PhUSE 2017

Paper DH07

Robust Study Database Build: An Effective Start Towards Creation of Quality SDTMs

Onkar Kajarekar, Tata Consultancy Services, Mumbai, India

ABSTRACT

CDISC standards have become an indispensable part of clinical data across all the pharmaceutical companies. With the emerging trends, companies are exploring multiple possibilities to produce high quality SDTM data. However, with the focus on good quality SDTM domains it is also vital to understand and improvise the different methodologies used to build the database for the purpose of clinical data collection. This paper will give an idea about few key checks that can be programmed during the study build phase which will enable the accurate clinical data entry eventually leading to development of good quality SDTM domains. This is highly beneficial being a proactive step taken at an initial stage ensuring quality data capture. Moreover, this will also benefit at the Statistical Programming level to develop the precise statistical reports enabling the Statisticians to analyze the safety and efficacy of the study drug.

INTRODUCTION

Study build phase is one of the key phases in the lifecycle of any clinical study as it serves as the starting point for a study from the data standpoint. It mainly involves the development of database with a graphical user interface enabling the collection of clinical data in electronic format. Typically, in this phase we have a Data Management Plan (DMP) providing the specifications for the checks that needs to be configured in study build system for a study enabling complete and accurate data entry. This is a very crucial step as it will have a direct impact on the subsequent phases of the study. In this paper, I have covered few key checks that will ensure that the data is precisely populated in the domains enabling display of accurate information in summary reports. Moreover, it will also lead to reduction in overall turnaround time and efforts as the issues would be taken care of at the very initial stage of a study rather than at later phase during the QC of reports in the statistical programming phase.

Check 1: Discrepancy between disposition reason and outcome of AE in case of 'Death'

SCENARIO:

Outcome of AE is 'Fatal' and *Date of Study Completion or Discontinuation* is captured on Subject Disposition form, but *Reason for Study Discontinuation/Completion* on the Subject Disposition form is not entered as 'Death'.

IMPACT:

This check will have an impact on both AE and DS SDTM domains. Please see below the snapshots of the raw datasets containing adverse event and Subject disposition information:

Table 1 – Adverse Event raw dataset

STUDY	SITE	PT	AERAW	AEOUT	AESEV	AESD	AEST	AEED	AEET
Study1	101	1001	VOMITING	FATAL	SEVERE	02OCT2010	9:30	02OCT2010	10:30

Table 2 – Disposition raw dataset

STUDY	SITE	PT	DISCRES	DISCDT
Study1	101	1001	SUBJECT WITHDREW	02OCT2010
Study1	101	1002	COMPLETED	03OCT2010

Adverse event information from raw dataset is mapped to AE domain whereas the Disposition information is mapped to DS domain. Please see below the snapshot:

Table 3 – AE SDTM domain

USUBJID	AETERM	AEOUT	AESEV	AESTDTC	AEENDTC
STUDY1-101-1001	VOMITING	FATAL	SEVERE	2010-10-02T09:30	2010-10-02T10:30

Table 4 – DS SDTM domain

USUBJID	DSCAT	DSTERM	DSDECOD	DSDTC
STUDY1-101-1001	DISPOSITION EVENT	SUBJECT WITHDREW	WITHDRAWAL BY SUBJECT	2010-10-02
STUDY1-101-1002	DISPOSITION EVENT	COMPLETED	COMPLETED	2010-10-03

Ideally, for a particular adverse event if a subject has the outcome as 'FATAL' then the subject should have the discontinuation reason as 'DEATH' in DS domain. However, in aforementioned scenario the discontinuation reason is populated as 'SUBJECT WITHDREW' because it has been captured that way in the raw dataset. This can go unnoticed if there is no check implemented during the study build phase to ensure that the discontinuation reason is entered based on the outcome of the adverse event. Eventually, this incorrect data entry will be reflected at SDTM level in AE and DS domain. As a result, the counts for 'Death' in Subject Disposition table won't match with the data represented in Adverse Event listing. Below are the snapshots of patient disposition table and adverse event listing:

