Use CDISC SDTM as a data middle-tier to streamline your SAS® infrastructure

Kalyani Chilukuri, Clinovo, Sunnyvale, CA
Ale Gicqueau, Clinovo, Sunnyvale, CA

ABSTRACT

In many life science companies, SAS is used to streamline biometric processes. Some of these activities include but are not restricted to Edit Checks, study metrics, patient profile, analysis data sets, TLGs for submission or even CDISC conversion. Unfortunately, we have a tendency to reinvent the wheels for each new study and adapt the existing code to match the new study's data structure. This process is often tedious and can lead to errors, as we do not fully understand the assumptions of the modified legacy SAS programs.

The convergence of open innovations such as the CDISC SDTM standard and the free SAS-based CDISC Express has brought a better alternative by providing an easy way for sponsor companies and CROs to develop an internal SDTM middle-tier. CDISC SDTM has provided a common data structure for raw clinical data while CDISC Express makes it drastically faster and cheaper to convert any clinical database to the SDTM format using free software. Using this paradigm, companies are able to fully embrace the concept of global standards for all their SAS processes, and leverage their entire standard reporting SAS assets with minimum programming. As CDISC Express is SAS-based, the most complex SDTM transformation can be done in one macro call and seamlessly integrated with internal SAS programs.

There are many benefits to this approach:

- Code re-usability

- No need for re-validation of standard SAS programs

- Better documentation, reliability of SAS standard study programs

- Reduced programming errors

- Faster study data-cleaning and analysis

- Easier FDA submission

This session will demonstrate this concept by showing in practice how to:

- Easily format your study to SDTM by using CDISC Express

- Re-run SAS Edit Checks, Enrollment graph, Patient Profile, ADaM and TLG for a new study with no SAS programming

- Quantify the savings generated with this new approach

INTRODUCTION

In order to speed up FDA submission and FDA review, the pharmaceutical industry is adopting standards for clinical trial data submission. Since CDISC standards came into place, they are subject to constant improvement and have been evolving to meet the requirements of the different clinical trials, encompassing a wide range of therapeutic areas.

The mission of CDISC (Clinical Data Interchange Standards Consortium) is to develop global, platform-independent data standards to catalyze the flow of information through the entire clinical research process. CDISC has developed several standards for clinical trial data submissions including Study Data Tabulation Model (SDTM), Analysis Data Model (ADaM), Operational Data Model (ODM), Case Report Tabulation Data Definition Specification (CRTDDS) –

(define.xml), Clinical Data Acquisition Standards Harmonization (CDASH) and many others.

CDISC SDTM forms the framework for organizing the clinical trial information to be submitted to the regulatory authorities. CDISC CRT-DDS or Define.xml standards give the specifications for the XML-based content and format for the data definitions for CDISC SDTM datasets. The first version of the SDTM standards and SDTM Implementation Guide have been released in 2005. Later in 2008, CDISC released the new SDTM Standards Version 3.1.2. Since the release of the SDTM standards, the standards have been evolving based on the feedback from pharmaceutical industry players, to accommodate the wide variety of clinical trials in different therapeutic areas. The adoption of CDISC SDTM standards has proven to significantly reduce time and cost for drug development. Recently in Feb 2010 FDA's Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER) approving CDISC SDTM standards for electronic submissions. The use of CDISC SDTM standards improves the quality of submissions and speeds up regulatory reviews.

Hence, it is necessary to convert the clinical study data into the CDISC SDTM format. However, currently, many companies redundantly define new conversion rules for each study. To solve this problem and improve the efficiency of data conversion, Clinovo developed CDISC® Express, a SAS-based application that easily maps clinical trial data into SDTM format.

CDISC® EXPRESS

Clinovo released CDISC® Express in April 2011. CDISC® Express is the first open source software available for clinical data conversion to CDISC SDTM format. Mapping definitions to convert clinical data from the source datasets into the SDTM domains are done through an Excel spreadsheet which is created by the programmer. The Excel document is called the mapping file.

