PhUSE 2011

Paper CD02

SDTM Implementation Guide – Clear as Mud: Strategies for Developing Consistent Company Standards

Brian Mabe, UCB Biosciences, Raleigh, USA

ABSTRACT

Many pharmaceutical companies are now entrenched in various CDISC models with SDTM being the most defined. There is even a very detailed implementation guide. One may think, "How hard can it be implementing this?" However, there are many references to sponsor defined fields for many important variables within SDTM. Additionally, how can one report in standard domains using different study models (Double Blind vs. Open Label, Phase 1 vs. Phase 4, etc.). The more one gains experience with SDTM the more apparent that there are many different philosophies for defining the same type of variable. The aim of this paper is to explain the difficulties in defining these vague sponsor-defined terms while being compliant and ensuring consistency across all studies. To achieve this, one of the most efficient methods is to have a company interpretation guide for SDTM that offers solutions to these issues in a consistent manner.

INTRODUCTION

It is now well known that SDTM is the general framework for organizing clinical trials information that is to be submitted to a Regulatory Agency. It is usually described as the study source data. This data provides a standardized platform-independent mechanism for representing all of the essential information collected with intent to easily interpret, understand, and navigate. The main challenge, however, is to consistently interpret the often vague rules and definitions for many SDTM domain variables. This paper will discuss these challenges and offer robust solutions to ensure compliance and consistency within company standards.

STRATEGIES

When tackling the daunting task of implementing these interpretations, there are several challenges that must be addressed:

1) Back to Basics
2) Defining the function of SDTM data – Data in, data out
3) General interpretation rules to apply
4) Controlled Terminology
5) Taking a look at the specific domains
6) Thinking ahead – utilize the benefits of an interpretation of the SDTM implementation guide

BACK TO BASICS

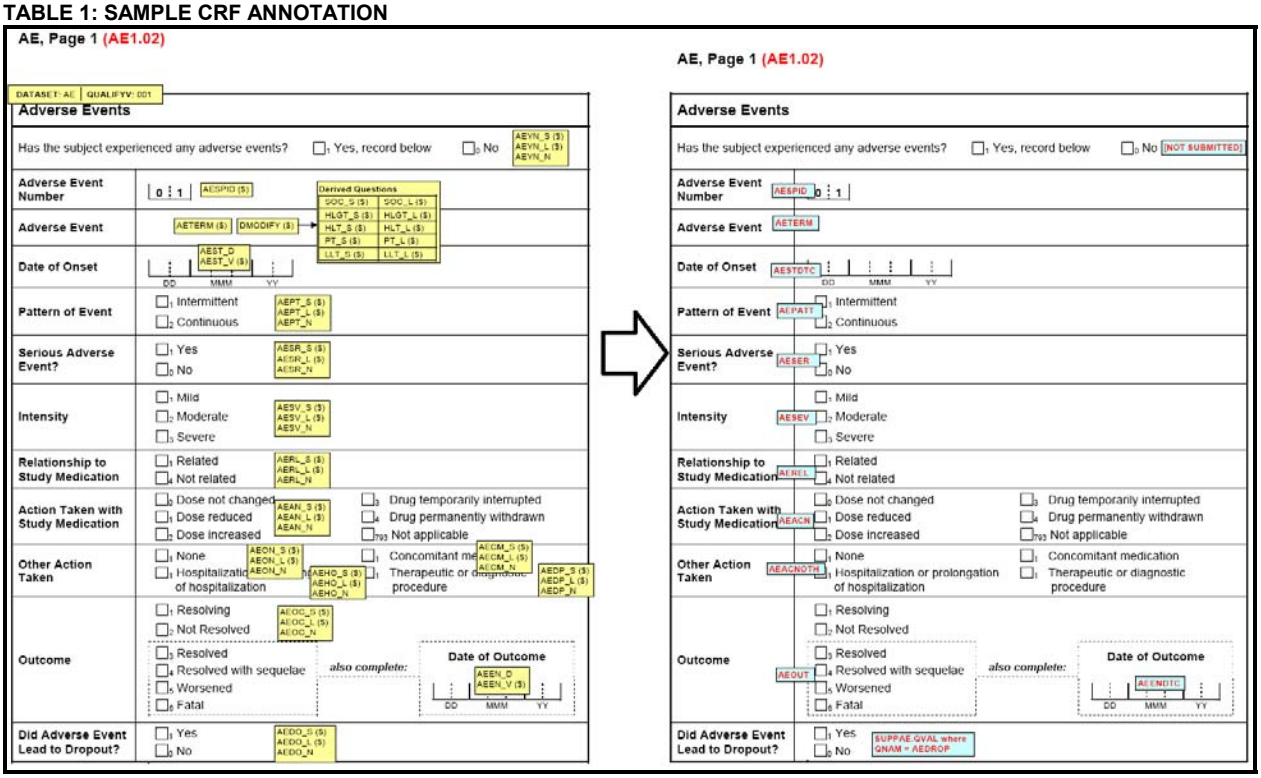
Before one starts to construct an interpretation guide for SDTM, there is the basic need to have a reliable source of reference data. This can only be achieved from a strong foundation: streamlined CRF annotation.

