PharmaSUG 2020 - Paper 45

Updates in SDTM IG V3.3: What Belongs Where – Practical Examples

Peng Du, William Paget, Lingyun Chen and Todd Case, Vertex Pharmaceuticals, Inc.

ABSTRACT

CDISC SDTM Implementation Guide (IG) Version 3.3 was released on November 20th 2018. New domains and implementation rules have been added to standardize SDTM implementation within the industry. A lot of information has been updated since the release of Version 3.2 five years previously and understanding all these updates presents a great challenge for people working in Pharma/Biotech. For example, what are the new domains and how should we use them? Furthermore, the same information could be mapped to different domains which have slightly different purposes, how should we decide which domain the information should go to?

Version 3.3 also provides additional guidance to resolve some known issues. For example, under Version 3.2 text variables with length greater than 200 characters have the additional text mapped to the corresponding supplemental (SUPP) domains but the labels for these SUPP variables vary within the industry. In addition, it was not clear how to populate the EPOCH variables in the Events, Findings and Interventions domains or how to deal with subjects in the DM domain who were randomized but never dosed.

In this paper, updates in the Version 3.3 IG will be highlighted and examples will be provided to address these questions.

INTRODUCTION

The SDTM IG provides an essential guideline for companies seeking market authorization, with detail on how to prepare the clinical trial tabulation datasets which are included in the submission package sent to regulatory authorities. The IG is prepared and maintained by the Clinical Data Interchange Standards Consortium (CDISC). The first version SDTM IG Version 3.1, was released in 2004. Since then the CDISC Working Group has released six additional versions of the IG, the latest, as mentioned above, being Version 3.3. Currently, many companies are still working with Version 3.2 which was released in 2013. At the meantime, CDISC are continually working with the pharmaceutical industry to add new therapeutic area standards and improve the current standards. The release of Version 3.3 in November 2018 was part of this collaboration.

As SDTM IG 3.3 is still relatively new to the industry, only a couple of papers have provided a very high-level review of the new updates. In this paper, specific examples are given to help readers gain a better understanding of these new updates.

Major updates in SDTM IG 3.3

There are a number of updates in the SDTM IG Version 3.3 and this paper has identified four major categories of updates that we will be focusing on:

1. A new section has been added: Section 9 Study References, which provides structures for representing study-specific terminology used in subject data. This section describes 3 domains, Device Identifiers (DI), Non-host Organism Identifiers (OI), and Pharmacogenomic/Genetic Biomarker Identifiers (PB).
2. New SDTM domains have been added, such as Meal Data (ML) and Procedure Agents (AG) in Interventions domains, Subject Disease Milestones (SM) in Special Purpose Domains, Functional Tests (FT) and a number of other Morphology/Physiology Domains in Findings Domains, and Trial Disease Milestones (TM) in Trial Design domains.
3. New variables have been added into some existing domains (see Table 1).
4. Details have been added clarifying the derivation of existing variables, for example adding more detail on the derivation rules for EPOCH and explaining how to handle more than 200 character strings in SDTM.

SDTM updates

Table 1. Selected List of New Variables Added to Existing Domains.

New variables	Label	Domain
ARMNRS	Reason Arm and/or Actual Arm is Null	DM
ACTARMUD	Description of Unplanned Actual Arm	DM
SPDEVID	Sponsor Device Identifier	EG
EGBEATNO	ECG Beat Number	EG

EGREPNUM	Repetition Number	EG
CMADJ	Reason for Dose Adjustment	CM
CMRSDISC	Reason the Intervention Was Discontinued	CM
MHEVDTYP	Medical History Event Date Type	MH
FOCID	Focus of Study-Specific Interest	MB
--LOBXFL	Last Observation Before Exposure Flag	All Findings domains
EPOCH	Epoch	All Findings and Events domains
TAETORD	Planned Order of Element within Arm	All Findings and Events domains

New Disease Milestone Domains

Two of the new domains that have been added are Subject Disease Milestones (SM) and Trial Disease Milestones (TM) and, in most cases, these two domains should be presented at the same time. SM is a Special Purpose domain and TM is a Trial Design domain. The purpose of these two domains is to highlight the study disease milestones and present information on each disease milestone at the subject level.

