Paper RS11

Data Conversion to SDTM: What Sponsors Can Do to Facilitate the Process

Fred Wood
Vice President, Data Standards Consulting
Octagon Research Solutions

ABSTRACT

An increasing number of sponsors are submitting clinical trials data to the FDA in the format of the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM). 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 that may be new to some sponsors. These requirements are described in the SDTM and the SDTM Implementation Guide (SDTMIG).

This paper discusses some of the steps sponsors can take to facilitate the conversion of data collected in their traditional format(s) to data in the SDTM format. For maximum efficiency, many of these steps should be undertaken at study setup or during the study (prior to database lock). There are, however, some actions that sponsors can take that will facilitate even late-stage (after database lock) conversions.

INTRODUCTION

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, eANDA, 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 must be eCTD.
- Announcements of a Notice of Proposed Rulemaking (NPRM) regarding 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 (4-6), most recently to September 2009 (7).
- The mention of the SDTM in the PDUFA IV IT Plan (8) as "the foundation for [the] standardized clinical content."

For a more detailed development and regulatory history of the SDTM, see Wood and Guinter (9).

Submission of data in the format specified by the SDTM (10), and described more fully in the SDTM Implementation Guide (SDTMIG) (11), 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 in order to facilitate collection, analysis, and/or reporting.
- 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 for the purpose of facilitating either study-level or integrated analyses.
- Data collection was initially performed for a non-U.S. application, and the submission of electronic data was not anticipated.
- The sponsor chooses to perform end-stage conversion on a study-by-study basis as part of their clinical data flow process, since this can often be more practical in terms of time, effort, and cost compared to building an entire infrastructure around the SDTM (or any other submission standard).

Even if a company has well-established and respected (i.e., adhered to) processes for standards governance, there will likely be at least a few challenges when converting to SDTM-compliant datasets. These include using standard variable names; adopting industry-wide controlled terminology; representing SDTM Findings data in a vertical, normalized format with added baseline flags; 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 creating the special-purpose Subject Elements and Subject Visits datasets. These challenges become greater, or course, for companies who did not maintain or enforce data standards.

This paper will focus on some of the things sponsors can do to facilitate the conversion of their data, whether it be legacy data or data they plan to convert in the future, regardless who performs the conversion activities – the sponsor or a third-party vendor. The recommendations are divided into three sections: The Data, Metadata and Supporting Documentation, and Other Considerations. At the end is a list of questions sponsors might want to ask potential outsourcing data-conversion partners to assess their qualifications.

THE DATA

Begin with Data that Are Complete. Any successful data conversion, or any successful submission for that matter, will begin with the quality of the source data. Sponsors should ensure that all the data specified in the protocol and represented on the CRFs are provided. The fact that some variables are listed as "Permissible" in the SDTMIG does not relieve the sponsor of the need to submit them if data were collected.

Begin with Data that Are Clean. Sponsors should not expect the company or individuals performing the data conversion to perform data cleaning months or years after database lock. It is the sponsor's responsibility to ensure that all necessary edit checks (also referred to as validation checks or discrepancy checks) have been performed prior to database lock. Edit checks include those required for data integrity, data consistency, and passing validation for SDTM compliance at the FDA. Examples include having a start date that's before the end date, and having one standard unit used consistently for each test or measurement. Companies that recognize these problems late may end up having to unlock the database so that data changes can be reflected in the audit trail.

Provide Data in an Electronic, Computer-Readable Format. The most preferred source-data format today is SAS® datasets, since many conversion utilities and/or programs use SAS® as their platform or their output format. Most companies involved in data conversion should, however, be able to create SAS® datasets from data supplied in Excel, CSV (comma-separated values), or tab-delimited files. If the creation of SDTM-compliant datasets requires the generation of electronic datasets from paper or PDF files, sponsors should recognize that timing and conversion costs will be increased.

Understand what Tabulation Data Are. Sponsors who are accustomed to following the 1999 Guidance may have analysis datasets as the only electronic data. 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 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 (12).

Minimize the Number of Necessary Dataset Splits, Merges, and Transformations. These programming actions are needed when source datasets contain either more information or less information than required to create compliant SDTM datasets, or they have a different structure. These problems are commonly seen more with vendor-supplied data than with sponsor-generated data. Recommendations for handling vendor data are discussed in a separate section later in this paper.

