Paper 237-2011

Using SAS® to Incorporate the SDTM for a Study Funded by the NIH

Fenghsiu Su MSBA[1], Bradford Jackson MPH[1], Kaming Lo MPH[1], Sumihiro Suzuki PhD[1],
Karan P Singh PhD[1], David Coultas PhD[2], Sejong Bae PhD[1]
[1]UNTHSC School of Public Health, Fort Worth, TX
[2]UTHSCT Department of Medicine, Tyler, TX

ABSTRACT

Since early 2000, the electronic submission for drug approval has become more common. Many pharmaceutical companies that were previously without standardized electronic databases started to adopt the Clinical Data Interchange Standards Consortium (CDISC) standard. The Case Report Form (CRF) data are converted as study data tabulation models (SDTM). The US Food and Drug Administration (FDA) and CDISC have begun to work together to standardize the format to improve the review and approval process of New Drug Applications (NDAs). The purpose of CDISC is not only to improve efficiencies in the pharmaceutical industry but all health-related fields.

The challenges to implement CDISC standards for National Institutes Health (NIH) based research are that people are not aware of what CDISC is and how to incorporate it into their study. This paper demonstrates how CDISC can be easily applied in an NIH-funded study by using SAS and comments on the benefits of using it.

INTRODUCTION

Because of the increased internet usage and environmental awareness, electronic submissions have become more acceptable and more efficient than paper formats. Recently, the FDA has requested all submissions via an electronic method. Many pharmaceutical companies have adopted the CDISC standard; in the past few years, CDISC has partnered with the FDA to standardize electronic submissions. It improves both the approval and review process and standardizes the submission package (Case report forms, raw data, analysis data, etc). By standardizing the submission package, a standard review application, such as customized software, can be built to strengthen the efficiencies by allowing different parties to verify the submission package. This expedites the approval/review process faster. The purpose of the CDISC is not only to improve the efficiencies in the pharmaceutical industry, but also to all health-related fields.

Pharmaceutical companies are moving in the direction of standardizing databases, and this paper recommends the utilization of CDISC in the non-pharmaceutical settings. However the main challenge in implementing the CDISC standards for NIH-based research is that people do not know what CDISC is, and consequently, they do not know how to incorporate it into their research. The purpose of this paper is to provide an introduction to what CDISC is, and to demonstrate, with an example, how it can be implemented in an NIH funded study with little adaptation.

STUDY DATA TABULATION MODEL

SDTM is a guideline for the organization, structure, and format of the tabulation of raw data[1]. Users can implement it as a data dictionary to manage the database. The CDISC SDTM implementation guideline was released in 2004[1] and has been reviewed and updated periodically (Currently version 3.1.2). The changes have been made to update the standards by incorporating different designs of study. The current CDISC SDTM classification scheme consists of 5 categories, and within each category are numerous domains. Some of these categories and a few of their domains are shown in table 1 below.

Category	Description	Domain
Special Purpose	Describe the subject's characteristics in the study	DM (Demography), CO (Comments), SE (Subject Elements), SV (Subject Visits)
Interventions and Observation Class	Any drugs/interventions that subjects are concurrently taking during the study period	CM (Concomitant Medications), EX (Exposure), SE (Substance Use)

Events	Any events happening to the subject during the study	AE (Adverse Events), CE (Clinical Events), DS (Disposition), DV (Protocol Deviations), MH (Medical History)
Findings	Any quantitative or categorical data measurements recorded for the subject during the study	DA (Drug Accountability), EG (ECG Test Results), IE (Inclusion/Exclusion Criteria Not Met), LB (Laboratory Test Results), MB (Microbiology Specimen), MS (Microbiology Susceptibility)
Trial Design	Describe the study characteristics so that the reader will understand the study design	TE (Trial Elements), TA (Trial Arm), TV (Trial Visits), SE (Subject Elements), SV (Subject Visits), TI (Trial Inclusion/Exclusion Criteria), TS (Trial Summary)

Table 1. SDTM Domain Category and Descriptions

The Case Report Form (CRF) data can be converted into a series of domains. For each main domain, two capitalized characters are used. Additionally, supplemental domains, which are titled “SUPP--", where “--“ is the two letter domain name, are also used. For example, for demographic information the domain name is “DM”, and the corresponding supplemental domain name would be “SUPPDM”.

