PharmaSUG2010 - Paper CD13

In-Depth Review of Validation Tools to Check Compliance of CDISC SDTM-Ready Clinical Datasets

Bhavin Busa, Cubist Pharmaceuticals, Inc., Lexington, MA
Kim Lindfield, 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.

In June 2008, we presented a paper at PharmaSUG titled - '*Validating CDISC SDTM-Compliant Submission-Ready Clinical Datasets with an In-House SAS® Macro-Based Solution'*. The paper described two options (SAS – PROC CDISC and Lincoln Technologies WebSDM™) available at that time along with an in-house SAS macro-based solution for validating the compliance of submission-ready clinical datasets. We concluded in our paper that the in-house SAS macro-based solution was more cost effective and provided greater flexibility than the other two options.

This paper will build upon the same topic and discuss additional validation tools (SAS - Clinical Standards Toolkit and OpenCISC Validator) that have since become available for checking the compliance of submission-ready clinical datasets. We will perform in-depth review of these tools and list pros and cons of each solution by implementing them on a real submission-ready clinical datasets.

INTRODUCTION

The FDA has endorsed CDISC SDTM as the preferred model for submitting clinical and bioequivalence data in the eCTD guidance [1]. Since then pharmaceutical organizations are quickly adopting the SDTM model both because of FDA submission requirements [2] and because of benefits to the industry in communicating and sharing standardized data [3]. 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.

In June 2008, we presented a paper [4] at PharmaSUG titled - '*Validating CDISC SDTM-Compliant Submission-Ready Clinical Datasets with an In-House SAS® Macro-Based Solution'*. The paper described two options (SAS – PROC CDISC and Lincoln Technologies WebSDM™) available at that time along with an in-house SAS macro-based solution for validating the compliance of submission-ready clinical datasets. As a part of introduction to this paper we will provide high-level schematic of SDTM data flow from Sponsor to FDA; list needs and requirements of SDTM validation tools; and briefly summarize three previous solutions described in the earlier paper. However, we do recommend readers to reference our earlier paper [4] for more details.

SDTM DATA FLOW FROM SPONSOR TO FDA

The FDA has developed a standards-based clinical data repository, Janus [3]. 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 [3, 6, and 7]. 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.

The schematic (Figure 1) describes a high-level flow of SDTM data from the sponsor to the FDA. In order to make an electronic submission in eCTD format using CDISC SDTM standards, the sponsor converts the CDMS data to the final SDTM structure. Once the sponsor submits both SDTM domains and define.xml, it gets loaded into the FDA Electronic Document Room (EDR) [3, 5]. A study submitted in SDTM format must include a define.xml file without which SDTM datasets will not be sent to the Janus staging area during the locating and extracting SDTM datasets from EDR. Upon receipt of all required submission files into the server, EDR notifies the Data Load and Validation staging area that the datasets are available for load into Janus.

Before the domains get loaded into Janus data warehouse it undergoes two layers of validation checks (WebSDM and Janus). These checks include set of validation rules that are run to check the compliance of each domain with the SDTM standard. The checks available include detection of structural and consistency errors which are rated by severity [For a detailed list of WebSDM and Janus validation checks – refer to references 6, 8]. The severity levels defines 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') [6]. If the SDTM domains and define.xml passes all the required WebSDM and Janus checks – it gets loaded into Janus data repository after which it becomes available to FDA for review using tools such as J-Review and WebSDM. However, if there are any discrepancies in the submission datasets (structural or content) it generates an error log and depending on the severity and impact on the Janus load, it might be returned back to the sponsor for corrections. This can cause significant delay to the submission review process, and may result in increased costs to the sponsor.

