Data Management Standards in Clinical Research

July 2009

Abstract

Use of standards has become increasingly widespread within clinical data management. Standards can reduce setup costs for a study, reduce conversion errors, and most importantly speed a medical treatment's path to market. This chapter discusses the importance of standards within clinical research, the history of standards used in health care delivery, some of the standards already commonly used, and future directions for standards within clinical research. The chapter also provides readers with an overview of standards relevant to clinical data management. Links are provided for more information about each standard, including downloads for most of the standards.

Introduction

Merriam-Webster defines the word "standard" as "something established by authority, custom, or general consent as a model or example." Within the context of clinical data management (CDM), standards are used to optimize the collection, transport and storage of data, and simplify the submission of data to regulatory bodies.

The advent of modern information technology has enabled widespread use of comprehensive standards. Today, standards encompass almost every part of data collection and handling. Although there are few regulatory mandates for using any particular standard, using standards in all areas of data collection and handling can greatly increase an organization's efficiency by shortening study setup time and incorporating effective and validated standards, thereby reducing overall time and expenses while maintaining consistency for data managers and those charged with collecting data at clinical sites. Most of the established standards currently in use are readily available and designed to be independent of any vendor or platform.

Scope

This chapter provides an overview of established standards commonly used within clinical studies. In addition to giving an overview of each standard and its purpose and scope, the chapter directs readers to where more information can be found about these standards. In most cases, links to the download of the standards discussed are included. Additionally, information is provided about emerging standards within clinical data management. For specific information about implementation of standards listed in this chapter, please follow the provided links to the standards development organizations.

Minimum Standards

- Use the most current version of any standard, if appropriate.
- Use standards required by regulatory agencies in the country where the study is conducted.
- Do not modify published standards.

Best Practices

- Use accepted standards whenever possible, and strive for interoperability.
- Use all standards recommended by regulatory agencies in the locale of the study.
- Review implementation guidelines for any standard having associated guidelines documents.

Purpose and Benefits of Standardization

The use of standards within clinical research involves using standardized names, codes, structures, and formats for data across different locations, studies, and organizations. Using the same formats, names, and codes for different studies can greatly decrease the time and money needed to set up a study, particularly in cases where similar studies have been conducted in the past. Standards provide benefits beyond study setup and can also help streamline processes for study conduct, data transfers, analyses, and

regulatory submissions. Ultimately, standards facilitate bringing safe and effective treatments to patients in a more timely and cost-effective fashion.

Although multiple standards exist for similar concepts, the ultimate goal is for researchers everywhere to use the same standards and naming conventions for their studies. This goal has not yet been realized, but the clinical research industry is trending in that direction. The US Food and Drug Administration (FDA) has strongly encouraged the use of the Study Data Tabulation Model (SDTM) for data submissions, and although this standard's use has not been mandated yet, it may become mandatory in the future. Data submissions in a standardized format allow the FDA and other regulatory bodies to expend fewer resources on their review of study data.

Another enormous benefit to standardization is that data can be more easily and accurately compared and combined across different studies. Although the Internet was originally created to promote sharing of scientific research data, the actual sharing of data has been somewhat limited, in large part due to researchers storing data in different file formats. Standards could potentially increase data sharing, as well as the compatibility of shared data. This increased data sharing could provide valuable benefits to science and humanity.

History of the Development of Standards Organizations

Before the advent of global communication tools such as the telephone and Internet, standards were typically limited to their locale of origin. As technological advances have sped globalization, organizations have emerged to promote standards for many industries. The following organizations have played integral roles in promoting the standardization of health care data used in clinical research.

International Organization for Standardization (ISO)

ISO was created in 1947 after delegates from 25 countries met to discuss the creation of an international organization to create and maintain international standards for industry. From starting with 67 proposed ISO technical committees in 1947, ISO has developed over 17,000 standards encompassing the full spectrum of industries across the globe. In addition to standards

formulated for specific industries, ISO has created generic standards for product quality and management systems that are applicable to any endeavor.

