Current Status and Future Perspectives for Systemization of Clinical Study related the issues of CDISC in USA and other

ABSTRACT
The term "the CDISC standard" has been used incorrectly for a few years. The more accurate term would be "the CDISC standards" as there are a number of standards which the Clinical Data Interchange Standards Consortium (CDISC) has developed since its start in the late 1990's. Some of these standards have been adopted by the FDA as their current "specification" for the submission of clinical study tabulation data. These data are required in support of marketing applications submitted to the FDA. Some of these standards have been adopted by the bio-pharmaceutical industry (drug sponsors, partners, contracted services, etc.) as a direct result of the FDA publishing their specifications. Other CDISC standards involve the operational use of data not directly related to FDA submissions. In addition to this regulatory demand, industry is beginning to see the benefits of standardizing data content and format. This paper will describe how the industry got to this place, where it currently is and where it appears to be going in relationship to the CDISC standards.

INTRODUCTION

"The CDISC standard" has been used by many people to generally refer to the collection of individual CDISC standards which have been developed by the past few years by the members of working teams of the Clinical Data Interchange Standards Consortium (CDISC) or to refer to the most visible standard, the Study Data Tabulation Model (SDTM) These teams of individuals brought together members of the US FDA, pharmaceutical and biotechnology company employees, technology company employees, Clinical Research Organization (CRO) employees and other industry consultants and experts. The resulting standards represent a variety of uses of data collected, shared, stored, analyzed and reported from clinical trials in addition to non-clinical (animal) trials. CDISC has, as its one of its goals, to unify these individual standards and promote their use at all of stages of the clinical trial data where data management exists and afterward, in the analysis and reporting of these data.

As these standards have published, they were adopted by industry at varying speeds depending on many factors. Some of these factors included companies' comfort with existing internal systems and company standards which had taken years to develop. In addition to the reluctance to change internally, the standards continued to evolve and were enhanced with new versions. There was concern by early adopters that the newer versions would not be "backwards compatible" with older versions. These and other factors which will be explored in this paper caused the adoption of the standards to move forward in a less-than constant pace.

Without a motivating external need to adopt the CDISC standards their adoption might still be in its infancy. It took active participation by the FDA, in addition to the influence of internal FDA initiatives, to help motivate to industry in adopting these standards. This paper will look at some of these external factors and FDA initiatives and relate these events to how they influenced the standards, as well as influencing those adopting the standards.

Finally, various methods for adopting these standards will be mentioned throughout the paper. There will be discussion of the advantages and disadvantages to each approach.

FDA FIRST ELECTRONIC DATA SPECIFICATIONS AND GUIDANCE: A STRONG INDUSTRY MOTIVATOR

For many years, individual companies within the bio-pharmaceutical industry have had few incentives to standardize with other companies. Many decision makers within these companies believed that their methods for collecting, working with and reporting data to the FDA were as good as, or better than, any other company's methods (although some reviewers at the FDA may have disagreed). .

Throughout the 1990's the FDA received data via a variety of proprietary computer hardware and software systems. These were delivered to the FDA reviewers in support of the official paper marketing applications, particularly for New Drug Applications (NDAs). These proprietary systems created two important problems for the FDA.

First, each of these systems were different, often times even if they were provided by the same drug sponsor. This meant that the training of the reviewers and the support of theses systems were left in the hands of the drug sponsors.

Second, it meant that, if these systems and the data contained within them, were used in place of the official paper copy of the tabulation data, the NDA review results might have been different than if the paper copy were used. While there is no evidence that these systems contained different data than what was submitted in paper, it remained a risk that the two MIGHT be different. SInce the official paper copy of the study tabulation data was often not the copy which was being used to come to a conclusion about the meaning of the data the FDA was very concerned about this risk.

In 1997, the FDA published a new regulation, the *Electronic Records; Electronic Signatures* rule. This rule became very important as it allowed the FDA to officially accept documents and data in an electronic format without needing to have an accompanying paper copy as the "official copy."

Within a year of that rule's release, the FDA Center for Drug Evaluation and Research (CDER) the FDA Center for Biologics Evaluation and Review (CBER) each published draft guidance documents which shared with industry the FDA's ideas about how electronic documents and data were going to be accepted and archived. These draft documents described the format and structure of the files and folders within an electronic-only NDA or BLA (Biologics Licensing Application). These documents suggested Adobe PDF files and SAS XPT or "transport files" as the recommended file types for submitted tabulation data and the documentation which supported these datasets. These file types were recommended because they were non-proprietary or "open" standards which fulfilled the FDA's mandate to support these types of standards over proprietary standards which might financially benefit a single owner of the standard.

In 1999, the two FDA Centers published a shared guidance document *Providing Regulatory Submissions in Electronic Format — General Considerations* which described the types of files, PDF and XPT and left the specifics about the file organization and naming to other guidance to be provided by the individual Centers.

At the same time, CDER published its pivotal 1999 guidance *Providing Regulatory Submissions in Electronic Format — NDAs* which described the structure of the files in an electronic NDA. Within this document they described the Case Report Tabulation (CRT) data section in some detail.

