Paper 1451-2014

Converting Clinical Database to SDTM: The SAS® Implementation

Hong Chen, McDougall Scientific Ltd.

ABSTRACT

The CDISC Study Data Tabulation Model (SDTM) provides a standardized structure and specification for a broad range of human and animal study data in pharmaceutical research, and is widely adopted in the industry for the submission of the clinical trial data. Because SDTM requires additional variables and datasets that are not normally available in the clinical database, further programming is required to convert the clinical database into the SDTM datasets. This presentation introduces the concept and general requirements of SDTM, and the different approaches in the SDTM data conversion process. The author discusses database design considerations, implementation procedures, and SAS® macros that can be used to maximize the efficiency of the process. The creation of the metadata DEFINE.XML and the final STDM dataset validation are also discussed.

INTRODUCTION

The Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) was designed to provide a standardized structure and specification for a broad range of human and animal study data in pharmaceutical research. In medical and biopharmaceutical product development, the CDISC standards can be used to support the electronic acquisition, exchange, submission and archiving of clinical trials data. The tabulation datasets, which are electronic listings of individual observations that comprise the essential data reported from a clinical trial, are created based on data captured in the clinical database of a study. However, since SDTM requires additional (derived) variables and datasets that are not normally available in the clinical database, further programming is required to convert the clinical database into the SDTM datasets. The following sections will introduce the concept and general requirements of SDTM, illustrate the different approaches in the SDTM data conversion process, and discuss the SAS implementation considerations in order to maximize the efficiency of the process. SAS programming tips and the creation of the metadata DEFINE.XML and the final STDM dataset validation will also be discussed.

SDTM BASICS

SDTM represents the data tabulation datasets that contain data collected in the human and animal studies; these can be thought of as electronic by-subject data listings. The fundamental element of the SDTM is the concept of observations and variables, consisting of information collected during the study.

VARIABLES

Each SDTM variable, corresponding to a column in a dataset, represents a discrete piece of information or measurement. SDTM classifies all variables into the following five roles:

- Identifier variables, which identify the study, the subject (individual human or animal) involved in the study, the domain, and the sequence number of the record.
- Topic variables, which specify the focus of the observation (such as the name of a lab test), and vary according to the type of observation.
- Timing variables, which describe the timing of an observation (such as start date and end date).
- Qualifier variables, which include additional illustrative text, or numeric values that describe the results or additional traits of the observation (such as units or descriptive adjectives).
- Rule variables, which express an algorithm or executable method to define start, end, or looping conditions in the Trial Design model.

OBSERVATIONS

Each SDTM observation, corresponding to a row in a dataset, is presented as a series of variables. Observations are divided into three general classes:

- The Interventions class, which captures investigational treatments, therapeutic treatments, and surgical procedures that are intentionally administered to the subject (usually for therapeutic purposes) either as specified by the study protocol (e.g. exposure), or preceding or coincident with the study assessment period (e.g., concomitant medications).
- The Events class, which captures planned protocol milestones such as randomization and study completion (disposition), and occurrences or incidents independent of planned study evaluations occurring during the trial (e.g. adverse events) or prior to the trial (e.g. medical history).
- The Findings class, which captures the observations resulting from planned evaluations to address specific questions such as observations made during a physical examination, laboratory tests, histopathology, ECG testing, and questions listed on questionnaires.

DOMAINS

A logical collection of observations forms a dataset, or domain in SDTM terminology. SDTM domains are identified by a two-letter name, and are further classified according to the purpose:

- Standard Domains– these are the typical domains created for most clinical studies. Standard domains are pre-defined with fixed (reserved) names, and represent three classes:Interventions (CM, EX, SU), Events (AE, MH, DS, DV), and Findings (LB, VS, EG, IE, SC, PE, QS, DA, PC, PK).

- Additional Domains – these includes domains for Special Purpose (DM, SE, CO, SV), Relationship (SUPPQUAL, RELREC), and Trial Design (TA, TE, TV, TI, TS).

