Paper RS07-2008

Validating CDISC SDTM-Compliant Submission-Ready Clinical Datasets with an In-House SAS® Macro-Based Solution

Bhavin Busa, Independent Consultant
Sheila Vince, Independent Consultant
Jameelah Aziz, Cubist Pharmaceuticals, Inc., Lexington, MA

ABSTRACT

Pharmaceutical organizations are proactively adopting Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) because of a broad variety of benefits to the industry and the FDA's intention to mandate submissions of SDTM within the Electronic Common Technical Document (eCTD) structure. Tools are needed to check compliance and streamline operations for the preparation of data for FDA submission.

This paper describes two currently available options for validating the compliance of submission-ready clinical datasets, and presents an in-house SAS®-based solution to design and implement a SAS macro library for the same purpose. We present the advantages of the in-house SAS solution versus the other options in checking the domain structure, variable attributes and domain content per the SDTM Implementation Guide (SDTM IG), and provide examples of strategies to achieve this.

INTRODUCTION

FDA AND CDISC

The FDA has endorsed CDISC SDTM as the preferred model for submitting clinical and bioequivalence data in the eCTD guidance [1]. As one of the Critical Path initiatives [2], the FDA is actively collaborating with the CDISC team to develop and promote fast industry adoption of the SDTM, and has announced a forthcoming update to the regulation [3] that would mandate companies to use standardized data structure and terminology according to the CDISC SDTM guidance, with a period of time allowed for transition to the new standards.

The FDA has developed a standards-based clinical data repository, Janus [4]. This repository provides a data model to collect and analyze clinical trial data submitted by various pharmaceutical and biotechnology companies. Janus provides a central access to standardized data and creates an integrated platform for tools used in analysis and review [4, 5, and 6]. The standardization of submission study tabulation datasets will greatly facilitate the FDA's ability to process, review and archive data into the Janus data warehouse.

CDISC AND INDUSTRY

Pharmaceutical organizations are quickly adopting the SDTM model both because of FDA submission requirements [3] and because of benefits to the industry in communicating and sharing standardized data [4]. A recent survey (2007) reported that nearly 70% of sponsor and CRO companies are currently utilizing CDISC SDTM standards, 22% are piloting the standards and the remaining 8% are planning to implement the CDISC model to develop SDTM-compliant submissions [7]. As the industry becomes increasingly involved in developing efficient and cost effective ways to produce CDISC SDTM-compliant clinical trial domains, it is necessary to develop tools to check compliance and streamline operations for the preparation of submission-ready files in accordance with the most recent SDTM IG.

Sponsors have adopted various means to achieve the final SDTM structure which can be built into different phases of the clinical data life cycle from data collection to data submission [8, 9, and 10]. The three basic approaches that are widely used by the industry are:

- The SDTM structure can be built into the Clinical Data Management System (CDMS) directly, requiring no or very minimal transformation to program the SDTM domains. This entails implementation of SDTM standards in the front-end of the clinical data life cycle and minimizes the efforts to achieve the SDTM structure at a later stage. Because of the normalized structure of the SDTM domains, this approach is difficult, if not impossible, for at least some domains.
- The SDTM domains can be created by converting any 'raw' data export from the CDMS to the SDTM structure using SAS or other transformation tools. The 'raw' data can be in the format of the sponsor's internal proprietary data standards. This could require significant conversion on the back-end but allows

conversion of a variety of CDMS structures to the SDTM without requiring a change to data management structures and procedures in use.

- SDTM domains can be created starting from a “near-SDTM” or “SDTM-like” structure built within the CDMS. This approach, described as a hybrid approach by Jack Shostak [8], is widely used. The conversion time and effort to achieve the SDTM structure are reduced considerably compared to the second approach, and significant changes are not required to the CDMS design.

Whatever the approach used, sponsors and CROs will benefit from the validation of the compliance of data before a submission to the FDA, and will probably further benefit by performing the validation as early as possible in the clinical data life cycle. Errors in the SDTM submission may have an impact when loading the data into the Janus repository. Additionally, any inconsistencies or discrepancies in the data may delay the submission review process, and may result in increased costs to the sponsor.

To bring the validation process forward in the clinical data life cycle requires either the use of a validation tool that can be used on incomplete SDTM domains constructed directly from the CDMS (where derived or defaulted fields may not yet be populated), or use of a tool that can be customized to proprietary or “near-SDTM” datasets. The ultimate purpose of this validation tool is to check that domains are submission-ready; however any versatility in the tool could significantly enhance the efficiency of production of the final domains.

SDTM VALIDATION TOOLS

LINCOLN TECHNOLOGIES - WebSDM™

Lincoln Technologies (Phase Forward’s safety division) with the FDA, under a Cooperative Research and Development Agreement (CRADA), has developed a Web Submission Data Manager (WebSDM™) application [11]. This application tests the compliance of submission-ready files (in SAS V5 Transport format or Oracle® views) according to the SDTM IG [11]. The FDA piloted the use of WebSDM™ and has been using it for review of studies since 2004. Users load SDTM-compliant files into the tool, and can then check for errors or inconsistencies in the structure and content of the data. A screen shot of the error log generated by WebSDM during the data loading process was presented at the 2005 FDA public meeting on the review of SDTM domains at CDER by Armando Oliva [12].

The checks available include detection of structural and consistency errors rated by severity (high, medium and low) [13]. WebSDM Version 2.6 currently runs 109 checks, of which 23 are high severity, 61 are medium and 25 are low. These validation rules are derived from the SDTM IG (Version 3.1.1 dated August 26, 2005) and are applied to all domains, an individual domain, general category of domains (such as interventions), or a set of variables in a given role (such as timing variables). The detailed validation check specifications, with the description of each rule, its severity, error message, and impact, can be found at Phase Forward’s website [13]. The tool also provides the capability to browse data with a graphical display [using PPD® Patient Profile], and can create custom and predefined reports for review [11].