THE BENEFITS OF SDTM UTILIZATION

Institutional Benefits

Naming variable and dataset may vary from study to study and different time points within the study, even in the course of conducting a longitudinal study by the same investigator. If an investigator wants to conduct a study for the Integrated Summary of Safety (ISS) or the Integrated Summary of Efficacy (ISE), it will take a large amount of resources to reorganize the data if it data are in a standardized or consistent format. By adopting standards, it reduces the amount of time and resources needed to complete ISS or ISE studies.

Turnover of staff, students, or investigators during the course of a study may result in different methodologies for approaching the design of databases or implementation of nomenclature. Personal transition can cause challenges in data handling. Maintaining data consistent integrity, standardized methods of data storage reduces unnecessary staff training time, and also reduces time to lean and able to work with the data.

Investigator/End User Benefits

SAS programs can be built around established standards at an institute so that analysis can be consistent. If the data are standardized then the users will not have to acclimate themselves with the dataset and variable names each time they are presented with a new study. Therefore these standardizations will decrease data management time, increase data quality and speed up the review process. The users will have more flexibility to check data accuracy and ensure that the study results meet the data quality standards. This will also increase accuracy in the auditing process.

Once the SAS programs have been built, the end user can reuse the existing program to cut down the programming time. A macro program can also be developed for various studies with similar design. Furthermore, the existing SAS programs will help the new employees in becoming familiar with the study sooner, which would benefit the investigators by reducing the necessary training time.

CASE STUDY

We provide a six-step SDTM conversion guideline in three stages.

Stage	Step
Preparation	A. Become familiar with the source documents and data • Protocol • Case Report Form (CRF)/ Original CRF annotation • Raw data collected • SDTM Guideline
Conversion	B. Prepare the mapping documents will serve as a reference for programmer and ensure dataset/variables consistency across the study C. SDTM-specific CRF annotation D. SAS programming • Libname directory • Program contents E. SDTM data validation
Final	F. Create data definition files

