PhUSE 2015

Paper CD08

Improving CDISC SDTM Data Quality & Compliance Right from the Beginning

Bharat Chaudhary and Padamsimh Balekundri
Cytel, Pune, India

ABSTRACT

Pharmaceutical companies are proactively adopting CDISC SDTM because of a broad variety of benefits to the industry and to the FDA. There are some tools and applications available to validate the submission ready SDTM datasets and results of this submission level validation might require us to rework from the root level. So why can't we have some utilities to check the quality and compliance during the each stage of SDTM development process to avoid the rework?

This paper describes three utilities to improve the quality and compliance of CDISC SDTM data creation process at different stages with real life examples.

1) SDTM Spec Checker (SSC): Verifies the SDTM mapping specification against the SDTM IG standards especially at variable level attributes like variable name, core, label, length & CT etc.

2) SDTM SAS Dataset Checker (SDC): This helps in checking data issues beyond CDISC compliance to ensure overall quality of SDTM dataset.

3) OpenCDISC Report Reviewer (OCR): This helps in speedy and efficient review of OpenCDISC reports by summarizing the report results.

INTRODUCTION

REGULATORY, INDUSTRY AND CDISC

The regulatory agencies have endorsed CDISC SDTM as the preferred model for submitting Study data in the eCTD guidance. Regulatory institutes are actively collaborating with the CDISC team to develop and promote fast industry adoption of the SDTM, and have announced a forthcoming update to the regulation that would mandate Pharmaceutical organizations to use standardized data structure and terminology according to the CDISC SDTM IG.

Industries are quickly adopting the SDTM model both because of regulatory submission necessities and benefits to the industry in collaborating and sharing standardized data. Now industry becomes increasingly involved in developing well-organized and cost effective ways to produce CDISC SDTM-compliant clinical trial domains. It is essential to develop tools or utilities to check compliance and data issues for the planning of submission-ready files in accordance with the SDTM IG. Industries will benefit from the validation of the compliance of data before a submission to the regulatory, and will probably further benefit by performing the validation as early as possible in the SDTM development process. Non-compliance in the SDTM submission impacts when loading the data into the FDA site. In addition, any discrepancies or variations in the data may delay the submission review process, and may result in increased costs to the sponsor.

SDTM DEVELOPMENT PROCESS AND VALIDATION

Data collected from a clinical study is organized and presented using the Study Data Tabulation Model (SDTM), a data structure standard created by the Clinical Data Interchange Standards Consortium (CDISC). SDTM is beneficial for clinical data reviewers and analysts as it brings standardization in presenting the data. Analysis Data Model (ADaM) datasets, which support the safety and efficacy analysis of clinical studies, are created from SDTM data. ADaM data is used to produce tables, listings, and figures, which together with SDTM and ADaM datasets are submitted to regulatory bodies, such as the FDA, for use in evaluating new clinical trials.

To ensure the quality of SDTM submission ready datasets, a free application called OpenCDISC Validator is typically used to generate a report. This report focuses on the implementation of CDISC standards, such as whether the dataset contains SDTM compliant variable names, types, labels, and code lists at the time of submission. But this report does not check the results for outliers, unexpected results, or sites for reporting in the wrong unit, which might need rework from root level. However, if there are any discrepancies in the submission; then datasets are rejected and error report is shared with sponsor. This can cause significant delay to the submission review process, and may result in increased. To avoid rework we have created in-house utilities based on SAS ® and MS Excel that check compliance as well as metadata issues during each stage of SDTM development process.