SDTM annotated CRFs are the basis for proper SDTM domain generation. It is crucial that it demonstrates the connection between the CRF and the assigned SDTM variable. If the study design from study to study is similar, then the layout and annotations should also be related. As the foundation, if CRF annotation differs too much from study to study despite similar layouts, then having a streamlined process for creating SDTMs becomes that much more difficult.

Therefore, as a first step, CRF annotation templates should be done at a consistent level where possible. This allows template programs to be developed for the general SDTM rules as outlined in the implementation guide. Another added benefit to standardized CRF annotations is the growing knowledge base for programmers regardless of study. More programmers that are familiar with the streamlined technique can aid when necessary to get a study completed with a higher standard of quality.

Table 1 is an example of an SDTM annotated CRF Adverse Event template. Note that this is a general template that can be applied to a broad range of studies. Again, this helps ensure consistency and a streamlined programming method for efficient work with high quality. The source data annotation to SDTM annotation should almost always be translated the same every time.

TABLE 1: SAMPLE CRF ANNOTATION



DATA-IN, DATA-OUT

As a reminder, SDTM compliant data acts only as the standardized source data for a study. It is strictly for reporting the data, not correcting it; therefore, it is best not to add unnecessary imputations or algorithms that are not reflected on the annotated CRF (aside from the derivations allowed by SDTM and supplemental information). The mantra of "data-in, data-out" fits best here in adapting a consistent SDTM interpretation. It gives way to a simple and concrete approach to develop compliant rules that can be followed at the sponsor level.

GENERAL INTERPRETATION RULES

In developing a sponsor defined interpretation guide, there are several general concepts that can be defined to help clarify the implementation guide while still ensuring compliance.

First and most significantly, communication is the key to a successful interpretation guide. This guide should further clarify exact variable scenarios that are common in sponsor studies as they arise. It cannot happen without the dedication of all of the groups involved in the development of SDTM domains. The better the communication, the more improved the interpretation guide will evolve as experience is gained.

Also, a sponsor defined interpretation guide is not a replacement for the SDTM implementation guide. It is merely a tool to help define vague definitions while still being compliant. As such, the interpretation guide should in no way contradict or challenge the implementation of SDTM.

Another important aspect of the interpretation guide is to remember that SDTM variables generally follow this motto: "Same name, same meaning, same values " That is, SDTM variables that share the same variable names across domains must be identical in all attributes and values. For example, ARM and ARMCD must match exactly from the TA domain to the DM domain. ELEMENT and ETCD from the TA domain to the TE domain must exactly match as well. There are several others, but this is a simple rule that should not be dismissed. It is a good idea to outline and cross-reference these examples in the interpretation guide along with others as a reminder to stay consistent.

