How to easily convert clinical data to CDISC SDTM

Ale Gicqueau, Clinovo, Sunnyvale, CA
Miki Huang, Clinovo, Sunnyvale, CA
Stephen Chan, Clinovo, Sunnyvale, CA

INTRODUCTION

Sponsors are receiving clinical information of increased complexity, from multiple sources and different formats. As a result, clinical data submission has become more time-consuming, costly and error-prone. To tackle this challenge, CDISC® (Clinical Data Interchange Standards Consortium) has been establishing new non-proprietary clinical data standards to speed up data-review and improve clinical data exchange, storage and archival. Conforming to these recognized CDISC standards improves and significantly speeds up FDA submission and FDA review. In addition, converting clinical data to a standardized format will improve SAS code re-usability for many programs used in data management and biostatistics such as Edit Checks, Patient Profile, TLGs, and custom reports.

SAS is often used as an ETL tool to manually convert SAS extracts from a clinical database to SDTM format. While this is a reasonable approach, it can quickly become tedious, error-prone, and time consuming. CDISC Express is a powerful open source SAS®-based clinical data management system that automatically and systematically converts clinical data into CDISC SDTM using an Excel framework. All CDISC Express mapping definitions and rules are defined in Excel, which are dynamically converted into a SAS program that automatically performs the SDTM transformation and validation through a series of SAS macros

CDISC Express source code is freely available, well-documented and easily understandable; it can be easily modified by any SAS programmer to fit his company SAS infrastructure.

This paper will provide SAS programmers with an introduction to CDISC Express, and show how the SAS programs and configuration files are organized. We will also show how to create macros, and convert clinical data to CDISC SDTM domains.

CDISC EXPRESS APPLICATION

How to convert easily Clinical Data to CDISC SDTM domains

We are describing below the seven key steps used to convert clinical data to CDISC SDTM using CDISC Express:

I) Download and install CDISC Express to your computer
II) Create a new study folder (if needed)
III) Create a new mapping file template (if needed)
IV) Modify Mapping Files 'tmpmapping.xls'
V) Validate Mapping File 'tmpmapping.xls'
VI) Generate CDISC SDTM domains
VII) Generate define.xml file

I) Download and install CDISC Express

Prerequisites:

1. Windows XP
2. SAS version 9.1.3 or 9.2
3. Excel 2002 or above
4. Around 60 mb available on the hard drive for the installation
5. Internet Explorer preferred, as our web pages are best viewed in that browser.

Download and Install CDISC Express:

1. Visit http://www.clinovo.com/cdisc/download,

2. An email with a download link will be sent to the mailbox you provided in the short form
3. Follow the download link provided in the email and install CDISC Express on your computer
4. Save 'Clinovo_CDISC_Express.exe' to your hard drive
5. Double click 'Clinovo_CDISC_Express.exe' to start the installation wizard
6. Click 'Run' when prompted to execute 'Clinovo_CDISC_Express.exe'
7. Click 'Next>' from the 'Welcome to the Clinovo CDISC Express v1.0 Setup Wizard'
8. Check the box for 'I accept the terms of the License Agreement' and click 'Next>' to continue.
9. Choose a destination folder, such as 'C:\Program Files\CDISC Express' and click 'Install' to continue
10. Once the installation is complete, click 'Finish' to exist the installation wizard.

By selecting "Launch" from the Welcome menu, you can see how CDISC Express program and configuration files are organized (Figure 1).