Octagon has converted PK and lab datasets, in particular, where all of these additional activities were required. The structure of the main source PK dataset was one record per subject per time-concentration curve. This dataset, which contained columns for twenty-five nominal-time concentrations and twelve PK parameters, needed to be split to separate the concentration data from the parameter data. The two datasets resulting from the split needed to be transformed to create the SDTM-based PK Concentrations (PC) and PK Parameters (PP) datasets. The actual date/times for sampling were in a separate dataset, which then needed to be merged onto the PC dataset.

Octagon has also encountered PK data that were supplied by the vendor as one Excel spreadsheet per time-concentration curve per subject, requiring that all the individual-spreadsheet data be combined to create a study-level domain before one could even begin data conversion.

Define Merge Keys for Datasets that Will Need to Be Merged for Submission. The SDTM defines a number of datasets that represent what might have existed as separate datasets in the sponsor's systems. An example is the Laboratory Test Results (LB) domain, which often exists as a specimen dataset and a lab results dataset. If the sponsor intends for such data to be converted to SDTM, the appropriate merge keys need to be present and defined. Octagon has seen cases where the only timing information in the specimen dataset was the visit, and the only timing in the results dataset was the sample collection date. Without a common merge key such as the date of the visit and the specimen type, ensuring that the appropriate results are merged with the appropriate specimen may either require considerable detective work and consume costly resources, or may not be possible at all.

Develop Standard Transfer Specifications for Data from External Sources. When a study uses multiple laboratories and/or vendors, often due to the geographical location of the study sites, there should be a single data-transfer specification for the same data for the entire study. In other words, it should not be different for different clinical study sites, subjects, visits, or labs or vendors.

The specification should include not only the structure of the transfer file, but the content. The content includes the lab test names, lab test codes, the specimen, and the units for each test. If each lab or vendor provides data in their own format, the lab data will require multiple mapping specifications and added programming effort, including tailored lab-test conversion factors. As example of the latter, if serum glucose data are received in both mg/dL and mg/L, two separate conversion factors would be needed to express all results in the SI units of mmol/L.

The specification should also indicate how normal ranges will be provided. Ideally, these will be included with each lab-test record, since this is how they are submitted in the SDTMIG Laboratory Test Results domain. When normal ranges are provided by the lab as a separate file, an extra step is required to merge of the normal-range file onto the lab results dataset, possible using age and/or sex as additional merge keys.

Identify Foreign-Key Relationships. Capturing the relationship between a concomitant medication and an adverse event may be very important in some studies. This is frequently captured 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. The CDISC CDASH (Clinical Data Acquisition Standards Harmonization) (13) data-collection variable, CMAENO, is an example. For submission, however, the SDTM and SDTMIG specify the use of the RELREC dataset for representing these record-to-record relationships in a consistent manner. Sponsors should recognize that some amount of programming will be required to create RELREC from the collected information.

Expect Fixing Data Collected in Multiple Variables to be Arduous. Sometimes, due to poor CRF design, poor instructions, and/or lack of adequate data cleaning, data are collected in one field that should have been collected in two or more, or the collection of such data is inconsistent from record to record. This is frequently seen with concomitant medication data, where numeric dose information is present in different fields for different subjects or medications. For example, the dose may be in a numeric variable equivalent to the SDTM-based CMDOSE (numeric) for some records, while it may be concatenated with the units in a variable equivalent to the SDTM-based CMDOSTXT (character) for others. A third type of record may have these two fields blank, with data in a variable equivalent to the SDTM-based CMDOSTOT (numeric total daily dose). The sponsor should not expect the data conversion process to include a record-level mapping of data without incurring additional, often considerable, cost.

METADATA AND SUPPORTING DOCUMENTATION

Provide as Much Information Possible, as Soon as Possible. Not all datasets or studies are equal in terms of conversion effort. The more a conversion partner knows about a sponsor's data before the work begins, the better it will be able to assess both the timing and cost of the project. Data definition files, zero-observation datasets, and CRF books, when provided in advance, are extremely helpful in this regard.

Provide an Accurate Data Definition File. Complete and proper source metadata are critical in communicating the nature of the data to be converted. Ideally, the metadata will be provided in a computer-readable format rather than a non-machine-readable format such as a define.pdf. Some of the more common formats seen are Word documents and Excel spreadsheets. Expect questions from conversion partners when relying only on SAS-dataset metadata.

When provided, the dataset-level and variable-level metadata should be consistent with the data in the datasets. All variables in the datasets should be listed in the metadata file, and all variables listed in the file should be in the datasets. The order of the variables in the dataset and the metadata file should be the same. Included in the metadata should be the variable names, variable labels, data types, data formats, and data origin. It should be recognized that if any of these have to be deduced by the conversion partner, additional costs would be incurred.

