PhUSE 2007

Paper RA10

Considerations for CDISC Implementation

Robert T. Stemplinger, ICON Clinical Research, Redwood City, CA
Jackie Lane, ICON Clinical Research, Redwood City, CA

ABSTRACT

Implementation of the CDISC standards presents a number of challenges that can have a deep and far reaching impact on an organization. To mitigate that impact and avoid wasted effort and exorbitant costs, careful consideration should be given to a handful of fundamental issues. Once these issues have been embraced a proper direction can be taken to meet the needs of the organization as they have been defined. This paper will detail some of the various challenges, propose possible solutions based on real life examples, and discuss technological options from SAS and other providers that can be applied to facilitate implementation.

INTRODUCTION

Prior to any discussion related to the implementation of the CDISC standards, it is necessary to first define its scope. Of the many models that comprise the standards, perhaps the most relevant for Biostatistics and Data Management departments are the Operational Data Model (ODM), the Study Data Tabulation Model (SDTM), the Analysis Dataset Model (ADaM), and the Case Report Tabulation Data Definition Specification (CRT-DDS). Any full implementation would require the inclusion of at least these elements, though it is still possible for an abbreviated or study specific implementation employing only one or two. For the purposes of this paper, a full implementation is assumed.

CONSIDERATIONS

The fundamental issues when implementing the CDISC standards can be broken down into a few broader, more general questions. The manner in which each question is addressed determines the direction the implementation will take and the impact it will have on the organization. These general questions are: (1) at what point during the study life cycle the standards should be applied, (2) who should interpret and apply the standards, and (3) how the implementation should be performed. Further influencing each of these considerations are training and technology.

APPLICATION

At what point during the study life cycle to apply the standards is perhaps the most fundamental consideration. There are three possibilities:

- Defined in the clinical database management system (DBMS)
- As part of the manipulation of the data extracted from the DBMS (i.e., the “back-end”)
- A combination of DBMS implementation and the back-end programming, or a hybrid approach

Defining the CDISC standards in the DBMS provides obvious benefits. It enforces standardization of database designs and structures across studies and projects and allows the data to be housed in the format that it will be submitted. It builds efficiencies into the development process, since there is an increased potential for the reuse of code and structures. And it minimizes or in some cases eliminates the need for manipulation of extracted data, since the underlying database tables are already CDISC compliant, and are extracted in the format in which they will be analyzed and reported. But there are drawbacks as well. It could complicate data entry, depending on the database system employed, if the underlying table structure dictates a less than user friendly entry screen. It also could require a substantial increase in resources required to build databases, especially for less experienced staff unfamiliar with the standards. The impact of defining the standards in the database is set squarely on the creation of the database structures themselves. Taking a step back from these initial pros and cons, an organization would be well advised to carefully plan its objectives for implementation and perform an assessment of available resources. This will differ from one company to the next, and is obviously impacted by the size of the company, the resources available, and the sector in which it resides. For instance, large pharmaceutical companies may decide to implement the standards within the DBMS, and they certainly have the resources to do it. Smaller biotech companies can not due to cost and resource constraints. The matter is a bit more complex for Contract Resource Organizations (CROs), who may have a varying distribution of clients all with different needs. Some clients are eager to embrace CDISC, some are not, and still some request their own customized database structures based, in varying degrees of conformity, on the CDISC domains. It isn’t possible in this case to define a single standard and apply it to every study, so a measure of flexibility is required.

Implementing the CDISC standards at the “back-end”, as part of the manipulation of the data in analysis and reporting, has benefits and drawbacks as well. Because there is no involvement with the DBMS, the potential for data entry complications is

eliminated. Databases can be built and put into production more quickly, using existing standards and techniques. The drawbacks are obvious. Database structures would not be standardized or might be albeit to an internal, company specific standard. The reworking of data would be required when intuitively it would seem to make the most sense to build the database so that the data can be extracted in a format that is most useable for analysis. Here, the creation of the database structures themselves is not impacted. The impact falls entirely on the manipulation of the extracted data. While this appears to be an equitable trade-off between building the standards into the database versus building them into the extracted data sets, it is a bit more complicated. The skill sets required at each stage of the process differ. Database programming and design skills are required for database builds, whereas SAS® programming skills are require for manipulation of the extracted data. Whether or not these skills are present in database programmers depends on the organization, and of course would be a consideration when addressing this issue of where the standards should be applied. Again, the decision should be driven by the defined objectives of the organization and a realistic assessment of resources. As such, many of the points made above in reference to how the various organizations might implement the standards hold true. Larger pharmaceutical companies typically have more robust implementation objectives and more resources to expend in order to meet those objectives. Smaller companies simply do not, and CROs find themselves trying to offer services to their clients while maintaining costs and gaining efficiencies.

