

Building Efficiency and Quality in SDTM Development Cycle

Kishore Pothuri, Vita Data Sciences, Waltham, MA

Bhavin Busa, Vita Data Sciences, Waltham, MA

ABSTRACT

The typical cycle for generating compliant CDISC SDTM datasets is to develop mapping specifications, programming and QC of the domains, followed by checking these domains for compliance issues per the FDA requirements. In some cases, the compliance checks of the SDTM domains take place towards the end of the study (i.e. closer to or after the database lock). This seems like a logical approach to the programming of the datasets, however checking for compliance and addressing issues identified by Pinnacle 21 validator once the programming is complete could result in re-work. A programmer will come across three possible scenarios: 1) compliance issues that will require programmer to make update to their dataset programs, 2) compliance issues that will require update to the dataset mapping specifications, SDTM aCRF, and to their dataset programs, and 3) compliance issues that will not require any update but will need explanation in the SDRG. In scenarios 1 and 2, the back and forth process results in programmer spending more time with the SDTM dataset development than originally estimated. In addition, depending on the changes made to the SDTM datasets to address compliance issues, it also affects the quality and the timing of the datasets and analysis deliverables (ADaM and TLFs) down-stream. In this paper, we recommend programmer to understand the compliance requirements upfront and implement those during mapping specifications development. This will avoid scenarios 1 and 2 as explained above and will result in building efficiency as well as quality during SDTM development cycle.

INTRODUCTION

The standardized clinical study datasets will be required in submissions for clinical and non-clinical studies that start on or after December 17, 2016 [1]. As noted, it will be expected that all the trials conducted after that date must use study data standards that are listed in the FDA Data Standards Catalog (DSC). This means that all studies going forward must utilize CDISC SDTM and ADaM standards for their tabulation and analysis datasets respectively. It is understood by the industry that this will greatly facilitate the FDA's ability to process, review and archive data into the Clinical Trial Repository (CTR) data warehouse.

On November 18, 2014, FDA published its first set of official validation rules for the CDISC SDTM datasets. These rules cover both the conformance as well as the quality requirements for the submitted CDISC SDTM datasets. As specified in the FDA Study Data Technical Conformance Guide, "data validation is a process that attempts to ensure that submitted data are both compliant and useful. Compliant means the data conform to the applicable and required data standards. Useful means that the data support the intended use (i.e., regulatory review and analysis)".

It is important to understand that SDTM development is not just about implementation of the CDISC SDTM standards but it is also about the conformance of the datasets on the validation rules released by the FDA.

In addition, one should also understand what is expected from the FDA to stay current with the FDA specific validation rules. As specified in the FDA Study Data Technical Conformance Guide, "Sponsors should validate their study data before submission using the most recently published validation rules and either correct any validation errors or explain in the Reviewer's Guide (SDRG or ADRG) why certain validation errors could not be corrected. The recommended pre-submission validation step is intended to minimize the presence of validation errors at the time of submission."

With that said, FDA has been closely collaborating with the industry (Pinnacle21) to implement conformance rules in an open source tool (OpenCDISC validator) which can be used widely by the industry to check their CDISC SDTM and ADaM datasets for compliance. In this paper, we will talk about the impact of checking for conformance during the SDTM development life cycle and the need for the programmer to understand the compliance requirements upfront during mapping specifications and datasets development/QC.

TYPICAL SDTM DEVELOPMENT LIFE CYCLE

The Figure 1 below demonstrates typical SDTM development life cycle. The number annotated on each box represents the process and sequence in which they are developed and implemented. Upon development and QC of the SDTM datasets, a programmer passes the datasets through Pinnacle21 validator to check for the compliance. In cases where the study specific define.xml is available, it is also passed through the validator along with the SDTM datasets. This ensures the datasets and the define document are in sync and that both the components are checked at the same time for the compliance. Also one should note that every time SDTM datasets are generated during the

study period (interim data snapshot, blinded data review, dry run TLFs, DMC, etc.), one may or may not run their datasets through Pinnacle21 validator to check for compliance in an ongoing manner. However, if one chooses to run each time, this additional step of checking for the compliance adds up significant amount of time to the SDTM development life cycle.