Sequence of Applying Utilities in the SDTM Development

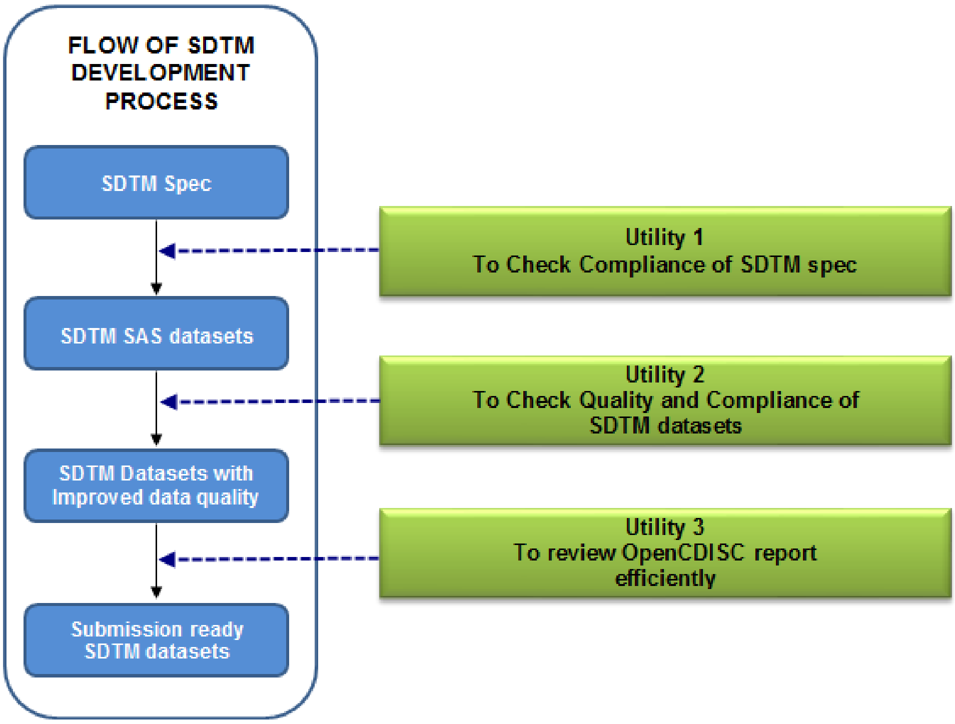

FLOW OF SDTM DEVELOPMENT PROCESS
SDTM Spec
SDTM SAS datasets
SDTM Datasets with Improved data quality
Submission ready SDTM datasets
Utility 1
To Check Compliance of SDTM spec
Utility 2
To Check Quality and Compliance of SDTM datasets
Utility 3
To review OpenCDISC report efficiently

Figure 1: Flow of SDTM Development process and In-house validation utilities

NEED OF SDTM VALIDATION IN-HOUSE UTILITIES

It will be beneficial to sponsors and CRO's to have a validation utility that sits right before or during the every steps of SDTM development process, which can avoid rework and produce better quality of SDTM data. Hence it is crucial to identify a validation utility upfront and chose the one that fits sponsor needs and necessities towards successful submission of a study in SDTM format. Below we have provided needs of a validation utility which can vary from sponsor to sponsor.

The need of validation utilities:

- Check SDTM compliance for domain datasets.
- Identify structural and consistency errors in SDTM domains at early stage.
- Identify outliers and unexpected results.
- Reduce risk of delay in the submission review process.
- Versatile for use throughout the SDTM development process.
- Reduce the time and cost of SDTM development process.

UTILITY 1: SDTM SPEC CHECKER

FRAMEWORK

A very crucial part of the entire process is defining the SDTM mapping specification; it begins with annotations of the CRF according to the SDTM structures by analyzing the raw data. This process is time consuming, doing it first time right will save time & efforts. This can be achieved by using the SDTM Spec Checker utility as it identifies the SDTM compliance issues in the spec level itself before we start off with programming for SDTM datasets.

The goal of this utility is to identify the SDTM variable level compliance issues in the first draft of SDTM specification itself to avoid errors/warning during/after programming the SDTM dataset and to avoid multiple versions of SDTM mapping specification and this can be achieved using the SDTM Spec Checker.

The advantage of this utility is, it will review the specification without actually editing the original specification and also generates compliance graph (Figure 3) and report (Figure 4), which will help us in identifying the SDTM compliance rate for each domain and entire specification.

As you can see in the below (Figure 2) of single window view of SDTM Spec Checker, all three sections are merged into single window so we can have better view of SDTM IG standards, actual spec data and results. This will helps us in understanding the spec issues at one go.

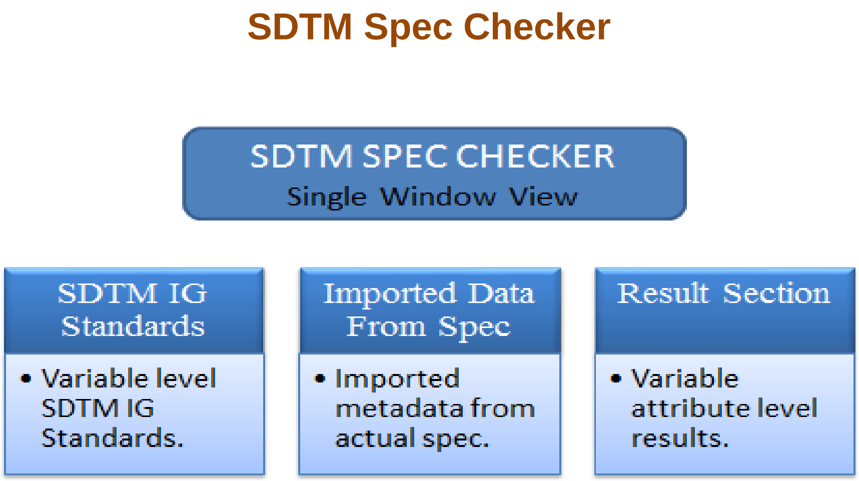

SDTM Spec Checker
SDTM SPEC CHECKER
Single Window View
SDTM IG Standards
• Variable level SDTM IG Standards.
Imported Data From Spec
• Imported metadata from actual spec.
Result Section
• Variable attribute level results.

Figure 2: Framework of SDTM Spec Checker

SDTM IG STANDARDS

The SDTM IG standards will be placed in the main window of utility to verify your results in the single window. Currently the utility is designed for SDTM IG version 3.1.2 and can be designed to all other versions if required. Also we can have SDTM specification verified against all versions at the same time.

