Paper FC02

SDTM, Plus or Minus

Barry R. Cohen, Octagon Research Solutions, Wayne, PA

ABSTRACT

The CDISC Study Data Tabulation Model (SDTM) has become the industry standard for the regulatory submission of clinical trials data. The CDISC Analysis Data Model (ADaM) is not far behind in emerging as the industry standard for the analysis SAS datasets based upon the collected data. Pharmaceutical companies have begun to implement SDTM and, to some extent, ADaM. As part of this process, companies are evaluating whether to use the SDTM standard only for the regulatory submission, for which it is designed, or incorporate it into their operational standard as well. Some companies that are adopting SDTM as their operational standard are finding that a near-SDTM standard works better than strict conformance to the SDTM standard. When a near-SDTM standard is used at various stages in the full business process, the data contains either less than exact-SDTM ("SDTM-minus" as it is sometimes called) or more than exact-SDTM ("SDTM-plus" as it is sometimes called) or a mixture of both. This paper will discuss the use of SDTM as an operational data standard, the use of a near-SDTM standard with its SDTM-plus and SDTM-minus aspects, and how this relates to the development of analysis (ADaM) datasets.

INTRODUCTION

The Electronic Common Technical Document (eCTD) guidance from the FDA (2005 and early 2006) includes a specification for data from clinical and preclinical trials to be submitted according to the SDTM standard[1, 2]. More recently (December 2006), the FDA published in the Federal Register a notice of "proposed rule making which would require that clinical study data be submitted electronically and also require the use of standardized data structure, terminology, and code sets contained in current FDA guidance (the Study Data Tabulation Model [SDTM] developed by the Clinical Data Interchange Standards Consortium)." [3] It is expected that there will be a two-year implementation period before the SDTM is required. As such, the industry today is developing its capabilities to make regulatory submissions of data according to the CDISC SDTM standard.

Some companies will continue to operate internally according to their own standard, and in a final step prepare the data tabulations according to the SDTM standard. The primary motivation for this approach is to meet regulatory requirements while minimizing disruption of internal operations. Other companies are considering, or have already committed to, implementation of SDTM as their new operational standard across their full clinical trials business process. In this approach, they will integrate SDTM throughout their business process instead of converting their data to SDTM as a final step for the regulatory submission. The primary motivation for this approach is to meet regulatory requirements while also (1) avoiding the ongoing cost of converting the data from every submitted study to SDTM, (2) maximizing efficiency by more easily interchanging data with vendors and partners throughout the business process, and (3) maximizing clarity in their process by having all people working with the data using a single standard.

One can probably expect that many or most companies will eventually integrate the SDTM standard into their full clinical-data life cycle instead of converting their submission data to SDTM as the last step in that cycle. This is because the benefits of doing this seem to outweigh the costs. However, companies working on this integration now are finding that full conformance with SDTM tends to make some operations more difficult than they were before. This is because SDTM is a standard for submission of data to the regulatory agency, and not an operational standard. So these companies are adopting a near-SDTM standard for their operations instead of exact-SDTM. The near-SDTM standard they develop allows the needed flexibility for operations during collection, management, warehousing, and analysis/reporting. Yet this near-SDTM standard is close enough to exact-SDTM such that the final conversion to SDTM before the regulatory submission is a relatively minor endeavor.

This subject will be explored further after a brief introduction to the three dimensions of the SDTM standard and to the stages of the clinical-data life cycle.

SDTM: STRUCTURE, NOMENCLATURE, AND CONTENT

The SDTM standard is comprised of structures, nomenclature, and content. Knowing and distinguishing all three is key to understanding both what is involved in converting clinical data to SDTM and where/when the SDTM standard and conversion to SDTM will be integrated into the stages of the clinical-data life cycle within an organization.

