PhUSE 2009

Paper CD01

SDTM Validation: How can we do it right?

Peter Van Reusel, Business & Decision Life Sciences, Brussels, Belgium
Nico De Leeuw, Business & Decision Life Sciences, Brussels, Belgium

ABSTRACT

The Janus data warehouse is used by the Food & Drug Administration (FDA) to store the submitted Clinical Data Interchange Standards Consortium (CDISC) SDTM clinical data. The WebSDM™ tool performs a set of CDISC SDTM validation checks before loading the clinical data into this environment. Clinical data with High severity errors will not be loaded into the Janus repository. This paper describes the design of an enhanced CDISC SDTM validation check application, using the published WebSDM™ validation checks as a starting point. An efficient SDTM validation check process is needed by any organization when validating FDA submissions.

INTRODUCTION

Pharmaceutical companies have always recognized the need for developing company data standards to streamline their clinical research activities. However, until 2004 there was no strong incentive for individual companies to adhere to an industry-wide clinical data standard. As of July 21, 2004, the FDA recommends the CDISC SDTM standard for all New Drug Applications (NDAs). This recommendation has pushed the industry towards the adoption of this industry-wide CDISC standard. CDISC is a global standards organization founded in 1999 with the objective of establishing standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC standards are vendor-neutral, platform independent and freely available via the CDISC website. The CDISC Study Data Tabulation Model (SDTM) defines a standard structure for human clinical study data tabulations for submission of a new product application to a regulatory authority such as the FDA. All SDTM data and metadata submitted to the FDA will eventually be loaded into the FDA's Janus data warehouse.

CDISC SDTM VALIDATION

CDISC SDTM SUBMISSION PROCESS

Janus is a standards-based data warehouse, more specifically, it is a relational database model based on the SDTM standard. In order for the SDTM clinical data to properly load into Janus, it must be accompanied with the SDTM annotated CRFs, datasets and Define.xml. The reviewer tools within Janus provide means for validation of the data and Define.xml. It also allows reviewing clinical data, producing spontaneous reports, performing cross-study analyses and facilitating communication of conclusions. [1] Final communication of the SDTM compliance issue report to the sponsor is handled by the Office of Business Process Support (OBPS) at the FDA. An overview of the CDISC SDTM submission process is shown in Figure 1.

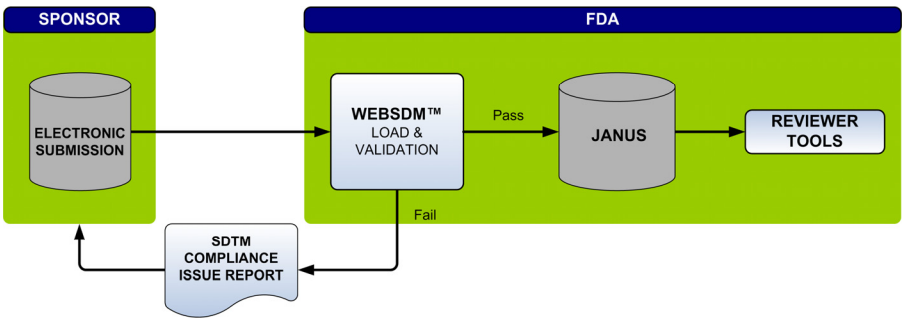


Figure 1: Overview of the CDISC SDTM submission process

CDISC SDTM COMPLIANCE

WebSDM™ verifies that the data conforms to the assumptions of the SDTM IG, and that the Define.xml conforms to the CDISC Operational Data Modeling (ODM) version 1.2.