Included in these details of the CRT section were three types of files. First was a Case Report Form (CRF) file named BLANKCRF.PDF which was to be annotated to describe the data collection points for the raw data. Second, a PDF file called DEFINE.PDF which contained a list of the datasets being submitted and tables for each dataset, describing details about them. These details included such things as the variable names, description, type of data, codes and decodes used, and "comments" for other important information about the data. And finally, the CDER described the datasets, with some important considerations for industry to use when creating the datasets.

Within the appendix to the CDER guidance there were examples of 12 safety-related "domains" or datasets which contained similar types of data, such as the "demographics" domain or the "adverse events" domain.

By the end of 1999, CBER published its guidance for submitted electronic BLAs *Providing Regulatory Submissions to the Center for Biologics Evaluation and Research (CBER) in Electronic Format — Biologics Marketing Applications*, which was very similar to the CDER guidance.

After publishing these guidance documents, the FDA told industry that they would no longer accept proprietary systems in support of NDAs or BLAs and that industry MUST submit electronically following these guidelines.

INDUSTRY'S RESPONSE TO THE FIRST GUIDANCE "STANDARDS"

Industry was now forced to give up the proprietary electronic systems and moved forward into this new reality of PDF documentation and XPT datasets in FDA's loosely defined "domains."

At the same time as the FDA was creating these guidance documents, some in industry were also thinking about standards.

In 1997 a group of individuals in industry came together at a meeting to discuss standards for sharing data within industry. These discussions were concerned with operational standards. They started looking at Operational Management Group (OMG) and a Industry-standards Glossaries. By 1998 they formed a Drug Information Association (DIA) Special Interest Area Committees (SIAC). This was the pre-cursor to the CDISC organization as we know it today.

This group was originally organized into two working teams; one for Modeling and one for Nomenclature. After splitting from the DIA, these two teams soon became five teams. The Nomenclature group later became the Glossary (or Terminology) team. The Modeling group was later split into four groups: the Submission Data Standards (SDS) team, the Analysis Data Model (ADaM) team, the Operational Data Management (ODM) team and a Laboratory Data (LAB) team.

Early CDISC meetings discussed alternative approaches toward defining standards, and a glossary group was established to define critical terminology relevant to clinical research, but little definitive progress was made toward defining actual data standards.

The first version of the CDISC Submission Metadata Model was presented at the DIA annual in 1999. They suggested that, while the FDA's 1999 guidance provided much detail about the document portion of a submission, it did not provide sufficiently detailed instructions for how to organize this data component. It only provided a requirement that data be submitted in a standard technical format, the open SAS V5 transport file format. It did, however, establish a precedent for submitting a PDF file (DEFINE.PDF) that would describe the contents and structure of the clinical data; i.e. its metadata. This presentation suggested providing more detail on how to provide this metadata.

By April of 2000, David Christiansen and Wayne Kubick had published version 1.1 of the CDISC Submission Metadata Model. This model, and its revised version (v2), became the basis for, and fundamental approach to, "establishing meaningful standards applicable to data submitted for FDA review." The metadata document describes the metadata or "data about the data" to describe

the data in the XPT domain files and place each variable within the context of the whole.

The authors of the metadata model soon established a team consisting of volunteers from several pharmaceutical companies to develop domain models that would show how to apply the metadata model concepts to specific datasets, a team that was soon joined by representatives from the FDA. This became the first CDISC data modeling team, the Submission Data Standards (SDS) team.

This metadata model was later supplemented by a collection of PDF files representing spreadsheets of metadata describing each CRT safety data domain. It also was accompanied by a PDF guide to formatting the descriptive spreadsheets for presenting these metadata. These spreadsheets also had CDISC notes describing "best practices" or comments about how each variable should be used. There was also a "CDISC Core Variable" designation assigned for each variable to define whether or not each variable should always be reported.

The value of this version was that it provided examples of "standard" domains with examples of metadata. It also clearly defined which variables should be in specific domains and which were not required for all submissions.

As this version provided clear examples of organizing the data for submission, some companies used this as a basis for submitting their data for NDAs. Unfortunately, the CDISC organization was very small at this time and very few companies new about this standardization initiative.

OTHER CDISC TEAMS: ODM, ADaM and LABS

The initial success of the metadata model and the SDS team attracted the interest of others who were more interested in standards that would support the data collection process as well as submissions. This soon resulted in the creation of the Operational Data Modeling (ODM) Team.

The ODM team was created in 1999 when CDISC invited a group of vendors of Clinical Data Management (CDM) systems to a meeting to discuss the possibility of creating a new standard for interchanging clinical data collected during trials. Over the summer, two separate CDM vendors approached CDISC in the hope that CDISC would consider supporting their proprietary data models as an industry standard.

The two companies, Phase Forward and PHT, Inc. had both developed models that would ideally be used for moving data from any data collection system to a Clinical Trial Sponsor’s central database, and they looked to CDISC as the best hope for getting such a standard adopted by industry. Both models were based on use of the Extensible Markup Language (XML), a new technical standard for

representing data and documents in a structured matter that was rapidly developing support among technology vendors, especially for e-Commerce.