As another example, Table 2 demonstrates the clarification of the SDTM origin classification for the respective Data Definition Tables (DDT). The SDTM interpretation guide covers this in Chapter 4 (Origin Metatdata for variables), but the table below simplifies these definitions and relates them to the study level design.

TABLE 2: SAMPLE OF GENERAL INTERPRETATION FOR ORIGIN CLASSIFICATION

CRF variables

Variables are considered to have orginated from the CRF if the value is not changed when it is added to the SDTM dataset, i.e. to assign a format like isodate or controlled terminology is applied.

Derived variables

A variable is derived if at least two values from the CRF or from any reference (e.g. upper limit, standardization) are included for generation, i.e. creation of ~DY variables because both RFSTDTC and ~DTC are used to create this variable.

Note: SDTM should contain as little derived values as reasonable.

Note: ~drvdfl variable marks a complete observation of a domain, i.e. a new record is derived for a dataset, and the source is not eDT. E.g. multiple readings at baseline are averaged to one baseline value in an additional record. (see, IG page 26 and 47)

Assigned variables

If pre-printed information on the CRF is not annotated, but is included in the data, then the origin could be "Assigned" or "Protocol". An example might be "Sitting" (VSPOS) or "Protocol xxxx" (STUDYID).

CONTROLLED TERMINOLOGY

Next, a particular difficult part of the SDTM implementation guide that needs clarification at an internal sponsor level is controlled terminology and the associated codelists. For non-extensible codelists, the sponsor level interpretation guide becomes vital in illustrating how to remap the source values into the SDTM compliant values. Due to different study designs, there will more than likely be more than one way to remap; however, one must know which method to use in each situation. Extensible codelist is even more of a challenge. These particular codelists must be maintained and communicated for every study. The best example of this is that of the variables LBTEST and LBTESTCD. Just about every study from phase 1 – 4 has more and more lab tests that have not been completely cataloged by CDISC. Therefore, they will need to be added by the sponsor. It is actually quite difficult to maintain this codelist as laboratory results could come from different vendors with different standards and values that refer to the same actual test. Therefore, a proper interpretation with accompanying programs needs to be in place to correct and harmonize these tests. This proves quite significant when one later needs to pool studies together for an integrated safety analysis or regulatory responses. Table 3 illustrates the need to capture these extensible codelists in a centralized source document. In this example, the blue highlighted fields are the sponsor defined codelists that are not provided by CDISC.

TABLE 3: SAMPLE EXTENSIBLE CODELIST FOR LBTEST, LBTESTCD

HEMATOLOGY (1) LABPANEL (101) SUBGROUP: WBC (3)				
Last LIC created:	**103263**	**NEUTRO_ABS_MAN_CALC**		
„Question	**„SAS**	**„Unit**	**SDTM Lbtestcd**	**SDTM Lbtest**
ANC_CALC	ANC	G/L	ANCCALC	Calculated Absolute Neutrophil Count
AUER_RODS	AUER	--	AUERRODS	Auer Rods
AUER_RODS_C	AUER_C	*sibling*		
AVG_CD3P	AC3P	%	AVGCD3P	Average CD3+
AVG_CD3P_ABS	AC3PA	cells/uL	AVGCD3PA	Average CD3+ ABS
BASOPHILS	BASO	%	BASOLE	Basophils/Leukocytes
BASOPHILS_ABS	BASOA	G/L	BASO	Basophils
BASO_ABS_MAN_CALC	BASOABMC	G/L	BASOABMC	Absolute Basophils Manual Calculated
BASOPHILS_ABS _O	BASOAO	G/L	BASO	Basophils
BASOPHILS_IMM	BASOIMLE	%	BASOIMLE	Immature Basophils/Leukocytes
BASOPHILS_IMM_ABS	BASOIM	G/L	BASOIM	Immature Basophils
BASOPHILS_MAN_DIFF	BASOMD	%	BASOMD	Basophils Manual Differential
BASOPHILS_MICRO	BASOM	%	BASOMCR	Basophils Microscopy
BASOPHILS_STIPP	BASOS	--	BASOSTP	Basophils Stippling
BASOPHILS_STIPP_QL	BASOSL	--	BASOSTPQ	Basophils Stippling (Qual)
BASOPHILS_STIPP_QL_C	BASOSL_C	*sibling*		
BLASTS	BLAS	%	BLASTLE	Blasts/Leukocytes
BLASTS_ABS	BLASA	G/L	BLAST	Blasts
BLASTS_UNDIFF	BLASUD	%	BLASTUN	Blasts Undifferentiated
CD14P	CD14P	%	CD14P	CD14+
CD14P_86P	CD14P86P	%	CD14CD86	CD14+ CD86+ Monocytes
CD16P_56P	C16P56P	%	C16P56P	CD16+CD56+
CD16P_56P_ABS	C16P56PA	cells/uL	C16P56PA	CD16+CD56+ ABS