In addition to general standards applicable to quality and management systems, there are multiple ISO standards specific to various processes involved with clinical research. More information about all ISO standards can be found at http://www.iso.org/iso/home.htm.

International Conference on Harmonisation (ICH)

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) began in 1990 as an effort to standardize pharmaceutical regulatory requirements in Europe, Japan, and the US. The ultimate objectives of ICH are to 1) maintain safety and quality while increasing efficiencies in the use of human, animal, and material resources, and 2) help eliminate unnecessary delays in bringing new medical treatments to market. To achieve these goals, numerous guidelines have been released by ICH since its inception. Many of these have had a strong impact on standards development, particularly in regard to regulatory submissions. More information about ICH can be found at http://www.ich.org.

Health Level 7 (HL7)

Founded in 1987, HL7 is a nonprofit Standards Development Organization (SDO) initially created to produce standards for hospital information systems. The organization's mission is to provide "…standards for interoperability that can improve care delivery, optimize workflow, reduce ambiguity and enhance knowledge transfer among all of our stakeholders, including healthcare providers, government agencies, the vendor community, fellow SDOs and patients."[1]

The following HL7 standards relate to clinical data management, and are discussed later in this chapter.

- Reference Information Model (RIM)
- Clinical Context Object Workgroup (CCOW)
- Clinical Document Architecture (CDA)

More information about HL7 standards can be found at http://www.hl7.org.

Clinical Data Interchange Standards Consortium (CDISC)

Unlike ISO and HL7, CDISC was formed solely to create standards for clinical research data. Their mission statement reads, "CDISC is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission 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 standards are vendor-neutral, platform-independent and freely available via the CDISC website."[2]

CDISC began in 1997 with a meeting of 25 people interested in standards creation for use within clinical research. Since that time, CDISC has grown exponentially and today has the support of over 200 member organizations from around the world. In addition to consulting with recognized leaders in the clinical research industry, CDISC works closely with other SDOs such as ISO and HL7 to improve interoperability between the various standards. As CDISC has grown in membership and acceptance, their scope has expanded as well. The following standards have been developed by CDISC and are currently published and available for use.

- Clinical Data Acquisition Standards Harmonization (CDASH)
- Laboratory Model (LAB)
- Operational Data Model (ODM)
- Study Data Tabulation Model (SDTM)
- Analysis Dataset Model (ADaM)

For more information about CDISC, visit http://www.cdisc.org/.

Standards for Clinical Research

The standards discussed in the remainder of this chapter are primarily those relating directly to CDM functions within clinical studies. The majority of the standards discussed come from CDISC, but CDM personnel should be aware of any new standards gaining traction within the industry, regardless of the

origin of the standard. Where possible, all standards employed by an organization should be sufficiently interoperable to allow for a comprehensive standard practice to effectively manage clinical data.

Clinical Data Acquisition Standards Harmonization (CDASH)

The Clinical Data Acquisition Standards Harmonization (CDASH) standard released October 2008 by CDISC, is intended to streamline and standardize data collection at clinical investigative sites. The development of CDASH was a global effort, with feedback provided from all three of the ICH regions (US, Europe, and Japan). The published CDASH standard consists of a basic set of data collection fields (variable name, definition, metadata) that apply to the majority of case report forms (CRFs), regardless of therapeutic area or phase of development. Sponsors are expected to make additions for therapeutic area-specific data collection fields, as well as other data collection fields needed for regulatory requirements. The CDASH standard also includes best practice guidelines, regulatory references, and information about the development of the CDASH standard.

In order to ensure harmonization between standards, recommendations are provided for mapping CDASH data collection fields (or variables) into the Study Data Tabulation Model (SDTM) submission structure.

CDASH Domains

The data collection fields, specified in CDASH, like SDTM, are divided into the following sixteen domains along with their associated codes.