MAPPING FILE FOR DATA CONVERSION

In order to perform the data conversion for a new clinical study, the first step is to create a new mapping file. As it only requires changing the definitions on the Excel sheet, it is easy to use and maintain.

The mapping file contains the mapping rules for performing the data conversion from the raw data sources into SDTM format. It consists of the following tabs (Figure 1.):

- StudyMetadata: To create the Define.XML file (will be discussed later in the paper).

- FORMAT: To define custom data formats in this tab.

- SDTM Domain tabs: Contains the mapping rules for the SDTM domains, eg: DM for Demographics, AE for Adverse Events, etc.

- SUPPQUAL Mapping rules for Supplemental Qualifier Domain.

- CO: Mapping rules for Comments Domain.

Figure 1. Mapping file showing different tabs

An example of SDTM Domain TAB for Medical History (MH) is shown in Figure 2:

Dataset	Merge Key	CDISC variable	Expression	Comments	Explanation
medhist	patid	USUBJID	%CONCATENATE(_variables=study sitecode patid)		
		MHTERM	meddiag		
		MHCAT	histtype		
		MHPRESP	%CONVERTIF(_if_variable=histtype,_if _value=TARGETED,_then_value=Y)		
		MHDTC	%FORMAT(_variable=formdat,_format =yymmdd10)		
		MHSTDTC	%FORMAT(_variable=histdat,_format= yymmdd10)		
		MHDY	%STUDYDAY(_date=histdat)		
medhistassess	patid	USUBJID	%CONCATENATE(_variables=study sitecode patid)		medhist and medhistassess are merged by patid as the merge
surgproc	ptnam	USUBJID	%CONCATENATE(_variables=study		
		MHTERM	surgproc		

Figure 2. Example of MH Domain tab

The Domain tab consists of the following columns:

- Dataset: Specifies the name of the source datasets used to create the SDTM domain.

- Merge Key: Specifies the variables needed to merge the datasets in the Dataset column. The unique subject identifier variable USUBJID is used by default.

- CDISC variable: Specifies the CDISC variables that will be created.

- Expression: Specifies the expression for the assignment statement of the SDTM variable in the CDISC variable column.

- Comments: It is used for documentation purpose and will appear in the column 'comment' of your define.xml when using CDISC® Express program to generate the define.xml of the study.

- Explanation: Provides additional details and explanation to help you create the mapping file for your study. It is not used by the CDISC® Express application.

Example of SAS Code generated for creating an AE dataset from a single dataset (Figure 3):

Dataset	Merge Key	CDISC variable	Expression	Comments
advevnt		STUDYID	study	
		DOMAIN	&domain	
		USUBJID	%CONCATENATE(_variables=study sitecode patid)	
		AETERM	aedesc	
		AEMODIFY	aecorr	
		AEDECOD	aept	

Figure 3. Example of AE Domain

```
data ae;
      set advevnt;
      studyid=study;
      domain=&domain;
      usubjid=%concatenate(_variables=study sitecode patid);
      …
run
```

The assignment for each expression for the CDISC SDTM variable can consist of either of the following

- One-to-one mapping of a variable from the source database.

- Constant expression

- Macro variable

- Macro call (eg: %CONCATENATE is a macro to concatenate the variables)

More complex transformation can be applied to get the exact conversion of the variables to the SDTM format. A detailed description of the CDISC mapping file and the interpretation of the mapping file have been described in an other paper published by Clinovo entitled *An Excel Framework to Convert Clinical Data to CDISC SDTM Leveraging SAS Technology* . Multiple source datasets can also be used to create the SDTM variables.

Some of the commonly used macros for standard data transformation are provided with the software. Additional custom macros can easily be developed by SAS programmers for a new clinical study. This method is highly efficient for performing data conversion for multiple clinical trials without reprogramming and revalidation.