The use of WebSDM provides sponsors with the same visualization and review capabilities that are already utilized by FDA reviewers. However, since the intent in creating the tool was to facilitate the review of submission-ready SDTM datasets by the FDA, it is not optimized for SDTM datasets “in-development” and cannot easily be customized to the proprietary or “near-SDTM” structure datasets. Without customization, the tool checks the metadata information of the SDTM datasets and sponsor defined custom domains against the define documentation (define.xml) pre-loaded into the system [14]. The structure and observation level checks on the SDTM datasets are programmed to pre-defined rules set forth by the tool [13]. These rules can be customized to the user requirements but this involves additional SQL coding in the WebSDM environment.

Since many drug development programs are terminated before submission of datasets to the FDA, the expense of licensing an additional validation tool for the entire duration of the clinical development program may be unfeasible for small or mid-size pharmaceutical companies. Nevertheless, the use of SDTM standards and timely validation of data may significantly reduce the overall cost of drug development. In the next sections we describe two SAS-based solutions; one is available as a free download from the SAS website and can be used with the SAS Base system (Version 8.2 and above); the other is an enhanced, fully customized SAS macro-based solution using any version of the SAS Base system developed in-house. Although some organizations may not have in-house SAS programmers to utilize or develop these alternatives, an internal cost analysis revealed that outsourcing the function to a contractor or CRO is still cost effective.

SAS - PROC CDISC

SAS is actively participating in CDISC initiatives and provides support and solutions to implement CDISC data standards [15]. SAS (Version 9.1.3 Service Pack 3 and above) includes a procedure called PROC CDISC that not only supports the import, export and processing of Extensive Markup Language (XML) documents that are in CDISC-defined formats (Operational Data Model - ODM 1.2), but also provides checks of data content against the domain definitions outlined in the SDTM IG (Version 3.1, dated July 14, 2004) [16, 17]. For SAS system (Version 8.2 and Versions between 8.2 and 9.1.3 SP3) a field response release is available for download from the SAS website [18]. The procedure is well summarized in the paper by Anthony Friebel from the SAS Institute [17]. For further information on the history and installation of the field response release, refer to the recently published paper by William C. Csont – "Taking a Walk on the Wildside: Use of PROC CDISC-SDTM 3.1 Format" [19].

The current SAS procedure PROC CDISC supports validation of 15 of the 23 domains outlined in CDISC SDTM version 3.1 [17]. It supports the interventions (CM-Concomitant Medications, EX-Exposure, SU-Substance Use), events (AE-Adverse Events, DS-Disposition, MH-Medical History), findings (EG-ECG Test Results, IE-Inclusion/Exclusion Exception, LB-Laboratory Test Results, PE-Physical Examinations, QS-Questionnaires, SC-Subject Characteristics, VS-Vital Signs), and special (DM-Demographics, CO- Comments) class of domains. Currently it does not support the trial design domains, custom-defined domains or other SDTM domains finalized since SDTM IG version 3.1 [17]. The syntax of the PROC CDISC to validate a domain in each of the four classes is outlined below.

```
** ‘Special’ class domain: DM **;

PROC CDISC MODEL=SDTM;
SDTM       SDTMVERSION= "3.1";
DOMAINDATA DATA=WORK.DM
           DOMAIN=DM
           CATEGORY=SPECIAL;
RUN;
```

```
** ‘Interventions’ class domain: CM **;

PROC CDISC MODEL=SDTM;
SDTM       SDTMVERSION= "3.1";
DOMAINDATA DATA=WORK.CM
           DOMAIN=CM
           CATEGORY=INTERVENTIONS;
RUN;
```

```
** ‘Events’ class domain: AE **;

PROC CDISC MODEL=SDTM;
SDTM       SDTMVERSION= "3.1";
DOMAINDATA DATA=WORK.AE
           DOMAIN=AE
           CATEGORY=EVENTS;
RUN;
```

```
** ‘Findings’ class domain: IE **;

PROC CDISC MODEL=SDTM;
SDTM       SDTMVERSION= "3.1";
DOMAINDATA DATA=WORK.IE
           DOMAIN=IE
           CATEGORY=FINDINGS;
RUN;
```

The DATA=parameter specifies the location of the SDTM domains in a SAS dataset format (.sas7bdat). In this example, we are assuming the domains are stored in a temporary SAS library ‘WORK’. The DOMAIN and the CATEGORY parameters tell the procedure which domain to check and the class of the domain respectively. The SDTMVERSION will depend on the version of the PROC CDISC installed on the SAS system. To verify which version of PROC CDISC is currently installed on the SAS system, use the syntax below and view the log. The version 2.15.49 of the procedure currently supports validation of the SDTM 3.1 domains.

```
** To check the PROC CDISC version **;

PROC CDISC VERSION;
RUN;
```

To demonstrate the usage of the PROC CDISC, a sample DM-Demographics SDTM domain dataset was created in Example 1 below. All examples presented in this paper use the DM domain. A thorough knowledge of this domain is recommended (see Recommended Reading).

