC-1	3	TRT2	Treatment 2			3	BLINDED TREATMENT
ABC	SE	ABC-1	4	FOLLOWUP	Follow-up			4	FOLLOW-UP

Figure 7. SE shell structure created from DM and TA.

The subject's progress into the study will need to be determined to identify which of the possible elements should be present for that subject. That can occur as the SESTDTC/SEENDTC values are derived for each element. If a subject does not enter an element, then that ELEMENT record would not be kept in SE for the subject. The final rows for each subject will be a set or subset of the ETCD and ELEMENT values defined in TA and TE. The exceptions to this are unplanned elements. If a subject has an

unexpected element, there will be an element present in SE that is not defined for that ARM in TA with ETCD set to UNPLAN and SEUPDES would contain a description of this unplanned element.

Now that the necessary columns and rows are identified, the timing of the elements needs to be derived and the SESTDTC/SEENDTC variables can be populated. This is where the bulk of the spaghetti mapping occurs in deriving these values.

SE DATE VALUES

Derivation of SESTDTC/SEENDTC starts with the need to identify what data points will be necessary to derive the dates in SE. TE should contain clear definitions of the rules that define the start of each element and then the rule for the end of the element or the expected duration of the element, but it is a good idea to cross check with the Protocol and Statistical Analysis Plan (SAP) for consistency and clarify of the details in the TE rules. From the element start and end rules in TE, the data points containing the dates necessary for the derivations can be identified. First, last and change of dose dates, informed consent, randomization, select visit dates, discontinuation dates are some of the date values that might be needed to derive the start and end of the elements.

The data points of interest may be found in both the raw data and other SDTM data sets. Using SDTM as input has the benefit of the values being in standard locations with standard format. However, using SDTM as an input to SE means data set dependencies and the consideration of which data sets to use as source will be impacted in part by the downstream use of the SE domain. Use caution that circular logic is avoided. SE may be used in the assignment of EPOCH across other domains. If this is the case, using SDTM as the source for the SE domain would not be possible since the SE domain would be needed before the SDTM that would be needed to derive the start and end dates in SE. For this example, SDTM is used only as a source for TA, TE and subject Treatment Arm from DM. Raw data sources are used for all other values.

In coding, all the date values needed for the derivations can be pulled together. This could be in one record for each subject such that dates can be compared on the same record or across multiple where a sort and select approach might be used. This example will use one row per subject so multiple dates may be compared on the record and it may be a little easier to include more involved derivations. The date values may need processing before being ready for use in SE. For example, all exposure dates may not be necessary, but first, last and change dates may be of interest. The first, last, and change dose dates would need to be selected from all the exposure records and then stored for use.

Once the available dates are together, the comparisons and derivations can be performed, and the derived element start date can be written out to the created SE structure. In general, the SE end dates will be equivalent to the start of the next element. This works nicely when the start of the next element is clearly defined. However, cases where there is not a clear start, the next element has not yet begun, or for the end of the last element, the SE end date would also be derived. Where SE start date may usually be defined by one or two specific occurrences like being either first dose or randomization, SE end date may differ a bit more in its definition. There may be cases of treatment or study discontinuation being the end of an element. An ongoing study may have a cutoff date ending the element. There may be cases where a visit occurring or some date + X days signifies the end of an element.

Programmatically, the derived dates could move from this one record per subject structure to the one record per element per subject structure in several different ways. An output statement from a data step could be used. A transpose could be done. One data set per element could be created and set together. Proc SQL could be used. It is a matter of preference and what works best for the data structure being used.

Once each SESTDTC has been defined and SEENDTC has been set as the start of the next element or defined itself, the unnecessary ELEMENT records per subject can be identified by those elements that did not begin.

CONSIDERATIONS AND CAUTIONS

SUBJECT IDENTIFIER

The target subject identifier that will exist in SE is USUBJID. While this will be present and consistent with the values in DM, it may not be the same for the other data sources. Since SDTM and raw are being mixed, use caution that the subject identifiers used to bring the data together are consistent.

PARTIAL DATES

Most dates that would be used in SE are important to the study and hopefully full and complete. However, watch out for partial dates. If they are present, handling would need to be considered. Is it better to use another date that would be more definitive? Should a window of X days from a known date be used instead? Is imputing appropriate?

TIME PRESENCE

To this point, wording used in this creation process has been in reference to date. Some data points that have been mentioned will be date only. But some of the data points that will feed into SE contain date as well as time. Be aware of the data that is being used and caution for situations such taking a minimum of 2 dates where date is present on one and date/time on the other. Depending on the method of comparison, the date alone value may be selected over the date/time value if the dates are the same. Consider if that is what is intended or the one with time should be selected as more complete. Also use caution that it may be intended for a date/time to be used, but if the time portion is missing on a record, the date/time variable may have a missing value even though date is available.

DATE/TIME FORMAT

Make sure that the same date or date/time format is used for any comparison of dates. For example, comparing a numeric date to a numeric date/time will not likely give the expected results and comparing a numeric date to a character date will not work. Additionally, check that the format makes sense for comparing. Character dates in a DD-MMM-YYYY format cannot be compared in that format. However, in some cases, using character YYYY-MM-DD may be able to be compared.

MULTIPLE SOURCES

Some variables, such as last dose date may have multiple potential sources. There may be dates in the exposure data and treatment discontinuation dates. Within the exposure data on the record containing the last date, there may be both a start date/time and end date/time of the dose. Consider the collection for the study, which is most appropriate, and which source of date may be most reliable as the actual end of dosing required for the algorithm.

DERIVATION INTRICACIES

For some derivations, there may be significant differences in meaning and value from very slight wording differences. Use caution in interpreting the TE and SAP wording.

The maximum of the last treatment date, study discontinuation and cutoff

The maximum of the last treatment date and study discontinuation or cutoff.

Study discontinuation or the maximum of last treatment date and cutoff.

The statements are very similar, but result in different code that could produce results that are quite different:

```
max(LAST_TRT, STUDY_DC, CUTOFF)
max(LAST_TRT, STUDY_DC), else CUTOFF
STUDY_DC, else max(LAST_TRT, CUTOFF)
```

CONCLUSION

Though SE is unique and takes a bit more planning and work to construct than other domain, at a high level it is first building the structure and then filing the start and stop dates by following the derivations. But by understanding where SE comes from, how it is built step by step, and caution areas, managing the sources and intricacies in the algorithms to generate it becomes a very achievable task.

REFERENCES

Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.2

CONTACT INFORMATION

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

Christine McNichol
Covance Inc.
Christine.McNichol@Covance.com