- Structure concerns specification of the set of domains, the set of variables that must exist in each domain or can optionally exist, the specification that no variables that are not in SDTM be added to a domain, and rules for creating new, un-modeled (custom) domains. It also concerns, for the domain variables, specification of their data types, lengths, and positions in the domain dataset. It also concerns the shape of the datasets (i.e., “long and narrow” or “short and wide”).
- Nomenclature concerns the standardization of names used in the SDTM model to identify SDTM domains (datasets) and items (variables) in SDTM domains. For example, the demography domain is named DM, not DEMOG, and the variable named SEX is used in the demography domain, not a variable named GENDER. DM and SEX are part of the standard nomenclature. SDTM nomenclature covers all domain and variable names and variable labels.
- Content concerns the standardization of the data values of certain variables in the SDTM model. For example, for the variable SEX, the standard values are “M”, “F”, “U”, not “Male”, “Female”, “Unknown”, and not “1”, “2”, “9”. As another example, the standard data values for Yes/No questions are “Y” and “N”, not “Yes” and “No”. Yet another example is that in a future release of the SDTM model, the values for some test codes will be standard. This standard content is also referred to as “Controlled Terminology” in the SDTM model.

Clinical data can be compliant with SDTM nomenclature and content without being compliant regarding structure. That is, clinical data can be stored in a different structure than the one defined by SDTM, such as in a Clinical Data Management System or in a Clinical Data Warehouse, and still use SDTM-compliant nomenclature and terminology. This fact is important to understanding how and where the SDTM standard will be implemented throughout the stages of the clinical-data life cycle in an organization.

CLINICAL-DATA LIFE CYCLE

Once clinical trials have been designed, the life cycle of the clinical-data business process has the following stages:

- Collection – The investigators in the field collect the data during this stage, either on paper-based Case Report Forms (CRF) or using eCRFs in an Electronic Data Capture (EDC) system. Either way, the CRFs are associated with data tables in a database environment where the data is processed (see Processing stage immediately below). Historically, data collection has been paper-based, but recently eCRF applications have emerged.
- Processing – The data is maintained and processed in this stage. All manners of processing needed to define study database tables, define CRFs, load the collected data to a database, and validate the collected data occur here. Data is stored here on an individual study basis, as opposed to trans-study as in a warehouse. Database systems are often used in this stage, generically called Clinical Data Management Systems (CDMS). If an organization is operating according to an internal data standard, then global libraries are built in the CDMS that hold the standard metadata for the CRF items used in the various studies.
- Storage – This is an emerging stage in the life cycle. Historically, sponsors have extracted the frozen data from the CDMS and stored it in raw SAS datasets for the downstream Analysis/Reporting stage. The storage repository has simply been an operating-system-controlled file system where the various datasets of individual studies are kept together. Limited consideration has been given to archiving according to a single data standard or to trans-study storage and trans-study analysis and reporting. But storage is changing today as companies begin to develop clinical data warehouses to archive their data in a single-standard, trans-study manner. They are planning to use the warehouse as the data source for the Analysis/Reporting stage for individual Clinical Study Reports (CSRs), as the data source for integrating the data for the Summary Reports, and for true trans-study data mining for a variety of purposes including clinical research, safety signaling, and others.
- Analysis/Reporting – The analysis datasets are built in this stage from the raw SAS datasets. Analysis and report programs are developed and executed in this stage, too, to produce the tables, listings, figures, and other analysis outputs for individual CSRs. This tends to be a SAS-centric stage in the life cycle. Data is also integrated across studies in this stage for Integrated Summaries. The efforts to integrate the data have historically been large since data has needed to be standardized across studies and since data standards and structures in the Storage stage have not necessarily afforded an easy integration of data across studies. The situation has improved somewhat over time, especially as sponsors have adopted and enforced internal data standards. However, many sponsors still struggle with this data integration process.
- Compilation/Submission – The data component of the Marketing Application is assembled and submitted in this stage. The data component includes both the collected (“tabulation”) data and the analysis data. The required metadata describing the submitted data is also prepared during this stage, as the “Define” document.