Summary 1: Patient Disposition

Category	Total (N=2) n (%)
Completed Study	1 (50%)
Discontinued from Study	1 (50%)
Withdrawal by Subject	1 (50%)
Lost to Follow-up	0
Death	0

Listing 1: Listing of Adverse Events

Site/Subject	AE term	Outcome	Severity	Start Date	End Date
101/1001	Vomiting	Fatal	Severe	2010-10-02T09:30	2010-10-02T10:30

However, this issue can be avoided with the implementation of mentioned check during the build phase itself. Once the check is implemented, the query will be automatically fired in case of incorrect data entry as depicted in the below snapshot:

Figure 1 - Subject Disposition form

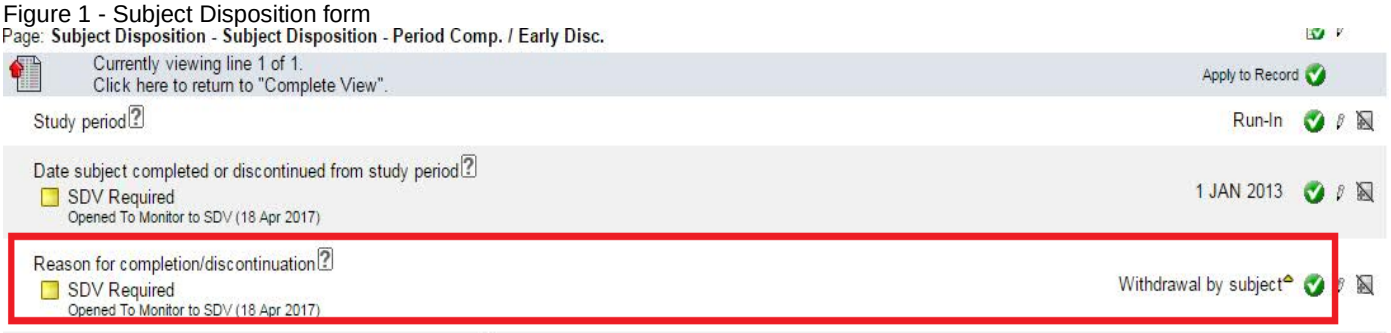
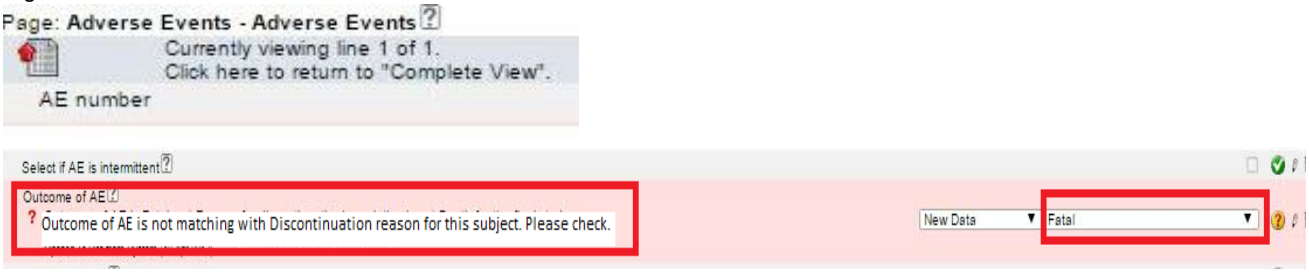


Figure 2 - Adverse Event form



This will enable the data entry team to query the issue to the corresponding site right in the beginning. The raw disposition dataset and corresponding DS domain will look as follows after accurate data entry:

Table 5 – Disposition raw dataset

STUDY	SITE	PT	DISCRES	DISCDT
Study1	101	1001	DEATH	02OCT2010

Table 6 – DS SDTM domain

USUBJID	DSCAT	DSTERM	DSDECOD	DSDTC
STUDY1-101-1001	DISPOSITION EVENT	DEATH	DEATH	2010-10-02

Additionally, following are the tables created on the updated data in which the data is clearly aligned:

Summary 2: Patient Disposition

Category	Total (N=2) n (%)
Completed Study	1 (50%)
Discontinued from Study	1 (50%)
Adverse Event	0
Lost to Follow-up	0
Death	1 (50%)

Listing 2: Listing of Adverse Events

Site/Subject	AE term	Outcome	Severity	Start Date	End Date
101/1001	Vomiting	Fatal	Severe	2010-10-02T09:30	2010-10-02T10:30

Check 2: Invalid date values across domains

SCENARIO:
At Screening, *Result Collection Date* for a specific parameter must be on or before the *First Study Drug Administration Date* recorded on Study Drug Administration form.

IMPACT:

Let's consider the example of Vital Signs parameter here. In this case, the check will have an impact on VS and EX SDTM domains. Please see below the snapshots of the raw datasets containing vital signs and study drug administration information:

Table 7 – Vital Signs raw dataset

STUDY	SITE	PT	SYSBP	DIABP	PULSE	TEMP	VTLDT	VTLTM	VISIT
Study1	101	1001	121	80	65	38	10OCT2010	9:30	SCREENING

Table 8 – Study Drug Administration raw dataset

STUDY	SITE	PT	TRT	DOSE	TSTDT	TSTTM
Study1	101	1001	Treatment1	100	02OCT2010	10:00
Study1	101	1001	Treatment1	100	03OCT2010	10:30
Study1	101	1001	Treatment1	100	04OCT2010	10:00
Study1	101	1001	Treatment1	100	05OCT2010	10:30

Vital signs information will be mapped to VS domain whereas Study drug administration information is mapped to EX domain. Additionally, first study drug administration date is captured in RFSTDTC variable in DM domain. Please see below the snapshots:

Table 9 – VS SDTM domain

USUBJID	VSTESTCD	VSTEST	VSDTC	VSDY	VISIT
STUDY1-101-1001	SYSBP	Systolic Blood Pressure	2010-10-10T09:30	8	SCREENING
STUDY1-101-1001	DIABP	Diastolic Blood Pressure	2010-10-10T09:30	8	SCREENING
STUDY1-101-1001	PULSE	Pulse Rate	2010-10-10T09:30	8	SCREENING
STUDY1-101-1001	TEMP	Temperature	2010-10-10T09:30	8	SCREENING

Table 10 – EX SDTM domain

USUBJID	EXTRT	EXDOSE	EXSTDTC
STUDY1-101-1001	TREATMENT1	100	2010-10-02T10:00
STUDY1-101-1001	TREATMENT1	200	2010-10-03T10:30
STUDY1-101-1001	TREATMENT1	300	2010-10-04T10:00
STUDY1-101-1001	TREATMENT1	300	2010-10-05T10:30

Table 11 – DM SDTM domain

USUBJID	ARMCD	ARM	RFSTDTC
STUDY1-101-1001	TRT1	Treatment 1	2010-10-02T10:00

In this scenario, the first treatment start date for subject 1001 is 02OCT2010. The Screening date should not be after first drug administration date however in the vital signs dataset we have the date incorrectly recorded as 10OCT2010 which is post first drug administration date. This is the common issue that we usually encounter in most of the studies. In any study, the dates are usually expected to be in chronological order for all the visits in that study. However, in the aforementioned scenario we can observe that there is an issue with the dates due to incorrect data entry. If we closely look at the dates entered we can see that this might have happened due to a typo while entering the date for SCREENING visit in vital signs data i.e. 01OCT2010 might have been accidentally entered as 10OCT2010. This eventually causes an impact at the SDTM level leading to incorrect Study day (VSDY) value derivation for screening visit. Ideally, the VSDY value for screening visit should not be >1 however due to incorrect data entry the value is >1. Moreover, this will also have an impact at the statistical programming level wherein the listing of Vital Signs parameter will display the invalid study day values.