Provide an Annotated CRF (aCRF). Having a high-quality aCRF facilitates the conversion process by representing how the data were collected. This is a very important tool in efficient data conversions because it unambiguously communicates the relationship between the data collected and the corresponding dataset. There are four major attributes of an aCRF that will affect the efficiency of data conversion:

1. Completeness: Sponsors should provide a complete CRF book, with pages or screen shots ordered by visits. There should be no missing pages. Providing only unique pages is not recommended because doing so will inevitably result in questions (and possibly delays) during the conversion process. For compounds that have been purchased or revived from the past, there may not be a complete CRF available.
2. Searchability: To facilitate navigation, annotation (if not performed by the sponsor), and the creation of Trial Design tables, a searchable PDF or Word version is preferable to a scanned-image version rendered in PDF. It is recognized that providing an electronic and/or searchable document may not be possible for studies using only paper, or with CRFs created in applications that were either homegrown or no longer exist.
3. Annotated electronically: Computer-generated annotations are preferable, but legibly hand-annotated CRFs are better than none.
4. Annotated by a knowledgeable person at the sponsor company: Annotations done without knowledge of the study have the potential to be inaccurate or misleading.

The absence of one or more of these will increase the time, effort, and/or cost of data conversion. It is best to provide information about the quality of this information to a third-party vendor or internal colleagues as early as possible, in order to facilitate planning and conversion-timeline development.

Provide a Copy of the Protocol. In data conversion, the protocol serves as the background for understanding the data that were collected on the CRFs. Sometimes, data necessary to populate the SDTM DM dataset are also provided in the protocol,

such as the investigator name(s) and country site(s). The protocol is frequently required in order to address questions that might arise from the CRF data or annotations (both those provided by the sponsor and those to be created by a business partner). The protocol should be complete, including all amendments. Ideally, it will be available in a searchable electronic format.

Without a protocol, even a complete CRF book doesn't often provide all of the information necessary to properly create SDTM datasets. Furthermore, Octagon has seen numerous instances where the CRF and the protocol are not in agreement, and the CRF cannot be relied upon to represent the protocol. As expected, such inconsistencies result in delays and added costs.

Provide Format Catalogs. Since the SDTM variables are intended to contain human-interpretable values, data collected using codelist codes must be converted to data containing codelists decodes. This requires that the sponsor provide the conversion partner with the format catalog for all codelists. Codes may be helpful in performing analysis of the data, and may exist in an operational dataset, but would not be part of the SDTM submission.

OTHER CONSIDERATIONS

Dictionary Coding. Sponsors will need to determine whether data conversion will involve upgrading to a different dictionary (e.g., from COSTART or WHOART to MedDRA) or a newer version of the current dictionary. All data within a study should be coded to the same version, and not a mix of versions. The SDTMIG provides no guidance for handling integrated databases, so sponsors should consult with the FDA review division to determine the best practice.

Sponsor Point(s) of Contact. To ensure an efficient data conversion, the sponsor should have a representative who is familiar with the protocol and the data. The lack of such a person results in the partner needing to play detective, which can lead to significant delays. Furthermore, if no one at the sponsor company understands the data, the sponsor cannot be confident that data have been converted correctly. The sponsor should also have an individual who is 1) capable of making decisions, 2) is empowered to gather together internal experts to make a decision, and 3) knows when to do each of these. Ideally, the roles of expert and decision maker would belong to individuals who are well coordinated or to a single point of contact.

The Submission of Screen Failures. The sponsor should have appropriate discussion with the review division to determine whether any data collected for screen failures should be submitted, regardless of the submission format. If screen failures are to be submitted, then these subjects should be clearly identified. Ideally, these subjects will have at least one documented inclusion or exclusion exception (which would be reportable in the SDTM Inclusion/Exclusion Exceptions domain) and should not have a record(s) indicating that the subject was randomized or treated. Poor clinical monitoring may have resulted in treatments being administered or assessments performed on screen failures. Representing such data will be challenging regardless of the submission format (not just SDTM) for the party converting the data, and would need to be handled on a case-by-case basis.