Example 1: Demonstrating the use of PROC CDISC on a special class DM-Demographics domain.

```
** Creating Special class domain: DM dataset **;
DATA DM (label="Demographics");
       STUDYID='STUDY ABC';
       DOMAIN='DM';
       USUBJID='ABC-001-0011';
       SUBJID='0011';
       RFSTDTC='2007-04-03T22:37:10';
       RFENDTC='2007-04-12T00:00:00';
       SITEID='001';
       INVID='XYZ';
       INVNAM='DR. BOND';
       BRTHDTC='1986-06-08';
       AGE=20;
       AGEU='YEARS';
       SEX='M';
       RACE='WHITE';
       ETHNIC='NOT HISPANIC OR LATINO';
       ARMCD='ABC';
       ARM='DRUG ABC';
       COUNTRY='USA';
       DMDTC='2007-04-03';
       DMDY=1;
RUN;

** PROC CDISC to validate SDTM compliance **;
PROC CDISC MODEL=SDTM;
SDTM       SDTMVERSION= "3.1";
DOMAINDATA DATA=WORK.DM
           DOMAIN=DM
           CATEGORY=SPECIAL;
RUN;
```

Log Contents for Example 1:

```
NOTE: DomainData DM data content is valid.
NOTE: PROCEDURE CDISC used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

After running the procedure PROC CDISC on the sample DM dataset, the log displays a note "DomainData DM data content is valid". This confirms that the dataset created in Example 1 is valid per the SDTM version 3.1.

To further demonstrate the functionality of PROC CDISC, structural and observation level changes were made to the example DM domain dataset. The same PROC CDISC statements from Example 1 were used after each change outlined below. The contents of the log window (output from PROC CDISC) are presented in each case.

Example 2: Structural level (domain content of the source) changes made to the sample DM dataset:

a) Dropped required variables (SITEID, ARM)

Log Contents for Example 2a:

```
ERROR: Required parameters not contained on DOMAINDATA(Domain=DM) statement.
       Required parameter SITEID not present.
       Required parameter ARM not present.
NOTE: The SAS System stopped processing this step because of errors.
NOTE: PROCEDURE CDISC used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

b) Added additional variable (RDTXGP)

Log Contents for Example 2b:

```
ERROR: Variable RDTXGP is not defined in this domain (DM) and must be dropped/deleted.
ERROR: SDTM-compliant datasets may contain only variables defined in the domain (DM).
NOTE: The SAS System stopped processing this step because of errors.
NOTE: PROCEDURE CDISC used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

c) Dropped expected variable (RACE), dropped permissible variable (ETHNIC), and changed the type of AGE from numeric to character

Log Contents for Example 2c:

```
ERROR: Variable AGE is not character type.
WARNING: Variable RACE is expected in this domain(DM), but is not present.
NOTE: Variable ETHNIC is permitted in this domain(DM), but is not present.
NOTE: The SAS System stopped processing this step because of errors.
NOTE: PROCEDURE CDISC used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

These examples demonstrate that PROC CDISC validates the SDTM domains on a structural level (domain content of the source) for:

- All the required variables in the domain are present (reports as an error, if not present)
- Any additional variables present in the domain but not defined by the SDTM model (reports as an error, if present)
- Any expected domain variables that are not present in the domain (reports as a warning, if not present)
- Any permitted domain variables that are not present in the domain (reports as a note, if not present)
- All domain variables are of expected data type (reports as an error, if not the same type)

Example 3: Observation level (domain data content of the source on a per observation basis) changes made to the sample DM dataset:

a) Required variable (SUBJID) set to missing, expected variable (RFENDTC) set to missing and timing variable (DMDTC) format changed from ISO 8601 to DATE9.

