

## Checking for SDTM Compliance: The Need for Human Involvement

Fred Wood and Adrienne Boyance  
Data Standards Consulting  
Octagon Research Solutions, Wayne, PA

### ABSTRACT

An increasing number of sponsors are submitting clinical trials data to the FDA in the format of the CDISC (Clinical Data Interchange Standards Consortium) SDTM (Study Data Tabulation Model). In many cases, however, the data were not collected, stored, or extracted from the database in the SDTM format. As a result, the data must be converted, and must meet a number of structure and content requirements described in the SDTM and the SDTM Implementation Guide (SDTMIG).

A number of companies/organizations have developed automated tools to check for what's commonly referred to as "SDTM compliance." Most of these tools check data as well as the define.xml file, and they play an important role in many organizations' validation processes. Because of the diversity of clinical trials data, however, even the best tools cannot identify all the areas where problems in SDTM compliance and accurate data representation may exist. This paper will provide some examples that illustrate the importance of human involvement in addition to the automated tools for assessing SDTM compliance.

### INTRODUCTION

#### SDTM BACKGROUND

Since the CDISC (Clinical Data Interchange Standards Consortium) Study Data Tabulation Model (SDTM) became a Study Data Specification in the FDA's eCTD Guidance (1) in 2004, pharmaceutical and biotechnology companies have increased their efforts to submit data to the Agency in this format. Interest in "submitting in SDTM" has been further increased by other FDA actions, which include the following:

- The withdrawal of the three Electronic Submission Guidances for eNDA (electronic New Drug Application), eANDA (electronic Abbreviated New Drug Application), and eAnnual Reports, announced on September 29, 2006 (2). This notice designates the eCTD as the "preferred format for electronic submissions," and notes that beginning January 1, 2008 any electronic submission going to CDER (Center for Drug Evaluation and Research) must be eCTD.
- Announcements of a Notice of Proposed Rulemaking (NPRM) regarding a requirement for the SDTM, the first of which appeared in December 2006 (3). The initial target date was March 2007, but this has been revised numerous times (4-7), most recently to June 2010 (8).
- The mention of the SDTM in the PDUFA (Prescription Drug User Fee Act) IV IT Plan (9) as "the foundation for [the] standardized clinical content."
- Recent statements in the CDISC blog (10) by Theresa Mullin and ShaAvhree Buckman, co-chairs of the FDA/CDER Computational Science Center.

For a more detailed development and regulatory history of the SDTM, see Wood and Guinter (11).

#### THE NEED FOR DATA-CONVERSION

The submission of data in the format specified by the SDTM (12), and described more fully in the SDTM Implementation Guide (SDTMIG) (13), will require at least some data conversion in many situations, including the following:

- Data was collected, stored, and extracted in a non-SDTM-compliant format.
- One or more studies in a submission began before the SDTM was accepted by the FDA, and the sponsor wishes to store or submit data from all studies in the same format.
- The submission of electronic data was not anticipated at the time of database design.
- A company merger or acquisition that drives the need to have a single standard for overall operating efficiencies.

## **DATA-CONVERSION CHALLENGES**

Even if a company has well-established and strictly followed processes for standards governance, there will be at least a few challenges when converting to SDTM-compliant datasets. Some of these include the following:

- Converting to standard variable names
- Ensuring that data is in the appropriate SDTM datasets (e.g., Weight in Vital Signs) irrespective of how data was collected
- Converting to industry-wide controlled terminology
- Representing non-standard variables in separate Supplemental Qualifiers datasets
- Adding SDTM-required Sequence Numbers (--SEQ variables);
- Representing foreign keys in separate relationship (RELREC) dataset
- Creating the Trial Design datasets and associated subject-level datasets

## **CHECKING FOR SDTM COMPLIANCE**

Given these challenges, several companies/organizations have developed automated tools to check for what's commonly referred to as "SDTM compliance." These checks are designed to identify some of the most common errors and inconsistencies with the SDTM and SDTMIG. It should be realized that "passing" these checks does not guarantee that datasets will be compliant with the SDTM/SDTMIG, or that they will be viewed correctly with the review tools used at the FDA. The following sections provide some examples of what were intended to be SDTM-compliant datasets whose errors could only have been identified by humans with some level of SDTM and clinical-data experience.