Since a chief operating principle of CDISC was to be vendor-neutral, the CDISC steering committee members instead invited both companies to join a new team that would also include other vendors, CROs and the companies that had developed the two largest CDM software packages: Clintrial and Oracle Clinical.

With work on two models proceeding simultaneously to address two separate needs within the agency (review and archiving), CDISC participants began to look at the world of clinical research as consisting of three distinct types of data, **Data Sources**, where the data was created, **Operational Data** where the data was collected, reviewed for consistency and managed to an acceptable standard of quality, and **Submission Data**, which is normally extracted from the operational data and sent to a regulatory agency.

Since the rules for submission data were set by the FDA, and these rules were felt as not being sufficient to meet the needs of data collection, CDISC believed it had to develop standards or models in two areas: Operational models to transfer data from the point or origination to a Sponsor's internal database, and Submissions models to transfer data from the sponsor to the FDA

In 2000, CDISC became formally established as a non-profit organization, and began to seriously ramp up its efforts to advance the movement to define industry-wide data standards.

Soon work was initiated on two new modeling teams: the LAB team to define Clinical Laboratory operational data interchange standards, and the Analysis Data Modeling team (ADaM) to define standard submission models for analyzing clinical data.

THE FDA MOVES FORWARD: THE PATIENT PROFILE VIEWER "CRADA"

In early 2001, the FDA had a public meeting to display an internally developed an electronic Investigational New Drug (IND) application viewer called CTOC (Cumulative Table of Contents) viewer. While this viewer demonstrated potential, the FDA decided that it should not be developing software.

About this time, the FDA started to more regularly use a process called a "CRADA" or Cooperative Research and Development Agreement." The CRADA is an agreement to work with software developers and technology providers to create useful software which the FDA might use. The advantage for the software developer is that they would retain the license, copyrights and own the intellectual property rights to the software. The developer could then sell copies of the software to interested parties such as bio-pharmaceutical sponsors or other partners.

One of these CRADA developments was the Patient Profile Viewer (PPV). In December 2001, the FDA published a notice that they were looking for a CRADA partner to create a PPV. This software was to be developed to generate and view patient profiles directly from CRT datasets. The FDA selected PPD Informatics to develop a module for its commercially available software "CrossGraphs".

This module was designed to open a collection of domain-structured datasets and convert the data (organized in a tabular format) a "patient profile" view. The patient profile views are defined by the FDA as "displays of study data of various modalities (e.g. from multiple domains) collected for an individual subject and organized by time." This organization provides a clear presentation of relationships between various events which are occurring in different domains at the same time or sequentially. An example of this might be if the patient was administered study drug (described in the "EXPOSURE" domain) and short time later displayed and adverse event (described in the "AE" domain).

In order for this tool to work successfully, the FDA stated that the use of standardized datasets and metadata would be needed as input to the tool. Standardized dataset and metadata would reduce the amount of preparation required by the reviewer to generate the patient profile and would eliminate the need for applicants to submit patient profiles in PDF. Patient profiles in PDF, while not always needed, could be requested by some FDA review divisions.

CDISC RESPONSE TO THE PPV: SDTM VERSION 2.0

The CDISC SDS team was aware of the FDA CRADA for the Patient Profile Viewer because there were FDA representatives on the SDS team. In order for the PPV to work correctly the data needed to be structured in a consistent manner.

By November 2001, Version 2.0 of the Submission Metadata Model was published to enhance the earlier version. In December 2001 an accompanying CDISC *Submission Data Domain Model v2.0* (SDDM) was published. This document put all of the example domain spreadsheets into one PDF document. It also added a list of assumptions and provided more information about which data was to be expected as well as other clarifications and enhancements.

The most significant difference in version 2.0 was that it introduced the option for sponsors to submit "vertical" or "more-normalized" datasets for the domains of "ECG" and "Vital Signs". The normalization of these domains would provide the FDA the ability to "pilot new database and data-viewing technologies." They would also provide greater flexibility in terms of data storage, retrieval and merging with other data for review.

The vertical models' development also led to some improvements in the horizontal or less-normalized versions of the ECG and Vital Signs domains. This version also introduced standardized LOINC (Logical Observations, Identifiers, Names and Codes) codes for LAB, ECG and Vital sign measurements.
At about this time, during the peak of the late 1990-2000 market, there were a number of larger pharmaceutical companies which started partnering with smaller start-up companies or bio-technology companies to develop drugs and bring them to market. Another trend was that smaller companies became the targets (or initiators) of takeovers and mergers. These partnerships and mergers demonstrated that industry standards could be beneficial, but few companies had yet to implement the CDISC standards.

With version 2.0 of the SDDM industry started looking at it more favorably and many CDISC sponsor companies started putting the standard in place within their submission preparation processes. Some companies even started implementing it with their Global Data Integration Databases (often in SAS). These database are also referred to as Submission DBs, Analysis DBs, Integration DBs, Global Integration and Analysis Databases (GIADB), or just the "Data Warehouse."