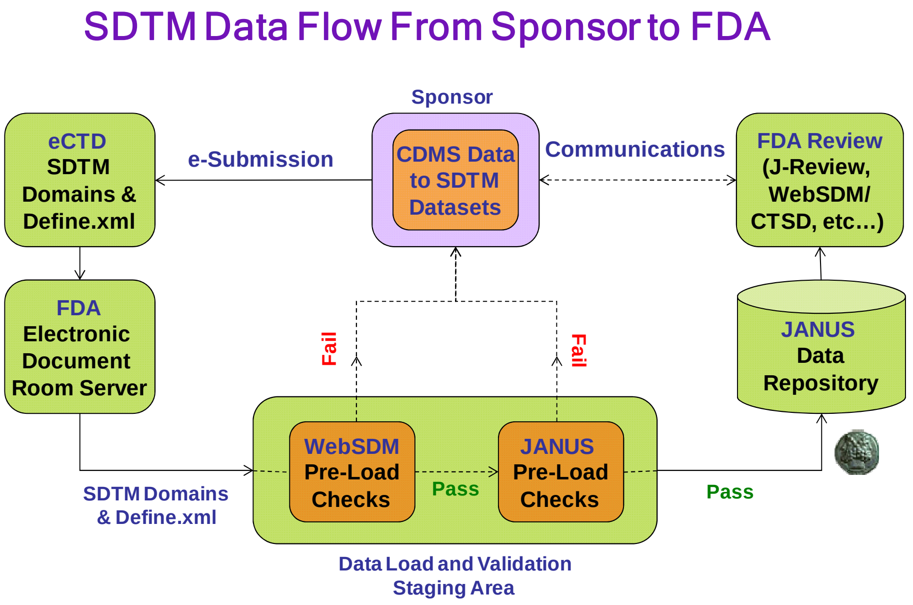

SDTM Data Flow From Sponsor to FDA
Sponsor
eCTD SDTM Domains & Define.xml
e-Submission
CDMS Data to SDTM Datasets
Communications
FDA Review (J-Review, WebSDM/ CTSD, etc…)
FDA Electronic Document Room Server
Fail
Fail
JANUS Data Repository
WebSDM Pre-Load Checks
Pass
JANUS Pre-Load Checks
Pass
SDTM Domains & Define.xml
Data Load and Validation Staging Area

Figure 1: Schematic of SDTM Data Flow from Sponsor to FDA.

NEEDS AND REQUIREMENTS OF SDTM VALIDATION TOOLS

In order to avoid any errors in the submitted datasets, sponsors and CROs will greatly benefit from the validation of the compliance of data before submission to the FDA. In other words, it will be beneficial to have a validation tool that sits right before the SDTM domains are submitted to the FDA. Therefore it is crucial to identify a validation tool upfront and chose the one that fits sponsor needs and requirements towards successful submission of a study in SDTM format. Below we have provided Cubist defined needs and requirements of a validation tool which can vary from sponsor to sponsor.

The needs of a validation tool are:

- Check the compliance of the SDTM domains for successful load into Janus.
- Reduce risk of delay in the submission review process.

Sponsor requirements of a validation tool are:

- Check the compliance of the SDTM domains as per the most recent SDTM Implementation Guide.
- Apply all WebSDM and Janus pre-load checks.
- Identify structural and consistency errors.
- Versatile for use throughout the Clinical Data Life Cycle.
- Flexible enough to handle variations across release and as new standards are developed and become available.

SDTM VALIDATION TOOLS

PREVIOUSLY EVALUATED 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 [9]. This application tests the compliance of submission-ready files (in SAS V5 Transport format or Oracle® views) according to the SDTM IG [9]. 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. We recommend referencing our earlier paper [4] for further information on this tool. Please note that the newer version of WebSDM 3.0 [8] is now become available and includes additional validation checks that covers SDTM IG 3.1.2 and includes the adoption of CDISC SDTM standard terminology.

- SAS – PROC CDISC:

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) [10, 11]. 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 [12]. The current SAS procedure PROC CDISC supports validation of 15 of the 23 domains outlined in CDISC SDTM version 3.1 [11]. 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. We recommend referencing our earlier paper [4] for further information on this tool. Please note that there has been a field response release 2.15.65 in June 2009 [12]. However, there has not been any change to its capabilities for validating CDISC SDTM SAS dataset.