SDTM DOMAIN VALIDATION

The SDTM domains created by the software are validated to report any issues including:

- Domain names not supported by CDISC

- Non-SDTM variable names

- Incorrect type, label of variables

- Missing expected and required CDISC SDTM variables in each domain

- Variable values missing but supposed to be non-null

In addition to the domains created by CDISC[®] Express, the SDTM domain validation component of CDISC[®] Express can also be used to validate the SDTM domains created by external sources.

DEFINE.XML

Define.xml is an XML document which provides information about the study metadata. It is requested by the FDA as part of the electronic submission of clinical trial data. The Define.xml file provides a concise roadmap regarding the SDTM domains, datasets, variable metadata and the codelists used in the study. Once the SDTM domains are created, CDISC® Express can automatically generate the Define.xml document for the clinical study.

More information about the contents and structure of the Define.xml can be obtained from the CDISC® website: http://www.CDISC.org/define-xml. Information about the XML elements (as described in the FDA guidelines) is present in the first tab STUDYMETADATA of the mapping file as shown below. The column 'Values' has to be updated based on the study. The study name, the name of the annotated CRF file and the Study description have to be changed accordingly.

XMLField	XMLElement	Status	Values	Comments
ODM Attributes	FileType	Required	Snapshot	
	FileOID	Required	quickstart	
	PriorFileOID	Optional	quickstart define.xml	Reference to the previous file (if any)
	ODMVersion	Required	1.2	
	Originator	Optional	Clinovo, Inc	The organization that generated the define.xm
	SourceSystem	Optional		The computer system, database management system, etc. that is the source of the define.xml
	SourceSystemVersion	Optional		The version of "SourceSystem" above
	CreationDateTime	Required		** Do not fill with any
ODM Child Element	Study	Required		
Study Attributes	OID	Required	quickstart	
Study Child Elements	GlobalVariables	Required		
	MetadataVersion	Required		
Global Variable Child	StudyName	Required	quickstart	Name of Study
	StudyDescription	Required	Study for testing	Description of Study
	ProtocolName	Required	quickstart	The Protocol Name
MetaDataVersion	OID	Required	CDISC.SDTM.3.1.1	
	Name	Required	CDISC SDTM for Study	
	def:DefineVersion	Required	1.0.0	
	def:StandardName	Required	CDISC SDTM	
	def:StandardVersion	Required	3.1.1	
MetaDataVersion	def:AnnotatedCRF	Optional	quickstart blank eCRF	
def:leaf blankcrf	ID	Required	quickstart blank eCRF	
def:leaf blankcrf	xlink:href	Required	quickstart blank eCRF.pdf	
def:leaf blankcrf	def:title	Required	Annotated Case Report Form	

Figure 4. Studymetadata tab in the mapping file

WORKFLOW OF CDISC EXPRESS

The basic steps to follow while using CDISC Express are:

1. *Create the mapping file for the clinical study*

   The mapping file is an Excel 2003 file that contains the mapping rules for performing the data conversion. It has to be created by the user according to their clinical study. CDISC Express supports both versions of CDISC SDTM 3.1.1 and 3.1.2. The application creates a mapping file template based on the version of SDTM version required. The working version of the mapping file called "tmpmapping.xls" has to be filled by the user.

   The expressions to create the CDISC variables from the source datasets should be entered by the user with the help of the study protocol and the structure of the source datasets. It is important for the user to understand the concept of SDTM standards. SDTM standards can be extended and adapted according to the clinical trial requirements. The SAS macros from the function library can be used, and this library can be further extended based on the requirements for the clinical study.

2. *Validate the mapping file*

   CDISC Express is equipped with the feature to check for logical and syntactical errors once the mapping file is created. To validate the mapping file before converting the data to SDTM, the application will check if the file meets the requirements. In case the validation fails, a list of errors will be displayed in the HTML page 'mapping_validation.html' located in the folder \CDISC Express\my study\results\Mapping Validation. You need to correct the mapping file and then validate it a second time. A message will be displayed on the HTML page to inform you that the validation is successful.