IMPORTED DATA FROM SPEC

This section will have the imported metadata from the actual specification and will be placed between SDTM IG standards and result section, so if there is any mismatch found you needn't to go back to verify values specified in the specification. You can just filter out the domain or variable you want to view and verify the results against actual spec metadata and SDTM IG standards.

RESULT SECTION

This section will have results for each SDTM variable and its attributes like label, length and type etc.

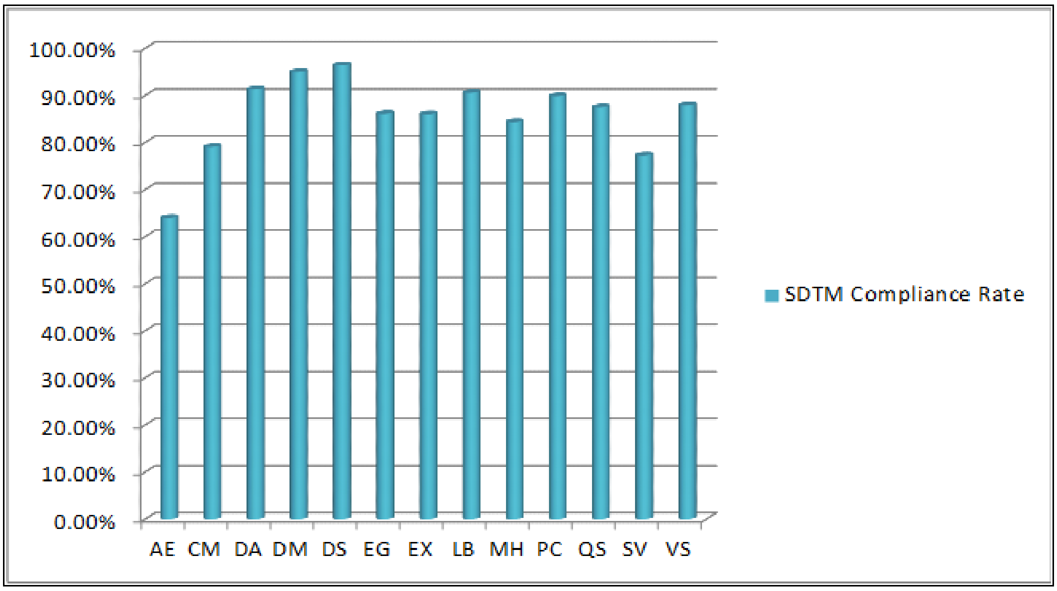
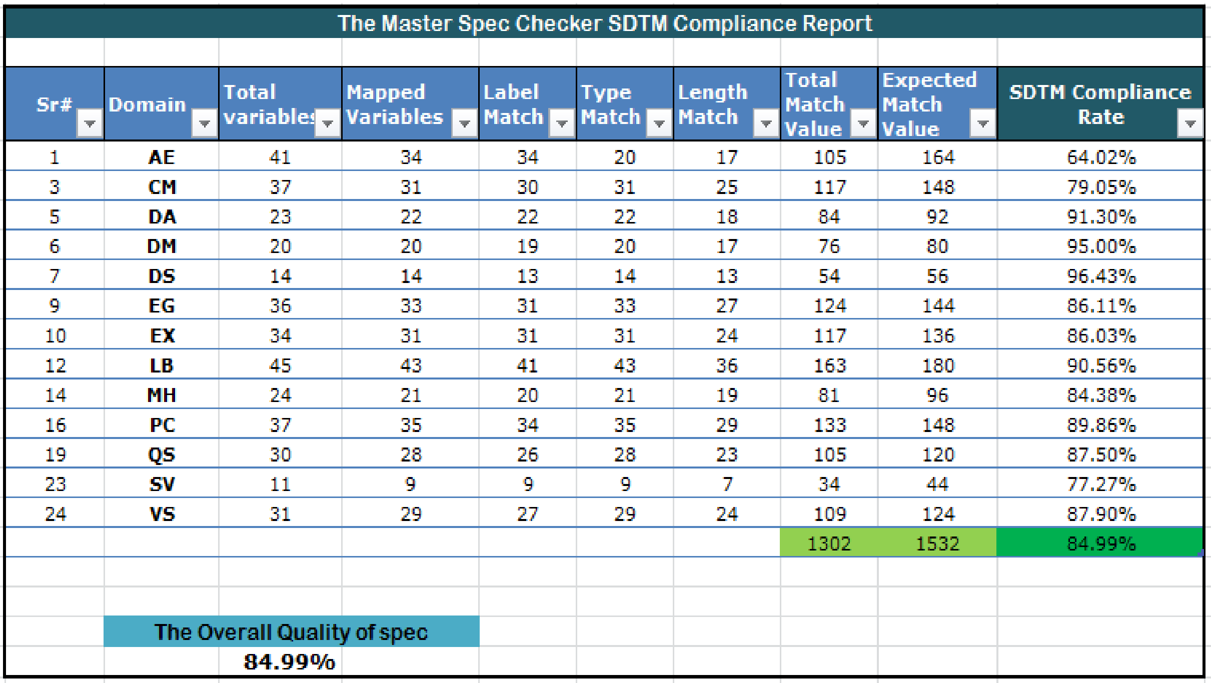