Log Contents for Example 3a:

```
ERROR: Required variable SUBJID has a MISSING value in observation 1
ERROR: Required variable RFENDTC has a MISSING value in observation 1
ERROR: Variable DMDTC has incorrect content in observation 1.
       Incorrect data is 25DEC2007
       Invalid characters in datetime expression.
       ISO-8601 document reference section: 5.4.1 - Complete representation.
ERROR: DomainData DM data content encountered prior errors.
NOTE: PROCEDURE CDISC used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

This example demonstrates that PROC CDISC validates the compliance of the SDTM domains on an observation level (data content of the source on a per observation basis) for:

- All the required variables in the domain do not contain missing values (reports as an error, if value missing)
- Any expected variables in the domain do not contain missing values (reports as an error, if value missing)
- Detects the conformance to ISO 8601 for any date, time, datetime, duration or interval type variables (reports as an error, if any do not conform)

Despite these demonstrated features, there are some noticeable limitations to using PROC CDISC:

- PROC CDISC is not available for SAS versions earlier then 8.2. SAS version 8.2 and later may require separate download and installation of the PROC CDISC engine components. Every field response release from SAS that upgrades PROC CDISC functionality and adds support for newer models (e.g., SDTM IG version 3.1.1 or version 3.1.2) would necessitate users to update their installation. That in turn might require qualified environment users to run SAS Installation Qualification (IQ) tests frequently on their system.

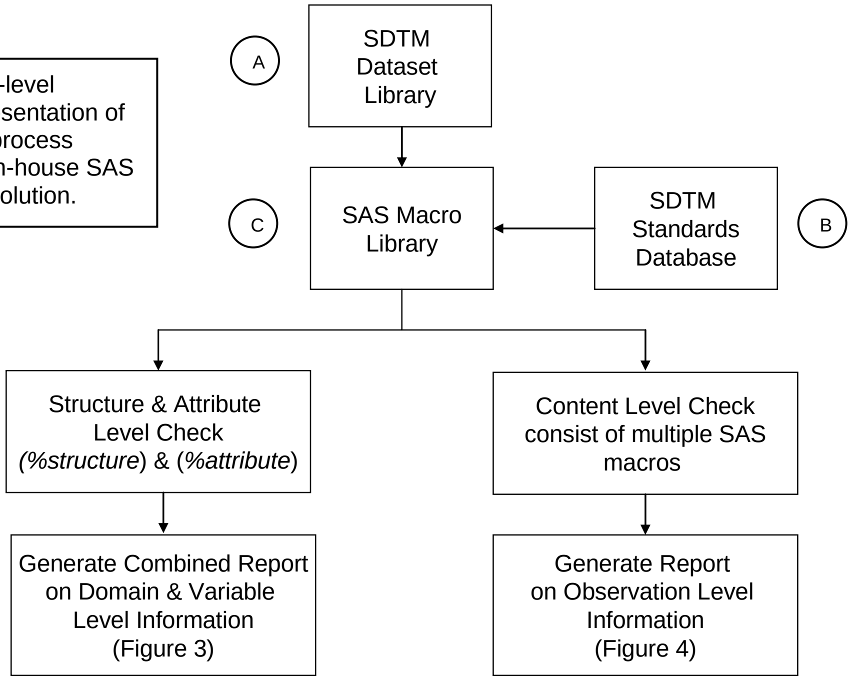
- The procedure does not create a report describing the type of checks and actions it runs on the SDTM domain upon completion of a successful validation. In addition, there is only limited documentation on the list of the checks and the assumptions made.
- The procedure generates reports, warnings, and notes of the discrepancies in the SDTM domain in the SAS log window with no direct option to save them in other more user-friendly output formats (e.g., Excel or Word or even SAS Output window).
- There appears to be a hierarchy for which errors are reported in the log after each run. For instance, if all of the inaccuracies from Example 2 and 3 are combined, the error log will first only indicate that the required variables are not present (Example 2a). Only after this condition is met, will the other inaccuracies be reported in the order demonstrated in Example 2b, 2c and 3.
- The current version (2.15.59) of the procedure does not validate the content of variables with controlled terminology against a list of acceptable values.
- The procedure does not provide a facility to check variable label information (demonstrated in Example 1 above where the PROC CDISC does not indicate missing labels in the log).
- The procedure currently does not support validation of custom-defined domains.

To circumvent the identified drawbacks of using PROC CDISC in the validation of SDTM datasets, we decided to build an in-house SAS macros-based solution that not only incorporates all the functionality of PROC CDISC but also includes many of the capabilities of the WebSDM tool used by the FDA.

IN-HOUSE - SAS MACRO BASED SOLUTION

The in-house SAS based solution includes a library of SAS macros that checks each SDTM domain for compliance with the latest SDTM IG (currently, Version 3.1.1). The checks in this solution are divided into three different categories: structure, attribute, and content level. A high-level graphical representation of the validation process utilized by this in-house SAS macro based solution is outlined below (Figure 1). The macros designed to run these checks access a SDTM standards database developed in-house using FileMaker® Pro. This database holds the domain, variable and format/controlled terminology level information from the SDTM IG (Figure 2). In addition, it also holds specifications for the user-defined domains and sponsor defined controlled terminology.

Figure 1: High-level graphical representation of the validation process utilized by an in-house SAS macro-based solution.


A
SDTM Dataset Library
C
SAS Macro Library
SDTM Standards Database
B
Structure & Attribute Level Check (*%structure*) & (*%attribute*)
Content Level Check consist of multiple SAS macros
Generate Combined Report on Domain & Variable Level Information (Figure 3)
Generate Report on Observation Level Information (Figure 4)

The detailed description of each block in Figure 1 is as follows:

A. SDTM Dataset Library: This is a library of SDTM datasets in SAS or XPT format for a study. If any dataset is in

XPT format, the SAS macro *(%xpt2sas*) is utilized from the SAS Macro library to convert the entire library of XPT files to a SAS datasets library.

