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

Fred Wood
VP, Data Standards Consulting
Octagon Research Solutions

CDISC U.S. Interchange
Baltimore, MD
November 2009
• Background
• The Data
• Supporting Documentation and Metadata
• Other Considerations
• Questions to Ask Potential Partners
Background: Interest in SDTM

- SDTM added as a Study Data Specification in the eCTD Guidance (July 2004)
- Withdrawal of Guidances for electronic submissions (September 29, 2006)
  - eCTD as the “preferred format for electronic submissions”
  - As of January 1, 2008, any electronic submission going to CDER must be eCTD.
  - SDTM would be required
- Mention of the SDTM in the PDUFA IV IT Plan (May 2008)
  - SDTM is “the foundation for standardized clinical content.”
Some reasons data may have been collected, stored, and extracted in a non-SDTM-compliant format:

- To facilitate collection, analysis, and/or reporting
- Studies began before the SDTM specification
- Submission of electronic data was not anticipated
- End-stage conversion is part of the clinical data-flow process
• Converting to SDTM-compliant datasets will present challenges, even for companies with well-established and followed processes for standards governance.
• These challenges become greater, or course, for companies who did not maintain or enforce data standards.
Challenges include the following:

- Using standard variable names
- Adopting industry-wide controlled terminology
- Adding SDTM-required Sequence Numbers (--SEQ variables)
- Representing SDTM Findings data in a vertical, normalized format with added baseline flags
- Representing non-standard variables in separate Supplemental Qualifiers datasets
- Representing foreign keys in a separate relationship (RELREC) dataset
- Creating the Trial Design datasets
- Creating the special-purpose Subject Elements and Subject Visits datasets.
The Data

- Tabulation Data in Electronic Format
- Complete and Clean
- Minimize Number of Splits, Merges, Transformations
- Define Merge Keys for Datasets Needing to Be Merged
- Determination of Study Anchors
- Have Standard Data-Transfer Specifications
- Identify Foreign-Key Relationships
- Avoid Splitting Data Across Variables
Tabulation Data
- A component of the Case Report Tabulations (CRT)
  - separate from Patient Profiles, Data Listings, and Analysis Datasets
- Data Collected on CRFs

Electronic Format
- SAS® datasets are preferred
- Excel, CSV (comma-separated values), or tab-delimited files also work
- The creation of SDTM-compliant datasets from paper or PDF files will affect timing and conversion costs
Data Cleaning

- Sponsor’s responsibility
  - Data integrity and consistency
  - Passing SDTM compliance checks at FDA
- Should occur prior to database lock
- May not be possible during conversion
  - Start date after the end date
- If possible during conversion, will be time consuming
  - Standard unit used in consistently for each test or measurement
- Possible need to unlock database so that changes can be reflected in audit trail
Issues more often seen with vendor-supplied data than with sponsor-generated data.

Real-life Case 1:
- Main source PK dataset included twenty-five nominal-time concentrations and twelve PK parameters per record
- Actual date/times for sampling were in a separate dataset.
- Actions needed
  - Split main dataset into PC and PK
  - Transform the two resulting datasets
  - Merge actual times onto PC dataset
Issues more often seen with vendor-supplied data than with sponsor-generated data.

Real-life Case 2:

- PK data supplied as one Excel spreadsheet per time-concentration curve per subject
- Actions needed
  - Combine all data in individual-subject spreadsheets
  - Convert to SAS
  - Transform dataset
The Data: Define Merge Keys for Datasets Needing to Be Merged

• The SDTM defines a number of datasets that might have existed as separate legacy datasets.
  – A specimen dataset and a lab results dataset.
• The appropriate merge keys need to be present and defined.
• Without these, merging lab results with the correct specimen may:
  – Require considerable detective work
  – Consume costly resources
  – Be impossible
• Real-life case:
  – Specimen dataset has a date with no visit.
  – Lab dataset has visit but no date
  – Dates were not collected for all visits
The Data: Determination of Study Anchors

• RFSTDTC and RFENDTC
  – Basis for Study Days (--DY) across all domains
  – Do not expect study-day derivations to be different in different domains

• --TPTREF
  – Reference point for PK and PD data
  – Each concentration or effect vs. time curve should have one
  – There should be an associated and matching --RFTDTC
  – If this is a dose, there should be an Exposure record

• --STTPT and --ENTPT
  – Sponsor defined, but actually determined by protocol and CRF
The Data: Have Standard Data-Transfer Specifications

• There should be a single data-transfer specification for the same data for the entire study
  – Across all study sites, subjects, visits, or labs, and vendors.
• It should include both structure and content
  – Lab test names and codes
  – Specimen naming
  – A unit for each test
  – Normal ranges
• Real-life case
  – Units for glucose coming in as mg/dL, mg/L, and mmol/L
  – Normal ranges in a separate file, needing to merged by age and sex
The Data: Identify Foreign-Key Relationships