## **REPRESENTATION OF NON-STANDARD VARIABLES (SUPPLEMENTAL QUALIFIERS)**

The Supplemental Qualifiers (SUPPQUAL) special-purpose dataset is used to submit values for variables not presently included in the SDTM. The SUPPQUAL dataset was created to enable efficient submission and data warehousing of what could be an unlimited number of variables existing across sponsors' clinical databases. Splitting what might be a single dataset within a clinical data management system in the non-SDTM world into two datasets for an SDTM submission can be challenging. It may not be as simple as converting all non-standard variables to SUPPQUAL QNAMs (variable names).

### **NOT USING STANDARD SUPPQUAL QNAMs**

Appendix C5 of the SDTMIG lists a number of standard values for Supplemental Qualifiers. The misuse of a sponsor's legacy variable names as QNAM values (e.g., using CLINSIG instead of EGCLSIG for a clinically significant ECG) would not be identified by many automated checks.

### **ADDING VARIABLES THAT REALLY BELONG IN A SEPARATE DATASET**

When one plans to put data with dates into Supplemental Qualifiers, this may be a signal for the creation of a new dataset. An example of such a case would be information collected about injection-site reactions for a vaccine. For reactions graded at least "MILD" an Adverse Event (AE) record was created. Measurements of the diameter of swelling and the diameter of erythema were collected until the reaction subsided.

Because Supplemental Qualifier records inherit the timing of the "parent" record, and there may be more than one measurement during the course of the AE, these data may be better represented in a separate dataset, rather than SUPPAE, where the dates and the corresponding measurements cannot be unambiguously associated.

### **ADDING DUPLICATE DATA**

Sponsors are often tempted to submit data in multiple representations "just in case the FDA might want to see them." Doing so may result in confusion during the review process, especially if the "duplicate" data are conflicting. Below are some examples we have seen.

**Numeric and Character Codes.** The SDTMIG states that controlled terminology should be used instead of arbitrary number codes. Because numeric codes may have been used to facilitate data entry, and may be available as extended attributes in some CDMS extracts, some sponsors feel this is justification to include them in SDTM datasets.

**CDISC Controlled Terminology.** When converting data to CDISC controlled terminology, cases have been seen where sponsors want to submit the originally collected (non-standard) data values in the SDTM datasets. Here is an example:

	Value in AE Dataset	Value in SUPPAE
AEACN	DOSE NOT CHANGED	NONE
AECONTRT	Y	YES
AEOUT	RECOVERED/RESOLVED	RESOLVED

**Dates and Times.** Sponsors who may be tempted to create specifications for an SDTM submission that include date/time information not only in the SDTM standard ISO 8601 format, but also as SAS® dates and the individual date and time components.

### NOT CREATING UNIQUE QNAMS FOR DATA RELATED TO THE SAME PARENT RECORD

Because viewing tools append the SUPPQUAL QNAM values as additional columns in displaying the data, the same “parent record” cannot have columns with the same name. The proper representation is reflected in the SDTMIG for examples of multiple values races and multiple locations. For example, if there are multiple races in SUPPDM, then the variable names must be unique, such as RACE1 and RACE2 rather than RACE and RACE.

### SUBMITTING ANALYSIS DATA

Sponsors who are accustomed to following the 1999 Guidance may only have data stored electronically as analysis datasets. When these datasets are provided as the source datasets to be converted to SDTM, extra time and effort are required to determine which data are truly tabulation data and which data were derived and/or imputed for analysis. Some sponsors might feel that the FDA reviewers would want the important analyses included in the SDTM datasets, and insist they be included. However, in all the interactions that the SDS (Submission Data Standards) Team has had with the FDA since 1999, this has never been requested. Therefore, the appropriate vehicle for the submission of analysis data is the analysis datasets, which would ideally be in the CDISC Analysis Dataset Model (ADaM) format (14).

### DECIDING BETWEEN SUPPLEMENTAL QUALIFIERS, FINDINGS, AND FINDINGS ABOUT

Information about an Event or Intervention that does not fit into standard variables can be represented by three mechanisms: Supplemental Qualifiers, Findings, and Findings About. The distinctions between these can be subtle, as shown in the table below, and choosing the appropriate one can be challenging.