- 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 2). 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. In addition, it also holds specifications for the user-defined domains and sponsor defined controlled terminology. The authors recommend referencing our earlier paper since for more details on this solution [4].

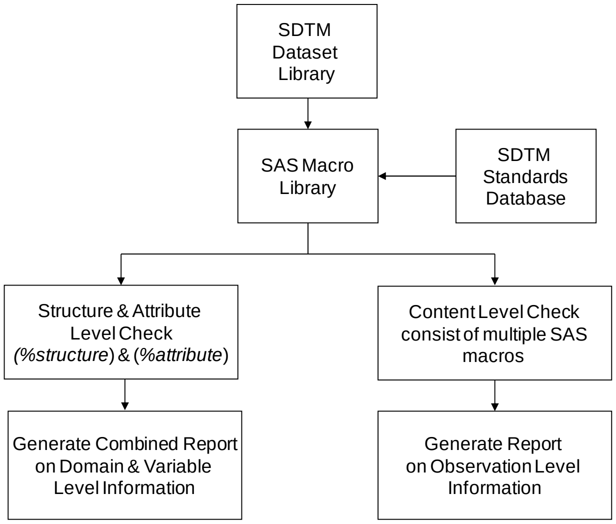

SDTM Dataset Library
SAS Macro Library
SDTM Standards Database
Structure & Attribute Level Check (%structure) & (%attribute)
Content Level Check consist of multiple SAS macros
Generate Combined Report on Domain & Variable Level Information
Generate Report on Observation Level Information

Figure 2: High-level graphical representation of the validation process utilized by an in-house SAS macro-based solution.

NEWLY AVAILABLE SDTM VALIDATION TOOLS

WebSDM and PROC CDISC were two tools available at that time for reviewing compliance of submission-ready clinical datasets. 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. We concluded in our earlier paper that the in-house SAS macro-based solution was more cost effective and provided greater flexibility than the other two options to generate customized checks and reports specific to user requirements, both for SDTM domains and for user-defined datasets.

Since June 2008 there are few more tools that have become available for validating the compliance of submission-ready clinical datasets. Peter Villiers from SAS Institute announced in PharmaSUG’08 [13] that SAS is working on Clinical Standards Toolkit (CST) which will be released as a download for SAS 9.1.3 and SAS 9.2 to the industry by end of 2008. Just few months later, November 2008, OpenCDISC [14] announced release of CDISC Validator. In this paper we will perform in-depth review of both these tools and list pros and cons of each solution by implementing them on a submission-ready clinical datasets.

- SAS – Clinical Standards Toolkit:

The SAS Clinical Standards Toolkit (CST) is a SAS macro based framework to check the compliance of clinical data and the metadata against a reference standard (i.e. SDTM 3.1.1). The toolkit (version 1.2) enables users to validate the structure and terminology of their submission-ready clinical datasets for regulatory submission purpose. It provides 143 unique validation checks which include both WebSDM and Janus rules as well as SAS custom checks. Furthermore, it is also design to produce CRT-DDS (define.xml) documentation file. The toolkit v1.2 is licensed as a part of Base SAS and is available as a download for SAS 9.1.3 and SAS 9.2 customers at no additional cost since July 2009 [15]. As described in the paper presented by Peter Villiers [16], the main objective of the SAS CST is to support the validation of standards used in regulatory submissions and at the same time provide functionality which is familiar to SAS programmers.

Installation: The toolkit v1.2 is available as a hot fix for SAS 9.1.3 and SAS 9.2 customers at no additional cost. Since SAS-CST is available as a hot fix - user will need to have BASE SAS installed on their machine prior to installing this tool. Prior to installation of SAS-CST, it is a pre-requisite to have Apache Ant V1.7.1 and Java environment (1.4 or later as a pre-requisite for Ant) on your system. Ant is a free java based tool that is used to move and edit SAS-CST installation files. The detailed installation instructions can be found from the SAS website [15].

Framework: SAS macro based framework consists of two distinct pieces [17]:
- The components that are installed as part of SAS foundation and shared files (SAS macros, Java JAR files, etc.)
- The global standards library that the CST framework macros operate on top of.