- Sponsor Defined Domains – sponsor defined domains are created (if necessary) to capture additional data/information captured in the study that are not covered in the above domains. Sponsors defined domains are to be created following the structure/specification in the Intervention/Event/Finding domains, and named with two-letters.

In preparing the SDTM data from a clinical database, the standards described above must be followed. Specifically, the following requirements are to be met:

- Domain name is represented by two letters
- Variable names are limited to 8 characters, with the first two letters representing for the domain name (except for DM and the relationship domains)
- Variable labels are limited to 40 characters
- Data values are not formatted; a variable is either Char or Num
- Datasets are saved as SAS V5 transport files
- Dataset Metadata (DEFINE.XML or DEFINE.PDF) is to be created to capture the attributes and description of each domain and variable.

CONVERTING CLINICAL DATABASE TO SDTM

Clinical study databases, either EDC or paper-based, are designed primarily to capture the study data collected during the study. They do not normally include derived variables or the relationship datasets. Given the standards and specific requirements of SDTM, it is not feasible to design a clinical database to output SDTM datasets. Additional programming is required to convert data captured in the clinical study database to SDTM. SAS becomes the logical choice for this task as it is widely used for programming and analysis in clinical research, and most pharmaceutical companies and CROs already have the technical resource (SAS programmers) and business process in place.

APPROACH

Depending on the business process of an organization, converting clinical database to SDTM usually follows one of the following approaches:

Forward Conversion: Database – SDTM - Analysis

In this approach, the database is converted to SDTM, and the SDTM datasets are in turn used for the analysis. The benefit of this approach is that the statistical analysis and tables, listings, graphs are based on the SDTM datasets – the same data used for submission. However, the drawback of this approach is the analysis can only be done after SDTM is created/finalized. Another major concern is programming efficiency – SDTM specifications may not be most efficient for analysis programming as formatted numeric variables are usually stored as character variables (e.g. SEX, RACE). In analysis programming, working with character variables is not efficient; very often this means that they will need to be converted back to numeric.

Backward Conversion: Database - Analysis – SDTM

With the linear backward approach, analysis datasets are created from the clinical database to include all derived variables required for the analysis. The SDTM datasets are then created based on the analysis datasets. The major advantage of this approach is that many derived variables that are not captured in the clinical database but required for the SDTM are available in the analysis datasets. This will save programming time and ensure consistency of the derived variables in the SDTM creation. While programming efficiency can be achieved with this approach, the concern is on the timeline as the process is dependent of the analysis datasets, meaning the SDTM datasets can only be completed after the analysis datasets are finalized. The process may face greater challenge if the SDTM conversion and analysis are not performed by the same programming unit.

Parallel Conversion: Database – Analysis, Database - SDTM

This approach provides the most flexibility from the process point of view, as the SDTM datasets and analysis datasets are created independently from the database. This parallel approach separates the SDTM dataset and analysis dataset programming tasks. Programming of these datasets can be performed by different programming units, or at different time. The benefit is obvious if timeline is of concern (which often is the case in clinical trial) because SDTM conversion and analysis programming can proceed simultaneously. While this approach is great from timeline and resourcing point of view, programming efficiency is decreased as the derived variables will have to be created twice in the process. Because of this shortcoming, extra efforts are required to ensure that the derived variables are created accurately and consistently in the SDTM and analysis datasets.

IMPLEMENTATION

No matter which approach is used, the SDTM conversion process and steps should be thoroughly analyzed to identify the areas where efficient programming techniques can be used to achieve efficiency and accuracy. The following technical considerations should be incorporated into the implementation process.

Database Design

A well designed database will make SDTM conversion much easier. As a general role, the database should be designed to have the data structure and variable attribute and naming convention mapped to the SDTM as much as possible. The CDISC's Clinical Data Acquisition Standards Harmonization (CDASH) compliant CRF, whenever possible, should be used as a basis for the clinical data collection. The benefit of using CDASH provides not only the standardized data structures and specifications over multiple studies, but also the direct mapping of many variables to the SDTM, significantly reducing the programming effort. Standard database structure and specifications across all studies also allow for the development and utilization of SAS macros.