Characteristic	SUPPQUAL	Findings About	Findings
CDISC Controlled Terminology	Limited number of QNAM values	None for --TESTCD and --TEST values other than that for those based upon SDTM Qualifiers (e.g., SEV, DUR, OCCUR)	--TESTCD and --TEST values for modeled domains
Timing	Same as the parent record	--DTC for --TESTCD/--TEST	--DTC for --TESTCD/--TEST
Uniqueness defined by other keys	No	Yes	Yes
Uses and Requires --OBJ variable	No	Yes	No
Ability to unambiguously relate multiple qualifiers (e.g., results and units) to each other	No	Yes	Yes
Relates to an Event or Intervention record	Always	Likely	Possible

### RELATING RECORDS (RELREC)

The Related Records (RELREC) dataset is used to describe collected relationships between records in two (or more) datasets. An example would be capturing the relationship between a concomitant

medication and an adverse event. This is frequently collected via the entry of an adverse-event line number on a concomitant medications page or a concomitant medication line number on an adverse-event page. A variable may exist in the clinical data management system. For submission, however, the SDTM and SDTMIG specify the use of the RELREC dataset for representing these record-to-record relationships.

RELREC is also used for dataset-to-dataset relationships. One area where communicating these relationships is critical is pharmacokinetics (PK), where multiple time-concentration curves (PK Concentrations; PC) and multiple sets of PK parameters (PK Parameters; PP) may exist for each subject (e.g., two dosing occasions and two metabolites). We have encountered cases where the only method of relating the PC and PP data consisted of reading the clinical study report (CSR). If these data had been converted to the SDTM format without using RELREC, it would be impossible to determine which parameters were associated with which curve.

Automated tools are not capable of recognizing when a record-to-record or dataset-to-dataset relationship is missing from RELREC.

## **TRIAL DESIGN AND RELATED TABLES**

An accurate representation of SDTM Trial Design tables, as well as the Special Purpose domains of Subject Elements and Subject Visits requires an understanding of the protocol, the CRF, and the SDTM/SDTMIG. Because of these requirements, it would be virtually impossible for automated checks to assess SDTM compliance. Certainly, tools could check to see that all of the necessary tables and SDTM Required variables are present, but our experience has been that accurately representing the design of a trial is a bigger problem. The complexities of Trial Design have been discussed by one of us previously (15).

## **OTHER PROBLEM AREAS**

### **POOR DATA QUALITY**

The quality of an SDTM conversion cannot be any better than the quality of the source data. In our data-conversion and consulting work, we continue to see data management issues that present significant challenges when creating compliant and meaningful SDTM datasets (15). Some initial reactions may be to “fix” the data, which is not really a solution, but a compounding of the problem. An example of a source-data deficiency that would cause a validation error would be a lab-test result with units that were not collected for some subjects.

### **NOT SUBMITTING ALL COLLECTED DATA**

Automated tools can only check the data that are present. Often, a thorough review of the source datasets is required to ensure that all the collected data were accounted for in the conversion. Some sponsors have erroneously interpreted that they had some discretion in submitting data for variables labeled as “Permissible” in the SDTMIG. Additional wording was added to SDTMIG v3.1.2 as follows:

- “As long as no data was collected for Permissible variables, a sponsor is free to drop them and the corresponding descriptions from the define.xml “
- “The sponsor does not have the discretion to not submit permissible variables when they contain data.”

### **SUBMITTING THE SAME DATA IN MORE THAN ONE STANDARD VARIABLE**

An example of submitting what represented a duplication of data in standard domain variables and SUPP-- domains was presented previously. This also occurs within standard-domain variables. One example seen was the copying of all data in CMDOSE into CMDOSTXT. The SDTM clearly states that --DOSTXT should not be populated when --DOSE is populated. There was also a case where the value of “NOT DONE” migrated to both --STAT and --REASND. It would be impossible for automated tools to predict all the possibilities of data being submitted in more than one variable.

### **SYNTHESIZING DATA TO AVOID VALIDATION ERRORS**

At least two commercial tools check for cases where neither --ENDTC or --ENRF is populated. In cases where the CRF does not collect information such as “Continuing” or “Ongoing,” and the absence of an end date is assumed to imply the event had not ended, neither would contain data. One sponsor chose to

populate --ENRF with "AFTER" in the submission dataset, even though was not traceable back to any collected data.

## **POPULATING NULL VALUES WITH CDISC CONTROLLED TERMINOLOGY OF U FOR UNKNOWN**

The controlled terminology of "U" (for Unknown) is intended to be used when a value of "Unknown" was collected on the CRF. It was not intended to replace data that was never collected. Data that was not collected should be represented as null values. When Findings results are not known, the --STAT variable should be populated with "NOT DONE".

