PharmaSUG 2011 - Paper HW03

Creating SDTM Datasets from Legacy Data

Fred Wood, Octagon Research Solutions, Wayne, PA

ABSTRACT

Many companies have been converting legacy data (data that are in a non-standardized, proprietary format) to the CDISC SDTM (Study Data Tabulation Model) format. Reasons may include the following: 1) anticipating an FDA need, 2) a sponsor need, such as when non-standard data are being prepared for the ISS/ISE, and 3) getting a specific FDA request. This paper will provide some background on the SDTM and SDTM Implementation Guide and discuss some of the issues facing companies who are performing such conversions.

While attendees are expected to be familiar with the SDTM and SDTMIG, this workshop will provide a brief overview of both. Attendees will be given examples of legacy data and be asked to represent it in an SDTM-compliant format. Legacy data will be provided in Excel as well as in SAS® datasets, so knowledge of SAS® programming is not required. A sharing and discussion of the produced SDTM datasets will follow.

INTRODUCTION

The SDTM and SDTMIG are the products of the CDISC Submission Data Standards (SDS) Team, a group of individuals from pharmaceutical companies, vendors, and CROs whose initial work product was known as the CDISC Submission Data Standards (Version 1.0 [October 2000] through v2.0 [December 2001]). Included in these versions were standardized representations of the safety domains listed in the CDER 1999 Guidance (1). In 2002, the Version 2.0 standards were used as a specification for the FDA Patient Profile Viewer Pilot Project. The success of the pilot led to the idea of expanding the standards to include all clinical trials data, not just safety data. At the same time, the SDS Team focused on better use of data-modeling principles. These ideas were crystallized in a concept proposed to the SDS Team by FDA liaisons in October 2002. After going through several different names, it finally became known as the Study Data Tabulation Model (SDTM).

The preliminary version of the SDTM concept was published in June of 2003 as the Submission Data Domain Models, Version 3.0. Because of the interest in balloting the SDTM as an HL7 (Health Level 7) standard in 2004, the model (which was envisioned to be more stable) was separated from the implementation guide (which was thought might change more frequently). Thus, the first version intended for implementation was published as two documents in June of 2004: the SDTM v1.0 (the model), and the SDTMIG v3.1 (the implementation guide). In July of 2004, the SDTM became a Study Data Specification referenced in the eCTD Guidance (2, 3). Version 1.1 of the SDTM was published in April of 2005, followed by Version 3.1.1 of the SDTMIG four months later. Version 1.2 of the SDTM Version 3.1.2 of the SDTMIG were both published in November of 2008.

The SDTM has been the basis for the Standard for the Exchange of Nonclinical Data (SEND), which now has its own implementation guide, the SENDIG, for the submission of pre-clinical toxicology and pharmacology data. SEND is also being tested in a pilot with the FDA Center for Veterinary Medicine. The history and basics of the SEND Model have been reviewed elsewhere (4).

SDTM and SDTMIG BASIC CONCEPTS

This paper provides a high-level overview of the SDTM and SDTMIG concepts. A more detailed summary of the basics of the SDTM and the SDTM Implementation Guide (SDTMIG) has been published previously (5, 6). The SDTM is built around several key concepts:

Domains: Domains are groups of related observations, which are grouped by topic in datasets. Datasets and domains are usually the same, but some domains may contain more than one dataset. Examples include split domains (for which the SDTMIG provides guidelines) as well as domains that need to utilize Supplemental Qualifiers (a topic which will be covered later). Each SDTM dataset is assigned a two-letter domain code that serves as a prefix to most of its variables and as the name of the SAS® transport file.

Observations: Observations are rows or records within a dataset, the information for which is contained in a series of variables (columns). Shown below are some example variables for a heart-rate observation in the Vital Signs (VS) domain.

STUDYID	USUBJID	VSTESTCD	VSORRES	VSORRESU	VSDY
ABC001	1234-0001	HR	100	BEATS/MIN	6

In this study ABC001, Subject 1234-0001 had a heart rate of 100 bpm on Study Day 6.