The SDTM IG describes that conformance with the CDISC domain models is indicated by: [2]

- Following the complete metadata structure for data domains
- Following SDTM IG domain models wherever applicable
- Using SDTM-specified standard domain names and prefixes where applicable
- Using SDTM-specified standard variable names
- Using SDTM-specified variable labels for all standard domains
- Using SDTM-specified data types for all variables
- Following SDTM-specified controlled terminology and format guidelines for variables, when provided
- Including all collected and relevant derived data in one of the standard domains, special-purpose datasets, or general-observation-class structures
- Including all Required and Expected variables as columns in standard domains, and ensuring that all Required variables are populated
- Ensuring that each record in a dataset includes the appropriate Identifier and Timing variables, as well as a Topic variable
- Conforming to all business rules described in the CDISC Notes column and general and domain-specific assumptions

The WebSDM™ application (as shown in Figure 2) was developed under a Cooperative Research and Development Agreement between the FDA and Phase Forward's Lincoln Safety Group and has been in use by the FDA since 2004. [3]

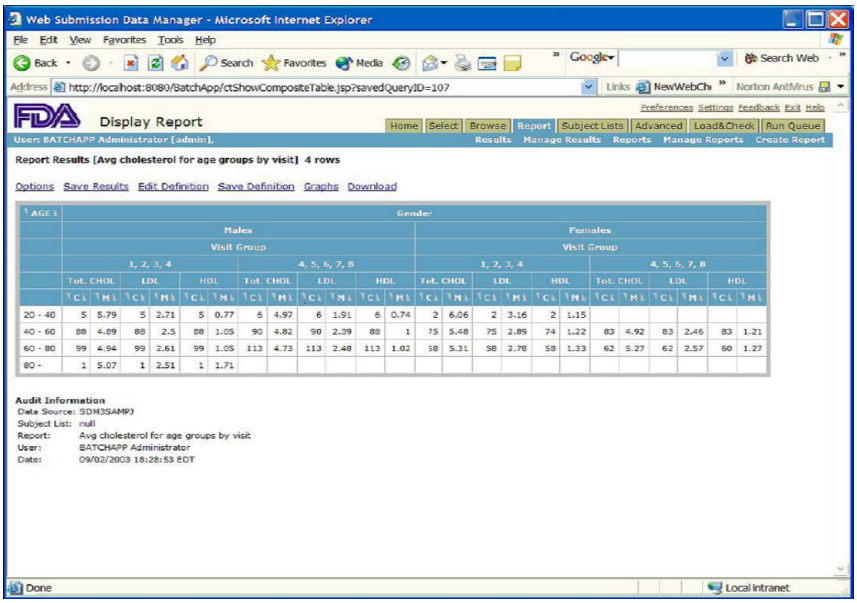


Figure 2: SDTM FDA tools: WebSDM™

In WebSDM™ there are checks for the existence of data anomalies, such as the Define.xml should correspond to the metadata of the datasets, values should correspond to the CDISC controlled terminology, and that there is cross-domain consistency. A particular validation check may apply to all domains, a general class of domains, a particular domain, a set of variables in a given role (such as timing variables), or a particular variable.

The list of anomalies produced by the validation checks is called "rule violations". WebSDM™ assigns rule violations a severity of High, Medium, or Low. The severity is the error's potential to affect the interpretation or use of the data for specific purposes. The study will fail to load into Janus if any of the rule violations identified by the FDA as having a High severity is triggered. Violations flagged as Medium are considered to impact the reviewability of the submission. While violations flagged as Low are considered to possibly impact only the reviewability.

The actual list of checks used by the FDA (in total 108 checks) will evolve over time to reflect input received from the FDA reviewers on the basis of accumulating experience with SDTM data submissions. [4]
Business & Decision Life Sciences has developed an enhanced SDTM validation check application including additional checks for the SDTM mapping, data and metadata. In total, 200 data and metadata checks for SDTM v3.1.1 and 250 data and metadata checks for SDTM v3.1.2 have been developed to ensure compliance with the latest Janus data warehouse requirements. In addition, 55 validation checks on the data mapping specifications have been developed to ensure the quality of the data conversion programs.

SDTM COMPLIANCE

The CDISC SDTM validation environment contains a metadata library of the current SDTM standards and the available CDISC controlled terminology.