Once the check is implemented it will result in issue getting flagged and queried at the data build stage itself. The resultant, corrected raw and SDTM vital signs data sets are as shown in table 12 and table 13 below:

Table 12 – Vital Signs raw dataset

STUDY	SITE	PT	SYSBP	DIABP	PULSE	TEMP	VTLDT	VTLTM	VISIT
Study1	101	1001	121	80	65	38	01OCT2010	9:30	SCREENING

Table 13 – VS SDTM domain

USUBJID	VSTESTCD	VSTEST	VSDTC	VSDY	VISIT
STUDY1-101-1001	SYSBP	Systolic Blood Pressure	2010-10-01T09:30	-1	SCREENING
STUDY1-101-1001	DIABP	Diastolic Blood Pressure	2010-10-01T09:30	-1	SCREENING
STUDY1-101-1001	PULSE	Pulse Rate	2010-10-01T09:30	-1	SCREENING
STUDY1-101-1001	TEMP	Temperature	2010-10-01T09:30	-1	SCREENING

Check 3: Incorrect dose values which are not aligned with the study design

SCENARIO:
Dose administered values entered on Study Drug Administration form should be based on the study design and should match with the pre-defined dose values mentioned in the protocol.

IMPACT:
This will have a direct impact on EX SDTM domain and the corresponding analysis datasets and tables. Consider a double-blind, parallel, Phase III multi-dose study which has 2 treatment arms viz. TRT-A and TRT-B. Each arm consists of single drug administration viz. DRUG-A and DRUG-B in TRT-A and TRT-B arm respectively. As per the study design, the expected dosing values for both the drugs are 100, 200 and 300 mg. During the study build phase, suppose free text is allowed to enter the dose values on study drug administration form instead of restricting the dosing values to 100, 200 and 300. In this case, please see below the snapshot of the raw dataset containing the study drug administration information and corresponding EX SDTM domain:

Table 14 – Study Drug Administration raw dataset

STUDY	SITE	PT	TRT	DOSE	TSTDT	TSTTM
Study1	101	1001	Drug A	100	02OCT2010	10:00
Study1	101	1001	Drug A	200	03OCT2010	10:30
Study1	101	1002	Drug B	3000	04OCT2010	10:00
Study1	101	1002	Drug B	3000	05OCT2010	10:30

Table 15 – EX SDTM domain

USUBJID	EXTRT	EXDOSE	EXSTDTC
STUDY1-101-1001	DRUG-A	100	2010-10-02T10:00
STUDY1-101-1001	DRUG-A	200	2010-10-03T10:30
STUDY1-101-1002	DRUG-B	3000	2010-10-04T10:00
STUDY1-101-1002	DRUG-B	3000	2010-10-05T10:30

In this scenario, the unusual values flow down to SDTM level eventually causing issues at reporting level. Therefore, it is very crucial to have the check implemented at the build level that will defy the entry of invalid dose values in the database. This is also essential as the dosing kits allotted to subjects in any study always contain the defined doses based on the study design thus eliminating the possibility of invalid dose administration to a subject.

Once the check is implemented at the study build phase the raw study drug administration dataset and corresponding exposure domain will look as follows:

Table 16 – Study Drug Administration raw dataset

STUDY	SITE	PT	TRT	DOSE	TSTDT	TSTTM
Study1	101	1001	Drug A	100	02OCT2010	10:00
Study1	101	1001	Drug A	200	03OCT2010	10:30
Study1	101	1002	Drug B	300	04OCT2010	10:00
Study1	101	1002	Drug B	300	05OCT2010	10:30

Table 17 – EX SDTM domain

USUBJID	EXTRT	EXDOSE	EXSTDTC
STUDY1-101-1001	DRUG-A	100	2010-10-02T10:00
STUDY1-101-1001	DRUG-A	200	2010-10-03T10:30
STUDY1-101-1002	DRUG-B	300	2010-10-04T10:00
STUDY1-101-1002	DRUG-B	300	2010-10-05T10:30

Check 4: Population of logical values in dependent fields based on the value populated in corresponding independent variable

SCENARIO:
This check can be implemented for multiple forms. Here we are considering an example of Alcohol use form wherein a subject doesn't have any current or previous alcohol use history however the number of units consumed by the subject per day or week are provided or vice versa.

