Paper 172-2009

Backpacking Your Way Through CDISC: A Budget-Minded Guide to Basic Concepts and Implementation

Rachel Brown, Westat, Houston, TX
Jennifer Fulton, Westat, Houston, TX

Abstract

By now, most people who work in the health sciences industry have at least heard the CDISC (Clinical Data Interchange Standards Consortium) acronym, but there is still a widely varying level of understanding of its concepts and with compliance among many in the field. Numerous helpful resources are available for CDISC training and implementation. However, many of these options can be costly for companies who want to work towards compliance but are not ready to make a large up-front investment. This paper offers some tips for how to approach the journey toward CDISC compliance without having to break the bank. Primary among these is that if you are a current SAS® user you may already have many of the supplies you need.

To prepare readers for the trip, the authors will introduce the components of CDISC standards while mapping out a plan for implementing the models using SAS and other tools already available within most clinical organizations. Then for those who are ready to pack, the paper will provide some resources and ideas for CDISC training that won't leave your pockets empty. We will also outline the benefits of CDISC compliance so the reader can decide if the journey is worth it. With SAS and a few other bare necessities in your backpack, your company can take the journey toward CDISC compliance without being left penniless.

Destination: CDISC

CDISC stands for Clinical Data Interchange Standards Consortium. It is an organization created to establish worldwide industry standards to support the electronic acquisition, exchange, submission and archiving of clinical trials data.[1] The CDISC strategy 2008+5 document[2] states, "The mission of CDISC is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare." CDISC recommendations are made up of a number of models and standards which will be explained in more detail in the next section of this paper.

CDISC is becoming the well-regarded standard for regulatory submissions. The Food and Drug Administration (FDA) issued a 2005 Guidance, revised in 2006, requiring the use of CDISC's "Study Data Tabulation Model" and strongly recommending CDISC's model for analysis data sets for eCTD (Electronic Common Technical Document) submission[3]. Also, the US Department of Health and Human Services is currently considering a proposed rule with regard to electronic submissions to the FDA that would "…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) to allow for more efficient and comprehensive data review."[3] Along with its potential to become a mandate by the FDA, companies who have implemented CDISC standards have seen significant reductions in time and cost in areas like development and review of data collection forms, database builds and testing, and working with Contract Research Organizations (CROs). And as is usually the case when standards are put in place in any capacity, the initial investment on a pilot project is expected to pay off significantly, as both time and effort are saved on subsequent projects. In other words, it's worth it to learn more!

Mapping Out Your Trip from Finish to Start

The indication from our research and training is that there is no one right way to implement CDISC. The approach that seemed to be the best fit for our typical projects was to implement CDISC from "finish to start", using SAS as our primary tool. While a clinical trial, for example, begins with protocol development and ends with submission of data to the FDA, our "finish to start" approach would facilitate starting with CDISC components that have been most extensively defined and piloted to date. We have also found that since the steps or models build on each other, it can aid understanding to describe them starting with the end product. This way, you will know where you are going before you get there.

Below we will map out our proposed implementation method, and we will provide a brief explanation of each model or step in the process. Starting at the end and working towards the start, the various "destinations" (CDISC components) are:

Destination #7: Metadata

The final product for CDISC-compliant projects is a metadata file. In CDISC terms, this metadata is a Data Definition Document named define.xml. This document provides "data about the data". It explains the various components of the study data in such a way that anyone could understand it using the explanations in define.xml.

The first part of the document provides the location of each dataset, along with its structure, purpose, and key variables. Then each dataset is broken down to provide information on each variable, including the label, type, format, origin, and role. CRF variables are linked to the actual annotated data collection form on which they appear, and derived variables are explained through comments.

The Data Definition Document is created in XML format. XML is a platform-independent language whose primary purpose is to facilitate the sharing of structured data across different information systems. Putting this document into XML allows others to view the document on their own systems without problems and provides a standard structure for archival and storage of study materials.

Destination #6: SDTM

One step before your metadata is a model called SDTM, which stands for Study Data Tabulation Model. This was the first model established by CDISC and is the model on which all other models are based. In simple terms, SDTM data is study data that has been formatted according to very specific standards. These datasets were created for submission to FDA, and reviewers would always know what your variables mean because they follow the same structure as in any other data submission.