Table 2. Six-Steps SDTM conversion guideline in three stages

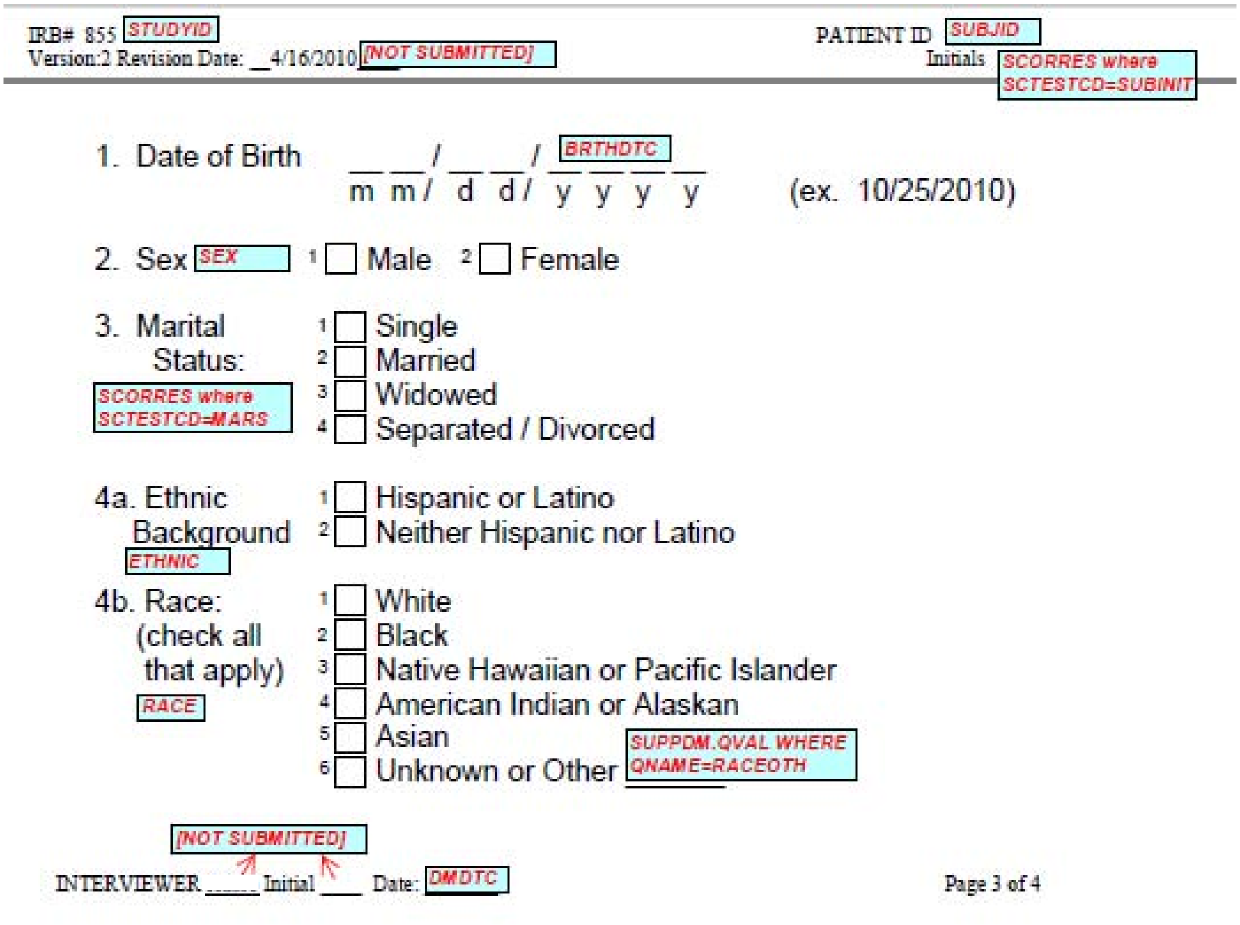
A. <u>Become familiar with the source documents and data</u>

The study protocol describes the study design, objective(s), methodology, the trial duration, number of visits and medication(s). The original case report forms and annotations help the users understand how the data was collected and stored. Based on the original CRF annotations, the user can explore the data.

Before starting a new SDTM project, the user first needs to identify the most updated implementation guide and download the corresponding documents from the CDISC website[2]. The implementation guideline contains information on how to implement the SDTM for study. The next step is to determine which SDTM domains will be created and the extent of SDTM compliance in the existing data[3].

For example, subject's demographic information was collected in the case report form (Figure.1). It is obvious that the DM (demographic) domain needs to be created; however, the structure of the DM is fixed and includes date of birth, age, sex, race, etc. It does not contain the subject's characteristics (e.g., subject's initials, marital status, income level, type of insurance, etc.) Subjects' characteristics are derived in the subject characteristics (SC) domain.

Based on the implementation guideline, race is required in the DM dataset, but 'race other' is not specified. The DM supplemental dataset (DMSUPP) is created to store the 'specify other race' option.



IRB# 855 STUDYID
Version:2 Revision Date: _4/16/2010 [NOT SUBMITTED]

PATIENT ID SUBJID
Initials SCORRES where SCTESTCD=SUBINIT

1. Date of Birth __ __/ __ __/ BRTHDTC __ __ __ __
m m / d d / y y y y (ex. 10/25/2010)

2. Sex SEX 1□ Male 2□ Female

3. Marital Status: SCORRES where SCTESTCD=MARS
1□ Single
2□ Married
3□ Widowed
4□ Separated / Divorced

4a. Ethnic Background ETHNIC
1□ Hispanic or Latino
2□ Neither Hispanic nor Latino