DESCRIPTION OF THE SDTM VALIDATION CHECKLIST

To ensure SDTM compliance, the list of available WebSDM™ checks was investigated and replaced by an enhanced list of validation checks. The current package of validation checks consists of three main categories; the mapping, metadata and data checks. With the introduction of the new SDTM IG 3.1.2 version, the validation checks were upgraded to ensure correct conversions to SDTM IG v3.1.2. In addition, validation checks for the data mapping specifications have been developed to ensure the quality of the data conversion programs. The list of validation checks is growing based on the additional needs and the feedback received.

CATEGORIZATION OF THE CHECKS

The SDTM validation checks have been categorized based on the following characteristics (as shown in Figure 3):

- Data, metadata or mapping sheet validation check
- Applicable class and domain
- CDISC Controlled Terminology and timing variables check
- SDTM validation check version

CHECKID	SUB CHECKID	CHECK DESCRIPTION	ERROR MESSAGE	CATEGORIZING							SDTM IMPLEMENTATION GUIDE VERSION	CHECK VERSION
				DATA	METADATA	MAPPING SHEET	CLASS	DOMAIN	CONTROLLED TERMINOLOGY	TIMING VARIABLES		
0001		Check if all the date variables (--DTC, --STDTC, --ENDTC, RFSTDC, RFENDTC) have a ISO 8601 standard format.	DATE VARIABLE does not have a ISO 8601 format.	x			ALL	ALL	N	Y	3.1.1	0.1
0002		Check per LBTESTCD if the LBSTRESU is consistent within the trial.	The STANDARD UNIT is not consistent per LAB TEST or EXAMINATION.	x			FINDINGS	LB	N	N	3.1.1	0.1
0003	1	Check if AESER ="Y" that [AESCAN="Y" or AESCONG="Y" or AESDISAB="Y" or AESDTH="Y" or AESHOSP="Y" or AESLIFE="Y" or AESMIE="Y" or AESOD="Y"] (Check only applies if serious criteria are completed).	The AE is serious, but none of the serious criteria is answered "Y".	x			EVENTS	AE	N	N	3.1.1	0.1
0003	2	Check if [AESCAN="Y" or AESCONG="Y" or AESDISAB="Y" or AESDTH="Y" or AESHOSP="Y" or AESLIFE="Y" or AESMIE="Y" or AESOD="Y"] that AESER ="Y".	The AE is not serious, but one of the serious criteria is answered "Y".								3.1.1	0.1
		Check if --STRESC <> NULL	A STANDARD RESULT is									

Figure 3: CDISC SDTM validation check categories

Based on this categorization, a subset of SDTM validation checks can be created. This functionality allows proper management of the validation checks that need to be submitted first. More specifically, a selection can be made based on the SDTM IG version (v3.1.1 or v3.1.2), check type (data, metadata of both) and/or the domain type (all, interventions, events, findings, special-purpose, and trial design).

PROCESS FLOW (AS SHOWN IN FIGURE 4)

The SDTM validation check application consists of different parts. The data and metadata environment contains:

- the SDTM metadata library: this includes the metadata definition of the SDTM IG (version 3.1.1 and 3.1.2)
- the metadata database
- the check scheduler
- the exception table

The process occurs as follows:

- A job definition is created in the application interface based on the selection criteria
- The job is executed and send to the check scheduler
- When the job executes it reads the folder with the SDTM datasets & Define.xml metadata file
- The SDTM validation checks (~250) verify for consistency between:
  - CDISC SDTM dataset definition
  - Metadata reported in the Define.xml
  - CDISC SDTM data standards in the metadata library
- Identified quality issues are added to an exception table
- Quality check reports are generated for further processing (in paper, electronic or .xml format)