In practice what is often observed is in fact a hybrid approach. While some of the CDISC domains lend themselves to implementation in the DBMS, some do not. The Vitals (VS) domain, for example, could be difficult to build in certain database systems, due to its hyper-normalized structure. The Supplemental Qualifiers (SUPPQUAL) domain is yet another example. Also, if the database structure is highly dependent on the layout of the CRF, it could prove to be exceedingly difficult and in some instances impossible to implement, unless the CRF is CDISC compliant. In each of these cases, then, there is additional programming required at the “back-end” to ensure the extracted data sets are fully CDISC compliant. Thus the natural tendency is to do as much as possible within the DMBS, because intuitively it is logical to do so, and then apply further manipulations to the extracted data using SAS. This approach could result in additional resource requirements for both DBMS development and manipulation of the extracted data, which indicates a less than optimal process. A further complication could also arise depending on the structure of the organization and the skill set resident in each department. If database programmers in Data Management are strictly responsible for database development and SAS programmers in Statistical Programming are responsible for data manipulation, then two groups are impacting by this hybrid approach.

It seems most logical, then, at least at this time when the CDISC standards have not yet had an appreciable impact on data collection, i.e., CRF design, to strongly consider implementation via “back-end” programming. Once the standards have propagated far enough along to the CRF, it follows naturally that database development will be facilitated. Then the resources required for implementation within the DBMS will be analogous to what currently exists for many organizations, and the additional time and knowledge needed to adapt the CRF to the standards will not be necessary. Until that time, a “back-end” approach eliminates potential data entry complications, keeps constant the time needed for database builds, and limits the impact of implementation to a single group within the organization.

INTERPRETATION

Perhaps the most difficult of all tasks related to implementing the CDISC standards is their interpretation. Prior to actually attempting this interpretation, however, a definition of the type of interpretation needed should be addressed. This should be defined as part of the objectives for implementation as set forth by the organization. As the initial overriding goal of CDISC was to establish a submission standard for medical review of clinical trial data at the FDA, it appears to make the most sense to adopt CDISC as an internal data standard. This would require a strict interpretation, one that is fully compliant. The benefits from the adoption of such a standard are numerous, and have been mentioned above. But data structured for medical review is not by rule optimally structured for analysis and reporting. In fact this is what is found in practice, where the structure and contents of the SDTM domains do not readily lend themselves to manipulation into tables, listings, and figures. From this, adopting CDISC as an internal standard does not appear to make the most sense, since it would result in a less than optimal data standard for all aspects of the organization. To overcome this limitation, some organizations have decided upon a variant of the standard domains. These variants are fundamentally structured in the spirit of the CDISC domains, but may differ slightly in a number of ways. Some may have additional variables specifically used for analysis, while others may have slightly different structures than the domain on which they are based. In any case, variants allow for less stringent data models and have wider application to all pertinent departments within the organization. They are not, however, immediately suitable for submission.

While the structure of the CDISC domains is clearly defined, mapping existing or legacy data structures is sometimes difficult and time consuming. An in-depth understanding of the standards is required, and there are a series of questions that arise that could affect and depend upon the overall plan for submission. Furthermore, some types of data map relatively easily, while others do not. Adverse event data is an example of data that tends to map well, since for the most part it is consistent from one study to the next, and one company to the next. Other historical or intervention types of data do not. More difficult is specialized data that is not common or common data collected in an unusual manner. For instance, child bearing potential information collected multiple times during the course of a trial presents difficulty since this type of information is usually present on the subject characteristics domain. Since that domain does not allow for characteristics information at multiple visits, it is not acceptable to place it there. These types of data usually require custom domains.

Clearly, then, who is best suited to interpret the standards is largely dependent on the goals of the organization. This is a familiar and consistent theme for the implementation process. A variant offers more flexibility in the creation of the data, while

a strict interpretation results in a fully conforming set of data for submission. Either approach requires the involvement of staff familiar with data standards and data manipulation. Staff experienced with submissions to FDA adds much value to the process as well. The most favorable approach would seem to be a cross functional team that possesses the skills outlined above, and this is in fact what is happening at most organizations. Participation from statistics, programming, and data management is required to fully implement the various models of the standard.

IMPLEMENTATION

Once it has been decided where the standards will be applied and the depth of their implementation, all that remains is to decide who will perform the implementation. That is, who it is that will program the data structures to conform to the standards. This is affected in many cases by where they are applied, and which models are utilized. If the application is within the DBMS, then it makes the most sense to have database programmers and data managers familiar with database design techniques execute the implementation. Implementation at the “back-end” requires resources skilled with data manipulation and restructuring using SAS, and in this case SAS programmers and biostatisticians are best suited for the task. There also is some delineation that can be made along the lines of the models themselves, with a natural tendency to have database programmers code SDTM and SAS programmers code ADaM (although this is by no means a hard and fast rule).