A note about these stages: eCRF applications have their own data repository that is separate from the database associated with the CDMS. These applications also have their own ability to do some of the processing heretofore associated with the

CDMS environment. At some companies, the data from the eCRF repository is still loaded to the CDMS database. However, other companies are now eliminating the CDMS environment by migrating some of their traditional Processing-stage activities into the eCRF application (i.e., into the Collection stage), and the remainder of them to the clinical data warehouse of the Storage stage.

A NEAR-SDTM STANDARD IN THE COLLECTION AND PROCESSING STAGES

Companies want to base the data in their clinical trials business process upon the SDTM standard for a variety of reasons, including:

- To easily produce data in the SDTM standard, which they expect will soon be the required standard for regulatory submission, instead of having to convert each study to SDTM
- To efficiently exchange data with external vendors (e.g., CROs, central laboratories) and business partners
- To maximize clarity in their business process by using a data standard with which everyone working on that process will be familiar
- To use a data standard that will be maintained by an industry consortium instead of having to continually maintain its own standard

However, as was mentioned above, SDTM is a data submission standard instead of an operational standard. There are places where use of the exact SDTM standard can make the business process more difficult than it had been previously with an internal data standard. This is particularly so in the closely connected Collection and Processing stages, and somewhat so in the Storage stage. So companies are adopting a strategy of using a near-SDTM standard in their operations instead of exact-SDTM. Their near-SDTM standard facilitates smoother operations, but it is close enough to exact-SDTM to make the final conversion to exact-SDTM at the needed point in their business process an easy task.

The basic reasons for utilizing a near-SDTM standard instead of the exact-SDTM standard in operational data are:

NON-STATISTICAL AND STATISTICAL SDTM DERIVATIONS

SDTM is a standard for submitting the tabulation data to the FDA. The tabulation data is one of the formats for the collected data. Nonetheless, the SDTM model does have a series of derived variables within it. The majority of these derivations are non-statistical in nature, meaning that they do not require the work of a statistician during the Analysis/Reporting stage to determine the correct derivation. A few the SDTM derivations are, however, statistical.

If a company bases the definition of the database tables in its EDC system or CDMS on the SDTM domains, it will not be able to populate the statistically derived items in these domains because these items cannot be derived until the activities of the Analysis/Reporting stage. Similarly, the EDC system environment or the CDMS environment is not typically the most conducive place to derive the non-statistically derived items. These typically are more easily derived after the data is extracted from the CDMS environment into SAS datasets. So an EDC system or a CDMS cannot easily produce fully SDTM-compliant data in that the data domains will be missing the statistical derivations and many/most of the non-statistical derivations.

Following are examples of the non-statistical derivations:

- AGE, in the DM domain
- ARMCD and ARM, which is unblinded treatment group that the subject is randomized to, in the DM domain
- Reference Start Date/Time (RFSTDTC) and Reference End Date/Time (RFENDTC), in the DM domain
- Actual Study Day (--DY), in most domains
- Date/Time and duration variables (--DUR, --ELTM) in ISO 8601 date-time format, for all domains

Following are examples of the statistical derivations:

- Population flags (in SUPPDM): Intent to Treat and Per Protocol Evaluability
- Derived records (--DRVFL), such as a computed baseline record that is the average of collected records in a domain, or a computed questionnaire score or subscore that is computed from the collected results using a particular algorithm
- Baseline flags (--BLFL): Indicators used to identify a baseline value. (These can sometimes be set without statistician input during Analysis/Reporting.)

HORIZONTAL STRUCTURE NEEDED FOR SOME FINDINGS DATA IN COLLECTION SYSTEMS

SDTM Findings domains have a vertical structure. On each record in a given Findings domain, the --TESTCD variable identifies the measurement, test, or examination, and the --ORRES variable holds the result or finding. In contrast, in a data collection system, (an EDC system or a CDMS), data that will be submitted in an SDTM Findings domain is sometimes much more easily collected and processed in a horizontal structure. An example would be a questionnaire where each individual question in the battery has its own unique response set. Data collection systems will need to store a unique set of rules for validating each of these questions and it would thus simplify the data checking if each question resides in its own variable in the database. This would necessitate a horizontal structure for this data.

