Some Limitations of CDISC - SDTM and ADaM as Operational Models

Mario Widel[1], Henry B. Winsor[2]
[1]Roche Molecular Systems, Inc., Pleasanton, CA
[2]WinsorWorks, Limited, San Mateo, CA

Abstract

Both SDTM (since 2005) and ADaM (more recently) have demonstrated distinct advantages to FDA reviewers when they receive data in these formats. Both SDTM and ADaM are mentioned in FDA guidance as highly desirable, so more and more sponsors are including SDTM and ADaM as an integral part of the NDA/BLA/PLA submissions.

While some sponsors create and submit these data sets as final products for review, others are trying to use these data sets internally as well as in submissions. While there are distinct advantages to the latter method (fewer steps, a more efficient organization, etc.), there are many challenges that arise and need to be addressed before a sponsor is going to be in a position to successfully use SDTM and ADaM internally.

This paper addresses these challenges and offers a few proposed strategies that can be used to overcome them.

Key Words: CDISC, SDTM, ADaM, Operational Model

Disclaimer

The opinions presented in this paper are solely those of the authors and not necessarily those of their respective employers or any CDISC team.

Introduction

In the late 1990’s, there were many pharmaceutical industry mergers, acquisitions and partnerships, requiring their employees to become proficient in bringing together data from multiple sources. Also, as regulatory agencies moved from paper to electronic sources of review material, it became evident that companies’ own internal standards were not adequate to easily interchange data. A multi-country, multi–company effort was necessary to bring some standards to the industry. In 1997, CDISC (Clinical Data Interchange Standards Consortium) was formed and began the effort of developing standards. CDISC was originally developed by industry professionals volunteering time in addition to their “day” jobs and most CDISC model development work is still done by volunteers from many countries.

Among the early standards are the two this paper considers, SDTM (Study Data Tabulation Model) and ADaM (Analysis Data Model), and evaluates their suitability as an operational model, i.e. their utility in defining the structure, rules and relationships of data elements that would allow them to be used as the framework of working data sources, enabling efficient clinical data review and manipulation in a regulated environment. Neither SDTM nor ADaM were initially intended for this purpose, with SDTM intended to submit Case Report Form data in place of the paper-based Case Report Tabulations and ADaM finally giving regulatory agencies direct access to the analysis data sets for the first time in a standard format. Both were intended to be submission products.

SDTM

In the beginning there was chaos. Each company had, if at all, their own data naming conventions, and sometimes many of those. SDTM offered consistency to the submission process by standardizing the data structure, defining standard data set and variable names, indicating what values from the CRF should be present and in conjunction with other CDISC initiatives, directed the user towards using controlled terminology as much as possible. Once submitted, SDTM it is used by FDA medical reviewers. Many sponsors have in the past, and some still do, create SDTM files after the study report, duplicating work. Other sponsors imbed SDTM within their operational process.

What Doesn't Work Well

However, we have identified some weaknesses in the design, many of which make it less than optimal for a working data source. The notion of the need for a place to put non-standard CRF items was necessitated by the rigid column structure that JANUS (FDA database) required. Instead of including all possible columns within a data domain, the SDTM design includes only those variables that were generally conceded to be in common use, and requiring all other items to go to a secondary structure, initially called SUPPQUAL, with the idea that the two pieces of the data could be merged together at some future time by some sort of a standard tool. Since SUPPQUAL was expected to contain data from all domains, the data set could be huge and difficult to update in a domain specific manner, so individual domain SUPPQUALs, SUPPxx for the xx domain, were later added to the standard. The problem with this is someone browsing the data set does not see all of the relevant data in one source.

A related problem comes from the SDTM support (or lack thereof) for key variables. Key variables in a data structure allow for easy identification of a unique record. An index variable, called xxSEQ in the xx domain, is the only key required below the level of USUBJID. While it is a unique value within the USUBJID's records, clearly identifying each record, it lacks any meaning that could be supplied by more meaningful key variables. For instance, if you used the variables LBDT and LBTESTCD as lower key variables in the LB domain, knowing that record where LBSEQ = 47 says nothing about the contents, while LBDT = 2009-07-10 and LBTESTCD = HGB indicates that the record contains the haemoglobin value from July 9th, 2009. Also, the xxSEQ value may change over time as data is augmented or reordered; it's not a problem for a final submission product, but liable to cause problems while the final data is in a preparatory stage. The xxSEQ variable is evidently required by the rigid structure, as there is no obvious place in the SDTM model to store more meaningful key variable information.

There are other problems within the SDTM standard. We find it problematic that comments are segregated from the other data into one domain, requiring a reviewer to review multiple data source objects to be able to see all of the data. The phrase "out of sight, out of mind" should come immediately to mind. Note also that all possible variations of partial dates are supported, even ones that are meaningless. Is there value to know that a concomitant medication was administered in the month of March when the year is not known, or on the first of the month, with both the year and month unknown? The ISO 8601 duration format, unlike the date format, does not naturally sort in a temporal sequence unless all values contain the same elements. For instance, duration values PT7H0M0S and PT12H0M0S are both legal, but it should be clear that the seven hour value will sort after the 12 hour duration. It is interesting that a variable derived from a Required Category variable is often not Required, such as the xxDY variable, although xxDTC is Required. Finally, there needs to be some clarification as to what is meant by the Permissible Category. Too many end users seem to think that a Permissible variable is optional instead of one that should be included if it is collected.