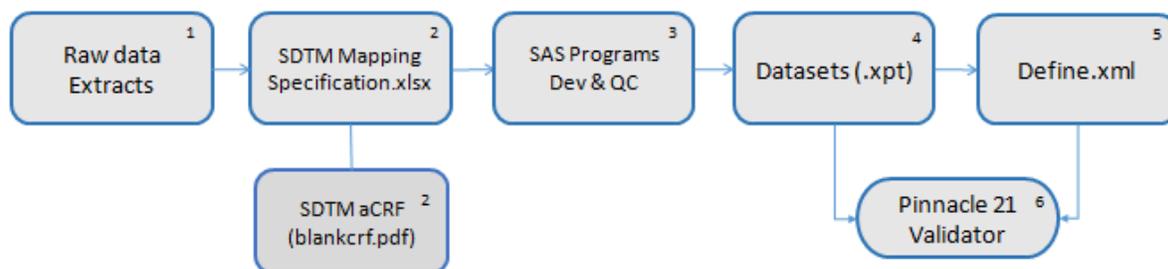


Figure 1. Typical SDTM Development Process Flow

DATASET COMPLIANCE SCENARIOS IMPACTING SDTM DEVELOPMENT LIFE CYCLE

In this section, we have provided three main scenarios that a programmer comes across during the validation and compliance checking process of the SDTM datasets using Pinnacle21 validator.

Scenario 1: Dataset compliance issues (errors/warnings) that require update to the dataset programming (dev & QC) but not the SDTM mapping specifications and/or the define.xml [2].

Scenario 2: Dataset compliance issues (errors/warnings) that require update to the mapping specifications and/or define.xml followed by dataset programming.

Scenario 3: Dataset compliance issues (errors/warnings) that require no update either to the mapping specifications, define.xml or dataset programming as these could be due to issues with the data.

SCENARIO 1

Issue Summary						
Source	Pinnacle 21 ID	Publisher ID	Message	Severity	Found	
DM						
	SD0006	FDAC113	No baseline result in VS for subject	Warning	18	
PE						
	SD1023	FDAC105	VISIT/VISITNUM values do not match TV domain data	Warning	2	
IE						
	SD0005	FDAC044	Duplicate value for IESEQ variable	Error	2	

Display 1. Examples for Scenario 1 (Update to the Development/QC Programs)

Below are the possible reasons for the Errors/Warnings:

1. No baseline result in VS for subject: All subjects who are not screen failures should have at least one baseline observation in VS domain.
2. VISIT/VISITNUM values do not match TV domain data: This warning is due to the mismatch of VISIT and VISITNUM with that of values present in TV domain.
3. Duplicate value for IESEQ variable: This Error is due to the Sequence variable having duplicate values. Sequence should have a unique value for each record present in the dataset.

In order to address the above scenarios, programming updates for both Development (Dev) and QC is required followed by generation of transport files before re-validation of the datasets using Pinnacle 21 validator. Below is the schematic representation of the above scenario:

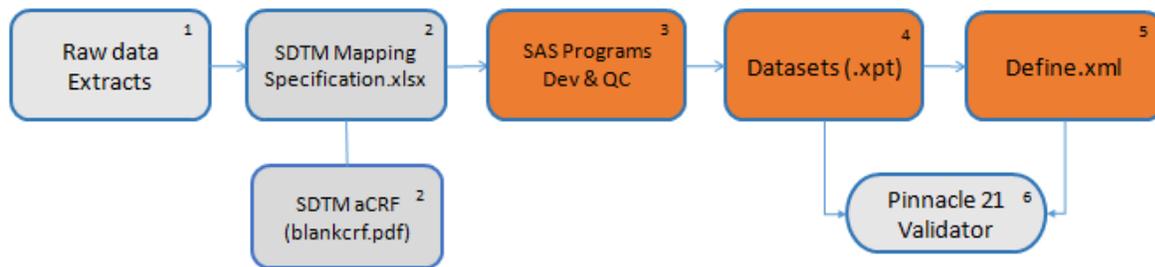


Figure 2. SDTM Development Process Flow for Scenario 1 where update to the development and QC programming is required (highlighted in orange)