Variable Metadata: The above example shows only variable names (which are limited to eight characters in the current FDA submission standard of SAS Version 5 transport files). Variables and the data therein can be further described by a number of attributes, or metadata, which are included in the define.xml file. Included are the following:

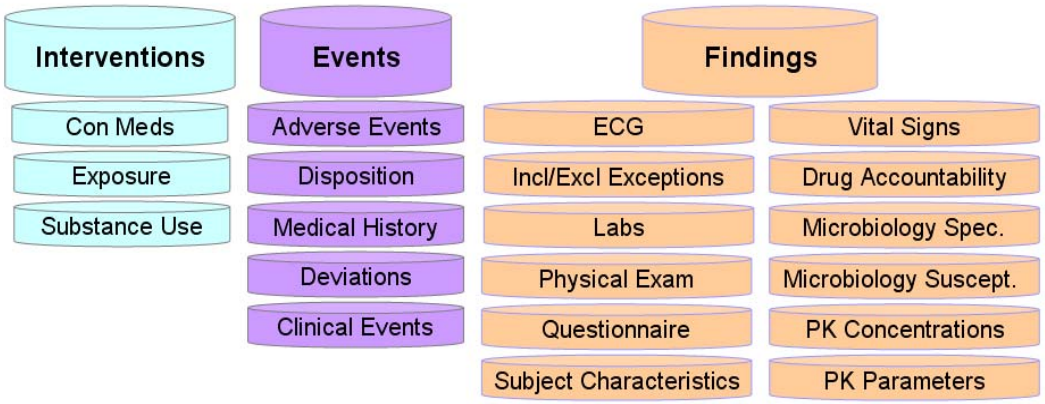
- The Variable Label, which is limited to 40 characters or less, unique for each variable in the dataset
- The Type, which describes whether the data value is in a character or a numeric format
- Controlled Terms, Codelist, or Format for the data values
- The Origin (or source of each variable), which is limited in the SDTMIG to five types: CRF (collected data), eDT (electronic data transfer), Derived, Assigned (data from dictionary coding), and Protocol (for data such as route of administration which might not have been collected on any CRF, but was obtained only from the protocol).
- The Role, which describes how the variable is used in the dataset. SDTM general-observation-class roles include the following: Identifier, Topic, Timing, and four different types of Qualifiers: Grouping, Record, Synonym, Variable, or Result. These are described in more detail in a separate section.

Observation Classes: Most observations can be classified as one of three major types: Interventions, Events, or Findings, described as follows:

- **Interventions:** investigational treatments, therapeutic treatments, and surgical procedures administered to or by the subject. The structure is one record per constant dosing or treatment interval.
- **Events:** planned protocol milestones, study completion (disposition events), or 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 structure is one record per event.
- **Findings:** observations resulting from planned evaluations (e.g. lab tests, ECGs, microscopic findings). The structure is one record per finding result or measurement.

Figure 1 shows how the domains in the SDTMIG v3.1.2 are organized into the general observation classes. The SDTM and the SDTMIG describe other domains that are classified as Special-Purpose Domains, including Demographics, Comments, Subject Elements, Subject Visits, Trial Design Model (TDM) tables, and Relationship tables. The latter are discussed in more detail below.

Figure 1. Fitting V3.1.2 Domains into Observation Classes



Variable Roles: Every variable has been assigned a Role that describes the type of information conveyed by each variable within an observation. Roles play an important part in how variables are used in the understanding and creation of SDTM-based domains. SDTM Roles include the following:

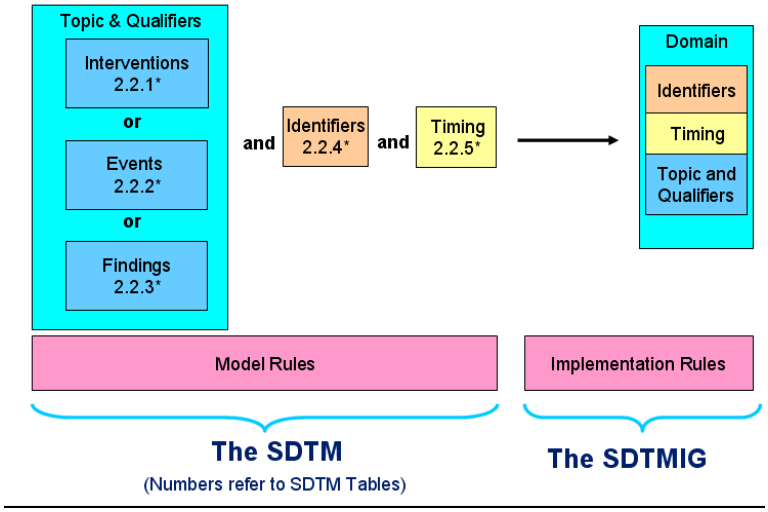
- **Topic Variable** - There is only one per dataset and it identifies the focus of the observation. Examples include AETERM (Adverse Event Term), CMTRT (Concomitant Medications Treatment), and LBTESTCD (Laboratory Test Short Name).
- **Identifier Variables** - The SDTM describes seven of these. Examples include those that identify studies (STUDYID), subjects (USUBJID), and domains (DOMAIN).
- **Timing Variables** - The SDTM describes more than twenty of these that describe date/times of observations; visits and time points, relative times, and study periods (SDTM Epochs).
- **Qualifier Variables** - These describe the attributes and results of an observation. These are the most numerous, and have been subdivided into Grouping, Result, Synonym, Record, and Variable Qualifiers.