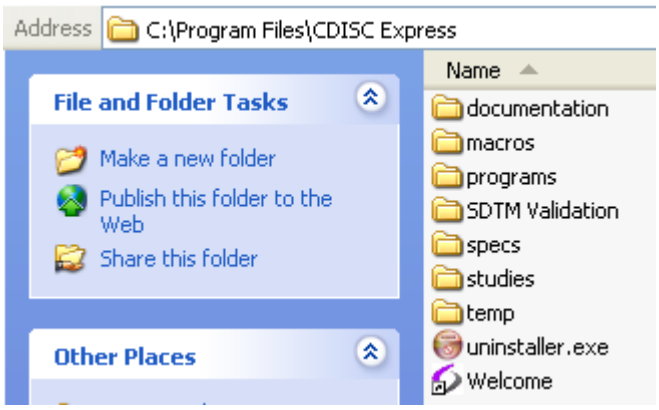

Figure 1. CDISC Express folder structure

- **welcome** – This shortcut displays the Welcome dashboard with useful links.
- **documentation** – This folder contains useful documentation: a Quick Start Guide, a User Guide, and FAQ, a Video Tutorial, an the User Agreement.
- **macros** – This folder contains all the macros.
- **macros\ClinMap** – This folder contains all the macros used by the core of the application.
- **macros\function_library** – This folder contains macros to map your data to SDTM domains.
- **programs** – This folder contains all the SAS programs that you can use with your studies.
- **SDTM Validation** – This folder is used to validate the SDTM domains.
- **specs** – This folder contains all the specification like SDTM terminology and LAB specs.
- **studies** – This folder contains all studies you want to map.
- **temp** – This folder contains a newly generated tmpmapping.xls file after executing 'generate_mapping_template.sas' file.

II) Create a new study folder (if needed):

1. Run 'create_new_study.sas' to create a new study folder with a specified study name. Once the study folder is created, it will create all the folder structure within the new study folder located at \CDISC Express\studies\<New Study Name>

III) Create a new mapping file template (if needed):

1. Once a new study folder is created, users can create a new mapping file with specified domain by running 'generate_mapping_template.sas' to create a new mapping file 'tmpmapping.xls' in the folder \CDISC Express\temp folder with 4 default sheets – Studymetadata, Format, DM, and SUPPQUAL.
2. Run %Createmapping.sas if a domain other than 'DM' is needed. Users will have an option to choose whether they like to have 'Required,' 'Expected,' and \or 'Permissible' CDISC SDTM variables by adjusting the parameters for 'Createmapping' macro.

IV) Modify the sample mapping file 'tmpmapping.xls'

The mapping file (Figure 2) is the heart of the system and contains all the mapping rules for the CDISC® variables. It is saved in the 'DOC' folder of the corresponding study. There are two sub folders:

- **'Mapping file - working version' folder:** This folder contains the working version of the mapping file (tmpmapping.xls). Any changes to the mapping rules should be done in this document.
- **'Mapping file - validated version' folder:** CDISC® Express has a program to validate the mapping rules in 'tmpmapping.xls.' After creating or updating the 'tmpmapping.xls' file in the 'Mapping file - working version' folder, a SAS program will validate the document by checking the syntax. If no issues are detected, the working file will be copied to the folder 'Mapping file - validated version'. It is important not to change this file. Only the working version of the mapping file should be updated by the users.

In this section, the user makes his necessary modifications to the 'tmpmapping.xls' file in the 'Mapping file - working version' folder. The validation of the tmpmapping.xls file will be done after the modification of this mapping file is complete. This mapping file is an Excel file in XML format with the following types of sheets:

- 'StudyMetadata' tab
- 'FORMAT' tab
- 'Domain' tabs (EM, EX, IE…etc)
- 'SUPPQUAL' tab

Figure 2. Mapping file structure

a) 'StudyMetadata' tab

The Studymetadata tab (Figure 3) contains the information to generate the Define.xml file. Information about the XML elements is present in the columns 'XMLField' and 'XMLElement.' You can update the 'Values' column to represent your study details. The column 'Comments' has some additional information to help you with understand each row of the 'StudyMetadata' tab.

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	

Figure 3. StudyMetadata tab of the mapping file

b) 'FORMAT' tab

All SAS formats can be used in the mapping file. You can also define custom formats and specify them in the FORMAT tab (Figure 4).

The FORMAT tab contains 3 columns:

- **format** – Defines the format name. It has to start with a $ sign for a text format and cannot contain blanks. Numeric formats do not need the $ sign.
- **from** – Defines the entry value that you want to apply the format to.
- **tovalue** – Defines the value that will replace the entry value.

For example, the first format is $sev. If you apply this format to a variable, the value '1' will be replaced by 'MILD.'

format	from	tovalue
$sev		
	1	MILD
	2	MODERATE
	3	SEVERE
	4	LIFE-THREATENING OR DISABLING
	5	DEATH-RELATED
$yn		
	YES	Y
	NO	N
	UNK	U

Figure 4. FORMAT tab of the mapping file

c) 'Domain' tabs (DM, TV, SV, AE, CM, MH, EX, VS, DS, LB, SC, IE, TI, CO..etc)

Each SDTM domain that will be mapped has to have its own tab. The name of the tab defines the SDTM domains that is created by the instructions contained in the tab.