Programming Considerations

While CDASH arms for the collection of the clinical data, the main objective of SDTM is for data reporting and exchange. As such, there are some unique SDTM requirements that are not met in the data collection process; they must be satisfied through additional programming. For example, ISO format are required for the date and time variables in SDTM; additional derived variables needs to be included in the datasets; formats are not normally applied instead formatted values are converted to characters. All these requirements are to be carefully studied so that SAS macros can be developed to simply the conversion process.

Perhaps the most difficult and time consuming task of converting a clinical database to SDTM is the creation of the SUPPQUAL and RELREC domains. SUPPQUAL contains additional variables that are captured but not stored in the originating SDTM domains, and RECREC captured the relationship of the variables in different domains. Intensive programming are usually required for the creation of these domains, although SAS macros may still be used in the process if the study data are collected in the standardized database (e.g. CDASH). However, extra validation effort and process should be in place to ensure these domains are created correctly according to the SDTM requirements.

It should be noted that a clinical database often records data that serves mainly for data monitoring purposes. Examples include "Was any AE reported?", "Was Diary Forms Provided to the Subject?", and so on. These data are not required by SDTM and should be excluded in the SDTM domains. Empty datasets without any data should also be excluded. To clarify the confusion on what variables are required to be included, the recent recommendations from the FDA state that SDTM variables classified as "expected" or "permissible" are generally required to be included in the SDTM domains.

Dataset Metadata

SDTM datasets should be accompanied by the dataset metadata, with description of the content, context, structure, and purpose of a database. There are two levels of the metadata: metadata definitions for domain datasets (Data Definition Metadata), and metadata definition for domain variables (Variable Definition Metadata).Metadata definitions are data about the data, and are presented in DEFINE.PDF or DEFINE.XML (recommended). Since the attributes of the SDTM datasets and variables are readily available once the domains have been created, SAS macros can be developed to take advantage of the attributes stored in the datasets (e.g. through PROC CONTENTS) to automate the creation of DEFINE.PDF or DEFINE.XML.

Validations

Once SDTM domains and metadata have been created, the last step is the validation against the SDTM requirements. SDTM creation process is not completed without proper validations. Validation macros may be developed based on the SDTM standards. However, commercial and open source validation tools are readily available and can be used for checking the SDTM specifications against the standards. Any issues identified during the validation process should be reviewed and resolved.

While the validation tools do a great job in identifying the problems or issues of SDTM standard compliance, it should be emphasized that they do not (and are not designed to) check the accuracy and consistency of the data contents (even though some packages may include data quality checks). Therefore, standard programming quality assurance procedures should be in place to ensure the SDTM data conversion is executed accurately.

CONCLUSION

With the SDTM standards, each dataset, observation, and variable are prepared following a set of pre-specified rules. This standardized structure and specifications make the review, exchange, and utilization of the study data much more effective. SDTM is now widely used in clinical trials and regulatory submission of the study results. Preparation of SDTM should follow the CDISC SDTM implementation guides and the FDA recommendations. Programming approach should be based on the business needs and process of the organization. CDASH standards should be followed for data collection and database design, and SAS macros are to be developed and utilized to achieve programming efficiency.

REFERENCES

- CDISC CDASH Team, 2011. Clinical Data Acquisition Standards Harmonization (CDASH).
- CDISC Submission Data Standards Team, 2013. Study Data Tabulation Model Version 1.4.
- CDISC Submission Data Standards Team, 2013. Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.2.
- Chen, H., 2008. Data Standards, Metadata and Regulatory Submissions. SSC Liaison, Statistical Society of Canada, Vol. 22, No.1.
- FDA (CDER and CBER), 2014. Guidance for Industry Providing Regulatory Submissions in Electronic Format – Standardized Study Data, February 2014.

CONTACT INFORMATION

If you have any comments and questions, please contact the author at:

Hong Chen
McDougall Scientific Ltd.
789 Don Mills Rd., Suite 305
Toronto, ON
Canada M3C 1T5

Email: hchen@mcdougallscientific.com
Web: www.mcdougallscientific.com