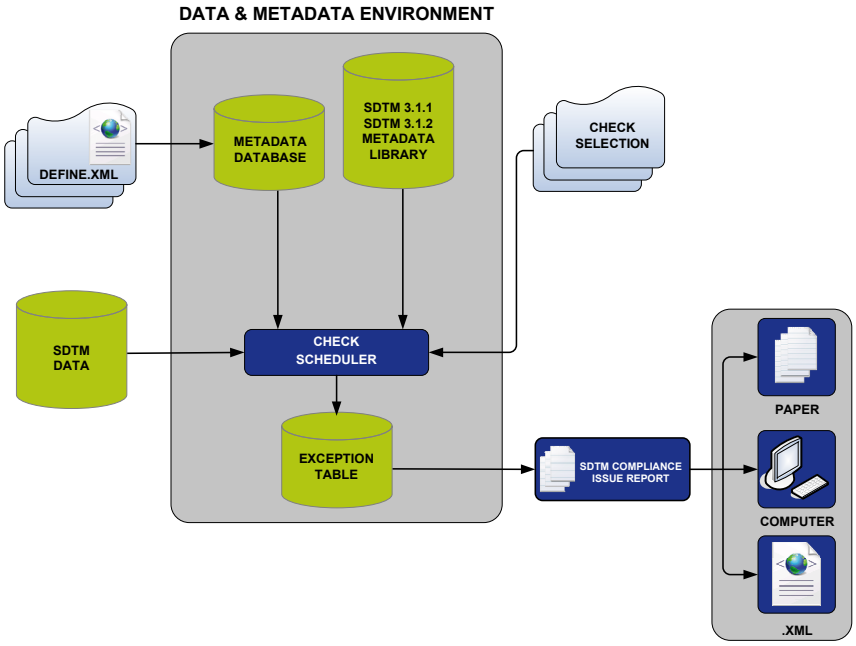


Figure 4: CDISC SDTM validation check application flowchart

CHECK SELECTION

An application interface has been built to enhance the user friendliness as shown in Figure 5.

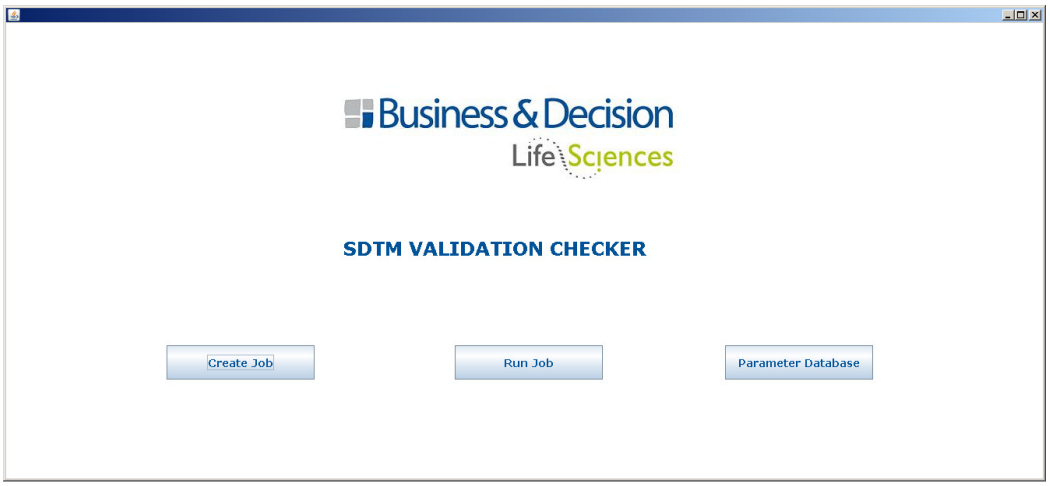


Figure 5: CDISC SDTM validation check application interface

There are three parts, i.e. a screen to create and run the validation checks; and a separate parameter database.

In the <Job Creation> screen, a tab of the check selection and a tab of the check description have been included. In the job creation screen, the user can select the project, study and list of checks to be included in the job as shown in Figure 6. The program allows to select all or individual checks, or a dynamically created subset of the checks based on the checks type, domains verified and SDTM version applied. Subsequently a job description is entered and jobID is created. The check description screen displays the list of checks included in the job together with the checks description.