A domain tab contains 6 columns (Figure 5). Users need to modify these columns in each domain tab to suit their clinical studies.

- Dataset – Specifies the source datasets that will be operated on, to create the STDM domains as defined by the name of the tab.
- Merge Key – Defines the variables that will be used to merge the datasets that are specified in the Dataset column. If this column is not empty, the application assumes that the variable USUBJID is to be used to merge.
- Join (optional) – Specifies whether an IN option should be employed in merging the datasets with a merge key.
- CDISC variable – Specifies the CDISC variables that will be created.
- Expression – Provides the detail on the assignment statement of the SDTM variable in the CDISC variable column. The expressions are to create the CDISC variables from the source datasets. Users fill this column out with the help of study protocol and the structure of the source datasets. 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.
- Comments – It is for documentation purpose and will appear in the column 'comment' of the define.xml of the study.
- Explanation – It provides additional details and explanation to help you with creating the mapping file for your study. It is not used by the CDISC® Express application.

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 5. Domain tab of the mapping file

Note that if you do not want to process a domain, you can add '-' before the tab name (Figure 6). The domains with a name starting by '-' are excluded from the mapping validation and the SDTM generation programs.

Figure 6. Excluded TV, SC, and AE domains with '-' prefix

d) 'SUPPQUAL' tab

The 'SUPPQUAL' tab defines the non-standard variables to be created that cannot be mapped to already defined SDTM variables. Because the CDISC SDTM does not allow the addition of new variables, it is necessary to represent the metadata and data for each non-standard variable/value combination in the SUPPQUAL dataset. Users need to fully define the metadata of the SUPPQUAL variables which include Domain Name, Variable Name, Variable Label, Type, Length, and Origin. The description of these 6 variables is as below:

- Domain – SDTM domain name.
- VariableName – Variable name which has to be uppercase.
- VariableLabel – Variable label.
- Type – Variable type which can be either Char or Num.
- Len – Variable Length.
- Origin – Variable origin which can be CRF or MACRO.

Note:

1) All data values are stored as characters, so that the type will always be a character, even if a numeric value is specified.
2) The length of the variable must be correctly specified to ensure no values are truncated.
3) The SUPPQUAL datasets are created for each domain, e.g. SUPPDM. These datasets may be transposed and merged back with the domain dataset, e.g. DM.
4) To distinguish SUPPQUAL variables from the Domain variables, the SUPPQUAL variables are prefixed with '~' in the Domain definition.

V) Validate the mapping file 'tmpmapping.xls'

Once the working version of the mapping file 'tmpmapping.xls' is completely filled, the file has to be checked for logical and syntactical errors by running the program, Validate_Mapping_File.sas,' before comforting the data to SDTM. This SAS program will check whether the tmpmapping.xls meets requirements.

A message will be displayed on the HTML page indicating the validation is successful. The temporary mapping file will be renamed as 'mapping.xls' and saved in the folder \CDISC Express\Studies\my

study\doc\Mapping file - validated version folder, and the previous validated mapping file will be archived by adding the current date and time to the file name and stored in the same folder.

If the validation fails, a list of error messages will be displayed in the HTML page 'mapping_validation.html' located in the folder \CDISC Express\Studies\my study\results\Mapping Validation. After reading the error message, user will correct errors in the mapping file and then validate it again until all errors are cleared.

As errors may occur in several domains, to be more efficient and focused, it is possible to comment out unnecessary domains by prefixing the sheet name with a dash in the 'tmpmapping.xls' file as below. However, a domain should not be commented if certain expressions require variables from other domains.

Below is the list of error handling codes that have been built into CDISC Express (\CDISC Express\specs\Mapping validation\validation_err.xls) with five error categories (Figure 7).

- Mapping file – Rules to check the mapping file structure
- FORMAT Tab – Rules to check the data entered on the FORMAT tab.
- CDISC mapping definition – Rules to check the mapping expression for the different domains
- SUPPQUAL domain – Rules for the SUPPQUAL domain
- CO domain – Rules for the CO domain

This spreadsheet is used by the validation program to interpret error codes with variable names, domain names, and/or type of errors. This list of error can be extended by adding new error codes and definitions. Once a new definition is added, the macro 'validatestudy.sas' should also be updated to test the mapping file for the presence of these new errors.