Figure 3: SDTM Compliance Rate



The Master Spec Checker SDTM Compliance Report

Sr#	Domain	Total variables	Mapped Variables	Label Match	Type Match	Length Match	Total Match Value	Expected Match Value	SDTM Compliance Rate
1	AE	41	34	34	20	17	105	164	64.02%
3	CM	37	31	30	31	25	117	148	79.05%
5	DA	23	22	22	22	18	84	92	91.30%
6	DM	20	20	19	20	17	76	80	95.00%
7	DS	14	14	13	14	13	54	56	96.43%
9	EG	36	33	31	33	27	124	144	86.11%
10	EX	34	31	31	31	24	117	136	86.03%
12	LB	45	43	41	43	36	163	180	90.56%
14	MH	24	21	20	21	19	81	96	84.38%
16	PC	37	35	34	35	29	133	148	89.86%
19	QS	30	28	26	28	23	105	120	87.50%
23	SV	11	9	9	9	7	34	44	77.27%
24	VS	31	29	27	29	24	109	124	87.90%
							1302	1532	84.99%

The Overall Quality of spec

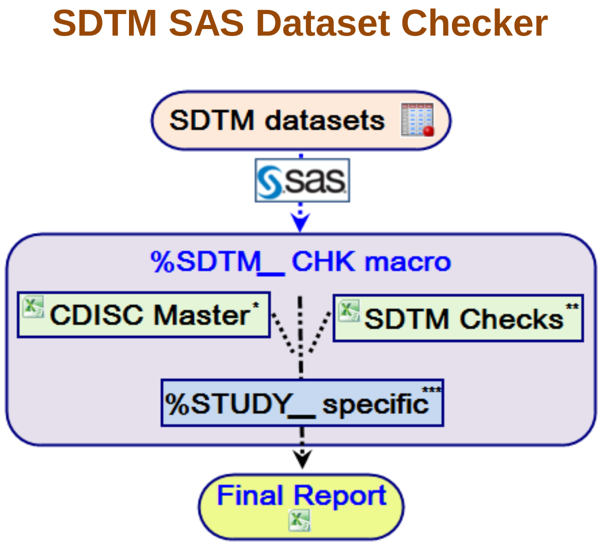
84.99%

Figure 4: SDTM Compliance Report

UTILITY 2: SDTM SAS DATASET CHECKER

FRAMEWORK

This utility has two step designs. The common SDTM checks are executed first to check basic SDTM compliance and non-compliance issue. As a second step, study /domain specific checks are executed. The advantage of the tool is in having customized study specific checks.



* CDISC Master Excel = CME, ** SDTM Checks Excel = SCE, *** %STUDY_specific = SSP

Figure 5: Framework of SDTM SAS dataset checker

The key points of the above flow chart are:

- %SDTM_CHK Macro**:** A master macro of this utility.
- CDISC Master Excel (CME) file**:** SDTM metadata of all versions from http://www.cdisc.org.
- SDTM Checks Excel (SCE) file **:** Having checks that will applies on SDTM domain, also user can add any number of new checks without editing %SDTM_CHK macro.
- Study Specific Macro "%STUDY_specific" (SSP)**:** Additional user-defined / study specific checks having complicated algorithm are possible to add in this macro.

%SDTM_CHK MACRO FLOW:

SAS macro program that performs following steps and generates final validation report:

- Running inbuilt checks in the macro
- Running checks mentioned in SCE file
- Study specific checks
- Final Report in Excel

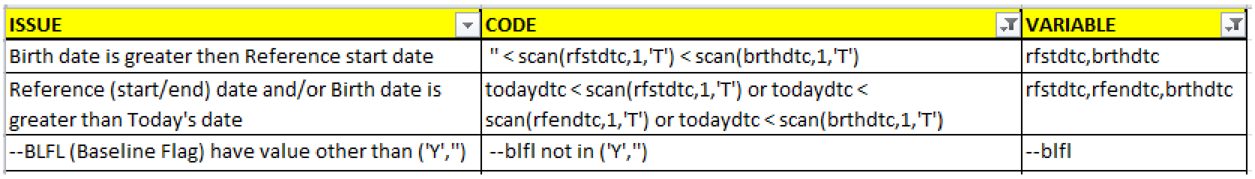
RUNNING INBUILT CHECKS IN THE MACRO

When user requests the SDTM sas dataset checker to check the SDTM domain, it will fetch the SDTM data from the location where it is stored. Then Metadata is extracted from it and will be checked against the inbuilt checks of the "SDTM_CHK macro" and core assessment criteria present in SCE file.

Although we have incorporated all possible SDTM compliance and non-compliance checks but still we may require considering more checks coming from every new study we work that needs to be listed in the SCE file.