CREATING DOMAINS BASED ON THE SDTM AND SDTMIG PRINCIPLES

General Observation Classes

Each observation class has its own defined Topic and Qualifier variables, described in Tables 2.2.1-2.2.3 of the SDTM. These, along with Identifiers (Table 2.2.4) and Timing Variables (Table 2.2.5), both of which can be used in all observation classes, are the building blocks for constructing SDTM domains. Figure 2 shows how domains are created using variables with various Roles. Specific business rules for each of the modeled domains are described in the CDISC Notes column for each variable, and in the Assumptions section for each domain. This same process is used for creating custom domains (i.e., those not modeled in the SDTMIG).

Figure 2. Modeling SDTM Domains



Supplemental Qualifiers

The Supplemental Qualifiers concept was created to address the need to submit non-standard variables (those not found in the tables in Figure 2). It consists of a normalized data structure designed to allow for efficient storage of what might be a wide variety of sponsor-specific variables. Supplemental Qualifiers was initially intended to be a single dataset for an entire study, and was given the name SUPPQUAL (SDTM v1.1 and SDTMIG v.3.1.1). This name has seemed to stick, despite the fact that the current recommendation is to submit a SUPP-- dataset (hyphens designating the two-letter domain code) for each domain that contains non-standard variables.

In the current versions of the SDTM and SDTMIG, Supplemental Qualifiers is one of two relationship tables (the other is RELREC, which will be discussed later). This is because they contain variables that point to (or relate to) one or more general-observation-class records (also referred to as "parent records" in the context of Supplemental Qualifiers). These variables are STUDYID (Study ID), USUBJID (Unique Subject ID), DOMAIN (SDTM Domain), IDVAR (Identifying Variable), and IDVARVAL (Identifying Variable Value). An example of the use of Supplemental Qualifiers in the form of an Adverse Event record and a series of SUPPAE records is shown below. Note that the first five columns in the SUPPAE dataset point back to the parent record.

The remaining five columns of Supplemental Qualifiers contain data and metadata. QNAM (Qualifier Name) is often the name of the variable in the sponsor's original dataset, QLABEL (Qualifier Label) can be used as the SAS label for the variable in QNAM, QVAL is the actual data value for each instance or record, and QORIG describes the origin of the value (a piece of variable-level metadata). In the case of subjective data, QEVAL (Evaluator) specifies the role of the individual who assigned the value, such as an Adjudication Committee or the sponsor. QEVAL finds its origins in the Findings general observation class variable, --EVAL, used when subjective evaluations are made by someone other than the investigator.

ae.xpt (partial)

STUDYID	USUBJID	DOMAIN	AESEQ	AESPID	AETERM	AESTDTC	AEENDTC
1999001	ABC-0001	AE	1	1	Nausea	2004-01-05	2004-01-12

suppae.xpt

STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
1999001	AE	ABC-0001	AESEQ	1	AELLT	Lowest Level Term	VOMITING	Assigned	
1999001	AE	ABC-0001	AESEQ	1	AEHLT	High Level Term	NAUSEA AND VOMITING SYMPTOMS	Assigned	
1999001	AE	ABC-0001	AESEQ	1	AEHLGT	High Level Group Term	GASTROINTESTINAL SIGNS AND SYMPTOMS	Assigned	
Relationship					**Data and Metadata**				

Related Records (RELREC)

The RELREC dataset is used to represent two different types of relationships: collected record-to-record relationships for a subject, and dataset-to-dataset relationships that exist for all subjects. In neither case should information in RELREC represent relationships established after the fact (e.g., as part of the analysis process).

Record-to-Record Relationships for a Subject. These are relationships that have been explicitly collected on the CRF, and typically managed in a clinical database through the use of foreign keys. An example is the CDASH variable, CMAENO (on the Concomitant Medications page), which contains an identifier for an adverse event that was the reason for a concomitant medication.

These types of relationships are documented by creating RELREC records that reference each of the related general-observation-class records, and then by linking them together by giving them the same relationship identifier. The reference to each of the observations is created by using the same keys used in SUPPQUAL (STUDYID, USUBJID, RDOMAIN, IDVAR, and IDVARVAL) to identify a record or group of records. RELREC uses an additional, unique variable, RELID (Relationship Identifier), which is the same for all related records. The value of RELID can be any value chosen by the sponsor. An example is shown in the table below, where the adverse event in the Supplemental Qualifiers section above was recorded on a concomitant medications page.

STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID
1999001	AE	ABC-0001	AESPID	1		AEHOSP1
1999001	CM	ABC-0001	CMSPID	3		AEHOSP1

Note that AESPID and CMSPID are used as the Identifying Variables (IDVAR), because this is the information collected on the CRF pages. The Sequence Number is usually created as part of a programming step on all datasets after the data are complete, and may have no documented relationship to any specific data records.

Dataset-to-Dataset Relationships for All Subjects. These are relationships that exist at the dataset level. An example is shown below for Microbiology Specimen (MB) and Microbiology Susceptibility (MS). In this case, the sponsor is taking steps to ensure that any organism identified in MB is assigned an MBSPID value, and that this value is carried into the MS dataset for that organism.

mb.xpt (partial)

STUDYID	DOMAIN	USUBJID	MBSPID	MBTESTCD	MBTEST	MBORRES
2007-001	MB	ABC-0001	ORG-001	ORGANISM	Organism	CLOSTRIDIUM DIFFICILE
2007-001	MB	ABC-0001	ORG-002	ORGANISM	Organism	HAEMOPHILUS INFLUENZAE

ms.xpt (partial)

STUDYID	DOMAIN	USUBJID	MSSPID	MBTESTCD	MSTEST	MSORRES	MSORRESU
2007-001	MS	ABC-0001	ORG-001	TETRACYC	Tetracycline	0.5	ug/mL
2007-001	MS	ABC-0001	ORG-001	ERYTHRO	Erythromycin	0.2	ug/mL
2007-001	MS	ABC-0001	ORG-001	MUPIRO	Mupirocin	1.0	ug/mL
2007-001	MS	ABC-0001	ORG-002	TETRACYC	Tetracycline	0.75	ug/mL
2007-001	MS	ABC-0001	ORG-002	ERYTHRO	Erythromycin	0.05	ug/mL
2007-001	MS	ABC-0001	ORG-002	MUPIRO	Mupirocin	2.0	ug/mL

In the RELREC example below, USUBJID is null because the relationship applies to all subjects. RDOMAIN describes the datasets that are related. IDVAR describes what is essentially a merge key in this case. IDVARVAL is not populated because this relationship applies to all values of IDVARVAL. In this example RELTYPE indicates that for each MB record with a unique MBSPID there are many records in MS with the same value in MSSPID. The same value in RELID establishes this as a relationship.

relrec.xpt

STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID
1999001	MB		MBSPID		ONE	MBMS
1999001	MS		MSSPID		MANY	MBMS

CHALLENGES TO IMPLEMENTING SDTM AND SDTMIG PRINCIPLES

There are a number of challenges many sponsors will face in implementing the SDTM and SDTMIG principles into their systems and/or submission processes. Some of the major ones will be described in the following sections.

1. All variable names and labels must follow the SDTM standards.

For the general observation classes, this means that only variables from Tables 2.2.1-2.2.5 in the SDTM can be used. All Topic and Qualifier variables require a two-letter domain prefix. Some Identifier and most Timing variables also require a prefix.

Variable labels for modeled domains (those in Figure 1) should not be changed. Variable labels for custom domains may be changed to better describe the data; the underlying meaning should not be changed.

2. **Non-standard variables (those not found in Tables 2.2.1-2.2.5 of the SDTM) cannot be submitted in SDTM parent domains.**

   They must be submitted as Supplemental Qualifiers, as described above. This means that what was one physical dataset in the sponsor's environment needs to be submitted as two. Many commercial viewing tools are capable of merging the SUPP-- datasets back onto the parent (main domain) records, so that QNAM values appear as columns containing data values from QVAL. Individuals with some knowledge of SAS programming can easily create the merged dataset as well.

3. **The SDTM/SDTMIG contain date/time variables that require the use of ISO 8601 date/times.**

   This format is YYYY-MM-DDThh:mm:ss, for years, months and days separated by hyphens, a "T" to indicate that the time component follows, with the hours, minutes, and seconds separated by colons. Partial information is represented via right truncation, omitting information that was not collected. This is different from how dates and times might have been collected (separately) or analyzed (SAS date/times).

4. **Use of CDISC Controlled Terminology**

   CDISC Controlled Terminology can be difficult to implement for several reasons:

   - The first, and most obvious, is that CDISC Controlled Terminology lists might contain CDISC Submission Values that are different from what a sponsor used in their study. At minimum, a one-to-one mapping schema is needed.
   - Sponsors may have more terms than permitted in a non-extensible list. For example, the CDISC list for Severity/Intensity Scale for Adverse Events has only three values: MILD, MODERATE, and SEVERE. Sponsors collecting AE severity with more granularity will have to determine the appropriate mapping strategy.
   - When sponsors add values to an extensible list, they should follow naming conventions that are already in place for existing CDISC Submission Values, rather than using their own values.