A decision will need to be put in place on the frequency of updating dictionaries like WhoDRUG and MedDRA. This also plays a part in properly maintaining datapools and preventing inconsistencies from one study to the next.

Finally, the interpretation guide will play a crucial role in defining and maintaining the sponsor defined codelists. In the SDTM Implementation Guide, the controlled terminology that is marked with a * are sponsor defined. Generally, with these codelists, all terms with the consistent writing style are allowed. Having a proper source of documentation that centrally houses these mapping values will also help in proper interpretations and consistency.

SPECIFIC INTERPRETATION RULES

Once the general guideline has been established for the sponsor defined interpretation guide, more work can be done on the specific issues that plague most studies in terms of consistency.

Key identifier variables can be easily defined in terms of attributes. They should be consistent across all studies. Examples of these types of variables are STUDYID, USUBJID, and the algorithms based on the sequence variables.

Aside from key variables, specific interpretations should be performed on the SDTM core domains. It is up to the sponsor to identify what it considers core, but these usually involve the most common domains found in every study. For example, an EX domain will need to be properly interpreted depending on the type of study design. A monotherapy study will have a different layout than a comparative treatment one, and so on. Developing a set of rules for each type of study design from the sponsor will again help ensure a consistent method of delivery and serve to help pool data together in an easier fashion.

There are many variables within a domain that need extra clarification dependent again on the phase or type of study design. Take a look at each domain individually and address the vague definitions while still staying within SDTM compliance. Having this interpretation guide will again make sure that everyone is on the right track and develops the domains in the most consistent fashion. Again, this is a living document that will continue to evolve as long as communication prevails. Table 4 takes a look at part of the AE domain and helps further define some interpretations to establish consistency.

TABLE 4: EXAMPLE OF SPECIFIC INTERPRETATION RULES: SNAPSHOT OF ADVERSE EVENTS

Variable Name	Variable Label	Type	Term. or Format	CDISC Notes	Core	Sponsor Interpretation	Origin
AESEV	Severity/ Intensity	C	(AESEV)	The severity or intensity of the event. Examples: MILD, MODERATE, SEVERE.	Perm	**Controlled terminology**	
AESER	Serious Event	C	(NY)	Is this a serious event?	Exp	**Controlled terminology:**	
AEACN	Action Taken with Study Treatment	C	(ACN)	Describes changes to the study treatment as a result of the event. AEACN is specifically for the relationship to study treatment. AEACNOTH is for actions unrelated to dose adjustments of study treatment. Examples of AEACN values include ICH E2B values: DRUG WITHDRAWN, DOSE REDUCED, DOSE INCREASED, DOSE NOT CHANGED, UNKNOWN or NOT APPLICABLE	Exp	**Controlled terminology** If AEOUT="DOSE CHANGED" (or something similar), then one must review the actual dosing data to determine if the value should be assigned to the proper controlled terminology of "DOSE INCREASED" or "DOSE REDUCED".	
AEACNOTH	Other Action Taken	C		Describes other actions taken as a result of the event that are unrelated to dose adjustments of study treatment. Usually reported as free text. Example: "TREATMENT UNBLINDED. PRIMARY CARE PHYSICIAN NOTIFIED."	Perm	**Sponsor terminology:** **NONE** **HOSPITALIZATION OR PROLONGATION OF HOSPITALIZATION** **CONCOMITANT MEDICATION** **THERAPEUTIC OR DIAGNOSTIC PROCEDURE** If more than one value was populated, then the value of AEACNOTH = "MULTIPLE" and separate SUPPAE records will need to be populated for each corresponding value Example: MEDICATION and THERAPEUTIC OR DIAGNOSTIC PROCEDURE is checked. In AE, the variable AEACNOTH="MULTIPLE". In SUPPAE, the corresponding values are QNAM="AEACNOT1" and QVAL="MEDICATION" and another record of QNAM="AEACNOT2" and QVAL=" THERAPEUTIC OR DIAGNOSTIC PROCEDURE"	