Disease milestones are observations or activities which are expected to occur in the course of the disease under observation, and these may occur before the study or during the study. For diagnosis information, SM could extract information from Medical History (MH) and in most cases, for disease milestones, information would be from the Clinical Event (CE) domain. SM and TM can be used to help regulatory agencies to clearly understand the time sequence of the milestone events for the disease, particularly if the protocol includes data collection that is triggered by a disease milestone. For example, SM and TM could potentially be used in oncology and rare disease studies such as Cystic Fibrosis (CF) studies.

I use an example CF study to highlight how these domains can be used in table 2. The below example also highlights the new variable added into the MH domain which is MHEVDTYP, labeled as "Medical History Event Date Type". This newly added variable is used to store the medical history diagnosis or symptoms date.

1. The first part of table 2 shows the TM trial domain dataset. In this study, the key disease milestones occur when the subject is first diagnosed with cystic fibrosis and when the subject has a pulmonary exacerbation event. Since the subject can only be diagnosed once TMRPT is set to "N" for diagnosis records. However, pulmonary exacerbation events can happen at anytime, and more than once, so TMRPT should be populated as "Y" for these records.
2. SM domain, as shown in the second part of the table, is the subject level domain. For each subject the dataset shows the occurrence of the events detailed in TM. The MIDS and MIDSTYPE variables identify the type of disease milestone (diagnosis or pulmonary exacerbation event) and provide the link between TM and SM. From the example, we can tell that MIDS and MIDSTYPE are not one to one related and the number in MIDS is used to show how many pulmonary exacerbation events occurred.
3. The milestones summarized in SM are identified from information originally presented in the MH and CE domains. The MIDS variable can be used to link the information in the MH and CE domains back to the SM domain.

Table 2. Example to show the relationship between Trial Disease Milestone (TM), Subject Disease Milestones (SM), Medical History (MH) and Clinical Event (CE).

ROW	DOMAIN	MIDSTYPE	TMDEF	TMRPT
1	TM	DIAGNOSIS	Initial diagnosis of cystic fibrosis, the first time a physician told the subject they had cystic fibrosis	N
2	TM	PULMONARY EXACERBATION	Pulmonary Exacerbation Event defined according to the Anthonisen definition	Y

ROW	DOMAIN	USUBJID	SMSEQ	MIDS	MIDSTYPE	SMSTDTC	SMENDTC	SMSTDY	SMENDY
1	SM	001	1	DIAG	DIAGNOSIS	2005-10			
2	SM	001	2	PUEX1	PULMONARY EXACERBATION	2013-09-10T11:00	2013-09-10T12:30	25	25

					EVENT				
3	SM	001	3	PUEX2	PULMONARY EXACERBATION EVENT	2013-09-24T11:00	2013-09-24T12:30	39	39
4	SM	002	1	DIAG	DIAGNOSIS	2010-05			

ROW	DOMAIN	USUBJID	MHTERM	MHEVDTYP	MHDTC	MHSTDTC	MHDY	MIDS
1	MH	001	CYSTIC FIBROSIS	DIAGNOSIS	2005-10			DIAG
2	MH	002	CYSTIC FIBROSIS	DIAGNOSIS	2010-05			DIAG

ROW	DOMAIN	USUBJID	CETERM	CECAT	CESTDTC	CEENDTC	MIDS
1	CE	001	PULMONARY EXACERBATION	PULMONARY SIGNS AND SYMPTOMS	2013-09-10T11:00	2013-09-10T12:30	PUEX1
2	CE	001	PULMONARY EXACERBATION	PULMONARY SIGNS AND SYMPTOMS	2013-09-24T11:00	2013-09-24T12:30	PUEX2

*Note: Some variables (e.g. STUDYID) are dropped to aid readability.

Updates to the DM Domain

Two additional variables have been added to Demography (DM) to provide more information on subjects who did not receive the expected study treatments. These two expected variables are ARMNRS and ACTARMUD, and they are listed in table 3.

Table 3. CDISC notes for ARMNRS and ACTARMUD.