In 2002, after the FDA entered into the CRADA agreement to develop its Patient Profile Viewer the FDA realized that it needed to have sponsors submit the data in a standardized structure which would be compatible with the viewer. The CRADA project published a set of PPP (Patient Profile Pilot) specifications for submitting data. The FDA invited members of the CDISC SDS team to submit data following the PPP specification to test the Patient Profile Viewer. It also would determine the compatibility between the SDDM v2.0 standard and the PPP specification.

During the summer of 2002, the CDISC SDS team was reviewing comments from the release of v2.0 for a release of v2.1 of the SDDM and started considering what needed to be accomplished with the next version. The SDS team was looking at new domains, extending the use of codes and increasing the existing domains' compatibility with the ODM message format standard for that had been published. They were also looking at possibly modifying the SDDM standard to make it more compatible with the PPP specifications. It was decided that the CDISC SDDM standards should incorporate the Patient Profile Pilot feedback in its next version.

The team also was looking at a broader CDISC initiative to start publishing their standards through the Health-Level 7 (HL7) organization as it was a member of the American National Standards Institute (ANSI), the US national standards setting organization. HL7 was developed to standardize data within the area of health care and medical provider reimbursements.

During an August face-to-face meeting of the SDS team the FDA representatives to the SDS team brought to the team the FDA plan to publish the Patient Profile Viewer Specifications as a data standard. The FDA proposed publishing the specification through the Health-Level 7 (HL7) organization, as it was a member of the American National Standards Institute (ANSI), the US national standards setting organization. HL7 was developed to standardize data within the area of health care and medical provider reimbursements. There was some justifiable concern that this could produce parallel or competing standards for the same objective.

There was also another set of needs of the FDA expressed to the SDS team during this meeting. The FDA was planning on a data warehouse to store all of the standardized data that was submitted. The hope was that this warehouse would allow the FDA to create software which could "mine" for data, particularly data which would indicate safety concerns. The expected advantage to the FDA data warehouse would be that it would allow the FDA to pool data for similar drugs or classes of drugs and then analyze this larger pool of data.

It was agreed at the meeting that the most pressing concern was correcting issues with v2.0 and publishing v2.1 by the end of October. Communications after the meeting made it clear that the FDA wanted to have a CRT standard which would be published by HL7 and made as a "normative" standard. This would allow the FDA to refer to "the standard" and not have to publish the PPV specifications as a standard.

The CDISC SDS team realized that the FDA was going to pursue having an HL7 standard so, after publishing the corrections in v2.1 (as draft and not for implementation) they embarked on then next version v3.0. The team realized, however, that many in industry were adopting v2.0 and v2.1 and there would be reluctance from industry to adopting a version which was expected to be quite different.

The LAB team, at this point published their version 1.0.0 of the LAB standards which described what data should be transmitted from central labs. It also provided an XML message standard for organizing these LAB messages.

The FDA was also pursuing a CRADA partner to create software for safety analysis, a data warehouse structure and for a data validation tool.

By March of 2003, the SDS team had created a new document *Version 3.0 Submission Data Standards - Review Version 1.0* which contained an introduction, the General Study Data Information Model and the CDISC Submission Domain Model. There were many changes from v2 to v3, but the goal was to provide compatibility for those in industry who had adopted SDDM v2.0. This was the review version for the HL7 organization to read and comment

upon prior to a final version which would be balloted and deemed the official normative standard.

After the HL7 members provided comments and changes were made to address important comments, the SDS team arrived at the wording for th CDISC *Submission Data Domain Models Version 3.0 - Final Version 1.2* which was balloted and approved by HL7 as the normative standard.

One of the exceptions to the compatibility was that horizontal (or non-normalized) domains in the SDDM v2 would NOT be able to be created in v3.0. This was because one of the most important goals of the standard was to provide a standard which would allow the FDA to use standard tools to convert the data from its CRT (vertical or normalized) presentation to a "listing" or horizontal presentation. While tools could be created that would convert the data from non-normalized to normalized, it would be more difficult and complicated. It was decided that it would be better to standardize on submitting normalized data.

Another advantage to normalized data was that it would be easier to validate for compliance to the standard and to import these data into a data warehouse.

Another aspect of the v3 standard was that it provided guidance for creating data in domains other than the safety domains originally described in v1 and v2. The standard opened the domains to all types of data. In order to provide some organization, the domains were classified as "findings," "interventions," "events," or "special purpose." The "special purpose" domains were clearly defined within the standard so new domains would have to be placed into one of the three remaining classifications.

In order to provide clear documentation for these domains, it was decided that a new standard should be created. In April 2003, a focus group was formed from members of the ODM team, the ADaM team and the SDS team to create a "DEFINE.XML" specification. A white paper was written to describe the requirements for this specification and its advantages. This specification was designed to use the ODM XML structure and formatting and apply the dataset, domain and variable information which would have traditionally been submitted in the DEFINE.PDF documentation file. This would provide an advantage to the FDA for loading data into their warehouse, as this would be a machine-readable file with standardized structure and formatting. The Case Report Tabulation Data Description Specification (CRT-DDS) was published and submitted to HL7. This standard went through many HL7 ballot cycles composed of submitting the specification, receiving HL7 member comments and re-submitting revisions. (The CRT-DDA v1.0.0 finally became official in February 2005.)