A further complication is in deciding the most efficient means of implementing the programming of the individual models. While it is certainly possible to program each model independently of the other, using the raw data source as input, it would seem that there should exist some sort of a relationship between SDTM and ADaM. It would be most beneficial, and efficient, to capitalize on that relationship. Unfortunately there are a number of complicating factors. The logical flow is to build the SDTM domains from the raw data source and then implement ADaM using the SDTM. The SDTM domains are not optimally structured to serve as input to ADaM, however, and present certain difficulties. An alternate approach would be to create an intermediate set of data structures, a variant of SDTM, which addresses the shortcomings of the fully compliant structures, and from that follow a dual path to compliant SDTM and ADaM data sets. This approach requires additional validation which makes it a less attractive option in some cases, since validation of the variant data sets is required in addition to validation of the compliant SDTM and ADaM structures. Thus, it would seem the linear approach is favorable, even though it may have a steeper learning curve.

TRAINING

Training is a critical consideration when planning implementation of the CDISC standards. Before any activities commence, a discussion on training, or training itself, should be held. Given the intricacy and nuances of the various elements of the standards, coupled with the fact that they present a structure that differs from what many in the industry are accustomed to, it is prudent to at least consider some kind of detailed training plan. This plan should include initial training designed to educate those who will begin immediate implementation of the standards. It does not make sense to offer training to staff who will not then use it, so timing is essential. Once sufficient experience is gained, follow-up training should be provided to ensure that skills and knowledge remain current.

TECHNOLOGIES

Implementation of the CDISC standards is an endeavor that is rich with opportunity for the application of technology provided that that application is not in place of well thought out processes and solutions for each of the considerations mentioned above. There are an abundance of external vendors with many offerings for mapping tools, XML generators, data conversion utilities, etc. Data review tools such as Integrated Review™ (iReview) or jReview™ work well with the standards, although don’t require them, and CDISC specific review tools such as WebSDM™ all provide mechanisms not only for checking the structure of the data sets but facilitate review of the contents as well. With a decision to implement at the “back-end”, and sufficient CDISC training and SAS expertise, a reasonably proficient SAS programmer could develop an internal toolkit to assist in transforming and verifying the data sets. As always, the defined objectives for implementation and available resources drive the application of technology. While large pharmaceutical companies are able to afford complex technologies, both financially and in terms of resources, smaller biotech firms can not. Worse yet is the case for CROs, whose clients may expect the efficiencies offered by the application of technology, but do not expect to have to share the cost. For organizations who simply can not afford the expense of such tools, realizing the maximum benefit of existing technology is paramount. Since a fair majority of companies utilize SAS software for analysis and reporting of clinical trial data, it is logical to at least investigate what solutions are available. SAS/DI® Studio offers a complete development environment for those organizations that might choose it. Its application is not limited to CDISC, but it does include transformation tools useful for manipulating the data to conform to the standards. PROC CDISC offers less functionality but is targeted specifically for CDISC and XML production, with future enhancements that make it even more attractive. In addition to checks to verify the structure of the domain data sets, future versions are purported to include functionality for custom domains, generation of XML code for CRTDDS (define.xml), and generation of XML code for ODM. Once this functionality is available, and given that the procedure is included as part of licensing SAS, it makes sense to employ it at least to some degree in the implementation process.

CONCLUSION

The steps for successful implementation of the CDISC standards are straightforward, even if their execution may not be. The first and most critical is to define the objectives for the organization. This definition should include more than just a general statement on the proposed approach for implementation, but rather a specific set of guidelines that govern the direction the implementation will take. Included should be the type of implementation required, whether it be a strict interpretation or some

variant of the standards, where and how it will be applied, and by whom. Training and technology considerations should be included as well.

Experience with the standards in a CRO setting has resulted in the following set of high-level objectives:

- Initial training by external vendor experienced in development of CDISC structures
- Standards applied as part of the manipulation of the data extracted from the DBMS (the “back-end”), with a task force to investigate the development of CDISC compliant CRFs
- Strict interpretation of the standards by staff experienced in data manipulation (database programmers for SDTM, SAS programmers for ADaM) and who have submission experience (statisticians, data managers)
- A linear approach to implementation of SDTM and ADaM models
- Internal development of tools using SAS, with evaluation of vendor products ongoing

CONTACT INFORMATION

Your questions and comments are valued and encouraged. Contact the authors at:

Robert T. Stemplinger
ICON Clinical Research
555 Twin Dolphin Drive, Suite 400
Redwood City, CA 94065
Phone: (650) 620-2165
Fax: (650) 591-0611
Email: stemplingerr@iconus.com

Jackie Lane
ICON Clinical Research
555 Twin Dolphin Drive, Suite 400
Redwood City, CA 94065
Phone: (650) 620-2139
Fax: (650) 591-0611
Email: lanej@iconus.com

SAS and all other SAS Institute Inc. products or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

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