Variable Name	Variable Label	Type	Role	CDISC Notes	Core
ARMNRS	Reason Arm and/or Actual Arm is Null	Char	Record Qualifier	A coded reason that Arm variables (ARM and ARMCD) and/or actual Arm variables (ACTARM and ACTARMCD) are null. Examples: "SCREEN FAILURE", "NOT ASSIGNED", "ASSIGNED, NOT TREATED", "UNPLANNED TREATMENT". It is assumed that if the Arm and actual Arm variables are null, the same reason applies to both Arm and actual Arm.	Exp
ACTARMUD	Description of Unplanned Actual Arm	Char	Record Qualifier	A description of actual treatment for a subject who did not receive treatment described in one of the planned trial Arms.	Exp

The new IG gives more detail on how to derive the ARM, ARMCD, ACTARM and ACTARMCD variables. In general, ARM/ARMCD and ACTARM/ACTARMCD in DM must only use ARM/ARMCD values that are defined in the TA domain, or they can be set to null. The one exception to this rule is for studies with multistage randomization where the subject will be randomized two or more times. If ARM/ARMCD or ACTARM/ACTARMCD are null, then ARMNRS must be populated with a reason. If both ARM and ACTARM are blank for the same reason, then it is assumed that ARMNRS explains why both are missing. Table 4 shows different scenarios on how to populate these variables.

1. Subject 001 was randomized to drug A and was dosed with drug A as expected.

2. Subject 002 was a screen failure so was not randomized and received no treatment.
3. Subject 003 passed screening but was not assigned treatment due to some other reason, for example maybe they were not randomized or lost contact etc.
4. Subject 004 was randomized (or if this was a non-randomized study, assigned) to treatment but not treated
5. Subject 005 was randomized to DRUG A but incorrectly received a different study drug, DRUG B. As 'DRUG B' is a study drug it is already included in the TA domain, therefore ACTARMUD can be null because the difference between ARM and ACTARM is enough to explain the situation.
6. Subject 006 was randomized to DRUG A but due to a dosing error they received the correct treatment for Week 1 but received DRUG B during the second week of dosing. As this is not a planned dosing regimen for the study this is not explained in the TA domain. In this case, ARMNRS must be populated as 'UNPLANNED TREATMENT' and ACTARMUD must be populated with a description of the unplanned dose.

Table 4 Example of how to populate ARM information in Demography (DM)

ROW	USUBJID	ARMCD	ARM	ACTARMCD	ACTARM	ARMNRS	ACTARMUD
1	001	A	DRUG A	A	DRUG A		
2	002					SCREEN FAILURE	
3	003					NOT ASSIGNED	
4	004	A	DRUG A			RANDOMIZED, NOT TREATED	
5	005	A	DRUG A	B	DRUG B		
6	006	A	DRUG A			UNPLANNED TREATMENT	WEEK1: DRUG A, WEEK2: DRUG B

*Note: Some variables (e.g. STUDYID and DOMAIN) are dropped to aid readability.

Another update added in V3.3 that affects demographics is that the study population flags (e.g. COMPLT, FULLSET, ITT and SAFETY) will no longer be included in the SDTM domain SUPPDM as these flags are already covered in the ADSL dataset.

Multiple ECG Measures

The ECG Test Results (EG) domain is used to document all ECG findings. Sometimes the protocol specifies that the ECG will be measured multiple times for a subject at the same visit/time point, for example the ECG may be measured in triplicate. To properly map this situation, EGREPNUM (Repetition Number) has been added to the EG domain and is illustrated in Table 5.

Table 5. Example of ECG Test Results (EG).

ROW	USUBJID	EGSEQ	EGTEST	EGORRES	EGORRESU	EGREPNUM	EGDTC
1	001	1	PR Interval, Aggregate	162	msec	1	2019-02-26T07:06
2	001	2	PR Interval, Aggregate	165	msec	2	2019-02-26T07:08
3	001	3	PR Interval, Aggregate	167	msec	3	2019-02-26T07:10

*Note: Some variables (e.g. STUDYID, DOMAIN, EGTESTCD, EPOCH, VISIT and EGDY) are dropped to aid readability.