IMPACT:
This will have an impact on SU SDTM domain. Please see below the snapshots of the raw datasets containing substance use information and corresponding SU domain:

Table 18 – Alcohol use raw dataset

STUDY	SITE	PT	ALCHIS	QTY	FREQ	DURATION	DURUNIT
Study1	101	1001	NEVER	50	Daily	1	YEARS

Table 19 – SU SDTM domain

USUBJID	SUOCCUR	QTY	FREQ	SUDUR
STUDY1-101-1001	N	50	QD	1

In this scenario, the variable ALCHIS specifying the 'alcohol use history' is an independent variable which will specify whether the subject had any history of alcohol use. Ideally, if the subject has not consumed alcohol in past or present then the variables capturing the quantity and frequency of consumption should be blank. However, in the aforementioned scenario the quantity and frequency values have been provided although the subject doesn't have any alcohol history. This will have an impact on 'Demographics and Baseline Characteristics' table in which we summarize the subject's alcohol consumption history. In the present scenario, if we implement the check it will ensure the valid data entry in two ways:

- If the subject has no alcohol consumption data in past or present (ALCHIS = 'NEVER') then it will not allow the data entry in the dependent fields i.e. quantity, frequency and so on.
- If the subject has alcohol consumption data in past or present (ALCHIS = 'CURRENT' or ALCHIS = 'PREVIOUS') then it will ensure that the data is provided for quantity and frequency of alcohol consumption.

Once the check is implemented at the study build phase the raw dataset containing substance use information and corresponding SU domain will look as follows:

Table 20 – Alcohol use raw dataset

STUDY	SITE	PT	ALCHIS	QTY	FREQ	DURATION	DURUNIT
Study1	101	1001	CURRENT	50	Daily	1	YEARS

Table 21 – SU SDTM domain

USUBJID	SUOCCUR	QTY	FREQ	SUDUR
STUDY1-101-1001	Y	50	QD	1

Check 5: Missing response despite of subject's availability at a particular visit

SCENARIO:
This check also can be implemented for multiple forms. Here we are considering an example of Laboratory Test Results form wherein at any given visit if result collection date is provided then response must be present for all the lab tests.

IMPACT:
This is the most common issue that we generally observe in any clinical study which will have an impact on LB SDTM domain. Please see below the snapshots of the raw datasets containing laboratory test results information and corresponding LB domain:

Table 22 – Laboratory Test Results raw dataset

STUDY	SITE	PT	COLDT	COLTM	SODIUM	POTASSIUM	CHLORIDE	VISIT
Study1	101	1001	10JUN2010	09:30				VISIT 1

Table 23 – LB SDTM domain

USUBJID	LBTESTCD	LBTEST	LBORRES	VISITNUM	VISIT	LBDTC
STUDY1-101-1001	SODIUM	Sodium		2	Visit 1	2010-06-10T09:30
STUDY1-101-1001	K	Potassium		2	Visit 1	2010-06-10T09:30
STUDY1-101-1001	CL	Chloride		2	Visit 1	2010-06-10T09:30

In this scenario, the lab collection date (COLDT) has been entered in the database which indicates that subject had appeared on Visit 1 in the laboratory to perform the necessary tests. Therefore, it is expected that the test result values should be non-missing for all the specified tests. However, since the check is not implemented at the study build phase it allowed the entry of missing values for the lab test results which is incorrect. This will have an impact on the summary of 'Laboratory test results' in which subject 1001 will not be counted at Visit 1 for any of the lab tests. However, with this check being implemented it will ensure that result values are strictly entered in the database. Moreover, it will also enable the investigator to assess the reason behind unavailability of result values and take appropriate actions.