## **MAPPING TO THE RIGHT DOMAIN**

There have been cases where sponsors think that their efficacy data are "special", and do not feel they should be submitted in one of the modeled domains in the SDTMIG. This tends to be most common when efficacy parameters are the same as safety parameters. An example would be serum glucose measurements for a drug that has insulin-like properties, where one sponsor chose to create a GL domain rather than including these data in the LB domain.

Another common mistake seen is migrating height and weight (and other non-PE data) to physical exam, rather than vital signs. CDISC-SDTM controlled terminology for common vital signs test and test-code values exists within the NCI-CDISC-SDTM controlled terminology list, and it includes height and weight. Choosing not to submit in the standard domains creates inconsistencies between Sponsors, which will lead to confusion during the review cycle.

Some sponsors confuse child-bearing potential with pregnancy lab tests and migrate all child-bearing and pregnancy test information to subject characteristics (SC). While child-bearing potential has been considered to belong in SC, pregnancy lab tests results, whether they be urine or serum, should be submitted in the LB dataset with all other lab tests.

Another area where automated tools would not be effective is when unlike data have been migrated into a single custom domain or into other standard domains in order to "cut down on the number of custom datasets that need to be created."

## **MISUSE OF STANDARD VARIABLES**

There are some variables that seem to be misused more than others. Commonly misused variables include --CAT, --SCAT, --GRPID, and the time-point variables. The misuse seems to be the result of two primary factors: not understanding how to use them and "hijacking" them. The latter results when sponsors have data that don't fit into any of the other standard variables, and rather than using SUPPQUAL, they decide to misappropriate a standard variable that's available. Examples are as follows:

**Category and Subcategory Variables.** The --CAT variable is intended to be used to group Topic-variable values. The --SCAT variable is intended to be a sub-categorization of the --CAT value. Only human understanding can assess whether this usage is correct. Consider the misuse scenarios where the --CAT values were as follows:

- 1) identical to the values in --TESTCD (the sponsor felt they had to put something in --CAT)
- 2) used as a key for multiple tests conducted at the same visit (e.g., REPEAT, REPEAT 2) rather than the time-point variables, which would have been more appropriate and better understood
- 3) a counter for the number of sites (e.g., SITE 1, SITE 2) examined for a particular condition

**--GRPID.** The --GRPID variable is intended to group records together within a subject. It has no inherent meaning across subjects. This means that a CMGRPID of COMBO1 for a combination of medications for one subject cannot be concluded to be the same combination in another subject. Sponsors will sometimes use --GRPID instead of --CAT, --SCAT, and/or timing variables such as time points and visits.

**Time-Point Variables.** The time-point variables, --TPT and --TPTNUM, were created to provide a way to convey uniqueness of measurements that might be collected within and across visits, usually within a single domain (since there is no SDTM concept of trial-level time points). When these are used, --TPTNUM is an SDTM Required variable, and is one of the keys. Time points are often based upon a reference event such as a dose, and have planned elapsed times (represented in --ELTM). Time points may also be used simply as a key in cases where there is no reference event. An example of a time

point where --TPTREF would not be populated would be when the protocol states that blood pressure should be measured three times during a visit.

Consider a study in which blood samples are to be taken every 12 hours after dosing, and the subject is allowed to go home at night. An example of time-point numbering that might be confusing is shown in the table below. In this example, the time points were reset at each visit. Keeping them in sequence (i.e., 1, 2, 3, 4, 5) would have allowed the data to be sorted logically using TPTNUM, and might have led to a clearer understanding of the data.

VISITNUM	VISIT	TPT	TPTNUM	--ELTM	--TPTREF
1	DAY 1	BASELINE	1	-PT10M	DOSE
1	DAY 1	12 HOURS	2	PT12H	DOSE
2	DAY 2	24 HOURS	1	PT24H	DOSE
2	DAY 2	36 HOURS	2	PT36H	DOSE
3	DAY 3	48 HOURS	1	PT48H	DOSE

Note also that in the above example that a value of "DOSE" in --TPTREF would provide little information if this sampling was repeated in another period (SDTM Epoch or Element) of the study two weeks later. Values such as "DAY 5 DOSE" and "DAY 19 DOSE" would have been more helpful. Once again, there are no automated checks that can determine whether the data representation is as meaningful as it could be.