4b. Race: (check all that apply) RACE
1□ White
2□ Black
3□ Native Hawaiian or Pacific Islander
4□ American Indian or Alaskan
5□ Asian
6□ Unknown or Other SUPPDM.QVAL WHERE QNAME=RACEOTH ______

[NOT SUBMITTED]
INTERVIEWER _____ Initial ___ Date: DMDTC

Page 3 of 4

Figure 2. Annotated Case Report Form with CDISC SDTM Domains

B. Prepare the mapping documents

Preparing a mapping document for each variable from original to destination dataset helps to write a SAS program efficiently. The user also needs to ensure all source variables can be mapped to the final dataset. SDTM and all documentations must be submitted to the agency. The SDTM datasets will be submitted as raw datasets; therefore, all source variables must be included.

Steps to prepare the mapping documents:

- Create an excel file and put all of the variable names, labels, and types required--specified in the SDTM guideline,
- Based on the original case report form and dataset; indicate the origin of each variable. (For example, STUDYID is from B_PT dataset),
- Clarify the derivation rules, such as the study rule for age calculation,
- Find the controlled term or codelist provided by CDISC such as SEX.

Variable Name	Variable Label	Type	Length	Format	Origin	Derivation/Comments
STUDYID	Study Identifier	Char	10		B_PT.STUDYID	IRB #855
DOMAIN	Domain Abbreviation	Char	2	DM	Hardcode	"DM"
USUBJID	Unique Subject Identifier	Char	15		Derived	Concatenation of STUDYID||'-'||PTID
SUBJID	Subject Identifier for the Study	Char	6		B_PT.PTID	
RFSTDTC	Subject Reference Start Date/Time	Char	19	ISO 8601	Derived	Merge with Randomization file and get the subject's randomization date. If the subject was not randomized, the subject's first telephone contact is used.
RFENDTC	Subject Reference End Date/Time	Char	19	ISO 8601	Derived	Subject's last contact in the study.
SITEID	Study Site Identifier	Char	2		Hardcode	SITEID='01';
INVNAM	Investigator Name	Char	15		Hardcode	INVNAM='Investigator'
BRTHDTC	Date/Time of Birth	Char	19	ISO 8601	Derived	PUT (B_PT. A02DOB, yymmdd10)
AGE	Age	Num	8		Derived	Complete years between DEMOG.RANDDATE and BRTHDTC AGE=int[(RANDDATE-B_PT.A02DOB+1)/365.25]
AGEU	Age Units	Char	5	Years	Hardcode	"YEARS"
SEX	Sex	Char	1	F, M	B_PT.A02SEX	If B_PT.A02SEX=1 then SEX='M'; Else if B_PT.A02SEX=2 then SEX='F'.
RACE	Race	Char	50	White, Black or African American, Asian, Nativ Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Other	B_PT.A02RACE	
ETHNIC	Ethnicity	Char	30	Hispanic or Latino, Neither Hispanic nor Latino	B_PT.A02ETH	
DMDTC	Date/Time of Collection	Char	19	ISO 8601	B_PT.SUDATE	Character value of Baseline Collection Date DMDTC=put(B_PT.SUDATE, yymmdd10.)
DMDY	Study Day of Collection	Num	8		Derived	DMDY=B_CM.SUDATE - numeric value of RFSTDTC

Table 3.Mapping Document

C. SDTM-specific CRF annotation

Annotation helps the reviewer find the data variable in each dataset. It can be prepared in a PDF format. Tips for annotation:

- When data is recorded on the CRF but not submitted, the variable should be annotated as 'NOT SUBMITTED',
- Only annotate the unique CRF page. Repetitive pages should refer to the first page that is annotated,
- If the variable itself does not have enough information to make annotations meaningful, it is better to include a description. See example below:
  For example, there are multiple characteristics (marital status, initials) for one subject so the user can add 'variable name (value) where category = certain character' ex: 'SCORRES where SCTESTCD=MARS.