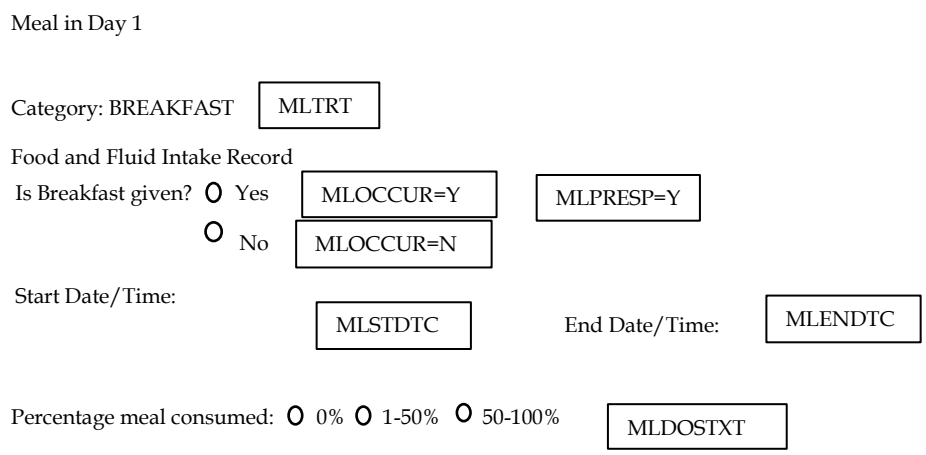
New Intervention Domains

Recording information on the meals that subjects received can be important for studies exploring the effect of food on the pharmacokinetic parameters. Alternatively, recording food taken may be useful for understanding the causes of hypoglycemic events for diabetic subjects. To record this information and cover other scenarios where recording food can be useful, the Meal Data (ML) domain has been added to the Interventions domains. The structure of this domain is one record per food product occurrence or constant intake interval (i.e. meal) per subject. ML should be used to document any fluid or food intake that is not recorded

elsewhere in the SDTM datasets. Information already recorded on the EC/EX, CM, SU or AG domains should not be repeated in ML.

The below example illustrates how the ML domain might be used in a food effect study.

Figure 1: Example annotated meal record CRF page



The associated SDTM dataset is presented below (Table 6). As the meals are pre-specified on the CRF (Figure 1), the variables MLPRESP and MLOCCUR are also included.

Table 6. Example of Meal Data (ML).

ROW	USUBJID	MLSEQ	MLTRT	MLPRESP	MLOCCUR	MLDOSTXT
1	001	1	BREAKFAST	Y	Y	1-50%
2	001	2	LUNCH	Y	Y	50-100%
3	001	3	DINNER	Y	Y	1-50%
ROW	**VISITNUM**	**VISIT**	**MLSTDTC**	**MLENDTC**	**MLSTDY**	**MLENDY**
1 (cont)	30001	DAY 1	2019-05-01T08:20	2019-05-01T08:40	1	1
2 (cont)	30001	DAY 1	2019-05-01T12:20	2019-05-01T12:45	1	1
3 (cont)	30001	DAY 1	2019-05-01T20:20	2019-05-01T20:50	1	1

*Note: Some variables (e.g. STUDYID, DOMAIN and EPOCH) are dropped to aid readability.

MLTRT should be taken verbatim from the meal detailed in the CRF, such as BREAKFAST, LUNCH, DINNER, SNACK. Some early phase studies may also need to distinguish between LOW FAT and HIGH FAT meals.

Procedure Agents (AG) is another new Interventions domain that is used to collect information regarding agents administered to the subjects other than drugs, medications and therapies administered with therapeutic intent. The example given below shows how AG could be used to record the sugar tolerance result for patients who are given a sugar drink. The test is not for therapeutic purpose but is used as a test to check the inclusion and exclusion criteria for each patient at the beginning of the study.

Table 7. Example of Procedure Agents (AG).

ROW	USUBJID	AGSEQ	AGTRT	AGDECOD	AGPRESP	AGOCCUR	AGSTDTC	AGENDTC
1	001	1	SUGAR DRINK	Glucose	Y	Y	2017-11-06	2017-11-08
2	002	2	SUGAR DRINK	Glucose	Y	Y	2017-12-01	2017-12-01

*Note: Some variables (e.g. STUDYID, DOMAIN, VISIT and EPOCH) are dropped to aid readability.