- Adverse Events (AE)
- Comments (CO)
- Concomitant Medications (CM)
- Demography (DM)
- Disposition (DS)
- Drug Accountability (DA)
- ECG Test Results (EG)

- Exposure (EX)
- Inclusion/Exclusion (IE)
- Laboratory Test Results (LB)
- Medical History (MH)
- Physical Examination (PE)
- Protocol Deviations (DV)
- Subject Characteristics (SC)
- Substance Use (SU)
- Vital Signs (VS)

An implementation guide is under development to accompany the standard, and is targeted for completion in the third quarter of 2009. Please see http://www.cdisc.org/standards/index.html for more information about the CDASH standard, including a link to download the most recent version of the standard.

Laboratory Model (LAB)

The CDISC LAB standard was initially released in 2002, and was designed to be a standard for the transfer of laboratory data. Other standards already existed for laboratory data, but those standards had limited applicability to clinical research. Use of the LAB standard is estimated to save 30% to 50% of laboratory costs, which has an enormous impact on overall costs considering that 60% to 80% of clinical data is estimated to come from laboratories.[3]

Data Field Levels

Data for this standard are categorized into the following 12 levels and associated data fields.

- Good transmission practice—version of LAB model used, local (and universal) date and time data file was created, identification code and name of organization that is the source of the data transmission

- Study—identification code and name of the study and whether the data transmission is incremental or cumulative
- Site—identification code of the site
- Investigator—identification code and name of the investigator
- Subject—identification code of the subject before and after randomization (and possibly an extra subject identifier code), subject initials, subject gender (and possibly gender code), subject date of birth, subject race (and possibly race code). Note: When collecting subject identification data, follow local regulations relating to subject privacy
- Visit—name of the visit, identification code or number of the visit, whether the visit was scheduled or unscheduled, and whether the visit was physician ordered, a retest, or early termination of the subject's involvement with the study
- Accession—name and identification code of the laboratory delivering the data, identification code of the kit used for the subject visit, local (and universal) date and time of the last modification made to the record
- Record extension type—specifies if any extension to the base LAB model was used, as described below
- Base specimen—identification code of an individual kit item used at the visit, actual and planned local date and time of specimen collection from the site, time discrepancy between planned and actual specimen collection, local (and universal) date and time of specimen receipt at laboratory, specimen condition, laboratory and investigator comments, specimen identification code, specimen name (e.g., blood, urine), subject age at collection, units of subject age at collection, fasting status of subject at collection
- Base battery—name and identification code of the battery, panel or group to which the test belongs
- Base test—name and identification code of laboratory, name and identification code of the test as defined by site, name and identification code of the test as defined by laboratory, LOINC (Logical Observation Identifiers Names and Codes) code and code list identifier, test status

(done, not performed or cancelled), test comments, local (and universal) date and time of testing, test type (study test, non-study test, unscheduled study test)

- Base results—this level contains 32 fields providing all test result names, codes, reference ranges, units, results, statuses, toxicity grades, flags, reporting time, and record type

Extensions

In addition to the LAB base model, the standard has several extensions designed for specialized laboratory data. The extensions currently published or in development include:

- Microbiology
- Pharmacogenomics
- Electrocardiogram (ECG) interpretation
- Specimen handling
- Edit/data query capabilities

Please see http://www.cdisc.org/standards/index.html for more information about the LAB standard, including a link to download the most recent version of the standard.

Operational Data Model (ODM)

The first ODM standard was released by CDISC in 2002 to address the structure of data rather than naming conventions. The ODM standard is designed to "…support the end-to-end data flow within clinical trials, from the operational database through analysis to regulatory submission. 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."[4] In addition to providing a standard format for transporting data, the flexibility of the ODM creates the possibility of automating creation of electronic CRFs used in an electronic data capture (EDC) system.[5]

The ODM uses the extensible markup language (XML) to create a file with the four following primary elements.

- Study information such as study name and metadata
- Administrative information such as users, sites, and authorizations for the study
- Reference data (e.g., normal ranges)
- Clinical data from the study [4]

Supported Data Formats

The ODM was designed to be vendor-neutral and platform-independent, and supports numerous data formats including integers, decimals, text strings, Boolean terms, hex binary, base 64 binary, dates and times, partial dates and times, intervals, durations, and more.