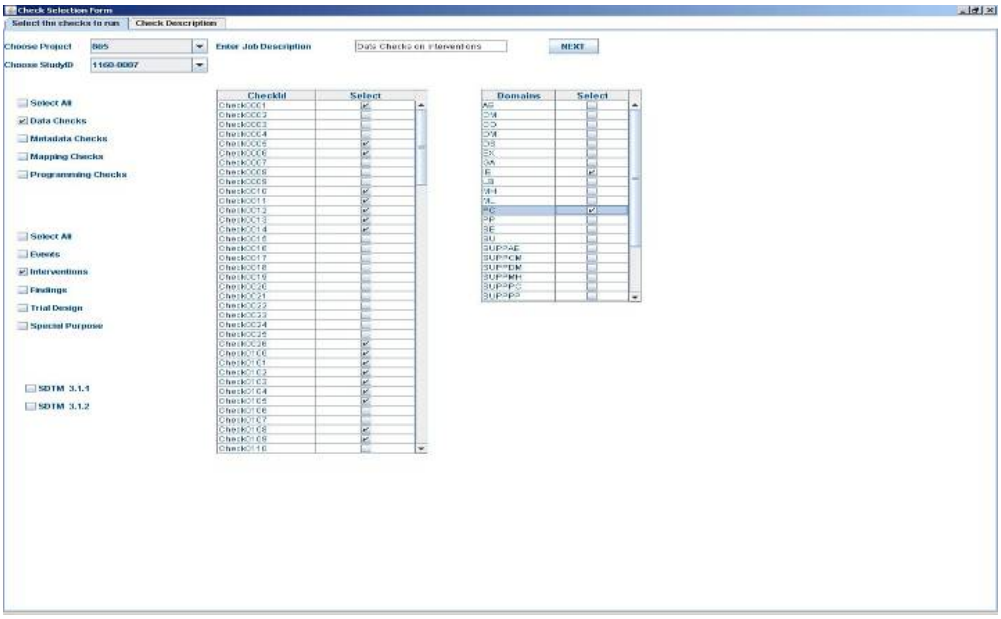


Figure 6: CDISC SDTM validation check application: check selection screen

The <Job Run> screen contains the job list, job description, job owner and creation date as shown in Figure 7. The job list allows the user to select one or more jobs to be submitted with the data. The job description lists the project, studyID and checks to run for the selected job and the domains on which the selected checks run.

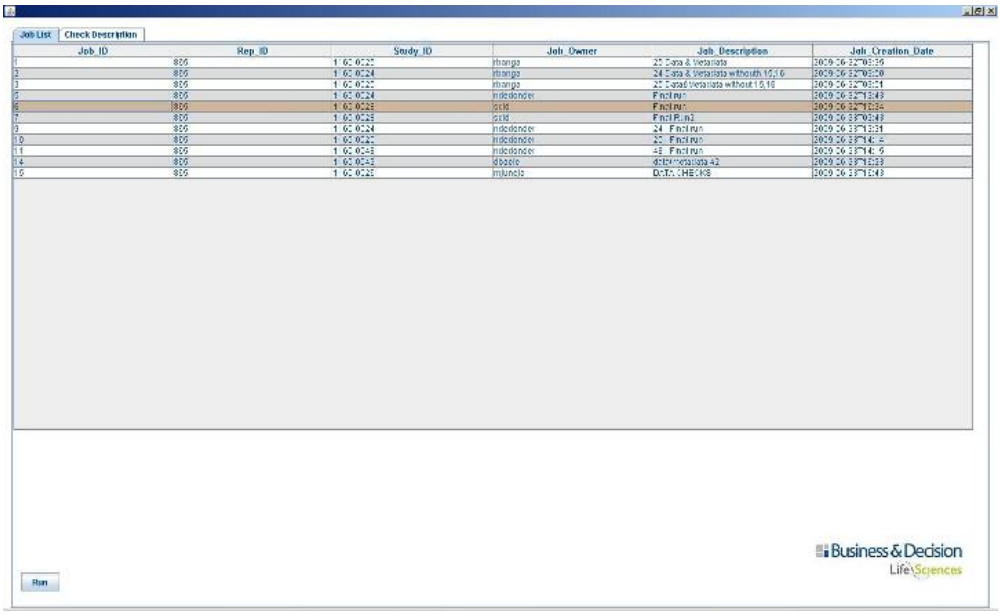


Figure 7: CDISC SDTM validation check application: job list screen

The parameter database is a central table that includes the project and study parameters.