Once the check is implemented at the study build phase the raw datasets containing laboratory test results information and corresponding LB domain will look as follows:

Table 24 – Laboratory Test Results raw dataset

STUDY	SITE	PT	COLDT	COLTM	SODIUM	POTASSIUM	CHLORIDE	VISIT
Study1	101	1001	10JUN2010	09:30	140	4	100	VISIT 1

Table 25 – LB SDTM domain

USUBJID	LBTESTCD	LBTEST	LBORRES	VISITNUM	VISIT	LBDTC
STUDY1-101-1001	SODIUM	Sodium	140	2	Visit 1	2010-06-10T09:30
STUDY1-101-1001	K	Potassium	4	2	Visit 1	2010-06-10T09:30
STUDY1-101-1001	CL	Chloride	100	2	Visit 1	2010-06-10T09:30

This way we can develop multiple checks at the Study Build level to ensure enhanced quality of clinical data entry. Here after I have explained few more checks in brief.

Check 6: Pregnancy test form should be available for data entry only for the female subjects

SCENARIO:
This is one of the basic check that needs to be implemented in every clinical study. For Male subjects, the pregnancy test form shouldn't be active for data entry as it is applicable for female subjects.

IMPACT:
This will lead to incorrect data being reflected in the laboratory table summarizing the 'Pregnancy' information.

Check 7: Informed consent date must be before the first Study drug administration date.

SCENARIO:
In any study, the study drug is administered to a subject only after the Informed consent is obtained from the subject and it also satisfies all the Inclusion/Exclusion criteria. Therefore. It is expected that the date when informed consent is obtained should certainly be before the first study drug administration date.

IMPACT:
If the given check is implemented, it will ensure that study drug administration date on every form is after the informed consent date.

Check 8: If subject discontinued from the study due to 'Adverse Event', then Reason for discontinuation entered in Subject Disposition form must be 'Adverse Event'

SCENARIO:
If a subject discontinued the study due to and AE then the study discontinuation reason for that subject must only be 'Adverse Event'. Similarly, if the study discontinuation reason for a subject has been entered as 'Adverse Event' then it will also ensure that subject has at least 1 corresponding record present in the Adverse Event dataset.

IMPACT:
This check will ensure the bilateral relationship between the Adverse Event and Study Discontinuation forms. It will guarantee that the data entered on both the forms correlate with each other.

Check 9: If the dose of study drug has been modified for a subject then reason for dose modification should be provided

SCENARIO:
Ideally, a subject should receive the study drug based on the planned doses of the treatment arm to which the subject has been randomized. However, during the course of the study a subject's dose value may be modified especially if a subject has experienced an 'Adverse Event' that needs dose reduction or subject has received an incorrect dose. Therefore, it is crucial to enter the dose modification reason that helps to identify the purpose of dose modification.

IMPACT:
This check will ensure the completeness of the dosing data for a subject.

Check 10: Medication Stop date on concomitant medication form is not provided and the flag indicating if medication is ongoing is also missing

SCENARIO:
In concomitant medications form, we typically have the fields to collect the start and stop dates of the medication and a field to check if the medication is ongoing. Usually, if a medication is still ongoing during the study then stop date should not be entered on the form and vice versa. However, at times we identify issues in data wherein both the fields have missing values which is incorrect.

IMPACT:
This check will ensure that at least one of the fields is populated in order to provide more clarity about the status of the concomitant medication administration.

CONCLUSION

Drug development process involves a huge financial investment by all the pharmaceutical companies globally. It is very certain that all the trials are closely monitored in order to ensure that accurate results are obtained from both efficacy and financial standpoint. Therefore, Data Quality apparently becomes most essential and most critical aspect of any clinical trial. This paper enables to understand some key checks which if implemented at initial stage will certainly contribute significantly to achieve the ultimate goal of achieving the enhanced data quality in efficient way.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank all my colleagues who have helped me in realizing my ideas and in refining this paper with their invaluable comments. My sincere gratitude to Neha Mohan, Usha Kumar and Kanchan Pawar for all their help.

CONTACT INFORMATION

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

Onkar Kajarekar
Tata Consultancy Services
Empire Plaza, 4th floor, Lal Bahadur Shastri Marg,
Chandan Nagar, Vikhroli West,
Mumbai, Maharashtra 400083
Work Phone: +91-22-67786057
Email: onkar.kajarekar@tcs.com
Web: www.tcs.com