Variable Name	Variable Label	Type	Term. or Format	CDISC Notes	Core	Sponsor Interpretation	Origin
AEREL	Causality	C	*	Records the investigator's opinion as to the causality of the event to the treatment. ICH E2A and E2B examples include NOT RELATED, UNLIKELY RELATED, POSSIBLY RELATED, RELATED. Controlled Terminology may be defined in the future. Check with regulatory authority for population of this variable	Exp	**Sponsor approved terminology**: **RELATED** **POSSIBLY RELATED** **UNLIKELY RELATED** **NOT RELATED** If the source data is missing (or NA, NR) then that value shall be retained in the SDTM instead of being mapped to one of the above choices.	
AERELNST	Relationship to Non-Study Treatment	C		Records the investigator's opinion as to whether the event may have been due to a treatment other than study drug. May be reported as free text. Example: "MORE LIKELY RELATED TO ASPIRIN USE.".	Perm		
AEPATT	Pattern of Adverse Event	C	*	Used to indicate the pattern of the event over time. Examples: INTERMITTENT, CONTINUOUS, SINGLE EVENT.	Perm	**Sponsor approved terminology:** **INTERMITTENT** **CONTINUOUS**	

BENEFITS OF DEVLOPING A SPONSOR DEFINED SDTM INTERPRETATION GUIDE

When a sponsor defined interpretation guide for SDTM has been developed, one will notice several "quick wins" as a result besides just having an SDTM compliant study with clarified rules and documented interpretation. First, one can develop and streamline programming techniques for many of the SDTM domains through macro or template programs. As experience grows and the documentation improves, the efficiency will as well. Aside from the programming benefits, the SDTM domains produced from these guidelines will be closer to compliance and ready for regulatory submission.

Even more importantly, if these sponsor rules as well as the SDTM implementation standards are followed, it will be theoretically simple to build a repository warehouse that is also SDTM compliant. From this repository, exploratory analysis can be performed as well as periodical safety updates. One can also quickly perform any ad-hoc analyses as well and answer regulatory questions that involve multiple studies knowing that the repository is consistent and with meaningful interpretations.

Having an interpretation guide will facilitate traceability with the source data that will lead into eventual integrated analyses (ISS, ISE). From this, standard variable names, structures, and terminology can be established and will greatly improve the ability to use standard analytical and reporting tools.

CONCLUSION

Although complete and comprehensive, the SDTM Implementation Guide still offers many challenges to the sponsor. However, through specific interpretation strategies, the difficulties in defining the vague areas while still being compliant and ensuring consistency across all studies will lessen considerably. By going back to consistent CRF design and annotation, developing general guidelines that support the implementation guide, and finally tackling the specific properties of the domains themselves, the sponsor can have a technical and consistent solution to the vague definitions while still being SDTM compliant. Having these tools will also aid in building comprehensive SDTM repositories that offer consistent interpretation and serve many functions that will greatly benefit the sponsor.

CONTACT INFORMATION

Brian Mabe
UCB Biosciences, Inc.
8010 Arco Corporate Drive, Ste 100
Raleigh, NC 27617
Phone: +1 919 767 2569
Fax: +1 919 767 2572
Email: brian.mabe@ucb.com
Web: www.ucb.com