RUNNING CHECKS MENTIONED IN SCE FILE

This stage of the utility is quite user friendly, as user can add any number of new checks.These newly added checks will be executed along with previously available checks in the SCE file. %SDTM_CHK macro will simply call those checks from the SCE file and apply that on to SDTM domain(s).



ISSUE	CODE	VARIABLE
Birth date is greater then Reference start date	'' < scan(rfstdtc,1,'T') < scan(brthdtc,1,'T')	rfstdtc,brthdtc
Reference (start/end) date and/or Birth date is greater than Today's date	todaydtc < scan(rfstdtc,1,'T') or todaydtc < scan(rfendtc,1,'T') or todaydtc < scan(brthdtc,1,'T')	rfstdtc,rfendtc,brthdtc
--BLFL (Baseline Flag) have value other than ('Y','')	--blfl not in ('Y','')	--blfl

Figure 6: SDTM Checks Excel File

The main purpose is to create SCE file is to reduce time behind validation of master macro every time for each update as it allows user to easily update or add extra checks without even disturbing the master macro. This SCE file contains three columns ISSUE (Description of the issue), CODE (SAS code that will identify that issue) and VARIABLE (List of variables used in the SAS code)

%SDTM_CHK macro will check availability (in SDTM domain) of all listed variable(s) from VARIABLE column and then executes the corresponding check. New checks and ideas can be incorporated each time in the SCE file when any team member come across any new possible compliance or non-compliance issues. This will be useful in creating exhaustive list of checks. Also there will be no restriction on number of checks to be added in the excel file and this will in turn enhance the quality of the SDTM checker utility.

STUDY SPECIFIC CHECKS

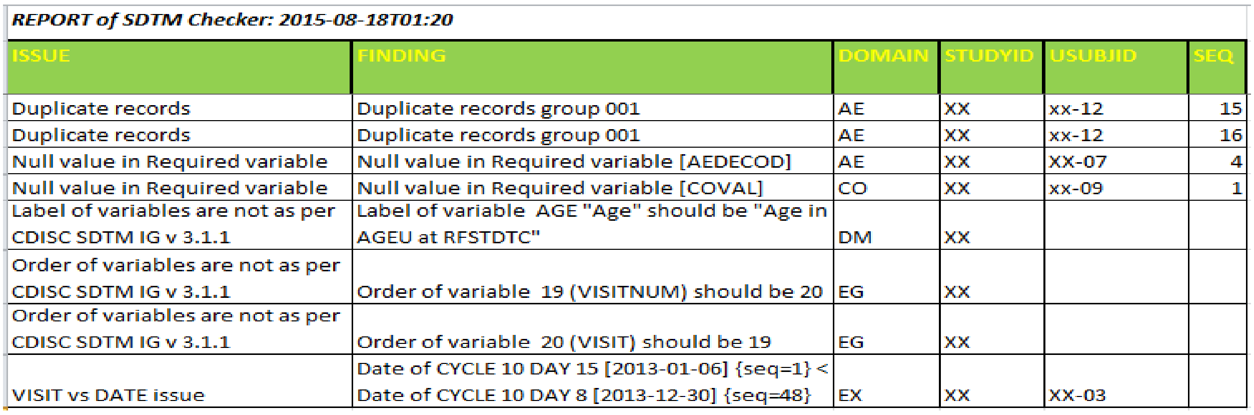
The SSP Macro can be updated by user as it is user-defined study specific checks which involves complicated algorithm. %SDTM_CHK macro will automatically calls this macro in the process of SDTM SAS dataset check.

User can write SSP Macro for different study and can apply checks on the SDTM domains. In doing so we can reduce the time for repetitive master macro validation process. Accumulation of various checks from the various SSP Macro from different studies can be transferred to the next version of the %SDTM_CHK macro.

FINAL REPORT

The final step of this macro is to generate violation dataset for each domain and all these datasets are combining to generate issue dataset which will process for the final report generation. So in the work library datasets will be like: work.check_xx (domain level) and work. Issue dataset (combined).

The difference between the issue dataset and the individual domain level datasets is that, domain level dataset will keep all the variables that are there in SDTM data set while the issue dataset will include only six variables: issue, finding, domain, studyid, usubjid, and seq. This domain level data will help study programmer to get full information about the finding with the observation and with full detail as in the real data. And that will reduce the time behind traceability the value in the domain.