Examples from Scenario 1	Domain(s) Affected	Type of Update	Number of Additional Hours to Address the Issue
Example 1: No baseline result in VS for subject	Demographics (DM)	Dev + QC Programming	3
Example 2: VISIT/VISITNUM values do not match TV domain data	All Findings domains	Dev + QC Programming	8
Example3: Duplicate value for IESEQ variable	Inclusion/Exclusion Criteria Not Met (IE)	Dev + QC Programming	2
Overall Additional Hours			13

Table 1. Matrix summarizing additional hours taken by programmers to fix the compliance issues for the 3 examples specified in Scenario 1 in addition to the time taken to program the domains

SCENARIO 2

Issue Summary					
Source	Pinnacle 21 ID	Publisher ID	Message	Severity	Found
TS	SD2017	FDAC115	Missing TSVAL value	Error	4
SE	SD1087	FDAC128	Missing SESTDY variable, when SESTDTC variable is present	Error	1
CM	SD1102	FDAC143	Missing CMENRTPT variable, when CMMENTPT variable is present	Error	1

Display 2. Examples for Scenario 2 (Update to the mapping specification, define.xml and Development/QC Programs)

Below are the possible reasons for the Errors/Warnings:

1. Missing TSVAL: TSVAL can only be missing when Parameter Null Flavor (TSVALNF) variable value is populated.
2. Missing SESTDY variable, when SESTDTC variable is present: SESTDY must be populated when SESTDTC is populated for a record.
3. Missing CMENRTPT variable, when CMMENTPT variable is present: This is due to having CMMENTPT variable without CMENRTPT variable in the specification.

In order to address the above scenarios, mapping specifications should be updated to capture the missing information, re-generation of define.xml, followed by programming updates for both Dev and QC before validation of the datasets using Pinnacle 21 validator. Below is the schematic representation of the above scenario:

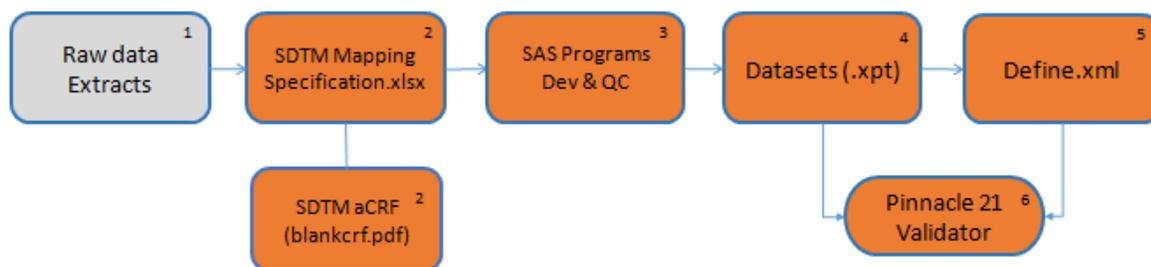


Figure 3. SDTM Development Process Flow for Scenario 2 where update to the mapping specifications, SDTM aCRF, Dev and QC programming, and define.xml is required (highlighted orange)

Examples from Scenario 2	Domain(s) Affected	Type of Update	Number of Additional Hours to Address the Issue
Example 1: Missing TSVAL	Trial Summary (TS)	Spec + define.xml + Dev + QC Programming	3
Example 2: Missing SESTDY variable, when SESTDTC variable is present	Subject Elements (SE)	Spec + define.xml + Dev + QC Programming	6
Example 3: Missing CMENRTPT variable, when CMENTPT variable is present	Concomitant Medications (CM)	Spec + define.xml + Dev + QC Programming	5
Overall Additional Hours			14