Some of the specifics required by SDTM are that every domain (or dataset) be named according to the model, and that every variable be named, labeled, structured, and formatted according to the model.

An extensive set of domains has been developed by the CDISC SDTM group, and more continue to be in progress. The goal is to eventually have domains set up to meet the needs of all studies. SDTM includes all CRF data as well as a few derived variables, such as population indicators.

Currently, SDTM data is submitted as SAS transport files along with the XML define file. However, it is expected that these datasets will eventually be submitted in XML as well.

Not all projects that include data preparation will be part of submission documents, so the two destinations just described may not always be incorporated. In some cases SDTM datasets and/or metadata would only be created if specifically needed.

Destination #5: ADaM

The next model is ADaM, the Analysis Data Model. This model provides standards for creating analysis datasets in SAS. The model contains variables from SDTM in addition to each and every derived variable that would be used in an analysis. The specifications for these datasets are not nearly as strict as those for SDTM, which is good since this allows you to create the variables you need that are unique to each study. The analysis data is often referred to as being one proc away from analysis, meaning that you can pull this data into SAS, and only perform one SAS procedure to get the results you want. There is some debate as to whether it is more efficient to create ADaM data before or after SDTM data. We prefer to make ADaM data first because, as mentioned above, sometimes we would stop there.

One way to add efficiency to this part of the process is to change your standard database setup to be based on SDTM standards. The more that is done up front (in database development), the less there is to be done later when SDTM and ADaM data are being created in SAS. And if raw data is standardized to SDTM specifications, anyone receiving data from your database can expect to see the same general data elements and structure for any study.

Destination #4: Lab Model

Since one of the largest components of clinical trial data is laboratory data, making the development of a model to support it is an important CDISC initiative. The Lab Group was created to develop a standard model for the acquisition and interchange of laboratory data between CROs, external vendors, and sponsors. In the absence of acceptable industry standards, each company has developed their own standards, specifically designed for their particular needs, which usually tend to be developed on a per-study basis. The Lab model defines a standard for the structure and content of lab data.

Destination #3: CDASH

Following a data standard would be difficult if the data collection instruments across studies didn't follow a standard

as well. CDASH is the Clinical Data Acquisition Standards Harmonization group. This group is developing standard data collection instruments for each of the SDTM domains. This process streamlines data collection to improve data interoperability throughout the biomedical research and product development process.

Destination #2: Terminology

We would be lost when developing standard data collection instruments without a standard set of terminology to put on these forms. A terminology group was developed under CDISC to define and support the terminology needs of CDISC models. The group's purpose is to publish standard codelists and vocabulary for variables in CDISC models.

For example, the relatedness of an adverse event to the study drug may be reported as either "No", "Unlikely", "Possible", or "Probable" in one study; "Not Related", "Possibly Related", or "Probably Related" in another; and simply "No" or "Yes" in yet another. The terminology group has the task of selecting one codelist for use in all studies.

The group creates terminology to suit international needs so that global organizations and projects can use the same terminology recommendations.

One model which falls under the ordinance of the Terminology group is the BRIDG model. BRIDG stands for Biomedical Research Integrated Domain Group. It was developed by CDISC in collaboration with the National Cancer Institute, HL7 (Health Level 7), caBIG (Cancer Biomedical Informatics Grid), and pharmaceutical and tech companies. Its purpose is to link the patient care world to the clinical research world in the form of a formal model of shared semantics of regulated clinical trials research. STDM is the first CDISC model being staged for harmonization under BRIDG.

Destination #1: PRG

The Protocol Representation Group has taken the charge of creating a model for developing a clinical trial protocol. The group developed a protocol elements spreadsheet that provides the arms, events, and visits for a protocol. This file is created in XML so that the information within can be used in later steps of the protocol. The group continues to work on a protocol model and has plans to develop a Stat Analysis Plan someday, and they are also working on UML (Unified Modeling Language) diagrams as part of the BRIDG Model.