Category	production	Code	Message	Comments
mapping file	ON	101	The mapping file is not in a valid format. Please upload again. Also make sure you saved it under Excel as an XML file.	
	ON	102	The mapping file has &n_bad_ws worksheet(s) that do not have a valid name: &bad_ws. Please correct these worksheet names.	
	ON	103	Please upload the mapping file in the right folder (\doc\Mapping file - working version) with the name tmpmapping.xls	
	ON	104	DM domain is mandatory. Please include the tab DM in the mapping file before validation.	
	ON	105	Welcome tab provides information about Clinovo and CDISC Express. Mapping file should contain the Welcome tab to proceed with mapping file validation.	
FORMAT tab	ON	201	The FORMAT tab has &columnvalue in &missingcolumn . Only the following columns are allowed: &permittedcolumnvalues	
	ON	202	Column ` &col1` value &col1val exists however column ` &col2` value &col2val must &negation exist then	
	ON	203	All lines from FORMAT should be filled but line number &linenumber is empty.	
CDISC mapping definitions	ON	401	The tab `&domain` has `&columnvalue` in &missingcolumn . Only the following columns are allowed: &permittedcolumnvalues	
	ON	402	Variable &cdiscvar from domain &domain needs to be uppercase	
	ON	403	Variable &cdiscvar from domain &domain has not been defined in the global SDTM specifications	
	ON	404	Variable &cdiscvar from domain &domain has not been defined in the SUPPQUAL tab of the mapping file	
	ON	405	The source dataset &srcds defined in the mapping definitions for domain &domain cannot be found	
	ON	407	Source variable &srcvar is not defined in source dataset &srcds in the mapping definition for &cdisc_variable in domain &domain	
	ON	408	Domain &domain, variable &cdiscvar - Macro variable &macrovar has not been defined	
	ON	409	The macro function &macrofunction used for the mapping definition of &domain and variable &cdiscvar does not exist in the macro library	
	ON	410	In the mapping definition of &domain and variable &cdiscvar, the macro function &macrofunction is missing a keyword parameter for value &parameter.	
	ON	411	The temporary variable `&tempvar` from domain `&domain` you defined in the mapping file already exist in the source dataset `&srcds`. Please give this temporary variable a different name.	
	ON	414	Line &line_number of domain `&domain` has a value for the `&filled_col` column but not for the `&empty_col` column. Please correct.	
	ON	415	Merge key `&merge_key` for data source `&srcds` and domain `&domain` appears as a single key however at least two instances of the same key `&merge_key` for two different datasets are necessary.	
	ON	416	The merge key `&merge_key` is invalid as it is not a variable of the source dataset `&srcds` for domain `&domain`.	
	ON	417	&cdiscvar is a required variable for domain &domain.. Please define it in the mapping file.	
	ON	418	Double quotes are not allowed in the definition of the dataset in the source dataset definition `&srcds` for domain `&domain`. Please use single quotes instead.	
	ON	419	Variable &cdiscvar from domain &domain has &lencdiscvar characters (over the limit of 8 characters) and therefore is not a valid CDISC variable name	
	ON	420	Merge key &mergekey for domain &domain should not contain special character &spechar	
	ON	421	Merge key &mergekey for domain &domain should not start with a numeric character	
	ON	422	The dataset name &dataset in domain &domain should not appear more than once	
	ON	423	A parenthesis is missing in the expression for the domain &domain and the CDISC variable &CDISC_variable	
	ON	424	SDTM variables cannot be used in the expression to map USUBJID for domain &domain.	
SUPPQUAL domains	ON	501	The SUPPQUAL domain has &columnvalue in &missingcolumn . Only the following columns are allowed: &permittedcolumnvalues	
	ON	502	SUPPQUAL variable &svar (&slabel) is associated with an unknown domain &invaliddomain. If it is a valid CDISC domain, contact clinical systems support to add it.	
	ON	503	SUPPQUAL variable &svar (&slabel) from domain &domain is invalid. It should be a valid variable name, uppercase and under 8 characters	
	ON	504	SUPPQUAL variable &svar (&slabel) from domain &domain has an invalid type `&svartype`. Only `Char` and `Num` are allowed	
	ON	505	SUPPQUAL variable &svar (&slabel) from domain &domain has an invalid length `&svarlength`. Only integer values between 1 and 32,767 are accepted	
CO Domain	ON	601	CO Domain should contain atleast one comment (#COMM) variable. If there are no comment variables to be defined, then delete the CO tab.	
	ON	602	CODTC variable is not defined for CO variable &srcvar from source dataset &srcds.	
	ON	603	all(stack) must be used to set the datasets if more than one comment is defined in the CO tab.	