Upon installation of SAS-CST, the tool creates a global standards library locally on the system. The global standards library consists of configuration directories and files for each standard (e.g. SDTM v3.1.1) that are referenced by CST macros. SAS-CST currently provides three modules – CDISC SDTM V3.1.1, CDISC CRT-DDS 1.0 (define.xml), and CDISC Terminology which are part of global standards library. Each module can be considered as a stand-alone and consist of datasets (control files, validation files, standards metadata files) and sets of standard specific macros.

The framework basic module handles the registration of standard-versions (e.g. SDTM v3.1.1) and has the capability to install new standard-version (e.g. SDTM v3.1.2) without affecting existing standards. Since SAS has used modular approach for this tool, user will not have to revalidate when the new standards becomes available. As the newer versions of the CDISC SDTM standards become available, SAS Institute will need to release new modules to register to the framework.

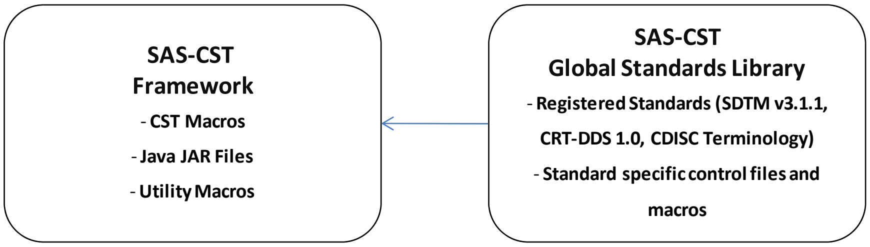

SAS-CST
Framework
- CST Macros
- Java JAR Files
- Utility Macros
SAS-CST
Global Standards Library
- Registered Standards (SDTM v3.1.1, CRT-DDS 1.0, CDISC Terminology)
- Standard specific control files and macros

Figure 3: High-level graphical representation of components of the SAS Clinical Standards Toolkit (CST).

<u>Pros</u>:

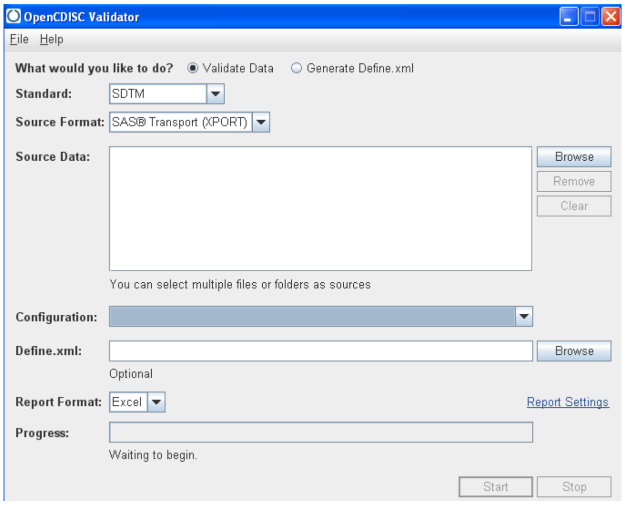
- Includes 143 unique checks that cover all WebSDM and Janus checks.
- SAS defined checks are included.
- User-defined checks are possible.
- Available with Base SAS (v9.1.3 and v9.2) at no additional cost.
- Flexible in upgrading to newly available standards (e.g. SDTM V3.1.2).
- SAS has used modular approach for this tool, no revalidation when the new standards become available.
- User can add their own standard module.
- Supports controlled terminology and CRT-DDS standards.
- Supports generation of Define.xml as part of this tool.
- IQ/OQ documentation to validate the installation of the SAS CST.
- SAS-CST user's guide available for download with the tool.
- Usage through SAS programming environment.
- Future expansion to CDISC ADaM and HL7.
- Sits on top of Base SAS – which already provides a compliant environment from a quality assurance perspective.
- SAS technical support available if needed.