The Plan

The initial implementation of all CDISC components can be an immense and daunting undertaking. Our suggestion for a budget-savvy approach is to set the short-term goal of implementing the most established, final deliverables (SDTM and ADaM datasets). We can then work toward a long-term goal of including CDISC in earlier steps, which will eventually improve efficiencies throughout the project.

Currently, we have developed, and continue to improve upon, processes for creating SDTM and ADaM datasets in SAS. And SDTM standards have been incorporated during database development for newer projects. This approach has allowed us to provide CDISC deliverables to clients without a large impact on the whole organization. Work is still needed to fully implement CDISC within the entire organization to take advantage of the efficiencies it can offer, but as you can see, this is a journey for which you can't pack lightly.

ODM

Now that we've checked out our 7 destinations, it's time to put them together to see the entire picture. This is done by use of the ODM, or Operational Data Model. ODM is a vendor-neutral, platform-independent format for interchange and archive of data collected in clinical trials. The model represents study metadata, data and administrative data associated with a clinical trial. The role of the ODM is to facilitate the movement of clinical data collected from multiple acquisition sources to an operational database, but it also has application in the subsequent exchange and archiving of such data. ODM essentially maps all of these components together and automates workflow from task to task.

The following is an example of an operational flow that could result once all components have been incorporated. Under the ODM, the starting point would occur during protocol development.

1. A table of elements and visits would be created in XML to serve two roles. First, it would become the schedule of assessments table in the protocol. Second, the XML data can be imported into an ODM-compliant EDC (Electronic Data Collection) system so that the basic study setup (visits, assessments, rules, etc.) is automatically generated in the system.
2. The ODM-compliant EDC system sets up annotated electronic CRFs (Case Report Forms) and structures data so that it follows SDTM variable-naming and format conventions and includes automatic edit checks, including across-field edit checks. Only more complicated or study-specific edit checks and rules need to be programmed.
3. Data is collected in the EDC system.
4. Data is extracted from the database and brought into SAS in a format very close to SDTM.

5. ADaM datasets are created in SAS by adding derived variables. These datasets are one proc away from most required analyses planned for the project.
6. Mapping tools in SAS map this data to fully-compliant SDTM data.
7. XML Metadata is created.
8. The study is archived using XML files which can include data, metadata, and audit trails (tracking who changed the data along with when and why).
9. Files are archived and delivered to our clients and government agencies, as needed.

This ideal CDISC implementation fully takes advantage of time-savings by implementing CDISC from the very beginning of a study, meaning starting with protocol development, and continuing through form setup, data capture, mapping, analysis, submission, and review.

The biggest advantage of this full implementation is that savings begin at step one and continue through each step – in protocol development, CRF development, database setup, dealing with external vendors, programming of submission data, and archival. This streamlining of information prevents problems with various study components getting out of sync. For example, if a visit is added to the study, this adjustment is made in step one, and the change is automatically carried through all other automatic steps (Schedule of assessments, eCRFs and EDC system are all updated with 1 change). This eliminates the "islands of execution" in a study (where you write the protocol, then rewrite information in CRFs, then rewrite information in the data collection system, etc.), and following CDISC standards allows for better reusability of information across studies.

Before You Travel, Get to Know the Landscape

So now that you know where you want to go, what steps should you take when you are ready to hit the road? The first part of that "initial investment" we mentioned earlier will be made in learning more about how CDISC works and how to implement the standards within your time and budget constraints. Becoming a CDISC member offers some benefits including access to the members-only area of the CDISC website and discounts on CDISC training and conferences.[4] Corporate sponsors and members are listed on the CDISC website as well. But CDISC membership can be quite expensive. Before you sign up for every conference, class, and webinar that is related to CDISC, consider a few of these lower-cost options:

1. One excellent (and free) source of a wealth of information is CDISC.org, the official website of CDISC. Here you will find the most current versions of the model standards and implementation guidelines, and descriptions of current and future pilot studies. You can sign up for CDISC's free mailing list which will inform you of upcoming conferences and trainings and other news stories related to CDISC and the industry. There are areas of the website that cannot be viewed by non-CDISC members, but we have found the free information to be helpful and informative. A notice in the February, 2009 CDISC newsletter indicated that implementation guides and standards documentation will be moving into the members-only area of the website some time in 2009, so you may want to download and save the most current version of these now while access is still free.