So in these instances, the EDC system or CDMS cannot operate with SDTM-compliant domains without having an impact on data processing. For optimal functionality, data will have to first be extracted from the collection system and then restructured.

OPERATIONAL DATA ITEMS ARE NOT IN SDTM DOMAINS

Companies sometimes include operational questions on CRFs that are valuable for their collection operations but do not belong in any SDTM domain. In these instances, the SDTM domains in the EDC or CDMS environment will not be SDTM-compliant. These operational data items will need to be dropped when the data is extracted from that environment. Example operational data items that will not migrate to SDTM include:

- “Did the subject have any adverse events? Check Yes or No. If yes, complete this page.”
- “Was a blood sample drawn for lab testing? Check Yes or No.”

SUPPLEMENTAL QUALIFIER DATA COMBINED WITH PARENT DOMAIN

The SDTM standard includes a rule that no new variables can be added to any data domain. If a user has additional data for a domain, which cannot be fit into the domain using the standard variables, they must use a Supplemental Qualifiers special-purpose dataset for this purpose. This is a separate dataset from the “parent” domain in question, and it has a vertical structure that allows the user to add any supplemental data in a “variable name – variable value” structure. However, in a collection system, it is sometimes difficult to store the supplemental data separately from the parent data domain being used to store and process the data. Rather, these supplemental variables are added to the parent domain in a horizontal structure. In these instances, the SDTM domains in the EDC or CDMS environment will not be SDTM-compliant. The data domain will need to be restructured into a parent domain and Supplemental Qualifier dataset when the data is extracted from the collection environment. Examples include:

- Description of “Other Medically Important Serious Event” – The description would usually be collected with the AE data collection module but will need to be submitted in the SUPPAE dataset.
- Treatment Emergent for an AE data collection module – This data point would also usually be collected with the AE data collection module and need to be submitted in the SUPPAE dataset.

A NEAR-SDTM STANDARD IN THE STORAGE STAGE

In the Storage stage, many of the operational constraints of collection and processing are lifted and the data standard can be closer to exact-SDTM. This is true whether the repository is a clinical data warehouse built in a database environment or a series of SAS datasets organized in the operating system’s native file system.

To consider the use of a near-SDTM standard in this stage, remember that the data is being stored initially for two activities: (1) Source data for building study-specific analysis datasets for the Analysis/Reporting stage, and (2) Source data for building the final study-specific SDTM datasets for the regulatory submission. The data is also being stored for two longer-term activities: (1) integrating data across studies for Integrated Summaries analyses/reports, and (2) A variety of warehouse-enabled activities such as Business Intelligence reporting and safety-signal analysis.

There are two basic approaches to the use of the SDTM standard in the Storage stage. In the first approach, the data is made fully SDTM-compliant. The tasks involved are:

- Complete the non-statistical derivations
- Transpose the Findings data from horizontal to vertical structure, where needed
- Drop the operational questions that are not part of an SDTM domain
- Build the domain-specific SUPPQUAL datasets, removing this data from the parent domains.
- After the Analysis/Reporting stage activities, load the statistical derivations into this data

The benefit of using the exact-SDTM standard in the Storage stage is twofold. First, the data will be submission-ready at this point. Second, the fully vertical structure is probably more amenable to storage in a data warehouse. This is especially true of a company that wants to replicate the Janus data warehouse that the FDA will use to store all the tabulation data it receives. The problem with using the exact-SDTM standard is that as a source for building analysis datasets, it would probably be easier to leave some of the Findings data in horizontal structure, and to leave some of the SUPPQUAL data as part of the parent domain, (which would thus also be in a more horizontal structure).

In the second basic approach, the data in the Storage stage remains in a near-SDTM standard. The tasks involved are:

- Complete the non-statistical derivations
- Drop the operational questions that are not part of an SDTM domain
- After the Analysis/Reporting stage activities, load the statistical derivations back into this data
- Leave the horizontal Findings data in that structure
- Leave the domain-specific SUPPQUAL data as part of the parent domain.