REPORTING

The system creates three different reports, the:

- Exception log
- Exception summary
- Exception report

The exception log report provides information on the actual job; if a job fails you can pick it up from here (as shown in Figure 8).

CHECK_LIST	ERROR_LEVEL	COMPLETION_TIME
Check0001	The Check Completed Successfully	2009-01-27T9-55-00
Check0002	The Check Completed Successfully	2009-01-27T9-55-00
Check0003	The Check Completed Successfully	2009-01-27T9-55-00
Check0004	The Check Completed Successfully	2009-01-27T9-55-00
Check0005	The Check Completed Successfully	2009-01-27T9-55-00
Check0006	The Check Completed Successfully	2009-01-27T9-55-00
Check0007	The Check Completed Successfully	2009-01-27T9-55-00
Check0008	The Check Completed Successfully	2009-01-27T9-55-00
Check0009	The Check Completed Successfully	2009-01-27T9-55-00
Check0010	The Check Completed Successfully	2009-01-27T9-55-00
Check0011	The Check Completed Successfully	2009-01-27T9-55-00
Check0012	The Check Completed Successfully	2009-01-27T9-55-00
Check0013	The Check Completed Successfully	2009-01-27T9-55-00
Check0014	The Check Completed Successfully	2009-01-27T9-55-00
Check0015	The Check Completed Successfully	2009-01-27T9-56-00
Check0016	The Check Completed Successfully	2009-01-27T9-56-00
Check0017	The Check Completed Successfully	2009-01-27T9-56-00
Check0018	The Check Completed Successfully	2009-01-27T9-56-00
Check0019	The Check Completed Successfully	2009-01-27T9-56-00
Check0020	The Check Completed Successfully	2009-01-27T9-56-00

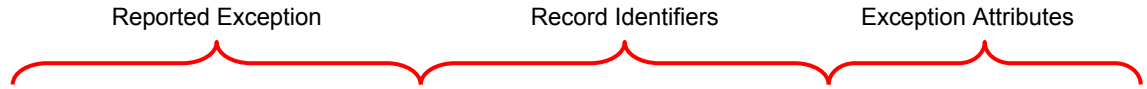
Figure 8: CDISC SDTM validation check application: exception log report

The exception summary provides an overview of the number of anomalies per check; it also lists immediately the type of check; data or metadata check (as shown in Figure 9).

LIST_OF_CHECKS	DOMAIN	VARIABLE1	EXCEPTIONS_COUNT	CHECK TYPE
Check0009	LB	LBORRES	73	DATA
Check0010	LB	LBORRESU	84	METADATA
Check0010	VS	VSORRESU	56	METADATA
Check0010	VS	VSSTRESU	56	METADATA
Check0015	CO	IDVAR	73	DATA
Check0100	IE	DOMAIN	1	DATA
Check0268	CM	COLUMN	1	METADATA
Check0271	VS	COMPALGO	5	METADATA

Figure 9: CDISC SDTM validation check application: exception summary report

The exception report contains all anomalies present in the data. The report consists of the reported exceptions, the record identifiers and the exception attributes. If more then 50 exceptions or anomalies per check per domain are retrieved then only three exceptions are reported (as shown in Figure 10). This increases the user friendliness of the report.