A second pilot was organized to test the v3.0 standard. This pilot would take place later in 2003. The results would be presented at an FDA public meeting in early October 2003. Eight companies participated in this pilot providing data in

version 2 format. One company mapped the legacy data into a V3.0 vertical submission. It was concluded that v3.0 could function to import data into a data warehouse. It was also concluded that v3.0 needed to be enhanced to provide greater clarity for those creating submission domains.

It was also agreed that this standard was not ready to be proposed to industry as one which they should implement into their submission preparation process. Even though this was the recommendation, a few companies did try to implement v3.0. Most companies, however, kept producing submissions compliant to v2.0 or v2.1 (even though v2.1 was only published as a draft version).

By June 2004, after reviewing the comments from the FDA pilot, pilot participant and the CDISC community, the SDS team had created a new version composed of two documents: the *CDISC Submission Data Tabulation Model version 1.0* (SDTM) and the *SDTM Implementation Guide V3.1* (I.G.).

The lessons learned from the pilot were published as a section of the appendix of the SDTM IG v3.1. It states "the number one learning from this pilot was that additional guidance and specifications are needed in order to reduce inconsistencies and increase comprehension of the models. Specifically, a detailed implementation guide is necessary to more clearly communicate the specifications, the rules, as well as to provide additional guidance through examples. Also, the team learned that the vertical nature of the datasets highlights the importance of and the need for specific controlled terminology and to be able to provide record level metadata (e.g., via define.xml)."

To provide more clarity the SDTM IG v3.1 included many more examples. This version also introduces domains for "Trial Design." These domains describe the arms of a trial (by defining the components of the arms and how they relate to the whole). They also describe how a subject is expected to be studied (such as which arm they are following in a cross-over study). These domains, and other subject-data domains may be compared to see that the subjects went through the trial as expected.

From August to December 2004, the ADaM team published 5 drafts and the final *Statistical Analysis Dataset Model: General Considerations Version 1.0* to describe the general structure, metadata and content typically found in Analysis Datasets. This guidance was built on the nomenclature of the SDTM v3.1, conformed to the *CDISC Submission Metadata Model* and referenced the "Define.XML" (CRT-DDA v1.0) as a mechanism for submitting analysis metadata in a machine-readable format.

THE FDA RECOGNIZES SDTM V3.1 IN ITS ECTD SPECIFICATIONS

The eCTD specification moved to step 5 (implementation) in November 2003. The FDA posted its interpretation of the eCTD guidance to its web site March 14, 2004.

In July 2004, the FDA published a *Study Data Specifications v1.0* which was a supplemental specification to its eCTD guidance for implementing the eCTD. This version of the Study Data Specifications referenced the CDISC SDTM v3.1 as the standard that should be followed when submitting Data Tabulation datasets to the FDA with in eCTD structured submission.

This reference added significant visibility to the CDISC organization as well as to the SDTM standard. By this one reference, the FDA told all of those working with study tabulation data which was going to be included in an eCTD format that the FDA felt CDISC was important. This reference also placed greater emphasis to industry that the SDTM v3.1 standard was to be adopted for future submissions.

Even though this eCTD data specification cited the FDA's preference for CDISC SDTM v3.1, in 2004 many companies were not yet ready to submit using the eCTD format. In fact, the FDA reported that in Fiscal Year 2004, 12 marketing applications (NDAs and BLAs), 2 INDs and more than 100 supporting submissions were received by CDER and CBER in the eCTD format. This is comparable to the totals of 137 original marketing applications and 81 re-submitted marketing applications reported in 2004. It appeared that companies had become comfortable with submitting applications using the 1999 electronic submission guidance and did not have much incentive to move forward toward eCTD. The FDA did not yet mandate that the eCTD be used as the required format for submissions. Companies were given the choice when submitting electronically to choose between the 1999 guidance format and the eCTD format.

Another development at the FDA was an initiative to require that clinical data that was to be submitted to the FDA be submitted in a standardized electronic format. In September 2003, the FDA published a Notice of Proposed Rulemaking (NPRM) in the Federal Register. This notice was one of the first steps in changing the regulations to require that data be submitted in a standardized electronic format. This NPRM was re-published in December 2004 with a proposed action date of June 2005. In May 2005 the FDA published its Federal Register, Unified Agenda it re-published the NPRM with an extended date of October 2005. In October 2005, in the Federal Register - Department of Health and Human Services (HHS) - Regulatory Plan, the HHS (the government department where the FDA resides) published its priorities for the year 2006 and cited the Submission of Standardized Clinical Data as one of its top seven priorities. The detail of the Notice of Proposed Rulemaking cited a timeframe of two years for implementing the rule change.

This notice also cites reasons for requiring data as making the review of the data more efficient and less prone to error which might happen if paper-supplied data were transcribed by hand to an electronic system within the FDA. Besides more efficient processing and review of data, the ability to archive the data more efficiently was cited as a benefit to this standardization.