5. **The Findings general-observation-class datasets are often more normalized than those that might have traditionally been extracted from the clinical database or used for analysis.**

   The classic example is vital signs data such as blood pressures, heart rate, and temperature being collected in separate columns. In an SDTM domain, these become records defined by the variable, VSTESTCD. The tables below show an example.

Sponsor Dataset

PATNO	VITDATE	SYSBP_MM	DIABP_MM	PULS_BPM	TEMP_C
ABC-0001	2003-02-01	120	80	65	37

SDTM-Based Dataset (Partial)

USUBJID	VSTESTCD	VSORRES	VSORRESU	VSDTC
ABC-0001	SYSBP	120	mmHg	2003-02-01
ABC-0001	DIABP	80	mmHg	2003-02-01
ABC-0001	PULSE	65	BEATS/MIN	2003-02-01
ABC-0001	TEMP	37	C	2003-02-01

6. **The Findings general-observation-class datasets contain additional variables for standard units.**

   The SDTMIG doesn't define what the standard units are for any given test. This is because the SDS Team could never predict what would be most useful to reviewers. These variables are expected in all Findings domains, and are shown in the last three columns in the table below.

USUBJID	VSTESTCD	VSORRES	VSORRESU	VSSTRESC	VSSTRESN	VSSTRESU	VSDTC
ABC-0001	SYSBP	120	mmHg	120	120	mmHg	2003-02-01
ABC-0001	DIABP	80	mmHg	80	80	mmHg	2003-02-01
ABC-0001	PULSE	65	BEATS/MIN	65	65	BEATS/MIN	2003-02-01
ABC-0001	TEMP	37	C	37	37	C	2003-02-01

7. **Adding the --SEQ Variable**

   The Sequence Number is usually created as part of a programming step on all datasets after the data are complete. In most implementations, it does not represent a collected data value. It was created in the SDTM/SDTMIG to serve as a surrogate key for a set of variables that comprise the natural key. It must be unique within study, subject, and domain. Its use, along with STUDYID, DOMAIN, and USUBJID, allows for the precise identification of a single record within a submission.

8. Relationships are represented using RELREC rather than foreign keys

This topic has been amply discussed above, but is mentioned here because it is a challenge that many sponsors face, and they frequently manage it poorly.

CONCLUSION

The SDTM provides a standard model and the SDTMIG provides a set of implementation rules for the creation of tabulation datasets. Unless sponsors have implemented CDASH standards in their data collection systems and SDTM-based domains in the data-management systems, there will always be a need for converting non-standard data into an SDTM-compliant format. The degree of difficulty in performing such conversions varies between sponsors. This paper describes some of the more common challenges sponsors may face in this process.

REFERENCES

1. Guidance for Industry: Providing Regulatory Submissions in Electronic Format — NDAs. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). IT 3, January 1999
2. Guidance for Industry: Providing Regulatory Submissions in Electronic Format -- Human Pharmaceutical Product Applications and Related Submissions using the eCTD Specifications. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Revision 1, April 2006.
3. Study Data Specifications. Current version: 1.5.1 January 1, 2010; Available via http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM199599.pdf
4. Wood, F., and Kramer, L. A. (2011) The Standard for the Exchange of Nonclinical Data (SEND): History and Basics. PharmaSUG 2011.
5. Wood, F., and Guinter, T. (2008) Evolution and Implementation of the CDISC Study Data Tabulation Model (SDTM). Pharmaceutical Programming 1 (1): 20-27.
6. Wood, F. (2008) The CDISC Study Data Tabulation Model (SDTM): History, Perspective, and Basics. PharmaSUG Proceedings, June 2008.

ACKNOWLEDGMENTS

The author would like to thank the following:

- Past and current members of the SDS Team, without whose commitment and dedication the SDTM and SDTMIG would not be possible.
- The leadership and members of the Clinical Data Management Department at Procter & Gamble Pharmaceuticals from 1998-2006, whose vision for data standards supported my involvement with various CDISC teams.
- The Senior Management Team at Octagon Research Solutions, who have long been advocates for CDISC standards development.

CONTACT INFORMATION

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

Fred Wood
Vice President, Data Standards Consulting
Octagon Research Solutions, Inc.
585 East Swedesford Road, Suite 200
Wayne, PA 19087
Work Phone: 484-881-2297
E-mail: fwood@octagonresearch.com