Please see http://www.cdisc.org/standards/index.html for more information about the ODM standard, including a link to download the most recent version of the standard.

Study Data Tabulation Model (SDTM)

The first implementation-ready version of the SDTM was released by CDISC in 2004, and was developed to provide a standard for the organization, structure, and format of tabulation data to be submitted to regulatory agencies. Tabulation datasets contain collected data from a clinical study, and should not be handled in the same manner as the other three types of data submitted to regulatory agencies (e.g., analysis datasets, patient profiles, and listings). The FDA has strongly recommended using SDTM for data tabulation submissions, but this has not been mandated.

Variable Classification Scheme

According to the SDTM, each variable, which normally corresponds to a column in a dataset, can be classified according to its *Role*. A Role determines the type of information conveyed by the variable in describing an observation. Variables can be classified into five major roles:

- *Identifier* variables—identify the study, the subject (individual human or animal) involved in the study, the domain, and the sequence number of the record.
- *Topic* variables—specify the focus of the observation (such as the name of a lab test), and vary according to the type of observation.
- *Timing* variables—describe the timing of an observation (such as start date and end date).
- *Qualifier* variables*—include additional illustrative text, or numeric values that describe the results or additional traits of the observation (such as units or descriptive adjectives). The list of Qualifier variables included with a domain will vary considerably depending on the type of observation and the specific domain.
- *Rule* variables—express an algorithm or executable method to define start, end, or looping conditions in the Trial Design model.[6]

*The SDTM further divides qualifier variables into five subclasses of grouping qualifiers, result qualifiers, synonym qualifiers, record qualifiers, and variable qualifiers. See the SDTM implementation guide for detailed descriptions of these qualifier variables.

Standard Domains

The SDTM contains the following domains and respective codes, which fall into six general categories.

- Special Purpose Domains
  - ☐ Demographics (DM)
  - ☐ Comments (CO)
  - ☐ Subject Elements (SE)
  - ☐ Subject Visits (SV)
- Interventions

- ☐ Concomitant Medications (CM)
- ☐ Exposure (EX)
- ☐ Substance Use (SU)

- Events
  - ☐ Adverse Events (AE)
  - ☐ Disposition (DS)
  - ☐ Medical History (MH)
  - ☐ Protocol Deviations (DV)
  - ☐ Clinical Events (CE)
- Findings
  - ☐ ECG Test Results (EG)
  - ☐ Inclusion/Exclusion Criterion Not Met (IE)
  - ☐ Laboratory Test Results (LB)
  - ☐ Physical Examinations (PE)
  - ☐ Questionnaires (QS)
  - ☐ Subject Characteristics (SC)
  - ☐ Vital Signs (VS)
  - ☐ Drug Accountability (DA)
  - ☐ Microbiology Specimen (MB)
  - ☐ Microbiology Susceptibility Test (MS)
  - ☐ Pharmacokinetic Concentrations (PC)
  - ☐ Pharmacokinetic Parameters (PP)
  - ☐ Findings About (FA)

- Trial Design Domains
  - ☐ Trial Arms (TA)
  - ☐ Trial Elements (TE)
  - ☐ Trial Visits (TV)
  - ☐ Trial Inclusion/Exclusion Criteria (TI)
  - ☐ Trial Summary (TS)
- Special Purpose Relationship Datasets
  - ☐ Supplemental Qualifiers (SUPPQUAL)
  - ☐ Related Records (RELREC)

SDTM Implementation Guide (SDTMIG)

CDISC has also released an implementation guide to augment the SDTM standard. This implementation guide is intended to guide the format, organization, and structure of tabulation datasets. Any organization using SDTM should also utilize this implementation guide.

See http://www.cdisc.org/standards/index.html for more information about the SDTM standard and implementation guide, as well as a link to download the most recent version of the standard and implementation guide.

Analysis Dataset Model (ADaM)

ADaM was initially released by CDISC in 2004 as a standard model to create analysis datasets for submission to regulatory bodies, and can be thought of as an extension to the SDTM standard. The ADaM describes the proposed content, structure, and metadata of analysis datasets, including analysis dataset metadata, analysis variable metadata, and analysis results metadata. The standard includes examples of datasets created using the ADaM.