<u>Cons</u>:

- Provided as a hot fix, so require BASE SAS license in order to use this tool.
- Installation of this tool for SAS 9.1.3 was a daunting task and required IT help in order to figure out all the steps towards successful installation.
- Needs pre-requisite tools such as Apache-Ant and Java environment for installation.
- Available for Windows platform only and not available for lower version of SAS.
- Build for SAS programmers and so limits the use of tool just for users with strong programming background. The tool cannot be used by non-SAS programmers (Standards Developer, Data Manager, etc.).
- As the newer versions of the CDISC SDTM standards become available, SAS Institute will need to release new modules to register to the framework. This means users will have to work around SAS release timeline.
- Validation results are presented as a SAS dataset – will need additional programming steps to generate results in a more user-friendly output (e.g. Excel or html).
- Not intuitive – steep learning curve to figure out implementation on real submission study.
- SAS macro based interface. Doesn't support GUI and is not a point and click tool.
- Staff will require training – which can contribute to additional cost.
- Although a quick start document for SDTM validation is available, clear examples are not provided.

- OpenCDISC Validator:

OpenCDISC validator is an open source java based project that provides validation of datasets against CDISC models (e.g. SDTM) [14]. The validation rules are defined using XML-based validation framework with capability to support wide range of rules and flexibility to support SDTM +/- datasets. It is available to download for free and requires no installation of the program on the computer but can instead be run off a USB flash drive. It includes both WebSDM and Janus checks and supports SDTM 3.1.1 and 3.1.2 standards.



What would you like to do? ● Validate Data ○ Generate Define.xml

Standard: SDTM

Source Format: SAS® Transport (XPORT)

Source Data:

You can select multiple files or folders as sources

Configuration:

Define.xml:

Optional

Report Format: Excel

Progress:

Waiting to begin.

Figure 4: Snap-shot of OpenCDISC Validator tool.

Installation: The OpenCDISC Validator 1.0.1 production version is available for download as a zip file on OpenCDISC website [18] at no cost. The validator requires Java Runtime Environment (JRE) version 1.5 or higher and a system with 2GB RAM. To install the validator simply unzip the files in the directory where you want the validator to reside.

Framework: The tool is an open source java based project which works on a XML-based validation framework. The tool architecture separates validation rules from application logic and comprises of two major components [14] – the validation engine and configuration files (xml format). The tool supports both SDTM v3.1.1 and v3.1.2 standards and provides configuration files to be able to use with the tool. The tool is flexible enough to be able to expand to standards such as ADaM, SEND and any others that becomes available in future.

Once the zip file is extracted on to the specified directory - it creates folders that hold configuration files, jar files, and reports. The application can be initiated by double-clicking on 'client.bat' file. The tool provides user with the graphical user interface (see figure 4) and also an option of using the tool in command line interface. The tool is also capable of generating Define.xml. By just providing location of the source data (XPT or delimited format), the tool will able to check the domains for any discrepancies using the standard configuration file provided with the installation. In addition, it is also capable of checking the submission domains using user input Define.xml file.

Pros:

- Includes checks that cover all WebSDM and Janus checks.
- OpenCDISC community defined checks are included.
- Additional user-defined checks are possible.
- Available for windows platform at no additional cost and can be deployed on server or hosted environment.
- Installation in minutes and can be run off a USB flash drive.
- Flexible in upgrading to newly available standards.
- Ready to use tool – no training required.
- The tool can be used by anyone (SAS programmers, Standards Developer, Data Manager, etc.) who has basic understanding of CDISC standards.
- The tool separates validation rules from application logic, so no need to update and revalidate.
- Support SDTM+ or near-SDTM standards.
- Supports generation of Define.xml as part of this tool.
- Supports controlled terminology and can check against Define.xml.
- Very intuitive - use through GUI.
- Generates validation reports in Excel, CSV, and HTML.
- Future expansion to CDISC ADaM and SEND.
- Does not require any additional tools and takes up <10MB space on the drive.
- Technical support available through OpenCDISC forum on their website.