### **NOT FOLLOWING CONVENTIONS ESTABLISHED FOR MODELED DOMAINS WHEN CREATING CUSTOM DOMAINS**

Examples seen include the following:

- A Findings dataset where --STRESC contained numeric data and --STRESN was not populated
- Not using title case in the variable labels of custom domains
- Choosing variable labels in custom domains that do not convey an accurate meaning
- An Interventions dataset where --DECOD is "Unspecified" and --TRT is the name of a treatment
- The addition of --DECOD to a domain, but not populating it for any records.
- The population of both --ORRES and --STAT in a Findings dataset
- The population of start dates in Findings domains

### **SEEKING HUMAN INTERPRETATION**

Once a sponsor has recognized that more than automated checks are needed in the assessment of SDTM compliance, they may choose to seek a third party to assist. Below is a list of questions that may help sponsors determine which vendors or service providers might be best qualified to provide support.

- How much experience (i.e., number of studies or submissions) do they have working with the following:
  - Assessing the compliance of legacy data conversion with and without the use of automated tools
  - Running and interpreting the output from automated SDTM-compliance checks
  - Providing solutions to compliance-check issues
  - Performing legacy data conversions
  - The proper creation of Supplemental Qualifiers and RELREC
  - The proper creation of Trial Design datasets at both the trial and subject level
  - The proper creation of custom domains not modeled in the SDTMIG
- What are the qualifications of the people working on this project in terms of the following:
  - Knowledge of the SDTM and SDTMIG
  - Knowledge of common problem areas in converting legacy data to the SDTM format
  - Level of involvement on the CDISC Submission Data Standards Team, which developed the SDTM
  - Therapeutic-area breadth and depth
  - Functional breadth and depth (e.g., data management, programming, statistics)

## CONCLUSIONS

Automated validation checks play an important role in assessing SDTM compliance; however, the variability in clinical trials data in general makes it virtually impossible for a finite number of checks to identify all potential compliance issues. As a result, there is no substitute for the involvement of experienced people in the compliance-assessment process. Humans are essential in determining appropriate action steps when issues are identified by automated checks.

## ACKNOWLEDGMENTS

We would like to thank our colleague, Mary Lenzen, for her thorough and expert review of this paper.

## REFERENCES

1. Study Data Specifications. Current version: 1.4. August 1, 2007; Available via <http://www.fda.gov/cder/regulatory/ersr/Studydata.pdf>
2. Guidances on Providing Regulatory Submissions in Electronic Format; Withdrawal of Guidances. Federal Register Vol. 71, No. 189, Page 57548. September 29, 2006.
3. Federal Register Notice (2006) Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics. Vol. 71, No. 237, Monday, December 11, 2006.
4. Federal Register Notice (2007) Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics. Vol. 72, No. 82, Monday, April 30, 2007.
5. Federal Register Notice (2007) Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics. Vol. 72, No. 236, Monday, December 10, 2007.
6. Federal Register Notice (2008) Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics. Vol. 73, No. 87, Monday, May 5, 2008.
7. Federal Register Notice (2008) Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics. Vol. 73, No. 227, Monday, November 24, 2008.
8. Federal Register Notice (2009) Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics. Vol. 74, No. 233, p. 64203, Monday, December 7, 2009 / The Regulatory Plan.
9. FDA PDUFA IV Information Technology Plan, May 2008. <http://www.fda.gov/OHRMS/DOCKETS/98fr/FDA-2008-N-0352-bkg.pdf>.
10. CDISC blog: <http://www.cdisc.org/fda-cber-cder>
11. Wood, F., and Guinter, T. (2008) Evolution and Implementation of the CDISC Study Data Tabulation Model (SDTM). Pharmaceutical Programming 1 (1): 20-27.
12. Study Data Tabulation Model (2008). Available at the CDISC website: <http://www.cdisc.org/models/sdtm/v1.1/index.html>.
13. Study Data Tabulation Model Implementation Guide: Human Clinical Trials (2008). Available at the CDISC website: <http://www.cdisc.org/models/sdtm/v1.1/index.html>.
14. Analysis Data Model (2006). Available at the CDISC website: <http://www.cdisc.org/adam>
15. Wood, F. (2009) Data Conversion to SDTM: What Sponsors Can Do to Facilitate the Process. PharmaSUG, June 2009.

## 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  
610-535-6500 x5418  
fwood@octagonresearch.com

Adrienne Boyance  
Senior Consultant, Data Standards Consulting  
Octagon Research Solutions, Inc.  
585 East Swedesford Road, Suite 200  
Wayne, PA 19087  
610-535-6500 x5565  
aboyance@octagonresearch.com

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