REPORT of SDTM Checker: 2015-08-18T01:20

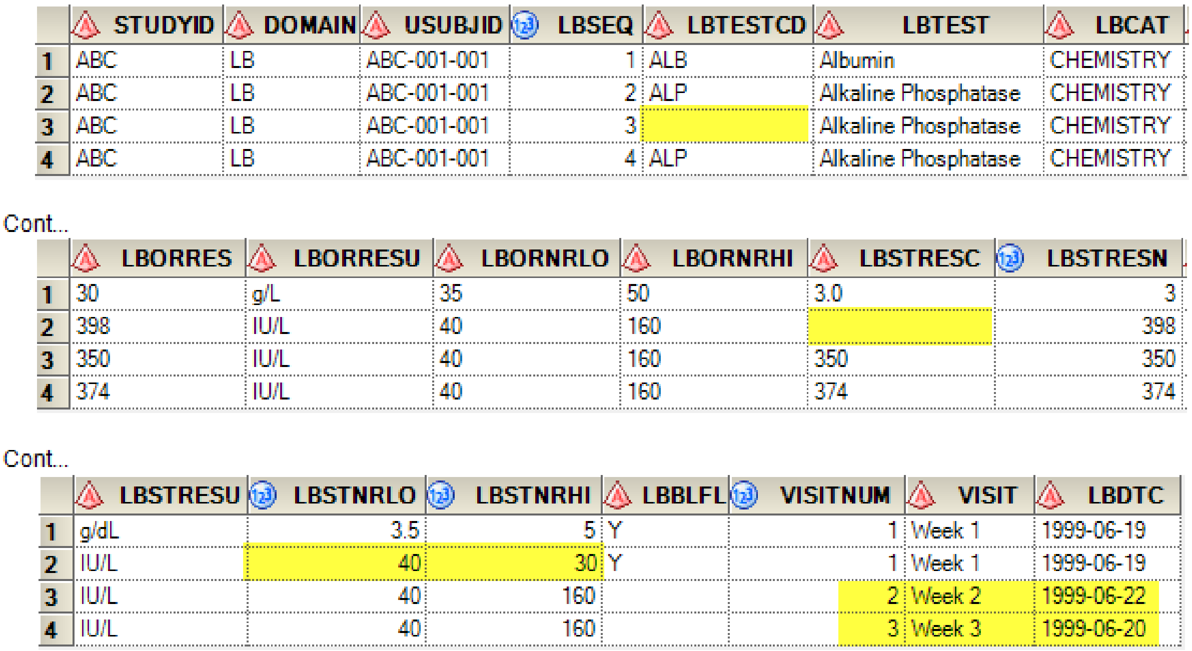
ISSUE	FINDING	DOMAIN	STUDYID	USUBJID	SEQ
Duplicate records	Duplicate records group 001	AE	XX	xx-12	15
Duplicate records	Duplicate records group 001	AE	XX	xx-12	16
Null value in Required variable	Null value in Required variable [AEDECOD]	AE	XX	XX-07	4
Null value in Required variable	Null value in Required variable [COVAL]	CO	XX	xx-09	1
Label of variables are not as per CDISC SDTM IG v 3.1.1	Label of variable AGE "Age" should be "Age in AGEU at RFSTDTC"	DM	XX		
Order of variables are not as per CDISC SDTM IG v 3.1.1	Order of variable 19 (VISITNUM) should be 20	EG	XX		
Order of variables are not as per CDISC SDTM IG v 3.1.1	Order of variable 20 (VISIT) should be 19	EG	XX		
VISIT vs DATE issue	Date of CYCLE 10 DAY 15 [2013-01-06] {seq=1} < Date of CYCLE 10 DAY 8 [2013-12-30] {seq=48}	EX	XX	XX-03	

Figure 7: SDTM SAS Dataset Checker Final Report in Excel

The final report is generated in excel format. The report is generated in a user friendly manner so that easy navigation and filtering can be applied for better understanding of compliance issues.

EXAMPLE:

To demonstrate the usage of the utility, a sample LB “Laboratory Test Results” SDTM domain is given for example as below with compliance and non-compliance issues.



	STUDYID	DOMAIN	USUBJID	LBSEQ	LBTESTCD	LBTEST	LBCAT
1	ABC	LB	ABC-001-001	1	ALB	Albumin	CHEMISTRY
2	ABC	LB	ABC-001-001	2	ALP	Alkaline Phosphatase	CHEMISTRY
3	ABC	LB	ABC-001-001	3		Alkaline Phosphatase	CHEMISTRY
4	ABC	LB	ABC-001-001	4	ALP	Alkaline Phosphatase	CHEMISTRY

Cont...

	LBORRES	LBORRESU	LBORNRLO	LBORNRHI	LBSTRESC	LBSTRESN
1	30	g/L	35	50	3.0	3
2	398	IU/L	40	160		398
3	350	IU/L	40	160	350	350
4	374	IU/L	40	160	374	374