B. SDTM Standards Database: This is a database containing information about the SDTM domains from the latest SDTM IG (currently, version 3.1.1). It is built in FileMaker Pro and consists of tables that hold domain (Figure 2: Table 1), variable (Figure 2: Table 2) and format/controlled terminology (Figure 2: Table 3) level information based on the SDTM IG or defined by sponsors for the custom-defined domains. The tables in this database can be easily updated to include additional information from the newest release version of the SDTM IG (upcoming version 3.1.2) or to edit information on the existing custom-defined domains. Additionally, the database can even be expanded to include Analysis Dataset Model (ADaM) standards information.

The standards database could also be build using other database (e.g., Microsoft® Access) or spreadsheet (e.g., Microsoft® Excel) software.

Figure 2: SDTM Standards Database (FileMaker Pro) Tables that hold domain, variable and format/controlled terminology level information. These tables are accessed by the SAS Macro Library to validate the compliance of the SDTM domain datasets at the structure, attribute and content level.

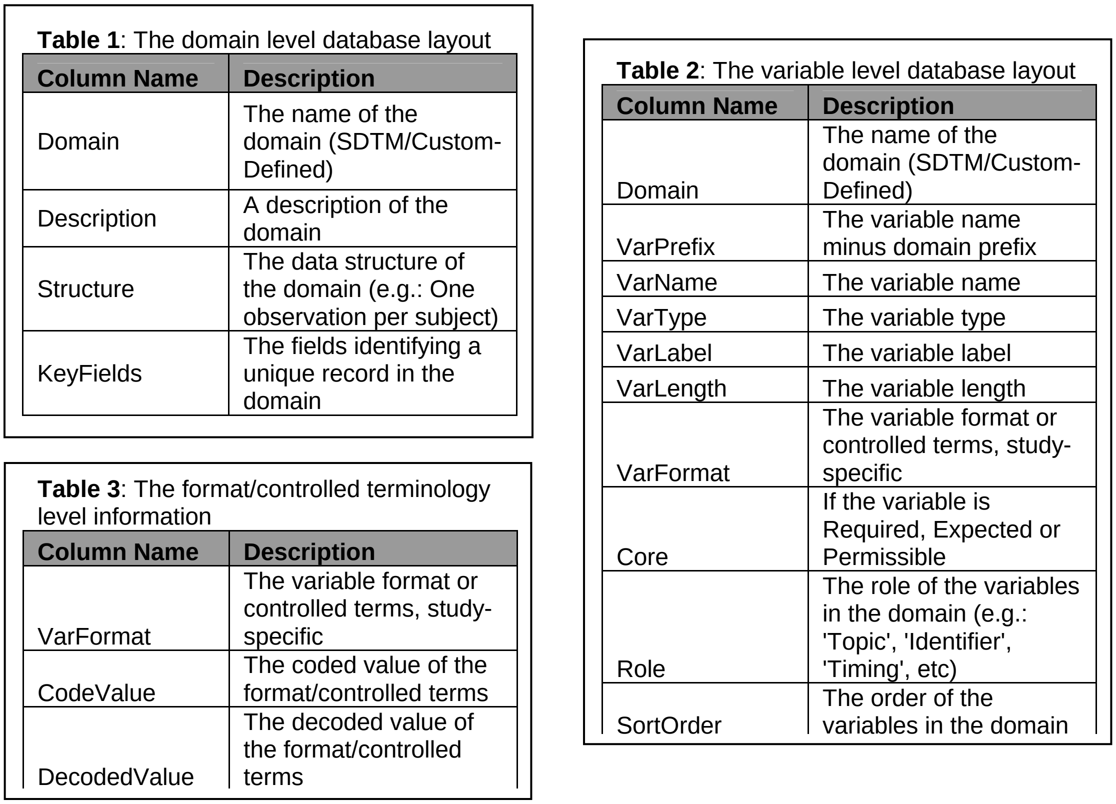


Table 1: The domain level database layout

Column Name	Description
Domain	The name of the domain (SDTM/Custom-Defined)
Description	A description of the domain
Structure	The data structure of the domain (e.g.: One observation per subject)
KeyFields	The fields identifying a unique record in the domain

Table 2: The variable level database layout

Column Name	Description
Domain	The name of the domain (SDTM/Custom-Defined)
VarPrefix	The variable name minus domain prefix
VarName	The variable name
VarType	The variable type
VarLabel	The variable label
VarLength	The variable length
VarFormat	The variable format or controlled terms, study-specific
Core	If the variable is Required, Expected or Permissible
Role	The role of the variables in the domain (e.g.: 'Topic', 'Identifier', 'Timing', etc)
SortOrder	The order of the variables in the domain

Table 3: The format/controlled terminology level information

Column Name	Description
VarFormat	The variable format or controlled terms, study-specific
CodeValue	The coded value of the format/controlled terms
DecodedValue	The decoded value of the format/controlled terms

C. SAS Macro Library: This consists of multiple macros that are categorized into structure, attribute and content level checks. The structure and attribute level checks consist of two SAS macros (*%structure* and *%attribute*). While the content level checks consist of multiple macros that validate the compliance of the SDTM datasets on a per observation basis.

➢ Structure and Attribute Level Check:

%structure: This SAS macro checks the structural level information such as domain name, domain description (label), variable name, and variable order in all of the SDTM datasets in a library versus the standards database. The information from the SDTM datasets can be obtained by using procedure PROC CONTENTS and using OUT= option to output the metadata information for all of the datasets in a library ('StudyABC') into a single SAS dataset ('*metadata*').

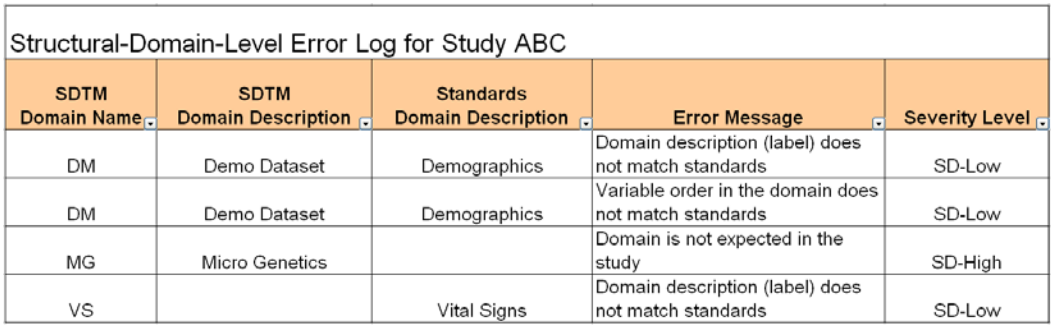
```
** PROC CONTENTS to get metadata information of all SDTM SAS datasets in a library **;
PROC CONTENTS DATA=STUDYABC._ALL_ OUT=METADATA;
RUN;
```

%attribute: This SAS macro checks the attribute level information such as variable label, variable type, and variable length of the common variables present in both the SDTM datasets library and standards database. The attribute information from the SDTM datasets can be obtained from the same '*metadata'* dataset created for the earlier SAS macro.

The *metadata* dataset will have information about all the SDTM datasets present in the library at the domain and variable level. This dataset, while retaining the variable of interest and re-structuring, can be divided into three separate datasets: one that will hold structural-domain-level information used for comparison with the standards database Table 1; the second and third that will hold structural-variable-level and attributes-variable-level information respectively used for comparison with Table 2. The individual standards table from the FileMaker Pro database can easily be converted to a Microsoft Excel or Comma Separated Value (CSV) file for import to a SAS dataset.