Also in March 2005 the FDA published revised eCTD *Study Data Specifications v1.1*. These revised specifications continued to reference the CDISC SDTM v3.1 for Clinical trial data standards but added other CDISC standards.

In the area of tabulation data, the new eCTD data specification referenced the CDISC Standard for Exchange of Nonclinical Data (SEND) which had been developed to conform to the Clinical SDTM v3.1 model but applied to specifics of animal toxicology data. This standard had been developed from the CDISC SEND team which had formed in 2002 and had developed this guidance in parallel to the SDTM developments. The SEND team had published its latest version 1.7.5 in December 2004 and an implementation guide in March 2005.

In the area of documentation (data definition file), the new eCTD data specification referenced the CDISC CRT-DDS (define.xml) as the preferred method of providing this metadata within an eCTD submission.

In the area of analysis data, this eCTD data specification did not yet reference CDISC as the only published guidance was the general consideration document which was seen as not specific enough to use as a comprehensive specifications document.

HOW IS INDUSTRY IS IMPLEMENTING THE STANDARDS

This proposed rule change to require the submission of clinical trial data electronically, in addition to public announcements that the eCTD standards would be replacing the 1999 guidance standard for submissions, motivated industry in 2005 to start taking seriously the need to adopt the CDISC SDTM standard. In September 2005, the number of eCTDs had risen in such a way that the totals from 2003-2005 had increased to 46 unique NDAs (totaling 588 submissions), 11 unique BLAs (totaling 233 submissions) and 43 INDs (totaling 234 submissions). (The final totals for fiscal year 2005 will be reported in June 2006.)

The increase in eCTD submissions is paralleled by an increasing the activity in implementing the CDISC standards. From 2003 through early 2005, many companies were planning conversion strategies which used their legacy internal standards and then applied a mapping to the CDISC SDTM data standard when preparing the data for submission.

Companies found that this strategy ran a risk that CDISC required or expected data may not have been considered when designing the study, designing the case report for or collecting the data. If this was the case, data might be missing which the standard designated as being required or expected, which might reflect poorly on the study.

Many companies are starting to look at implementing "CDISC-like" or "CDISC-friendly" variables within the Data Management Systems (DMS) which collect and verify CRF data. These data variables would be a subset of the CDISC SDTM v3.1 standard and would use the SDTM v3.1 variable names within the DMS. Other examples of CDISC-friendly variables would be unique variables which represent collected data, but the data would need to have some formula applied to the data to arrive at the true SDTM variable. These variables may have names which are similar to the SDMT V3.1 variable names but might have some distinct difference to differentia them from true SDTM V3.1 variables.

For instance, the SDTM variable of "AGE" is rarely collected on the Case Report Form. More typically, the "date of birth" and "randomization date" are collected and the "AGE" is derived by subtraction. In many companies, many of these derivations are performed by the biostatistical programming departments rather than in the data management department. It has been done this way in many companies to make certain that the appropriate algorithm is used consistently and that the data is then applied appropriately to analyses.

This division of responsibility also has implications for when the data is prepared into a submission-ready SDTM V3.1 format and what data is used for the source data for preparing the Analysis datasets. A few approaches to this process have been suggested by Susan J. Kenny and Jack Shostak in Pharmaceutical SAS User Group (PharmaSUG) papers and are being used be companies.

Three approaches to creating SDTM variables have been described by Mr. Shostak. These are as follows:

1) Build the SDTM entirely in the DMS (front-end preparation)
2) Build the SDTM entirely in SAS (back-end preparation)
3) Build the SDTM using a combined, hybrid approach (front end + back end)

Many early efforts at implementing SDTM were done using back-end preparation (method 2) as it used the flexibility of SAS to map data from more rigid or proprietary DMS structures to the SDTM structures. This also kept the SDTM variables out of the DMS. This was seen as important as the SDTM model was being revised rapidly from version 2 through version 3.1 but is now seen as less important as version 3.1 has stabilized and is being broadly adopted.

The biggest disadvantage of the front-end preparation (method 1) is that it requires extra variables to be created within the DMS which have not traditionally

been created there. This means extra work and extra overhead for these systems. In addition, the biostatistical programming departments may be reluctant to accept the results of these extra variables which they have traditionally prepared.

The the hybrid approach (method 3) is becoming the most generally accepted approach. This approach uses "CDISC-friendly" naming conventions within the DMS for "raw" or collected variables which would traditionally be collected, verified and stored there. The data is then exported and SAS is typically used to generate the rest of the CDISC SDTM variables as well as analysis datasets.

Some CDISC variables cannot be created within a DMS without significant planning. An example of this would be the SDTM variable for the "Unique Subject Identifier" which has the variable name of "USUBJID". This variable is to be unique for each person within the drug program. In the case where the program has subjects who proceed from one study into follow-on studies, a scheme for making the USUBJID unique would have to be determined before the studies in order to be created in the DMS.