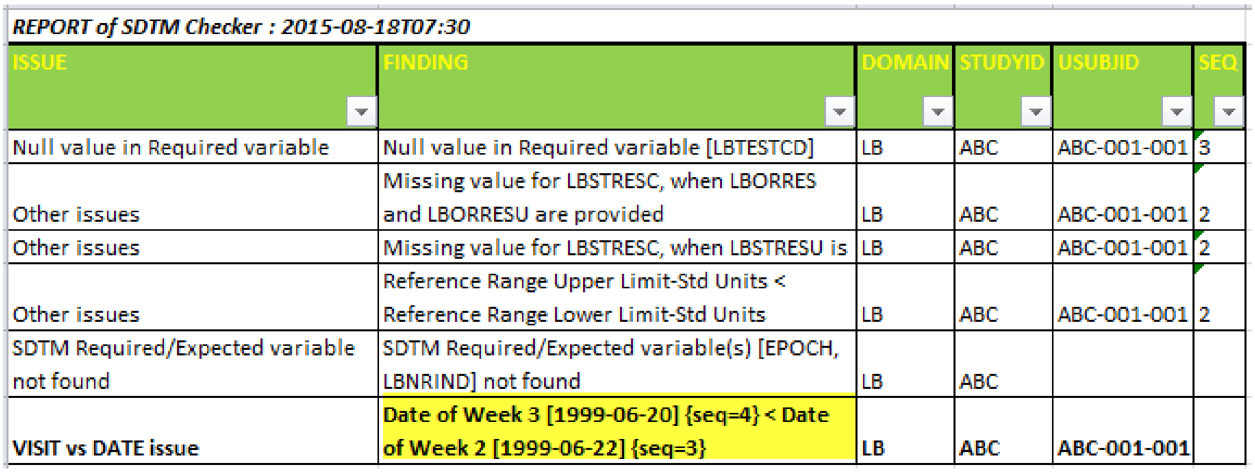
Cont...

	LBSTRESU	LBSTNRLO	LBSTNRHI	LBBLFL	VISITNUM	VISIT	LBDTC
1	g/dL	3.5	5	Y	1	Week 1	1999-06-19
2	IU/L	40	30	Y	1	Week 1	1999-06-19
3	IU/L	40	160		2	Week 2	1999-06-22
4	IU/L	40	160		3	Week 3	1999-06-20

Figure 8: LB domain with Compliance and Non-Compliance issues for demonstrate

To find out the all compliance and non-compliance issues in this LB domain (SDTM.LB), user needs to call the master macro like: %sdtm_chk (lib=sdtm, check=lb, sdtm_v=3.2);

After complete run of this macro user will get the final report as below with all compliance and non-compliance findings.



REPORT of SDTM Checker : 2015-08-18T07:30

ISSUE	FINDING	DOMAIN	STUDYID	USUBJID	SEQ
Null value in Required variable	Null value in Required variable [LBTESTCD]	LB	ABC	ABC-001-001	3
Other issues	Missing value for LBSTRESC, when LBORRES and LBORRESU are provided	LB	ABC	ABC-001-001	2
Other issues	Missing value for LBSTRESC, when LBSTRESU is	LB	ABC	ABC-001-001	2
Other issues	Reference Range Upper Limit-Std Units < Reference Range Lower Limit-Std Units	LB	ABC	ABC-001-001	2
SDTM Required/Expected variable not found	SDTM Required/Expected variable(s) [EPOCH, LBNRIND] not found	LB	ABC		
VISIT vs DATE issue	**Date of Week 3 [1999-06-20] {seq=4} < Date of Week 2 [1999-06-22] {seq=3}**	**LB**	**ABC**	**ABC-001-001**	

Figure 9: SDTM SAS Dataset Checker Report for above Example

In this example, above report confirms that the given sample LB dataset is not valid as per the requested version of SDTM IG 3.2. There are compliance and non-compliance (highlighted in yellow) issues present.

User can run this utility for all the domain(s) present at given library with just keeping macro parameter check is blank like : %sdtm_chk (lib=sdtm,sdtm_v=3.2) and get the final report for all the compliance and non-compliance issues present in all the domain present in the give library.

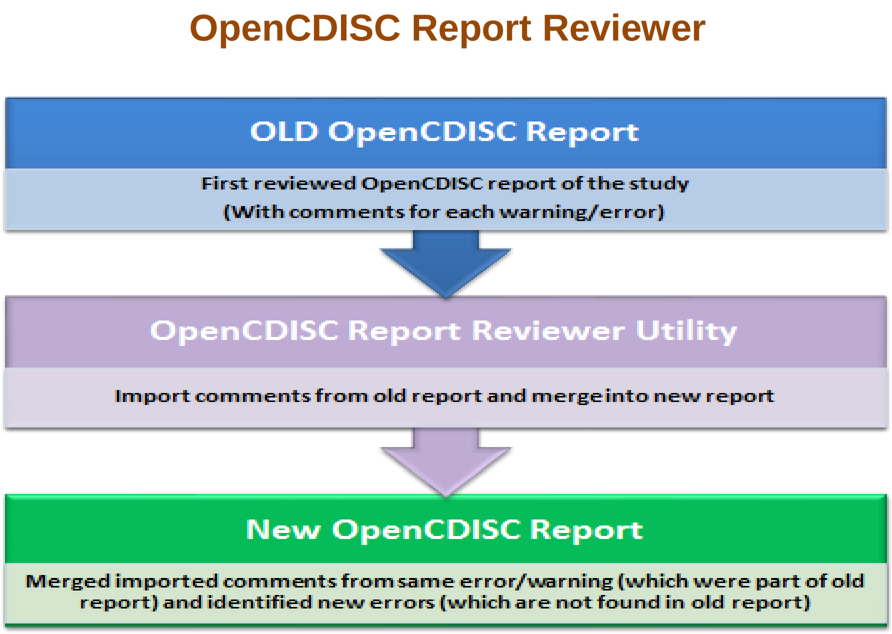
UTILITY 3: OPENCDISC REPORT REVIEWER

FRAMEWORK

It's easy to generate an OpenCDISC report for any study to validate your SDTM datasets however reviewing the report is time consuming tasks and sometimes repetitive as well; when you fix issues found or run it for new extracted data. For each run of OpenCDISC validator you will see thousands of error and warning messages and it becomes really difficult to address each message and write your comments for it.