   During the validation, the application checks the mapping file and will show an error message if the mapping file doesn't meet the requirements. For example in Figure 5, a CDISC variable for the domain DM has been misspelled. In this case you need to correct the mapping file and then validate it again.

Figure 5. Results of mapping file validation

The error codes and the validation rules are documented in an excel spreadsheet. CDISC Express checks for any error in the format of the excel file, the structure of the tabs in the mapping file, or the CDISC mapping definitions. This spreadsheet is used by the validation program to interpret the error codes, display the variables and/or domains and the type of errors. This list of errors can be extended by adding new error codes and definitions.

3. *Create the CDISC SDTM Domains & Validate the SDTM domains*

The mapping file becomes the source code that converts the raw clinical trial data into SDTM domains. As described earlier in the paper, SDTM domains are validated by CDISC Express to display any discrepancies with the standards. (Figure 6).

Figure 6. Results of SDTM domain validation

4. *Generate the CDISC Define.xml and the .XPT SAS transport datasets*

CDISC Express automatically creates the Define.xml file for the clinical study. The SDTM domain datasets are converted to the transport file format .XPT which is the required format for FDA submission. Define.xml contains several tables containing the metadata about the SDTM domains.

There is a separate SAS program to run within CDISC Express to execute each of the steps above. These programs reference the SAS macros which form the core of the CDISC Express application.

ADVANTAGES OF USING CDISC® EXPRESS

As an open source software, CDISC Express is easy to adopt without significant costs. In order to promote global standards, CDISC Express is open to the community and evolves based on the users' feedback. It is highly extensible: Users can define macros according to their clinical study requirements and these can be added to the pool of existing macros for further use with other similar studies.

The code for data conversion can be reused and adapted for multiple clinical studies , thus saving both time and cost. A clinical study which took about 3 months and three programmers could be performed by one developer in about 2 weeks. Not only do CDISC standards speed up regulatory review, but they also help act as a data middle-tier.

CONCLUSION

This paper illustrates how to use of CDISC SDTM standards as a data middle tier to speed up regulatory review and improve the quality of data submission. There are many benefits of adopting CDISC standards, but it is tedious and time-consuming for many companies. CDISC Express is an open source SAS-based application which provides an easy-to-use tool to efficiently perform clinical data conversion to CDISC SDTM format. This paper demonstrates how to perform the data conversion and gives an overview of the advantages of using CDISC Express.

REFERENCES

- Sophie McCallum, Stephen Chan, PharmaSUG 2011, An Excel Framework to Convert Clinical Data to CDISC SDTM Leveraging SAS Technology http://www.lexjansen.com/pharmasug/2011/ad/pharmasug-2011-ad08.pdf

- CDISC website www.cdisc.org

- CDISC SDTM Implementation Guide (Version 3.1.1), Feb 2005

- CDISC SDTM Implementation Guide (Version 3.1.2), Nov 2008

- Case Report Tabulation Data Definition Specification (CRT-DDS, also called define.xml) Final Version 1.0, Feb 2005

- Robert W. Graebner, SAS Global Forum 2008, Practical Methods for Creating CDISC SDTM Domain Data Sets from Existing Data http://www2.sas.com/proceedings/forum2008/207-2008.pdf

ACKNOWLEDGMENTS

CONTACT INFORMATION

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

Kalyani Chilukuri
Clinovo
1208 E. Arques Avenue, suite 114
Sunnyvale, CA 94085
Phone: +1 800 987 6007
E-mail: Kalyani@clinovo.com

Web: http://www.clinovo.com/

Ale Gicqueau
Clinovo
1208 E. Arques Avenue, suite 114
Sunnyvale, CA 94085
Phone: +1 800 987 6007
E-mail: ale@clinovo.com
Web: http://www.clinovo.com/

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

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