Four Key Principles for Analysis Datasets

The ADaM standard is based on the following four general principles.

- Analysis datasets should facilitate clear and unambiguous communication
- Analysis datasets should be useable by currently available software applications
- Analysis datasets should be linked to machine-readable metadata
- Analysis datasets should be analysis-ready

ADaM Implementation Guide (ADaMIG)

As with the SDTM standard, CDISC has released a draft implementation guide to augment the ADaM standard. This implementation guide is intended to guide the format, organization, and structure of analysis datasets. Any organization using ADaM should also utilize this implantation guide.

See http://www.cdisc.org/standards/index.html for more information about the ADaM standard and implementation guide, as well as a link to download the most recent version of the standard and implementation guide.

Electronic Common Technical Document (eCTD)

The eCTD standard was developed by the ICH to provide a standardized format for submitting files from pharmaceutical studies to regulatory bodies. Unlike some standards used in clinical research, eCTD focuses more on data and file structures than naming conventions. The eCTD relies heavily on the Document Type Definition (DTD) specification of the XML markup language. These DTDs are used to create a detailed hierarchical folder structure for each eCTD.

In addition to the structure of an eCTD, the standard is designed to support high-level functional requirements. Some of these functional requirements include the ability to copy and paste, view and print documents, have annotated documentation, and export to databases. An eCTD should also allow users to search both within and across applications and allow navigation throughout the eCTD and any subsequent amendments or variations.[7]

eCTD Modules

Every eCTD consists of five modules, four of which are common to all countries and regions. The first of the following five modules may vary between different ICH regions.

1. Regional Administrative Information and Prescribing Information—Module One contains administrative information and forms that may vary between countries and regions.

2. Common Technical Document Summaries—Module Two contains summaries of the information contained in Modules Three, Four, and Five.

3. Quality—Module Three provides detailed information about the treatment being studied and details of the product's development and manufacturing processes.

4. Nonclinical Study Reports—Module Four provides detailed pharmacological, pharmacokinetic and toxicological information.

5. Clinical Study Reports—Module Five contains the results of the study, including data related to background and development rationale, efficacy, safety, benefits and risks.

The eCTD has become the recommended format for regulatory submissions in the European Union, US, Canada and Japan, and may become mandatory in time. Many companies sell eCTD submission solutions, but more free information about eCTD can be found at http://www.fda.gov/cder/Regulatory/ersr/ectd.htm.

HL7 Standards

Although HL7 does not design standards specifically for use within clinical research, the increased use of electronic health records within hospitals gives CDM personnel an incentive to become familiar with the following HL7 standards.

- Reference Information Model (RIM)—This standard provides structure, naming and coding conventions to be used among disparate organizations and platforms.

- Clinical Context Object Workgroup (CCOW)—This is a vendor-independent standard designed to enable different computer applications to communicate with each other effectively.

- Clinical Document Architecture (CDA)—This standard is based on the RIM, and uses the XML markup language to specify the coding, structure, and semantics of clinical documents to be exchanged.

See http://www.hl7.org/ for more information about HL7 standards, including links to downloads of the most recent versions of the standards.

Future Directions

With the numerous standards that currently exist, the ultimate goal is to make these standards interoperable to the degree that any health-related data can be easily shared between different researchers and institutions. The US National Cancer Institute (NCI), FDA, HL7, and CDISC are all collaborating to create the Biomedical Research Integrated Domain Group (BRIDG) model. This standard is being designed to integrate HL7 and CDISC standards, which will reduce potential errors and streamline the flow of data from health care providers to clinical researchers.

CDISC is also creating a standard called the Protocol Representation Model (PRM), which identifies, defines and describes over 300 common protocol elements and maps those elements to elements within the BRIDG model. The PRM model is intended as a standard to be used in designing a study, selecting investigative sites, developing data collection tools, and describing an analysis plan and study procedures.