Figure 7. Error handling codes table

VI) Generate CDISC SDTM domains

Once the validation of the mapping file is successful, we can create CDISC SDTM domains by running 'generate_SDTM.sas' from \CDISC Express\Programs folder. This program will generate all the SDTM domains based on the specifications defined in the mapping.xls file. The generated SDTM domains will reside in the \CIDSC Express\studies\<Study Name>\results\SDTM folder.

After each run of generating CDISC SDTM domains, the message 'SDTM tables were successfully generated for study <Study Name>' will appear on your browser with hyperlinks to access the SDTM

generated information such as a list of domain created, SDTM terminology issues, and SDTM validation issues (Figure 8).

The SDTM domains can be created using the following mechanism:

1) From a single dataset
   - By putting only one source dataset in the Dataset column, the domain will be created from a single dataset

2) By stacking multiple datasets from a source dataset
   - By using several datasets in the 'Dataset' column and use the term 'all(stack)'
3) By merging multiple datasets using the same merge key
   - By leaving the merge key column blank, datasets will be merged using the default variable
4) By merging multiple datasets using different merge keys
   - By specifying different variables in the merge key column

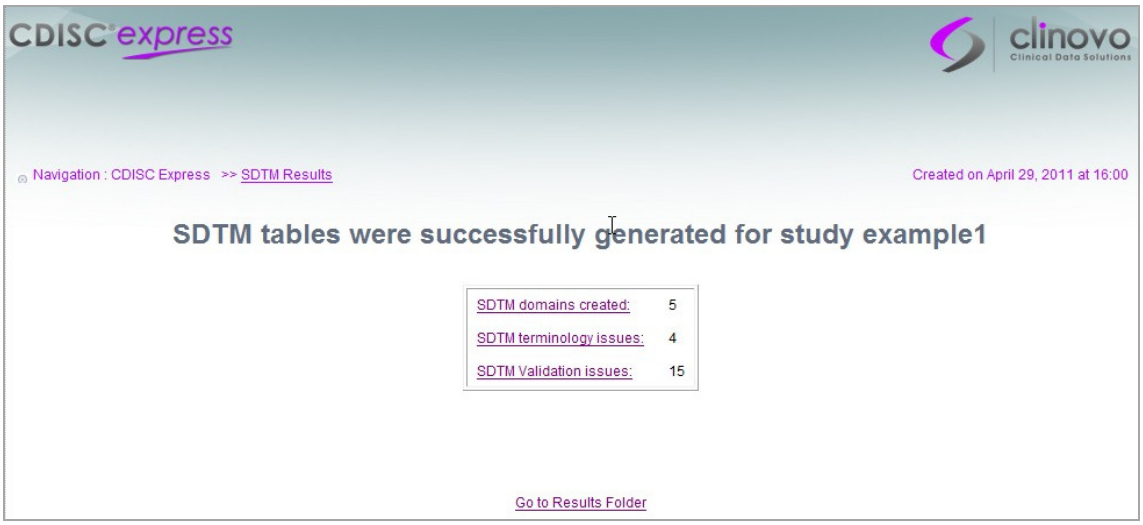

Figure 8. CDISC SDTM generation summary report

VII) Generate Define.xml file

After CDISC SDTM domains are generated, we can create CDISC SDTM domains by running 'generate_Definexml.sas' from \CDISC Express\Programs. This program will create a report 'definexml.html' in \CIDSC Express\studies\<Study Name>\results\reports\definexml folder.

CONCLUSION

CDISC is a matured clinical data standard that helps manage clinical data in a standardized and uniform way. It is strongly recommended by FDA; therefore, complying with this format significantly improves the quality of FDA submission and accelerates the FDA review, resulting in a reduced time to market. Once clinical data is converted to CDISC, SAS code can be re-used for clinical data management and biostatistics activities, as well as for cross study comparisons. CDISC Express is a powerful tool that streamlines complex data mapping and SDTM conversion through the use of an easy-to-understand Excel-based framework. The Excel mapping file can serve as a specification document and source codes, as it is automatically converted to SAS codes by macros during the conversion.

ACKNOWLEDGMENTS

We thank Kalyani, Romain, Leila, Megha, and Gaetan, who were involved in the development and release of CDISC Express application.

CONTACT INFORMATION

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

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/

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

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