Cons:

- No user's guide or validation documentation provided with the installation.
- Requires user to create their own configuration files if additional sponsor defined checks are to be included.
- Since configuration and validation rules are both in xml format – it requires prior knowledge of the language to be able edit or generate these files.
- As the newer versions of the CDISC SDTM standards become available, OpenCDISC will need to make the configuration files available to the user. This means users will have to work around OpenCDISC timeline.

CONCLUSION

Tools are required by the pharmaceutical industry to validate submission datasets according to the CDISC SDTM standards. Ideally these tools may be customized to also validate in-development SDTM, near-SDTM or custom domains and even check domains for controlled terminology.

In our earlier paper [4] we described two options (SAS – PROC CDISC and Lincoln Technologies WebSDM™) available at that time along with an in-house SAS macro-based solution for validating the compliance of submission-ready clinical datasets. We concluded in our paper that the in-house SAS macro-based solution was more cost effective and provided greater flexibility than the other two options.

In this paper, we evaluated two additional validation tools (SAS CST and OpenCDISC Validator) that have since become available for the industry. We introduced readers to these tools and provided some insight on the framework, installation and their pros & cons. The review of these tools was under way at the time of submitting this paper. The authors plan to use both these tools on a test SDTM datasets provided by CDISC SDTM/ADaM pilot project and provide an in-depth side-by-side comparison with real case scenarios at PharmaSUG (May' 10) 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] 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.

[3] 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

[4] Bhavin Busa, Sheila Vince, Jameelah Aziz (2008), Cubist Pharmaceuticals, Inc., '*Validating CDISC SDTM-Compliant Submission-Ready Clinical Datasets with an In-House SAS® Macro-Based Solution',* proceeding of the PharmaSUG 2008 conference. http://www.lexjansen.com/pharmasug/2008/rs/rs07.pdf

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

[6] Janus Operational Pilot, US Food and Drug Administration. http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM190628.pdf

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

[8] Phase Forward, White paper on WebSDM™ v3.0 Edit Checks, "*Validation Checks Performed by WebSDM (version 3.0)*", https://www.phaseforward.com/products/cdisc/

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

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

[11] 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

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

[13] Peter Villiers (2008), SAS Institute, "*SAS, CDISC and Clinical Data Integration*", proceeding of the PharmaSUG 2008 conference. http://www.lexjansen.com/pharmasug/2008/sas/sa11.pdf

[14] Max Kanevsky (2008), "*Validating SDTM, an open source solution",* proceeding of the CDISC Interchange 2008.

[15] SAS Clinical Standards Toolkit 1.2 for SAS 9.1.3, Installation instructions and IQOQ validation documents can be downloaded from SAS website - http://ftp.sas.com/techsup/download/hotfix/12clintlkt.html

[16] Peter Villiers (2009), SAS Institute, "*Supporting CDISC Standards in Base SAS Using the SAS Clinical Standards Toolkit*", proceeding of the NESUG 2009 conference. http://www.nesug.org/Proceedings/nesug09/ph/ph14.pdf

[17] SAS Clinical Standards Toolkit 1.2 User's Guide can be downloaded from SAS website - http://support.sas.com/documentation/onlinedoc/clinical/toolkit_ug12.pdf

[18] OpenCDISC Validator download tool, instructions and configurations files from OpenCDISC website at: http://www.opencdisc.org/download

RECOMMENDED READING

We recommend the following documents:

- SAS Clinical Standards Toolkit 1.2 User's Guide - http://support.sas.com/documentation/onlinedoc/clinical/toolkit_ug12.pdf

CONTACT INFORMATION

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

Bhavin Busa
Sr. Statistical Programmer
Cubist Pharmaceuticals, Inc.
65 Hayden Avenue,
Lexington, MA, 02421
781-860-8534
bhavin.busa@cubist.com

Kim Lindfield
Sr. Manager, Biostatistics
Cubist Pharmaceuticals, Inc.
65 Hayden Avenue,
Lexington, MA, 02421
781-860-8371
kim.lindfield@cubist.com