In the new IG, there are 7 Interventions domains (AG, CM, EC, EX, ML, PR and SU). Except for PR, all these domains are used to record food or agents taken by the subjects so care should be taken when selecting the best way to process your data. Each of these

domains is designed to cover different scenarios, depending on the purpose of the administered agent. Some examples from IG include:

1. Investigational nutritional products represented in EC/EX
2. Food or drink used to treat hypoglycemic events represented in CM
3. Glucose given as part of a glucose tolerance test represented in AG
4. Caffeinated drinks represented in SU
5. Meal information used to explore the food effect on drug adsorption represented in ML

Changes to the EPOCH variable

In IG 3.3, EPOCH has been added to most of the SDTM domains except the Special Purpose domains. The one special purpose domain where EPOCH is still defined is the SE domain. This is because in most studies, SE is the source data to derive EPOCH within each SDTM domain. Further information on how to derive the EPOCH variable can be found in Section 4.1.3.1 of IG 3.3.

In general, EPOCH should be derived based on –DTC variables in Findings class domains and --STDTC variables in Interventions or Events class domains. An exception to this rule can be made when a finding based on a specimen collection starts in one epoch and continues into another. In this case it may be more appropriate to assign EPOCH based on the end date. An example where this rule may be used is when urine is immediately collected for PK analyses starting immediately before dosing and continuing after the dose was received. In this case, PCENDTC should be included in the PC domain with the information of urine collection end date/time, and EPOCH should be derived based on PCENDTC, not PCDTC. In addition, EPOCH should be blank if the observation cannot be accurately assigned to an EPOCH or if the observation occurred before the subject started in the study.

Table 8 shows the example of how to derive the epoch in PC (Findings) domain.

1. The blood sample was collected 10 mins before dosing in day 1. EPOCH should be SCREENING. The SCREENING epoch is defined as running from the time ICF is signed to the first dosing.
2. The blood sample was collected 2 hours after dosing in day 1. EPOCH should be TREATMENT.
3. The urine sample collection started immediately before the first dose and continued to 4 hours post dose. EPOCH is derived as "TREATMENT" as PCENDTC is used to calculate the epoch rather than using PCDTC.

Table 8. Example of EPOCH in PC domain.

USUBJID	PCTESTCD	PCORRES	PCORRESU	EPOCH	PCDTC	PCENDTC	PCTPT	PCEVLINT
001	DRUG1	0	ng/ml	SCREENING	2019-02-18T09:50		PREDOSE	
001	DRUG1	8.21	ng/ml	TREATMENT	2019-02-18T12:00		2 HOURS POSTDOSE	
001	DRUG1	4.37	ng/ml	TREATMENT	2019-02-18T09:55	2019-02-18T14:00	PREDOSE	PT4H

*Note: Some variables (e.g. STUDYID, DOMAIN, PCSEQ, PCTEST, PCSTRESN, PCSTRESC, PCSTRESU, PCSPEC, PCDY, PCENDY, PCTPTNUM, PCTPTREF and PCRFTDTC) are dropped to aid readability.

Another update which is not mentioned in Section 4.1.3.1 is that in IG 3.2, EPOCH should be null for records with a DSCAT of "PROTOCOL MILESTONE" in the DS domain. However, in IG 3.3, EPOCH should be populated for every row in the DS domain (Table 9).

Table 9. Example of EPOCH in DS domain.

SDTM IG 3.2							
ROW	USUBJID	DSSEQ	DSTERM	DSDECOD	DSCAT	EPOCH	DSSTDTC
1	001	1	INFORMED CONSENT OBTAINED	INFORMED CONSENT OBTAINED	PROTOCOL MILESTONE		2001-01-12
2	001	2	COMPLETED	COMPLETED	DISPOSITION EVENT	SCREENING	2001-01-14
SDTM IG 3.3							
3	001	1	INFORMED CONSENT OBTAINED	INFORMED CONSENT OBTAINED	PROTOCOL MILESTONE	SCREENING	2001-01-12
4	001	2	COMPLETED	COMPLETED	DISPOSITION EVENT	SCREENING	2001-01-14

*notes STUDYID='XYZ' and DOMAIN='DS' are dropped to aid readability.