Minimize Creativity. Many sponsors, departments, or individuals feel that data standards stifle creativity. Some see the CRF as a vehicle for expressing such creativity. By the time the conversion partner has the data in house, it's too late to rein in the artistic endeavor, so they could be relegated to creating mapping specifications 1) for each study, even when the primary and secondary endpoints are the same, and/or 2) to a level as low as each visit when the same information is collected differently at each visit within a study.

Trial Design. Whether or not a sponsor intends to submit trial design tables should be discussed ahead of time. The retrospective creation of these datasets can take considerably longer than the creation of safety-domain datasets, depending upon how well the protocol is written and/or how well it was translated into CRFs. An accurate creation of several Trial Design tables requires an accurate recording of data upon which the start and end rules will be based. If a treatment period (SDTM Element) is intended to start or end at the occurrence of some other event (e.g., the start of a visit or the start or end of a dose), then capturing the timing of that event is critical in establishing the Element start and end rules in the Trial Elements table. Without the proper definition of rules, it will be difficult or impossible to create an accurate Subject Elements table.

Likewise, the timing for the start and end of a Visit should be captured to so that unequivocal Visit start and end rules can be created in the Trial Visits table. These rules are critical in creating an accurate Subject Visits table. An example of a problem is when the protocol states that a visit begins with admission to the study site, but the CRF does not collect date and time of admission to the site. The person preparing the Trial Visits table must then look to see if any alternate collected date/time (e.g., a pulse measurement) could serve as an anchor to which to tie a Visit start rule. If, however, this activity is not consistently the first activity that took place at the visit for all subjects at all sites, there will be no way to populate a visit start rule that would apply to all subjects.

QUESTIONS TO ASK POTENTIAL PARTNERS

The sections above discuss the potential challenges sponsors and data-conversion partners face in creating SDTM-compliant datasets from legacy data. The list below, which summarizes many of these, can be used to assess the qualifications of potential partners for legacy data conversion.

- How much experience (i.e., number of studies or submissions) do they have working with the following:

  - No or limited metadata
  - No aCRFs
  - Data that are not in electronic format
  - Electronic data that are not tabulation data
  - Data that require splits, transformations, and merges, sometimes without the merge keys being clear
  - Performing multiple mappings for lab and/or vendor data in a variety of formats
  - The proper creation of Supplemental Qualifiers
  - Creating RELREC from foreign-key variables
  - The creation of Trial Design datasets at both the trial and subject level
  - Custom domains not modeled in the SDTMIG
  - Sponsors who might have limited knowledge of their data, as well its origins
  - Working with sponsors to unlock a database and make necessary changes
  - Running and interpreting the output from automated SDTM-compliance checks
- What are the qualifications of the people working on this project in terms of the following:
  - Number of years of SDTM conversion experience
  - Level of involvement on the CDISC SDS Team, which developed the SDTM
  - Number of studies converted
  - Number of datasets converted
  - Number of complex datasets converted
  - Therapeutic-area breadth and depth
  - Functional breadth and depth (e.g., data management, programming, statistics)

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. FDA PDUFA IV Information Technology Plan, May 2008. http://www.fda.gov/OHRMS/DOCKETS/98fr/FDA-2008-N-0352-bkg.pdf.
9. Wood, F., and Guinter, T. (2008) Evolution and Implementation of the CDISC Study Data Tabulation Model (SDTM). Pharmaceutical Programming 1 (1): 20-27.
10. Study Data Tabulation Model (2008). Available at the CDISC website: http://www.cdisc.org/models/sdtm/v1.1/index.html.
11. 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.
12. Analysis Data Model (2006). Available at the CDISC website: http://www.cdisc.org/pdf/ADaMdocument_v2.0_2_Final_2006-08-24.pdf
13. Clinical Data Acquisition Standards Harmonization (2008). Available at the CDISC website: http://www.cdisc.org/standards/cdash/downloads/CDASH_STD-1_0_2008-10-01.pdf

SOME OF THE ACRONYMS USED IN THIS PAPER

aCRF	Annotated CRF
ADaM	Analysis Data Model
CDASH	Clinical Data Acquisition Standards Harmonization
CDISC	Clinical Data Interchange Standards Consortium
CRF	Case Report Form
CRT	Case Report Tabulations
eCTD	electronic Common Technical Document
HL7	Health Level 7
NPRM	Notice of Proposed Rulemaking

SDS	Submission Data Standards
SDTM	Study Data Tabulation Model
SDTMIG	SDTM Implementation Guide: Human Clinical Trials
SEND	Standard for Exchange of Nonclinical Data
XML	eXtensible Markup Language

1. CONTACT INFORMATION

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