SUBJID	SCTESTCD	SCORRES
101-0001	SUBINIT	A-A
101-0001	MARS	Single

D. SAS programming

- Specify the original and final dataset directory.
- Read in all necessary datasets.
- Create the designated variables and derive values.
- Use PROC SQL to arrange the variable order and save in a permanent location.

```
OPTION VALIDVARNAME=UPCASE; *** CREATE UPPERCASE VARIABLE NAME ***;
LIBNAME FINAL 'C:\COPDSMART\DM\DATA\FINAL' ACCESS=READONLY;
PROC FORMAT;
VALUE RACE 1='WHITE' 2='BLACK' 3='NATIVE HAWAIIAN OR PACIFIC ISLANDER'
            4='AMERICAN INDIAN OR ALASKAN' 5='ASIAN' 6='UNKNOWN OR OTHER';
VALUE ETHNIC 1='HISPANIC OR LATINO' 2='NEITHER HISPANIC NOR LATINO';  RUN;

*== READ IN EXTERNAL RANDOMIZATION FILE ==*;
DATA _RAND;
       SET A_RAND.A_RAND;
       LENGTH SUBJID $6 RFSTDTC $19 ARMCD $5 ARM $50;
       SUBJID=TRIM(LEFT(PID));
       RFSTDTC=PUT(RANDOMIZED,YYMMDD10.);
       IF TREAT='CONTROL' THEN DO; ARMCD='PLB'; ARM='PLACEBO'; END;
       IF TREAT='TREATMENT' THEN DO; ARMCD='INT'; ARM='INTERVENTION'; END;
       KEEP SUBJID RFSTDTC ARMCD ARM;
       PROC SORT; BY SUBJID;
RUN;

*== READ IN TELEPHONE CONTACT FILE ==*;
DATA TEL;
       SET TEL.ALL;
       LENGTH SUBJID $6;
            SUBJID=PID;
       KEEP SUBJID COLDAT;
       PROC SORT; BY SUBJID COLDAT;
RUN;

*** GET THE FIRST/LAST TELEPHONE CONTACT ***;
DATA _FIRST(RENAME=(COLDAT=FCONTACT)) _LAST(RENAME=(COLDAT=LCONTACT));
       SET TEL;
       BY SUBJID COLDAT;
       IF FIRST.COLDAT THEN OUTPUT _FIRST;
       IF  LAST.COLDAT THEN OUTPUT _LAST;
RUN;

*== READ IN SUBJECT SUMMARY DATA ==*;
DATA _SBSM;
       SET FINAL.SBSM;
       LENGTH SUBJID $6 SBSMDT 8;
       SUBJID=TRIM(LEFT(PID));
       IF SBSMDATE^=. THEN SBSMDT=SBSMDATE;
       IF SBSMDATE=. THEN SBSMDT=LASTCON;
       KEEP SUBJID SBSMDT;
       PROC SORT; BY SUBJID;
RUN;
*===============*;
* FINAL DATASET *;
*===============*;
DATA _FIN;
       MERGE FINAL.B_PT(IN=A) _RAND _FIRST _LAST _SBSM;
       BY SUBJID;
       IF A;
       LENGTH STUDYID $10 DOMAIN $2 USUBJID $15 RFENDTC $19 SITEID $2 INVNAM $15
       BRTHDTC $19 AGE 8 AGEU $5 SEX $1 RACE $50 ETHNIC $30 COUNTRY $3 DMDTC $19 DMDY
       8;
       STUDYID='IRB # 855';
       DOMAIN='DM';
       USUBJID=TRIM(STUDYID)||'-'||TRIM(SUBJID);
       IF RFSTDTC='' THEN RFSTDTC=PUT(FCONTACT, YYMMDD10.);
       RFENDTC=PUT(SBSMDT, YYMMDD10.);
       IF RFENDTC='' THEN RFENDTC=PUT(LCONTACT, YYMMDD10.);
       SITEID='01';
       INVNAM='INVESTIGATOR';
       BRTHDTC=PUT(A02DOB,YYMMDD10.);
       AGE=INT[(INPUT(RFSTDTC,YYMMDD10.)-A02DOB+1)/365.25];
       AGEU='YEARS';

       IF A02SEX=1 THEN SEX='M';
       ELSE IF A02SEX=2 THEN SEX='F';
       RACE=PUT(A02RACE,RACE.);
       ETHNIC=PUT(A02ETH,ETHNIC.);
       COUNTRY='USA';
       DMDTC=PUT(SUDATE,YYMMDD10.);
       DMDY=SUDATE-INPUT(RFSTDTC,YYMMDD10.);

       LABEL
       STUDYID='Study Identifier'
       DOMAIN='Domain Abbreviation'
       USUBJID='Unique Subject Identifier'
       SUBJID='Subject Identifier for the Study'
       RFSTDTC='Subject Reference Start Date/Time'
       RFENDTC='Subject Reference End Date/Time'
       SITEID='Study Site Identifier'
       INVNAM='Investigator Name'
       BRTHDTC='Date/Time of Birth'
       AGE='Age'
       AGEU='Age Units'
       SEX='Sex'
       RACE='Race'
       ETHNIC='Ethnicity'
       ARMCD='Planned Arm Code'
       ARM='Description of Planned Arm'
       COUNTRY='Country'
       DMDTC='Date/Time of Collection'
       DMDY='Study Day of Collection';
       PROC SORT; BY USUBJID;
RUN;

PROC SQL;
       CREATE TABLE SDTM.DM(LABEL='DEMOGRAPHICS') AS
       SELECT STUDYID, DOMAIN, USUBJID, SUBJID, RFSTDTC, RFENDTC, SITEID,
                      INVNAM, BRTHDTC, AGE, AGEU, SEX, RACE, ETHNIC, ARMCD, ARM,
                      COUNTRY, DMDTC, DMDY
       FROM _FIN;
QUIT;
```