Text strings with more than 200 characters in SDTM

The new version of IG also gives further guidance in Section 4.5.3.2 for text strings greater than 200 characters in length. These rules include (Table 10):

1. When splitting a text string into several records, the text should be split between words to improve readability.
2. The SUPP domain QLABEL should be the same value as the main domain variable label.
3. For the CO and TS domains, COVALx and TSVALx variables should be added into the main domain and the labels of the added COVALx and TSVALx variables should be the same as the associated COVAL/TSVAL variable.
4. IETEST should be re-written to reduce the length to below 200 characters. Care should be taken to ensure that the criteria are still clear and that no pertinent information is lost.

Table 10. How to populate the information longer than 200 characters.

General Observation Class & Supplemental Qualifier Variables	**CO. COVAL**	**TS. TSVAL**	**TI. IETEST and IE. IETEST**
The first 200 characters of text should be stored in the variable and each additional 200 characters of text should be stored as a record in the SUPP—dataset.	The first 200 characters of text should be stored in COVAL and each additional 200 characters of text should be stored in COVAL1 to COVALn.	The first 200 characters of text should be stored in TSVAL and each additional 200 characters of text should be stored in TSVAL1 to TSVALn.	If the inclusion/exclusion criteria text is >200 characters, put meaningful text in IETEST and describe the full text in the study metadata.
When splitting a text string into several records, the text should be split between words to improve readability.	When splitting a text string into several records, the text should be split between words to improve readability.	When splitting a text string into several records, the text should be split between words to improve readability.	Not applicable.
The value for QLABEL should be the original domain variable label.	The variable labels for COVAL1 to COVALn should be "Comment".	The variable labels for TSVAL1 to TSVALn should be "Parameter Value".	Not applicable.

New baseline variable added into the Findings domains

In order to ensure a consistent definition of baseline a new –LOBXFL baseline variable has been added in IG 3.3. It is defined as "last non-missing value prior to RFXSTDTC". If –LOBXFL provides a sensible baseline flag, then –BLFL will no longer be needed (Table 11). For example:

- In a simple parallel study –LOBXFL may already provide a sensible definition of baseline so –BLFL would not be included.
- In a drug-drug interaction study where some study drugs are started on Day 1 but others are started later –BLFL may be required for some datasets to provide different definitions of baseline.

Neither --BLFL nor –LOBXFL can be used for analysis purpose (as ABLFL should be used for analyses) but they can aid traceability when they are brought to the ADaM datasets.

Table 11. Example from IG shows baseline flag change in SDTM Findings domains.

Variable	**Structure Where it is Defined**	**Requirement in That Structure**	**Definition**	**Intended Use**
--LOBXFL	SDTM Findings	Expected or Permissible	Last non-missing value prior to RFXSTDTC (Operationally derived)	Consistent pre-treatment reference value baseline for use across all studies and sponsors.
--BLFL	SDTM Findings	Permissible (formerly expected in some domains)	A baseline defined by the sponsor (Could be derived in the same manner as --LOBXFL or ABLFL, but	Any sponsor-defined baseline use

			is not required to be)	
ABLFL	ADaM BDS	Conditionally Required	Flags the record that is the source of the baseline value for a given parameter specified in the Statistical Analysis Plan (May differ both across and within studies and datasets)	Baseline for ADaM analysis as specified in the Statistical Analysis Plan

CONCLUSION

The updated SDTM IG provides a clear guideline to standardize the SDTM domains across most common therapeutic areas and provides many examples under each domain specification to clearly explain each scenario. This paper is intended to highlight some of the changes and help the readers to better understand the updates in IG 3.3. In addition, it summarizes some of the new variables added to existing domains, the new domains and new rules in the updated SDTM IG.

REFERENCES

CDISC SDTM IG : https://www.cdisc.org/standards/foundational/sdtmig

Fred W. (2019) What's New in the SDTMIG v3.3 and the SDTM v1.7 PharmaSUG Proceedings,-Paper 311
https://www.lexjansen.com/pharmasug/2019/DS/PharmaSUG-2019-DS-311.pdf

Contact Information

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

Peng (Lucas) Du

Vertex Pharmaceuticals, Inc

PENG_DU@VRTX.COM

William Paget

Vertex Pharmaceuticals, Inc

WILLIAM_PAGET@VRTX.COM

Lingyun Chen

Vertex Pharmaceuticals, Inc.

LINGYUN_CHEN@VRTX.COM

Todd Case

Vertex Pharmaceuticals, Inc

TODD_CASE@VRTX.COM