CHECKID	ERRORMSG	STUDYID	DOMAIN	USUBJID	KEY1CD	KEY1	VAR1CD	VAR1	VAR2CD	VAR2
9	An ORIGINAL RESULT is completed but the STANDARD RESULT (CHARACTERISTIC) is missing.	1234-5678	LB	1234-5678-00002	LBSEQ	29	LBORRES	32,4	LBSTRESC	
9	An ORIGINAL RESULT is completed but the STANDARD RESULT (CHARACTERISTIC) is missing.	1234-5678	LB	1234-5678-00002	LBSEQ	13	LBORRES	19	LBSTRESC	
9	An ORIGINAL RESULT is completed but the STANDARD RESULT (CHARACTERISTIC) is missing.	1234-5678	LB	1234-5678-00001	LBSEQ	52	LBORRES	1,1	LBSTRESC	
9	TOO MANY FAILURES (73)									
10	The value cannot be found in the codelist attached to the variable.	1234-5678	LB	1234-5678-00016	LBSEQ	55	LBORRESU	10^6/uL		
10	The value cannot be found in the codelist attached to the variable.	1234-5678	LB	1234-5678-00027	LBSEQ	54	LBORRESU	10^6/uL		
10	The value cannot be found in the codelist attached to the variable.	1234-5678	LB	1234-5678-00008	LBSEQ	55	LBORRESU	10^6/uL		
10	TOO MANY FAILURES (84)									
10	The value cannot be found in the codelist attached to the variable.	1234-5678	VS	1234-5678-00024	VSSEQ	16	VSORRESU	kg		
10	The value cannot be found in the codelist attached to the variable.	1234-5678	VS	1234-5678-00020	VSSEQ	16	VSORRESU	kg		
10	The value cannot be found in the codelist attached to the variable.	1234-5678	VS	1234-5678-00012	VSSEQ	16	VSORRESU	kg		
10	TOO MANY FAILURES (56)									

Figure 10: CDISC SDTM validation check application: exception report

RESOLVING ISSUES AND INTERPRETATION

The exception report also contains a status column in which the exceptions are flagged and potentially documented. Possibilities are a status flag of “resolved” or “pending” or “a valid explanation why the reported exception is not a true error”. These are a few examples:

- The exception report indicates a variable is required but no value is completed. The underlying reason is that the source data is missing. If the source data is available, the value should be uploaded and the error will be resolved. If not available, the error will be documented.
- The exception report indicates when a record is duplicated. After verification it becomes obvious that there was a data conversion program error that resulted in a duplicate record. Because of this exception message, we can easily resolve the error and rerun the validation check. Once resolved, the exception should no longer appear on the exception report.
- The exception report indicates a “date/time of collection is available, but the study day of visit /collection/ exam” is missing. In some cases it is valid (if the study day can be calculated); in others cases it might be invalid i.e., if the study day is based on the start date of the study drug, but the patient did not take any study drug.

At the end of the review cycle you should receive a report with only valid exceptions; e.g. anomalies based on permanently missing source data.

CONCLUSION

The CDISC SDTM datasets and Define.xml should be tested for CDISC SDTM compliance before being sent to the FDA, resulting in a submission that will load successfully into the Janus data warehouse.

The SDTM validation check application described before is designed to identify and resolve efficiently anomalies at all levels of the conversion process:

- Once the mapping specifications are completed, the mapping specification checks are submitted through the SDTM validation check application. Identified exceptions are verified, documented and resolved before transmitting the mapping sheet to the programmers.
- The SDTM validation check application runs the metadata validation checks to verify that all SDTM specific metadata validation rules are met.
- The SDTM validation checks are run against the target datasets. After verification and resolution, the datasets are ready for final FDA release.

REFERENCES

1. Experiences Submitting CDISC SDTM and JANUS Compliant Datasets, Carol R Vaughn and Greg Ridge (Sanofi-Aventis), 2008
2. CDISC SDTM Implementation Guide v3.1.2, pg. 20, November 12, 2008
3. http://www.phaseforward.com/products/clinical/ads/default.aspx
4. Experiences Submitting CDISC SDTM and JANUS Compliant Datasets, Carol R Vaughn and Greg Ridge (Sanofi-Aventis), 2008

CONTACT INFORMATION

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

Peter Van Reusel
Business & Decision Life Sciences
Sint-Lambertusstraat 141
1200 Brussels
Work Phone: +32 2 774 11 00
Fax: +32 2 774 11 99
Email: peter.vanreusel@businessdecision.com
Web: www.businessdecision-lifesciences.com