2. CDISC has also announced the impending publication of what they call "the only standards primer developed by CDISC". It is a book titled Introducing the CDISC Standards, New Efficiencies for Medical Research.[5] It can be purchased for about $30 on CDISC.org and is anticipated to be released for printing in February, 2009. It will include a thorough explanation of the CDISC models and components, and a section on how to implement CDISC for your own data.

3. SAS has pledged support of CDISC standards and is a registered solution provider[6], which means CDISC has recognized SAS as a qualified consultant with sufficient knowledge and experience implementing the various CDISC standards. The SAS website can also direct you to SAS-initiated webinars, whitepapers, and articles. If you are a current SAS user, you should explore what SAS 9.2 has to offer before investing in the numerous other vendor options that are popping up to answer the needs of companies looking to become CDISC compliant. SAS has a number of products to consider for data mapping including:
    - XML (Extensible Markup Language) Engine allows for loading of XML files into SAS and export to XML files from SAS.
    - Proc CDISC
        - validates against the CDISC SDTM model,
        - verifies all required SDTM variables are present,
        - identifies dataset variables not defined in a domain,
        - identifies missing, but expected domain variables,
        - identifies missing, but permitted domain variables,
        - verifies variables are of expected type and length,
        - identifies missing, but expected, controlled terminology (formats),
        - verifies that required variables do not contain missing values,

        - identifies expected variables that contain missing values,
        - verifies conformance of all ISO-8601 specification assigned values,
        - notes correctness of yes/no and yes/no/null responses.
    - SAS Data Integration Tool – does not know CDISC but helps with mapping
    - SAS Clinical Data Integration Tool – knows CDISC and SDTM
        - includes the SDTM Model and allows user to drag and drop 1:1 mappings. Code is needed for more complicated mapping.
        - creates a model discrepancy report

4. Conferences and webinars are another source of information. Both vary in convenience and price, of course, so here are some details that may help you make the choices that best fit your budget and needs:

   We attended a CDISC International Interchange and found it to be well worth the trip. We were able to attend workshops and hear from others in the industry who shared their experiences and tips as they had worked through the implementation process. We also heard from FDA representatives on important topics such as how FDA validates SDTM data and FDA's expectations for CDISC in the future. And there were numerous presentations and booths displaying the plethora of external vendor options.

   Thanks to SAS's commitment to CDISC, SAS user group meetings often offer a number of papers, presentations, and training sessions with valuable training offered by not only industry experts but members of the CDISC standards teams. Two such conferences that have been of particular help to the authors in terms of CDISC information are PharmaSUG and SAS Global Forum.

   Other industry conferences, including DIA (Drug Information Association) and SCDM (Society for Clinical Data Management) are also offering specific types of CDISC training and presentations at meetings.

   Of course webinars offer the convenience and cost-savings of training without ever leaving home. As mentioned above, if you sign up on the CDISC website for the mailing list, you will receive frequent announcements of the latest webinars being offered. CDISC has also recently begun to offer its own online training courses, or "v-Learning", which can be ordered through CDISC.org. The cost for these v-Learning courses varies by course and by membership level, and while they are not cheap, they may be a good lower-cost alternative to traveling for training. BetterManagement.com has also offered some valuable free training webinars on CDISC topics that we have attended in the past year. You'll want to keep an eye out for similar webinars being presented by different sources for different costs. We had one case last year where two nearly identical talks were offered and one was free and one was around $400.

Is It Worth It To Take This Journey?

In many cases, implementing CDISC standards represents a costly and somewhat overwhelming investment of time in the early stages. But given CDISC's links to the FDA and its widespread infusion into the fabric of the clinical trials industry already, most companies agree they cannot afford NOT to begin the process.

A proposed rule by the US Department of Health and Human Services regarding electronic submissions to the FDA requires the use of SDTM, and this rule is expected to become a regulation. Since this department governs other groups besides FDA, it seems reasonable that the regulation could eventually be adopted elsewhere too.

Several companies who have implemented CDISC standards reported at conferences we attended that they have seen significant reductions in time and cost in areas like CRF development and review, database builds and testing, and working with CROs.