After running the comparison between the metadata information obtained from the SDTM datasets versus the standards database, the output from the SAS macros *%structure* and *%attribute* are combined into one multi-sheet Excel workbook (Figure 3) [20] consisting of identified structural or attribute level discrepancies divided into three separate tabs (Figure 3a, 3b, and 3c). Note: the severity level assigned to each check is based on the information obtained from the Janus Operational Pilot - SDTM Validation Specification, v1.0, November 2007 (draft) [5]. The severity levels define the impact of errors on the official receipt of SDTM submissions for Janus processing ('high – indicates the error is serious and will prevent the study data from being loaded into the Janus repository, medium – the error may impact reviewability of the submission, but will not impact on loading the study data into Janus repository, low – the error may or may not impact reviewability or the integrity of the submission and will not have an impact on loading the study data into the Janus repository') [5]. If the check is not defined by the specification from Janus Operation Pilot, we mark it as 'Sponsor Defined' (SD) and based on our internal processes and standards, assign it as either a SD-low, SD-medium or SD-high severity level.

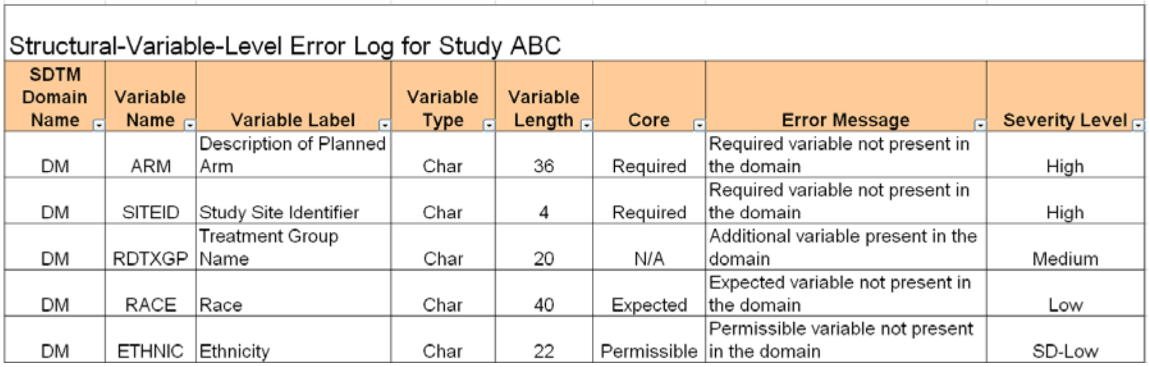
Figure 3: Snap-shots of the multi-sheet Excel workbook output that consist of structural and attribute level error log generated by the SAS macros *%structure* and *%attribute* on the SDTM Dataset Library for Study ABC.



Structural-Domain-Level Error Log for Study ABC

SDTM Domain Name	SDTM Domain Description	Standards Domain Description	Error Message	Severity Level
DM	Demo Dataset	Demographics	Domain description (label) does not match standards	SD-Low
DM	Demo Dataset	Demographics	Variable order in the domain does not match standards	SD-Low
MG	Micro Genetics		Domain is not expected in the study	SD-High
VS		Vital Signs	Domain description (label) does not match standards	SD-Low

Figure 3a: Snap-shot of the Structural-Domain-Level Error Log Output

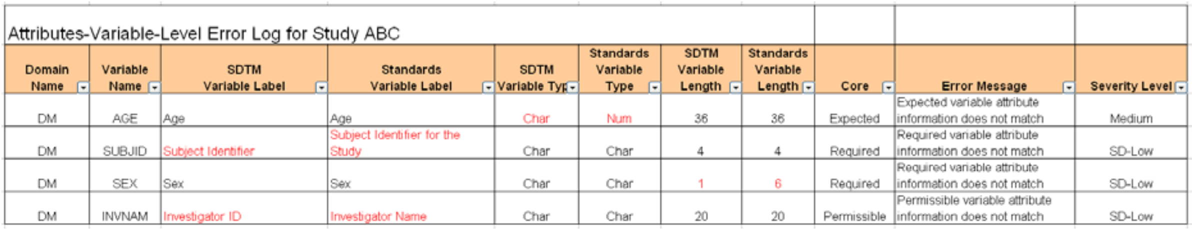


Structural-Variable-Level Error Log for Study ABC

SDTM Domain Name	Variable Name	Variable Label	Variable Type	Variable Length	Core	Error Message	Severity Level
DM	ARM	Description of Planned Arm	Char	36	Required	Required variable not present in the domain	High
DM	SITEID	Study Site Identifier	Char	4	Required	Required variable not present in the domain	High
DM	RDTXGP	Treatment Group Name	Char	20	N/A	Additional variable present in the domain	Medium
DM	RACE	Race	Char	40	Expected	Expected variable not present in the domain	Low
DM	ETHNIC	Ethnicity	Char	22	Permissible	Permissible variable not present in the domain	SD-Low

Figure 3b: Snap-shot of the Structural-Variable-Level Error Log Output

Figure 3: Continued



Attributes-Variable-Level Error Log for Study ABC

Domain Name	Variable Name	SDTM Variable Label	Standards Variable Label	SDTM Variable Typ	Standards Variable Type	SDTM Variable Length	Standards Variable Length	Core	Error Message	Severity Level
DM	AGE	Age	Age	Char	Num	36	36	Expected	Expected variable attribute information does not match	Medium
DM	SUBJID	Subject Identifier	Subject Identifier for the Study	Char	Char	4	4	Required	Required variable attribute information does not match	SD-Low
DM	SEX	Sex	Sex	Char	Char	1	6	Required	Required variable attribute information does not match	SD-Low
DM	INVNAM	Investigator ID	Investigator Name	Char	Char	20	20	Permissible	Permissible variable attribute information does not match	SD-Low

Figure 3c: Snap-shot of the Attributes-Variable-Level Error Log Output. The SDTM column represents attribute information for the SDTM submission datasets while the standards column represent the information present in the study-specific standards database tables.

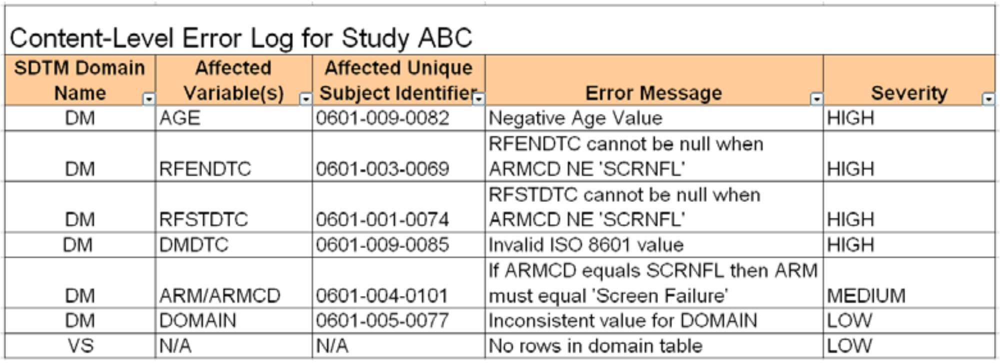
The checks performed by SAS macros *%structure* and *%attribute* include:

- All the required SDTM domains are present in the library and any additional domains that are present in the library but not in the standards database are reported (Sponsor Defined Severity: High)
- The dataset description (label) for each domain is accurate (Sponsor Defined Severity: Low)
- The order of the variables in each dataset is as defined by the standards (Sponsor Defined Severity: Low)
- All required and expected variables are present in the dataset (Severity: High and Low respectively)
- Any variables in the dataset that are not defined in the standards database are reported (Severity: Medium)
- Any permitted variables that are not in the dataset are reported (Sponsor Defined Severity: Low)
- All the variable labels (description) and variable lengths in the dataset match the in-house standards database (Severity: Sponsor Defined: Low, for both)
- All the variable types match with the standards database (Severity: Medium)

➢ Content Level Check:

The content level section consists of multiple SAS macros that are designed to check the SDTM dataset library against the pre-defined set of rules that are based on the Janus Operational Pilot SDTM validation specification. For the scope of this paper, authors have listed (Table 4, next page) several utility SAS macros developed in-house that are limited to 11 check rules applicable to either ALL domains or specific to the Demographics (DM) domain. This is not an extensive list of checks and is not adequate to validate the compliance of all the SDTM submission-ready datasets in the library. The macro library can be expanded to have all the rules on the content (observation)-level information described in the specification document provided by Janus Operational Pilot to validate the SDTM submissions for Janus processing. Sample output from these macros is illustrated in Figure 4.

Figure 4: Snap-shot of the Content-Level Error Log Output. The SAS-macro library validates the compliance of the SDTM domain for the check rules listed in the Table 4 above. The output from each macro is combined to create a single file that list all the discrepancies found for each domain.



Content-Level Error Log for Study ABC

SDTM Domain Name	Affected Variable(s)	Affected Unique Subject Identifier	Error Message	Severity
DM	AGE	0601-009-0082	Negative Age Value	HIGH
DM	RFENDTC	0601-003-0069	RFENDTC cannot be null when ARMCD NE 'SCRNFL'	HIGH
DM	RFSTDTC	0601-001-0074	RFSTDTC cannot be null when ARMCD NE 'SCRNFL'	HIGH
DM	DMDTC	0601-009-0085	Invalid ISO 8601 value	HIGH
DM	ARM/ARMCD	0601-004-0101	If ARMCD equals SCRNFL then ARM must equal 'Screen Failure'	MEDIUM
DM	DOMAIN	0601-005-0077	Inconsistent value for DOMAIN	LOW
VS	N/A	N/A	No rows in domain table	LOW

Table 4: An example of SAS macros that validate the compliance of the SDTM datasets (limited to 11 check rules) at the content level. The severity associated with each macro is assigned based on the information provided in the validation specification by Janus Operation Pilot.

Macro Name	Purpose (Check Rules)	Applicable Domain(s)	Applicable variable(s)	Severity
%zerorows	reports domain table that has zero observations and contains no data	ALL	N/A	Low
%domain	identifies records where the value in the Domain Abbreviation column (DOMAIN) does not match with the domain name	ALL	DOMAIN	Low
%nullreqvar	reports a missing value found in a column where the Core attribute in the standards database is Required	ALL	ALL Required variables	Medium
%iso8601	identifies records where the value for a date does not conform to the ISO8601 standard	ALL	ALL Date variables	High
%nullrfdt	identifies records where - Reference Start Date/Time (RFSTDTC) EQ “ “ and Planned Arm Code (ARMCD) NE ”SCRNFAIL” or Reference End Date/Time (RFENDTC) EQ “ “ and Planned Arm Code (ARMCD) NE ”SCRNFAIL”.	DM	RFSTDTC/ RFENDTC	High
%armscrnfail	identifies records that violate the condition – if Planned Arm Code (ARMCD) =”SCRNFAIL” then Description of Planned Arm (ARM) =”Screen Failure”, and vice versa	DM	ARM/ARMCD	High
%agechk	reports records with negative Age (AGE) value. Missing AGE is ignored.	DM	AGE	High
%ageunit	identifies records with missing Age Units (AGEU) where Age (AGE) is not missing	DM	AGEU	Medium
%ageucodelist	identifies records where value for Age Unit (AGEU) can not be found in the codelist (AGEUNIT) for AGEU variable from the standards database Table 3, limited to records where AGEU is not missing.	DM	AGEU	High
%sexcodelist	identifies records where value for Sex (SEX) can not be found in the codelist (SEX) from the standards database Table 3.	DM	SEX	High
%dupusubj	identifies records where Unique Subject Identifier (USUBJID) variable is not unique. This is applicable only when USUBJID is not missing.	DM	USUBJID	High

Note: At the time of writing this paper, the draft version of the validation specification document was used. Any changes in the final version of this document may necessitate changes.

CONCLUSION

Tools are required by the pharmaceutical industry to validate submission datasets according to the CDISC SDTM. Ideally these tools may be customized to also validate in-development SDTM, near-SDTM or custom domains. WebSDM and PROC CDISC are two tools available for this purpose. WebSDM is used by the FDA but is not optimized for SDTM datasets “in-development” and requires customization for proprietary or “near-SDTM” structured datasets. PROC CDISC is a low cost solution but has limitations and cannot be customized.

The in-house SAS macro-based solution described in this paper is a cost effective means to provide greater flexibility to users by providing options to generate customized checks and reports specific to user requirements, both for SDTM domains and for user-defined datasets. The implementation of this solution was under way at the time of submitting this paper. The authors plan to present additional details at this year’s PharmaSUG (June’ 08) proceedings.

REFERENCES

[1] US Health and Human Services - Food and Drug Administration, "Guidance for Industry: Providing Regulatory Submissions in Electronic Format – Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications", Revision 1 issued April 19, 2006. http://www.fda.gov/cder/guidance/7087rev.pdf.

[2] FDA Critical Path Opportunities Initiated During 2006, http://www.fda.gov/oc/initiatives/criticalpath/opportunities06.html