The choice of approach will probably be driven by how a company wants to design their data warehouse. If the needs of the data warehouse require the exact-SDTM standard to be used in the Storage stage, then there will be some additional work involved in building analysis datasets, but less additional work in compiling the submission-ready tabulation data.

"SDTM-PLUS" and "SDTM-MINUS"

The near-SDTM data domains used in an EDC system database or CDMS database are sometimes referred to as "SDTM-minus" data because they do not yet have the required SDTM derived items and records. This EDC- and CDMS-based data is also sometimes referred to as "SDTM-plus" data because some of the Findings data are still horizontally structured and thus have extra variables, and/or because SUPPQUAL data is still in the parent domain and thus is extra data in that domain, and/or because it has operational data that is not part of the SDTM domains.

There are in fact no official, exact definitions for the terms "SDTM-plus" and "SDTM-minus", even though they are gaining traction in the industry. People use these terms, in general, to indicate that their data domains in a given stage of the clinical-data life cycle are based on a near-SDTM standard instead of the exact-SDTM standard, and the deviation from exact-SDTM stems from the data domains having less than all the required items and records, and/or having additional, non-compliant data.

A near-SDTM standard is likely to be most SDTM-compliant regarding SDTM nomenclature and content (i.e., controlled terminology), and somewhat less SDTM-compliant regarding SDTM structure.

A near-SDTM standard used in the Storage stage is likely to be closer to exact-SDTM than it would be in the closely associated Collection and Processing stages.

CONCLUSION

The CDISC SDTM is the new industry standard for the regulatory submission of tabulation data. Companies have begun to incorporate this standard into the operations of their clinical-data life cycle, particularly into the activities of the Collection, Processing, and Storage stages, and to some extent into the building of analysis datasets in the Analysis/Reporting stage. However, SDTM is a submission standard and not an operational standard. Using the exact standard in operations can lead to certain processing difficulties that do not exist with an internal operational standard. Thus, companies are developing near-SDTM standards instead of using the exact-SDTM standard in order to avoid these undesired impacts on operations. The near-SDTM standards being used are close enough to exact-SDTM to make the final conversion to SDTM for the regulatory submission a relatively minor endeavor.

The near-SDTM standards being developed are sometimes called "SDTM-plus" or "SDTM-minus". These two terms have gained traction in the industry but they do not have precise definitions. In general, they refer to a standard that produces data domains that either are missing some items required in SDTM (SDTM-minus), or have additional items not compliant with SDTM (SDTM-plus), or both. The deviation of near-SDTM standards from the exact-SDTM standard tends to concern SDTM structure more than nomenclature or content (i.e., controlled terminology).

REFERENCES

1. Guidance for Industry Providing Regulatory Submissions in Electronic Format — Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications, Revision 1, April 2006 - 16 pages http://www.fda.gov/cber/gdlns/esubapp.pdf

2. Study Data Specifications, Version 1.2, 2006-03-04 – 5 pages http://www.fda.gov/cder/regulatory/ersr/Studydata-v1.2.pdf

3. Federal Register Vol. 71, No. 237 - 2006-12-11, Regulatory Action Plan, Sequence Number 36

Susan Kenny and Michael Litzinger: “Strategies for Implementing SDTM and ADaM standards”, Paper FC03 in Proceedings of the PharmaSUG 2005 Conference, May 2005. < http://www.pharmasug.org/2005/FC03.pdf >

ACKNOWLEDGEMENTS

The author would like to thank his colleagues Dr. Fred Wood and Ms. Mary Lenzen, Principal SDTM Consultants at Octagon Research and members of the CDISC SDS team (that developed the SDTM), for their thoughtful review of this paper and their insight throughout the year into the details of SDTM and its use in the industry.

CONTACT INFORMATION

Barry R. Cohen
Director, Clinical Data Strategies
Octagon Research Solutions, Inc.
Wayne, PA 19087
610 535 6500 x5635
bcohen@octagonresearch.com