Some of the benefits of CDISC standards (in part as listed on the CDISC website) include:

1. Compliance with FDA guidance/future regulation and established standards
2. For CROs: Getting business and staying in commercial business
3. Reduced time for regulatory submissions; more efficient regulatory reviews of submission
4. Savings in time and money on data transfers among business partners[7]
5. More efficient archive and recovery procedures[7]
6. More accessible information[7]
7. Better communications amongst team members[7]
8. Beneficial to all types of health research studies (even studies not under FDA regulation) since it creates a company standard for all projects[7]

Are we there yet?

How do you know if you have "arrived", meaning your company or application can advertise itself as "CDISC compliant"? CDISC has already implemented a certification program for applications utilizing the ODM, and according to their website they are working toward a more comprehensive program that would certify that an application is compliant with one or more specific CDISC standards. As for measurable goals in the way your company collects, stores, and analyzes clinical trials data, currently only SDTM has been mentioned by the FDA as a possible mandate in the near future, and only SDTM has not only a set of defined standards but also published guidelines and the WebSDM software for validating data against the CDISC standards. Guidelines for SDTM metadata are also in development and the draft version is available on the CDISC website. As of the writing of this paper, implementation guidelines are in development for ADaM as well. In the meantime, CDISC standards continue to be a work in progress. Each company will need to develop their own processes and techniques, and measure their data against the most current published CDISC standards documents. The reaction from vendors, clients, or FDA reviewers will also provide valuable indicators as to whether you are on the right track.

One interesting option CDISC offers to companies who feel they have attained a certain level of understanding about one or more components of CDISC is to apply on the CDISC website to become a "Registered Solutions Provider". CDISC says they are looking for "qualified consultants, system integrators, and subject matter experts believed by CDISC to have sufficient knowledge and experience implementing the various CDISC standards" to assist organizations who are working toward implementing CDISC standards. This offers a potential marketing opportunity, including the right to display the CDISC Registered Solutions Provider symbol on your company website. SAS is one such provider.

CDISC repeatedly advertises that it needs input from throughout the industry in order to reach its goals. Over time the hope is that the standards will become widely recognized and accepted as the norm with the long-term goal of an FDA warehouse repository of data from all clinical trials that will greatly increase the efficiency of review and approval of drugs and devices.

Ultimately whatever approach to the implementation of CDISC you choose, it will be a long journey, requiring some up-front strategizing to map out the best route for your needs and means. At least reading this paper didn't cost you anything so hopefully now you feel like you can travel the road to CDISC compliance without having to sleep under the stars or bathe in a lake along the way!

References

1. "Mission Statement" http://cdisc.org/about/index.html
2. "Strategy 2008 (+5y) Document". http://cdisc.org/about/about_strategy.html (Oct. 2008)
3. "FDA Information/Documents Relative to CDISC" http://www.cdisc.org/publications/fda.html
4. "Become a Member or Sponsor of CDISC" http://cdisc.org/membership/index.html
5. "Publications – CDISC Primer". http://cdisc.org/publications/cdisc_primer.html (Feb. 2009)
6. SAS Institute Inc. "SAS pioneers support for CDISC standards to access electronic health record (EHR) data". http://www.sas.com/news/preleases/061807/news2CDISCStandards.html (Oct. 2008)
7. "Frequently Asked Questions" http://cdisc.org/faq/index.html#question23

Acknowledgments

We would like to thank our supervisors who allowed and encouraged us to research and present our findings at SAS Global Forum.

Contact Information

Your comments and questions are valued and encouraged.

Rachel Brown, MS
Westat
5615 Kirby Drive, Suite 710
Houston, TX 77005
Work Phone: 713-353-7916
Fax: 713-529-4924
Email: RachelBrown@westat.com

Jennifer Fulton, MS
Westat
5615 Kirby Drive, Suite 710
Houston, TX 77005
Work Phone: 713-353-7925
Fax: 713-529-4924
Email: JenniferFulton@westat.com

Disclaimer

The contents of this paper are the work of the author and do not necessarily represent the opinions, recommendations, or practices of Westat.