- Often seen with the entry of an adverse-event line number on a concomitant medications record.
- Need to be clearly identified by sponsor.
- The SDTM and SDTMIG specify the use of the RELREC dataset for representing these record-to-record relationships in a consistent manner.
- Some amount of programming will be likely required to create RELREC from the collected information.
The Data: Avoid Splitting Data Across Variables

- The problem often occurs as a result of poor CRF design, poor instructions, and/or lack of adequate data cleaning.
- Frequently seen with concomitant medication data.
- Numeric dose information may be present in different fields for different subjects or medications.
  - In a numeric variable equivalent to the SDTM-based CMDOSE
  - Concatenated with the units in a variable equivalent to the SDTM-based CMDOSTXT (character)
  - In a variable equivalent to the SDTM-based CMDOSTOT (numeric total daily dose)
- A record-level mapping of data may be costly in terms of time and money.
• Data Definitions
• Annotated CRF
• Protocol
• Format Catalogs
• Not all datasets or studies are equal in terms of conversion effort.
• Early communication results in more accurate assessment of timing and cost.
• Examples of useful information:
  – Data definition files
  – Zero-observation datasets
  – CRF books
• 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.
• When provided, the dataset-level and variable-level metadata should be consistent with the data in the datasets.
• Include variable names, variable labels, data types, data formats, and data origin.
• Having a high-quality aCRF facilitates the conversion process by representing how the data were collected.

• Major attributes of an aCRF that will affect the efficiency of data conversion:
  – Completeness
  – Searchability
  – Annotated Electronically
  – Annotated by a Knowledgeable Person at the Sponsor Company
• It serves as the background for understanding the data that were collected on the CRFs. Ideally, it should be:
  – Complete, including all amendments
  – In a searchable electronic format
• 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.
Data collected using codes must be converted to data containing codelists decodes to meet SDTMIG requirements.

- The sponsor should 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.
• Dictionary Coding
• Submission of Screen Failures
• Trial Design
• Sponsor Points of Contact
• Minimize Creativity
Other Considerations: Dictionary Coding

- All data within a study should be coded to the same version, and not a mix of versions. Sponsors may need to decide whether to:
  - Change to a different dictionary
  - Upgrade to a newer version of the current dictionary
- The SDTMIG provides no guidance for handling integrated databases, so sponsors should consult with the FDA review division to determine the best practice.
• Discussions with the review division should 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 in the source data.

• Ideally, these subjects will:
  – Have at least one documented inclusion or exclusion exception
  – Not have a record(s) indicating that the subject was randomized or treated.

• Octagon has seen cases where screen failures were randomized and or treated. This is challenging.
Other Considerations: Trial Design

• Submission of TD tables should be discussed with FDA ahead of time.
• Retrospective creation can take considerably longer than the creation of safety-domain datasets.
• The rules for creating the start and end of treatment periods and visits are critical:
  – Ability to know what treatment a subject was on at any point in the trial
  – Ability to know at what visit any data was collected
• Rules are often based upon other study anchors. If these anchors do not have collected times, there will be problems.
• Real-life case:
  – 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 sponsor should have a representative who is familiar with the protocol and the data.

• The sponsor should also have an individual who is:
  – capable of making decisions
  – empowered to gather together internal experts to make a decision
  – 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.
Other Considerations: Minimize Creativity

• Faulty beliefs:
  – Data standards stifle creativity
  – The CRF is a vehicle for expressing creativity
• Conversion partner may be relegated to creating mapping specifications:
  – For each study, even when the primary and secondary endpoints are the same
  – To a level as low as each visit when the same information is collected differently at each visit within a study
Questions to Ask Potential Partners (1)

• How much experience (i.e., number of studies or submissions) do they have working with data:
  – Not in electronic format and/or not tabulation data
  – Requiring splits, transformations, and merges, sometimes without the merge keys being clear
  – Requiring conditional mapping (record dependent)
  – Having no or limited metadata, aCRFs, or protocol
• How much experience do they have creating:
  – Supplemental Qualifiers
  – RELREC from foreign-key variables
  – Trial Design datasets at both the trial and subject level
  – Custom domains not modeled in the SDTMIG
Questions to Ask Potential Partners (2)

- How much experience do they have:
  - Working with sponsors who might have limited knowledge of their data, as well its origins
  - Running and interpreting the output from automated SDTM-compliance checks
- What are the qualifications of the personnel:
  - Number of years of SDTM conversion experience
  - Level of involvement on the CDISC SDS Team, which developed the SDTM
  - Number of studies and datasets converted
  - Number of complex datasets converted
  - Therapeutic-area breadth and depth
  - Functional breadth and depth (e.g., data management, PK, programming, statistics)
Conclusions

• A successful data conversion is dependent upon several key factors:
  – the quality of the data
  – the quality and extent of the supporting documentation
  – an understanding of the study by someone at the sponsor company
  – The experience of the conversion partner