SUPPQUAL Remediation

Limiting the problems stemming from the need to store data in a supplementary qualifier data set can be handled in two ways. First, always use a domain specific SUPPQUAL. The presence of a separate SUPPxx data set will alert any reviewer that there is supplemental data for the domain. Secondly, consider using an augmented version of the domain data set, so that the data set contains all of the variables. While this is a violation of the SDTM standard (you cannot just add new variables to the standard set), it is not necessary to have submission data sets throughout the operational process. At the end of the process, when submission data set are necessary, simply split the data set into two pieces, the domain data set XX and the supplemental qualifier data set SUPPxx. It is further suggested that a standard tool be developed to do the splitting action, one that is fully tested and always splits the operational data set into the correct two pieces. Relying on someone’s open code to do this at crunch time is a recipe for getting the data sets back from the FDA and getting to redo them..

Actual Key Variables Remediation

This is a more difficult problem to fix, as there is not a requirement within the SDTM standard for a DOMAINS data set, which would be an overarching one that lists all data sets in the submission and would thus provide a place to contain domain specific metadata like the actual data set keys used to order the data set. There is a place in the define.xml file for this metadata, but the define.xml standard references more items than merely SDTM data sets.

One possible solution is the use the group ID variable (xxGRPID) within each data set to store the actual key values, either with the variable name identified (USUBJID = 17-3-491 LBDT = 2009-07-25), or if space is a particular problem, just the values (17-3-491 2009-07-25). Please note that if you do this and have numeric key values, the numeric values should be sized properly so they will sort appropriately when contained within a character variable.

Remediation of Other SDTM Issues

There is little that can be done with the comments data, unless one is willing to add them to the augmented domain data sets, much like we have proposed that you add the SUPPQUAL data, and splitting off the comments data when it comes time to create submission ready data sets. Again, developing a tool to do this easily and consistently is preferable to expecting someone to do it in open code, especially since the comments data could come from multiple domains.

Making sure that all duration values sort properly is relatively easy. All that is necessary is implementing an internal standard that all duration values contain all of the same elements and each element value contains the same number of places, even if leading zeroes are required. It is suggested that partial dates that do not contain the meaningful sequence of elements are probably better treated as comments, rather than forced into a column otherwise full of values with actual date/time information.

Finally, addressing the last two items are really more of an internal standards issue than anything else. Some sponsors would greatly benefit from realizing that supplying the FDA with all possible Permissible variables is more likely to ease and speed the review process and that the negligible amount of time and effort saved by not including variables like study days is counterproductive. Taking the stance that if a variable is collected or derived from collected data, then it should be reported, will only aid a reviewer in doing his/her job, and if these variables are in the operational data, it is no effort to pass them on in the submission data sets.

ADaM Issues/Remediation

Like the SDTM model, the ADaM model was never intended as an operational model of any sort, and unlike the SDTM model, the ADaM model is not easily extensible into one. The ADaM model data sets are focused on delivering the data to support specific analyses, and augmenting them to be comprehensive would render them huge, unwieldy and minimize their utility. Think of ADaM data sets as intermediate deliverables to facilitate the analyses and don’t abuse them too badly. It is one thing to add a few variables to a data set to handle the few variables otherwise contained in a SUPPQUAL or Comments data set; it is another thing to add a few hundred or thousand variables to an ADaM data set, just so it is comprehensive.

Conclusion

If you work at a company that maintains their own internal data standard and is quite happy with it, this paper may have limited interest. Once data is mapped to a standard, it’s pretty easy to map it to another, so your company may see no gain from adopting a tweaked version of SDTM for operational use. However, if your company either doesn’t have a standard or has multiple, non-shareable standards, then your company will probably benefit from adopting a single standard for working data sets. The SDTM model, although needing some additions, has the advantage of being a publicly available standard with which most of your employees will already have experience. Conforming your data to the standard is work that will need to be done eventually if your research program is successful. Doing the conversion early in your process is not wasteful even if you never submit your data, if you are willing to select a single standard and develop tools to take advantage of the standard.

The authors believe that SDTM could be more than a filing requirement. It has the potential to be a data source that is useable by many users before, during and after the completion of the study report.

Acknowledgements

The authors would like to thank their respective employers, both past and current, for the opportunity to learn the contents of this paper.

Also, we would like to thank the FDA for the jobs that we have.

.

References

CDISC Study Data Tabulation Model (SDTM) v1.3 and Study Data Tabulation Model Implementation Guide (SDTMIG) v3.1.3
http://www.cdisc.org/sdtm

Analysis Data Model (ADaM) Implementation Guide
http://www.cdisc.org/adam

FDA. CDER Common Data Standards Issues Document
http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM254113.pdf

FDA CDER,CBER,CDRH. Providing Regulatory Submissions in Electronic Format — Standardized Study Data.
http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf

International Standards. ISO 8601:2004 - Data elements and interchange formats — Information interchange — Representation of dates and times.
http://www.iso.org/iso/catalogue_detail?csnumber=40874

Fred Wood, Tom Guinter. Evolution and Implementation of the CDISC Study Data Tabulation Model (SDTM). PharmaSUG2006.
http://www.lexjansen.com/pharmasug/2006/fdacompliance/fc03.pdf

Barry R. Cohen. SDTM, Plus or Minus. PharmaSUG 2008.
http://www.lexjansen.com/pharmasug/2008/rs/rs01.pdf

Sandra Minjoe. Implementing CDISC When You Already Have Standards: A Case Study. SAS Global Forum 2009.
http://support.sas.com/resources/papers/proceedings09/162-2009.pdf

Mario Widel, Henry B. Winsor. Good Versus Better SDTM -- Date and Time Variables. PharmaSUG2011
http://www.pharmasug.org/proceedings/2011/CD/PharmaSUG-2011-CD23.pdf