A hybrid approach to this variable is to use a "CDISC-Friendly" variable such as SUBJID within the DMS. The use of this variable would indicate that this subject identifier is unique only at the study level. This variable would then be used exported, and within SAS, be used to map a SUBJID to a USUBJID at the program level. This USBUJID would then be used in any analysis or exporting of submission-ready SDTM datasets.

This hybrid approach still requires communications and agreement between those in data management who are building the DMS and the biostatistical programmers working in SAS. Agreement must be reached in advance on who is responsible for creating each variable and which CDISC-friendly variables are passed directly from the DMS and which are to be used for creating other, final SDTM v3.1 variables.

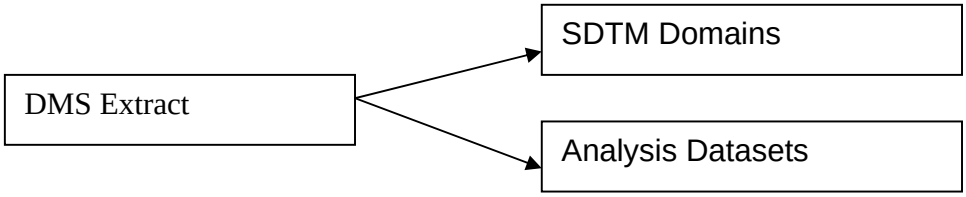
ANALYSIS PREPARATION

In April 2005, the ADaM team released three draft specifications for submitting Change from Baseline Analyses, Categorical Data Analyses and Subject-Level Analyses. In order to prepare these (and other) analyses, Susan Kenny has suggested that one of four approaches might be used. The four methods are:

1) Parallel Method
2) Retrospective Method
3) Linear Method
4) Hybrid Method

These approaches or methods are designed to define the relationship between the DMS, the Analysis datasets and the SDTM data domains.

The Parallel approach (method 1) is described be the diagram below:



This method uses the data exported from the DMS as the source data for creating SDTM data as well as for the Analysis data, but the two are generated separate from one another. This may allow for two different teams to do the work (one for the SDTM processing and a second for the Analysis processing). These teams may be in-house teams or the work may be outsourced.

The most significant disadvantage to this method is that the Analysis data does not use the SDTM variables as source data. FDA statisticians, who receive only the SDTM data and the analysis datasets, may have difficulty in reproducing the analyses should they want to do so. In addition this parallel approach requires a high degree of agreement and communications between the two teams to maintain consistency between the two types of data being submitted. Inconsistencies may lead to significant questions by the FDA reviewers which would delay an approval.

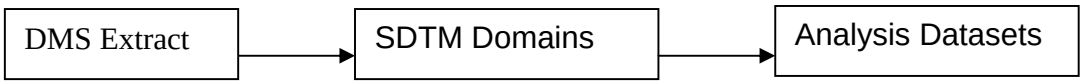
The Retrospective approach (method 2) is described be the diagram below:

This method uses the data exported from the DMS as the source data for the analysis. The SDTM data would then be created after the analyses are complete.

The most significant advantage is that, if the analyses indicate failure and a submission does not occur, there is no need to generate the SDTM data.

There are many disadvantages to this approach. Like the parallel approach, the FDA statistician would not have the source data for the analyses. If the DMS is not CDISC friendly, variables which would be used in the analyses and pushed forward into the SDTM domains would need to be converted to SDTM variables for the analyses. Imputed dates or other types of coding performed in the analyses would need to be undone for the SDTM to represent the original data as it was collected. This method appears to be very inefficient.

The Linear approach (method 3) is described be the diagram below:



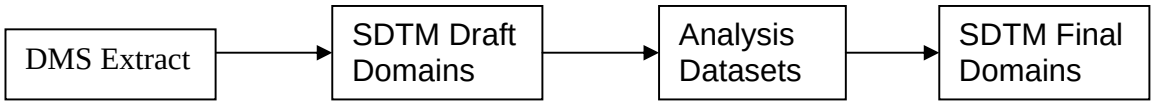
This method uses the data exported from the DMS as the source data for the SDTM preparation. The SDTM data would then be used as the source data for the analyses.

This method appears to be one of the best approaches. If the DMS is not CDISC Friendly, then there may be much effort required to convert the DMS to SDTM domains. This step may slow down the overall process as the Analyses may not begin until the SDTM data domains are completed. In addition, the SDTM are done for all studies even if the analyses do not show a positive result.

The biggest advantage is that the FDA reviewers have the source data for the analyses and may recreate the results using the same programs and metadata as provided with the submission.

This method does show a need for clear communications between data management and biostatistics, particularly if any part of the data management or analysis is outsourced.

The Hybrid approach to SDTM and Analysis preparation (method 4) is described be the diagram below:



This method exports the data from the DMS and creates draft SDTM data as the source data for the Analysis datasets. The SDTM submission domains are then finalized after the Analyses are complete.

This appears to assume that the DMS is not CDISC friendly. DMS extracted data would be converted, only as necessary, to create SDTM domain data sufficient as source for the analyses. If the analyses confirm that the program is to go to submission, then the final SDTM domains are created.

The advantages and disadvantages of this approach are very similar to the linear approach. The distinct advantage is that the final SDTM domains would not need to be created for every study: only those programs where the analyses deem it appropriate for submission.