E. Check the dataset created

Currently, the FDA has proposed a validation check for the SDTM datasets[4]. The validation check provides a detailed guideline to validate whether the SDTM dataset has been created correctly or not. The following items should be checked thoroughly:

- All required variables have values,
- Subject's age meets the study design's requirement,
- Any value truncation; especially the free text variables,
- The treatment planned arm (ARM- treatment group) for each subject is consistent with trial design dataset,
- All visit dates are in sequential orders for each subject.

F. Create data definition file

This paper focuses on the NIH funded projects so the final definition document in xml format is not introduced. However, the template for a regular definition file is provided. The definition file is very similar to the mapping document, except for the origin columns.

Variable Name	Variable Label	Type	Length	Controlled Term or Format	Origin	Comments
STUDYID	Study Identifier	text	6		CRF Page 1	
DOMAIN	Domain abbreviation	text	2		Derived	Assigned as 'DM'
AGE	Age	integer	8		Derived	Complete years between subject's randomization date and birth date. AGE=integer[(Randomization Date-Date of Birth+1)/365.25]
Sex	Sex	text	6	M, F	CRF Page 3	

CONCLUSION

This paper has demonstrated how to implement CDISC for data standardization using SDTM. The various benefits of implementing these standards have also been commented on from the perspective of the investigator as well as the end user. We have provided: a sample code, a template, an example of an annotated document, and some logic checks in order to help you start implementing CDISC in your work.

REFERENCES

1. Kenny SJ, Litzsinger MA, Strategies for Implementing SDTM and ADaM Standards, retrieve from www.pharmasug.org/2005/FC03.pdf
2. www.cdisc.org/
3. Graebner RW, Practical Methods for creating CDISC SDTM Domain Data Sets from Existing Data, SAS Global Forum 2008.
4. http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM190628.pdf

ACKNOWLEDGMENTS

Completion of the manuscript was funded, in part, by a grant from the National Heart, Lung, and Blood Institute of NIH, Grant number R18 HL092955.

CONTACT INFORMATION

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

Name: FengHsiu Su
School: University of North Texas Health Science Center
Address: 3500 Camp Bowie Blvd
City, State ZIP: Fort Worth, TX 76107
Work Phone: 817-735-5196
E-mail: fesu@live.unthsc.edu