The FDA is piloting a program for a standard known as the Summary Technical Document (STED), which is a harmonized format for medical device regulatory submissions that is already accepted by multiple regulatory bodies worldwide. For information about the pilot program, see http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/PremarketSubmissions/SummaryTechnicalDocumentSTEDPilotProgram/default.htm. The FDA has already released eSubmitter, a

standardized tool that is part of an electronic submissions program originated in the Center for Devices and Radiological Health (CDRH). The eSubmitter program evolved from two very successful pilot programs (eLaser and Turbo 510(k)) at CDRH. FDA eSubmitter is an improved and expanded package for a variety of submission types and is now available for voluntary use by sponsors and manufacturers in certain device and radiological health and blood regulated industries. Like other attempts to standardize, the goal is to improve efficiencies in the regulatory submission and review process. See http://www.fda.gov/ForIndustry/FDAeSubmitter/default.htm for more information.

Another interesting and evolving initiative is the National Cancer Institute's cancer Biomedical Informatics Grid (caBIG®), which is intended to simplify collaboration by leveraging shared expertise and large multidisciplinary data collections to speed many of the processes of cancer research. The four key principles of caBIG®—open access, open development, open source, and federation—have guided the development of interoperable software tools, data standards, and a computing infrastructure conceived to advance basic and clinical research. Originally designed solely for cancer research, the caBIG® initiative may expand outside cancer research to serve as a model for improving collaboration, data sharing, and patient outcomes in other therapeutic areas in the future. For more information about caBIG®, see https://cabig.nci.nih.gov.

Standards across medical research are contributing to more efficient research activities. This success has spawned a mounting interest in standards development and resulted in an increasing number of new and revised standards. Staying abreast of standards that affect CDM is a challenge. One effective strategy is to visit the Web sites of organizations that have been involved in standards development to keep informed of their progress.

Recommended Standard Operating Procedures

- CRF Design
- Database Design
- Medical Coding

- Data Transfers
- Regulatory Submissions

References

1. HL7 (Health Level 7) Web site. Available at: http://www.hl7.org/. Accessed November 24, 2008.
2. CDISC (Clinical Data Interchange Standards Consortium) Web site. Available at: http://www.cdisc.org/about/index.html. Accessed November 24, 2008.
3. Bassion S. The CDISC Lab Model: Saving Users Time and Money. *Applied Clinical Trials*. Walnut Creek, CA: The Biotech Communications Group; 2005.
4. Iberson-Hurst D. The CDISC Operational Data Model: Ready to Roll? *Applied Clinical Trials*. Walnut Creek, CA: The Biotech Communications Group; 2004.
5. Aerts J. CDISC Standards Energize EDC. *Applied Clinical Trials*. Walnut Creek, CA: The Biotech Communications Group; 2006.
6. CDISC Submissions Data Standards Team. *Study Data Tabulation Model: Version 1.1.* Austin, TX: Clinical Data Interchange Standards Consortium; 2005. Available at http://www.cdisc.org/models/sdtm/v1.1/index.html. Accessed November 24, 2008.
7. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH M2 EWG Electronic Common Technical Document Specification*, V. 3.2, Feb 2004.

Further Reading

CDISC Analysis Dataset Modeling Team. *Analysis Data Model: Version 2.0*. Austin, TX: Clinical Data Interchange Standards Consortium; 2006. Available at http://www.cdisc.org/models/adam/V2.0/index.html. Accessed November 24, 2008.

CDISC CDASH Core and Domain Teams. *Clinical Data Acquisition Standards Harmonization (CDASH).* Austin, TX. Clinical Data Interchange Standards Consortium; 2008. Available at. http://www.cdisc.org/standards/cdash/index.html. Accessed November 24, 2008.

CDISC Operational Data Modeling Team. *Specification for the Operational Data Model (ODM) Version 1.3.* Austin, TX. Clinical Data Interchange Standards Consortium; 2006. Available at http://www.cdisc.org/models/odm/v1.3/index.html. Accessed November 24, 2008.

Chapter Revision History

Publication Date	**Comments**
July 2009	Initial publication.

This page is intentionally blank.