The goal of this utility is to save time, efforts and avoid rework by importing comments from previous report, align them with new report and by summarizing the OpenCDISC review results.

As shown in the below (Figure 10) it imports the comments from the old OpenCDISC report and merge it with each matching error/warning messages found in the new report, so you need not spend much time on reviewing the same issues which were found and addressed by you in the old report.



OpenCDISC Report Reviewer

OLD OpenCDISC Report

First reviewed OpenCDISC report of the study (With comments for each warning/error)

OpenCDISC Report Reviewer Utility

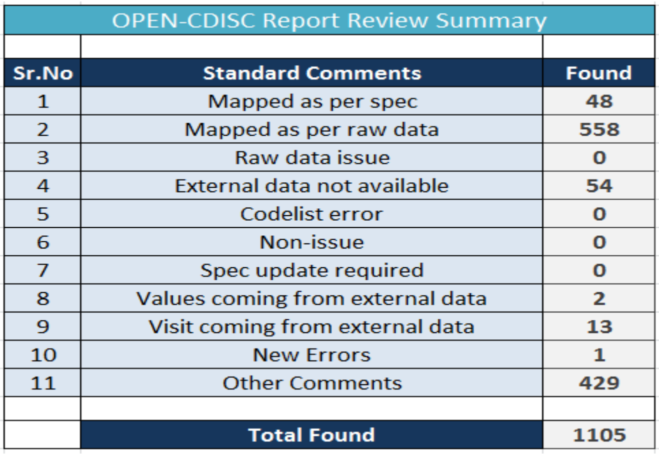
Import comments from old report and merge into new report

New OpenCDISC Report

Merged imported comments from same error/warning (which were part of old report) and identified new errors (which are not found in old report)

Figure 10: Frame work of OpenCDISC Report Reviewer

It also summarizes standard comments and produces the report as shown in below figure.



OPEN-CDISC Report Review Summary

Sr.No	Standard Comments	Found
1	Mapped as per spec	48
2	Mapped as per raw data	558
3	Raw data issue	0
4	External data not available	54
5	Codelist error	0
6	Non-issue	0
7	Spec update required	0
8	Values coming from external data	2
9	Visit coming from external data	13
10	New Errors	1
11	Other Comments	429
	Total Found	1105

Figure 11: Open CDISC Report Review Summary

CONCLUSION

The SDTM (Study Data Tabulation Model) plays a vital role in the clinical trial data process as it defines a standard structure for data tabulations and for nonclinical study data tabulations that are to be submitted as part of a product application to a regulatory authority such as FDA. In order to produce the SDTM compliant datasets we do have some validation tools and programs available in the industry like WebSDM, OpenCDISC and other SDTMIG custom checks tool. However all these tools will work on submission ready datasets and produce the reports with huge number of compliance issues, which require us to rework on programs.

These in-house solutions explained in this paper make sure the quality and compliance of SDTM standards right from the beginning and that will reduce time & efforts in rework from root level. These utilities are under review at the time of submission of this paper.

REFERENCES

http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm

http://www.cdisc.org/system/files/members/standard/stdmig_md_v_1_0.pdf

http://www.cdisc.org/system/files/members/standard/sdtmig_203_2_20release_20package_today.zip

http://www.cdisc.org/system/files/members/standard/sdtmigv3.1.3.zip

http://www.cdisc.org/sdtm-v1-1---sdtm-ig-v3-1-1

http://www.cdisc.org/sdtm-v1-2---sdtm-ig-v3-1-2

http://www.cdisc.org/metadata-submission-guideline-%28msg%29-package-preface

http://evs.nci.nih.gov/ftp1/CDISC/SDTM/SDTM%20Terminology.xls

http://www.opencdisc.org/projects/validator/cdisc-sdtm-3.1.2-validation-rules www.opencdisc.org/

SAS® 9.2 Macro Language: Reference. Cary, NC: SAS Institute Inc.

SAS® 9.2 Language Reference: Dictionary. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

I would like to thank all my colleagues at Cytel who shared their experiences, read my drafts and gave me the valuable and constructive feedback.

CONTACT INFORMATION

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

Author Name: Bharat Chaudhary and Padamsimh Balekundri
Company: Cytel
Cytel Statistical Software & Services Pvt. Ltd., Pune, India
T +91(20)6709-0175|0142 | M +91.96650.08352, +91 99023 25353
Time Zone : UTC+05:30
Bharat.chaudhary@cytel.com | Padamsimh.Balekundri@cytel.com