Table 2. Matrix summarizing additional hours taken by programmers to fix the compliance issues for the 3 examples specified in Scenario 2 in addition to the time taken to program the domains

SCENARIO 3

Issue Summary					
Source	Pinnacle 21 ID	Publisher ID	Message	Severity	Found
AE	SD0002	FDAC018	NULL value in AEDECOD variable marked as Required	Error	1
EG	SD1124	FDAC176	Missing value for EGREASND, when EGSTAT is 'NOT DONE'	Warning	1
QS	SD0048	FDAC179	Value for QSORRES is populated, when QSSTAT is 'NOT DONE'	Warning	42

Display 3. Examples for Scenario 3 (No Update to the mapping specification, define.xml and Development/QC Programs)

Above are some of the issues which need to be explained with a proper logical justification and documented in the Study Data Reviewer's Guide (SDRG). As and when the data issues are resolved (if at all), no update is required to the mapping specifications and CRF annotations.

Examples from Scenario 3	Domain(s) Affected	Type of Update	Number of Additional Hours to Address the Issue
Example 1: NULL value in AEDECOD variable marked as Required.	Adverse Events (AE)	Justification in Pinnacle21 Report / SDRG	1
Example 2: Missing value for EGREASND, when EGSTAT is 'NOT DONE'	ECG Results (EG)	Justification in Pinnacle21 Report / SDRG	1

Examples from Scenario 3	Domain(s) Affected	Type of Update	Number of Additional Hours to Address the Issue
Example 3: Value for QSORRES is populated, when QSSTAT is 'NOT DONE'	Questionnaires (QS)	Justification in Pinnacle21 Report / SDRG	1
Overall Additional Hours			3

Table 3. Matrix summarizing additional hours taken by programmers to fix the compliance issues for the 3 examples specified in Scenario 3 in addition to the time taken to program the domains

CONCLUSION

In scenarios 1 and 2, the back and forth process results in programmer spending more time with the SDTM dataset development than originally estimated. As noted in Tables 1 and 2, addressing compliance issues results in additional hours in the SDTM development life cycle. In the 9 examples provided across all 3 scenarios, programmers spent 30 additional hours on top of time spent initially to program these domains. If one extrapolates this to a real case scenario for a Phase 3 study, addressing compliance issues could add up significant time to the project and affect overall project timeline and cost.

In addition, depending on the changes made to the SDTM datasets to address compliance issues, it also affects the quality and the timing of the datasets and analysis deliverables (ADaM and TLFs) down-stream. Based on our assessment, we recommend programmer to understand the compliance requirements upfront and implement those during mapping specifications and datasets development stages. This will avoid scenarios 1 and 2 as explained above and will result in building efficiency as well as quality during SDTM development cycle.

REFERENCES

- [1] Ron, Fitzmartin and Ginny, Hussong. "Required Electronic Submissions to CDER / CBER." fda.gov. 10-08-2015 Available at <http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/SmallBusinessAssistance/UCM467501.pdf>
- [2] Sakampally, Vara Prasad and Busa, Bhavin. "Consider Define.xml generation during development of CDISC dataset mapping specifications – AD18." Proceedings of PharmaSUG 2016 Conference.

RECOMMENDED READING

- FDA Data Standards Catalog (DSC)
- FDA Data Standards Strategy

CONTACT INFORMATION

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

Name: Kishore Pothuri
 Enterprise: Vita Data Sciences
 Address: 281 Winter St. Suite 100
 City, State ZIP: Waltham, MA 02451
 Phone: 818-536-4612
 Fax: 781-466-9681
 E-mail: kpothuri@softworldinc.com
 Web: www.softworldinc.com

Name: Bhavin Busa
 Enterprise: Vita Data Sciences, Division of Softworld, Inc.
 Address: 281 Winter St. Suite 100
 City, State ZIP: Waltham, MA 02451
 Work Phone: 781-373-8455
 Fax: 781-466-9681
 E-mail: bbusa@softworldinc.com
 Web: www.softworldinc.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.