[3] The Regulatory Plan, Item 36, "Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics". Federal Register Vol. 71, No. 237, Page 72784. December 11, 2006.

[4] Elise Blaese (2007), IBM Healthcare and Life Sciences, "Janus Clinical Data Architecture, Introduction and Overview" with updates from Jay Levine (FDA), and Wayne Kubick (Lincoln Technologies). http://gforge.nci.nih.gov/docman/?group_id=142

[5] Janus Operational Pilot, US Food and Drug Administration. http://www.fda.gov/oc/datacouncil/janus_operational_pilot.html

[6] Janus and NCI GForge, http://gforge.nci.nih.gov/projects/janus, Janus and NCI CRIX, http://crix.nci.nih.gov/projects/janus/

[7] Ken Getz (2007), Tufts CSDD, CDISC 2007 International Interchange, October 2007, "Study on the Adoption and Attitudes of Electronic Clinical Research Technology Solutions and Standards". Session 8 – Business Case/Metrics. http://www.cdisc.org/publications/interchange2007/session8/TuftsGetzCDISCInterchangeOct2007.pdf

[8] Jack Shostak (2005), "Implementation of the CDISC SDTM at the Duke Clinical Research Institute". Proceedings of the PharmaSUG 2005 conference. http://www.lexjansen.com/pharmasug/2005/fdacompliance/fc01.pdf

[9] Susan J. Kenny, Michael A. Litzsinger (2005), "Strategies for Implementing SDTM and ADaM Standards". Proceedings of the PharmaSUG 2005 conference. http://www.lexjansen.com/pharmasug/2005/fdacompliance/fc03.pdf

[10] Edelbert Arnold, Ulrike Plank (2005), "Customer oriented CDISC implementation", proceedings of the Phuse 2005 conference. http://www.lexjansen.com/phuse/2005/cd/cd10.pdf

[11] Phase Forward, WebSDM™ data sheet, https://www.phaseforward.com/products/clinical/ads/

[12] Armando Oliva (2005), FDA, "The Review of SDTM Datasets at CDER: A Clinical Reviewer's Perspective", proceedings of the FDA public meeting held on Feb 1, 2005 at Rockville, MA. http://www.fda.gov/oc/datacouncil/presentations.html

[13] Phase Forward, White paper on WebSDM™ v2.6 Edit Checks, "Validation Checks Performed by WebSDM (version 2.6) on SDTM version 3.1.1 Datasets", https://www.phaseforward.com/products/safety/

[14] Sally Cassells (2007), Phase Forward, WebSDM and Janus presentation, proceeding of the DC area CDISC User Networks, Dec 6, 2007

[15] SAS® and the Clinical Data Interchange Standards Consortium (CDISC), http://www.sas.com/industry/pharma/cdisc/

[16] CDISC Procedure for the CDISC SDTM 3.1 Format, http://support.sas.com/rnd/base/xmlengine/proccdisc/cdiscsdtm.html

[17] Anthony Friebel, Thomas Cox, Edward Helton (2005), SAS Institute, "SAS® Dataset Content Conversion to CDISC Data Standards", proceeding of the PharmaSUG 2005 conference. http://www.lexjansen.com/pharmasug/2005/sasinstitute/sas04.pdf

[18] PROC CDISC Field Response Release for SAS 9 and SAS 8.2, http://support.sas.com/rnd/base/xmlengine/proccdisc/index.html

[19] William C. Csont (2007), "Taking a Walk on the Wildside: Use of the PROC CDISC-SDTM 3.1 Format", proceeding of the PharmaSUG 2007 conference, http://www.lexjansen.com/pharmasug/2007/cc/cc23.pdf

[20] Vince DelGobbo (2007), SAS Institute, "Creating Multi-Sheet Excel Workbooks the Easy Way with SAS®", proceeding of the PharmaSUG 2007 conference, http://www.lexjansen.com/pharmasug/2007/hw/hw09.pdf

ACKNOWLEDGMENTS

The authors would like to acknowledge Cubist Pharmaceuticals, Inc. for giving us the opportunity to work on this topic and the paper. Although, the authors (first and second) on this paper are Independent Consultants in their respective field, they do represent Cubist Pharmaceuticals, Inc. and were supported for their services on the project by Cubist. We would also like to express gratitude to our internal reviewing team for providing us significant comments on the paper. We are grateful to have Sally Cassells (Vice President of Clinical Systems at Lincoln Technologies) review the Lincoln Technologies – WebSDM section of the paper and greatly appreciate her feedback. A special thanks to Kim Lindfield, Conni Otradovec, Cassandra Pauyo and other group members for giving us valuable feedback.

RECOMMENDED READING

We recommend the following documents that are extensively referenced in this paper:

- Janus Operational Pilot - SDTM Validation Specification, v1.0, November 2007 (draft), for more information on the topic follow link: http://www.fda.gov/oc/datacouncil/janus_operational_pilot.html
- CDISC SDTM Implementation Guide Version 3.1.1, September 2005, for more information on the topic follow link: http://www.cdisc.org/models/sdtm/v1.1/index.html

CONTACT INFORMATION

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

Bhavin Busa
SAS Programmer
Independent Consultant
Waltham, MA, 02452
631-220-5446
bhavinbusa@gmail.com
bhavin.busa@cubist.com

Jameelah Aziz
Sr. Manager, Clinical Data Management
Cubist Pharmaceuticals, Inc.
65 Hayden Avenue,
Lexington, MA, 02421
781-860-8432
jameelah.aziz@cubist.com