These four methods provide some valuable insight into the processes employed by the industry in preparing SDTM and analysis data. In addition it points out advantages for adopting SDTM or SDTM friendly variables as early in the data collection process as is feasible.

NEXT STEPS

CDISC has been working with partners such as HL7 and the FDA to discover how data consumers use the data. As stated throughout this paper, the FDA is looking at standardized electronic data for more efficient receipt, better review and re-use within a data archive or data warehouse. Other data consumers are also being looked at by CDISC and initiatives have been moving forward to define standards which meet their needs.

HL7 and CDISC have been working on a Protocol Representation (PR) project to define a machine-readable protocol. The Trial Design component of the SDTM is also a sub-group of this PR group. Some consumers of the PR data have been identified, such as the WHO, the U.S. National Cancer Institute (NCI) and other trial registries. These registries would like to be made aware of details of studies so they may inform possible trial candidates of the potential benefits of their enrolment. Those conducting the trials may use these registries to "advertise" for subjects, especially when researching treatment for rare conditions.

Other consumers of PR data would be those who are conducting the trial and analyzing the data from the trial. In many cases the trial protocols are worded in ways which may be interpreted in multiple ways. These ambiguous protocol texts can be problematic when conducting the trial and particularly difficult to analyze their data. Clear, un-ambiguous wording or a machine-readable protocol could help in avoiding these issues.

Another pilot project that CDISC and FDA are currently conducting involves the use of data from a real study. The pilot's purpose is to create a submission of SDTM v3.1 datasets, ADaM analysis datasets and metadata (with an annotated CRF and either a DEFINE.PDF and/or a DEFINE.XML) to submit to FDA reviewers. This review will generate feedback for industry. The pilot team's intent is to write a white paper and present their findings to industry so that the feedback can help others create better submissions in the future.

Another project is to harmonize the CDISC standards into a single unified standard. This is a multi-year project which is based upon an HL7 and CDISC UML modeling effort called the BRIDG project. The BRIDG project is to model all of the functional processes involved in clinical trials from protocol design through study conduct and review. This model will be the basis for harmonizing the CDISC models and unifying them.

Other standards such as LAB and ODM are being updated occasionally. While the LAB model is uniquely designed to share data between sponsors and central labs, the ODM model has more potential. The ODM XML model is the basis of the DEFINE.XML in the CRT-DDS and may eventually become the format for datasets in FDA submissions. The potential exist for software sponsors to use ODM for data archives or data transfers. More software vendors such as SAS, Phase Forward (ClinTrial) and Oracle Clinical are talking about ODM imports and exports in the "near future."

CONCLUSION

The CDISC standards have been developing for many years. Industry did not embrace these standards and start adopting them until the regulatory authority in the US stated publicly that they wanted data in this format. Now that the FDA has not only stated that they want it in this format, but that soon they will not accept data in any other format, industry is moving rapidly to adopt the CDISC standards.

Industry is seeing advantages to using these standards. They are learning a common structure for talking about and sharing data. These communications may be between drug sponsors and the FDA or they may be between partner companies, researches or contract organizations. Whoever is communicating, these standards are becoming the language and should improve the collection, sharing, storage, analysis, reported and re-use of the data.

REFERENCES

FDA - Electronic Records; Electronic Signatures (ERES rule)
FDA - CDER/CBER - Providing Regulatory Submissions in Electronic Format — General Considerations
FDA - CDER - Providing Regulatory Submissions in Electronic Format — NDAs
FDA - CBER - *Providing Regulatory Submissions to the Center for Biologics Evaluation and Research (CBER) in Electronic Format — Biologics Marketing Applications (BLAs)*
CDISC Submission Metadata Model
CDISC *Submission Data Domain Model v2.0* (SDDM)
CDISC Version 3.0 Submission Data Standards - Review Version 1.0
CDISC Submission Data Domain Models Version 3.0 - Final Version 1.2
CDISC Submission Data Tabulation Model version 1.0 (SDTM)
CDISC SDTM Implementation Guide V3.1 (I.G.)
CDISC Statistical Analysis Dataset Model: General Considerations Version 1.0
CDISC Case Report Tabulation Data Description Specification (CRT-DDS) v1.0.0
FDA - eCTD - Study Data Specifications v1.0
FDA - eCTD Study Data Specifications v1.1
CDISC Standard for Exchange of Nonclinical Data (SEND) v1.7.5

PharmSUG 2005 Paper FC01 - Implementation of the CDISC SDTM at the Duke Clinical Research Institute, Jack Shostak, Duke Clinical Research Institute (DCRI), Durham, NC
PharmaSUG 2005 Paper FC03 - Strategies for Implementing SDTM and ADaM Standards, Susan J. Kenny, Maximum Likelihood Solutions, Inc and Octagon Research Solutions, Inc., Chapel Hill, NC
Michael A. Litzsinger, SCHWARZ BIOSCIENCES, Inc., Research Triangle Park, NC

ACKNOWLEDGMENTS
Wayne Kubick - Lincoln Technologies
Gary Walker - Quintiles