ISSN 2348 – 9928
IJAICT Volume 3, Issue 8, August 2016 Doi:01.0401/ijaict.2014.06.12 Published on 05 (8) 2016

DATA MANAGEMENT AND DATA ANALYSIS INTEROPERABILITY IN MEDICAL RESEARCH

Mr. R. Saravana Kumar
Assistant professor,
Computer Science and Engineering,
Jayam College of Engineering and Technology,
Dharmapuri, Tamilnadu, India

Dr. G. Tholkappia Arasu,
Principal,
AVS Engineering College,
Salem, Tamilnadu, India

Abstract— Clinical Data Management (CDM) is a critical phase in clinical research, which leads to generation of high-quality, reliable, and statistically sound data from clinical trials. This helps to produce a drastic reduction in time from drug development to marketing. Team members of CDM are actively involved in all stages of clinical trial right from inception to completion. They should have adequate process knowledge that helps maintain the quality standards of CDM processes. Various procedures in CDM including Case Report Form (CRF) designing, CRF annotation, database designing, data-entry, data validation, discrepancy management, medical coding, data extraction, and database locking are assessed for quality at regular intervals during a trial. In the present scenario, there is an increased demand to improve the CDM standards to meet the regulatory requirements and stay ahead of the competition by means of faster commercialization of product. With the implementation of regulatory compliant data management tools, CDM team can meet these demands. Additionally, it is becoming mandatory for companies to submit the data electronically. CDM professionals should meet appropriate expectations and set standards for data quality and also have a drive to adapt to the rapidly changing technology. This article highlights the processes involved and provides the reader an overview of the tools and standards adopted as well as the roles and responsibilities in CDM.

Keywords— Clinical Data Interchange Standards Consortium, Clinical Data Management Systems, Data Management, E-CRF, Good Clinical Data Management Practices, Validation.

I. INTRODUCTION

Clinical trial is intended to find answers to the research question by means of generating data for proving or disproving a hypothesis. The quality of data generated plays an important role in the outcome of the study. Often research students ask the question, "what is Clinical Data Management (CDM) and what is its significance?" Clinical data management is a elevant and important part of a clinical trial. All researchers try their hands on CDM activities during their research work, knowingly or unknowingly. Without identifying the technical phases, we undertake some of the processes involved in CDM during our research work. This article highlights the processes involved in CDM and gives the reader an overview of how data is managed in clinical trials.

CDM is the process of collection, cleaning, and management of subject data in compliance with regulatory standards. The primary objective of CDM processes is to provide high-quality data by keeping the number of errors and missing data as low as possible and gather maximum data for analysis.[1] To meet this objective, best practices are adopted to ensure that data are complete, reliable, and processed correctly.

This has been facilitated by the use of software applications that maintain an audit trail and provide easy identification and resolution of data discrepancies. Sophisticated innovations [2] have enabled CDM to handle large trials and ensure the data quality even in complex trials.

How do we define 'high-quality' data? High-quality data should be absolutely accurate and suitable for statistical analysis. These should meet the protocol-specified parameters and comply with the protocol requirements. This implies that in case of a deviation, not meeting the protocol-specifications, we may think of excluding the patient from the final database. It should be borne in mind that in some situations, regulatory authorities may be interested in looking at such data. Similarly, missing data is also a matter of concern for clinical researchers. High-quality data should have minimal or no misses. But most importantly, high-quality data should possess only an arbitrarily 'acceptable level of variation' that would not affect the conclusion of the study on statistical analysis. The data should also meet the applicable regulatory requirements specified for data quality.

II. TOOLS FOR CDM

Many software tools are available for data management, and these are called Clinical Data Management Systems (CDMS).

In multi centric trials, a CDMS has become essential to handle the huge amount of data. Most of the CDMS used in pharmaceutical companies are commercial, but a few open source tools are available as well. Commonly used CDM tools are ORACLE CLINICAL, CLINTRIAL, MACRO, RAVE, and eClinical Suite. In terms of functionality, these software tools are more or less similar and there is no significant advantage of one system over the other. These software tools are expensive and need sophisticated Information Technology infrastructure to function. Additionally, some multinational pharmaceutical giants use custom-made CDMS tools to suit their operational needs and procedures. Among the open source tools, the most prominent ones are OpenClinica, openCDMS, TrialDB, and PhOSCo. These CDM software are available free of cost and are as good as their commercial counterparts in terms of functionality. These open source software can be downloaded from their respective websites.

In regulatory submission studies, maintaining an audit trail of data management activities is of paramount importance. These CDM tools ensure the audit trail and help in the management of discrepancies. According to the roles and responsibilities (explained later), multiple user IDs can be created with access limitation to data entry, medical coding, database designing, or quality check. This ensures that each user can access only the respective functionalities allotted to that user ID and cannot make any other change in the database. For responsibilities where changes are permitted to be made in the data, the software will record the change made, the user ID that made the change and the time and date of change, for audit purposes (audit trail). During a regulatory audit, the auditors can verify the discrepancy management process; the changes made and can confirm that no unauthorized or false changes were made.

III. REGULATIONS, GUIDELINES, AND STANDARDS IN CDM

Akin to other areas in clinical research, CDM has guidelines and standards that must be followed. Since the pharmaceutical industry relies on the electronically captured data for the evaluation of medicines, there is a need to follow good practices in CDM and maintain standards in electronic data capture. These electronic records have to comply with a Code of Federal Regulations (CFR), 21 CFR Part 11. This regulation is applicable to records in electronic format that are created, modified, maintained, archived, retrieved, or transmitted. This demands the use of validated systems to ensure accuracy, reliability, and consistency of data with the use of secure, computer-generated, time-stamped audit trails to independently record the date and time of operator entries and actions that create, modify, or delete electronic records.[3] Adequate procedures and controls should be put in place to ensure the integrity, authenticity, and confidentiality of data. If data have to be submitted to regulatory authorities, it should be entered and processed in 21 CFR part 11-compliant systems. Most of the CDM systems available are like this and pharmaceutical companies as well as contract research organizations ensure this compliance.

Society for Clinical Data Management (SCDM) publishes the Good Clinical Data Management Practices (GCDMP) guidelines, a document providing the standards of good practice within CDM. GCDMP was initially published in September 2000 and has undergone several revisions thereafter. The July 2009 version is the currently followed GCDMP document. GCDMP provides guidance on the accepted practices in CDM that are consistent with regulatory practices. Addressed in 20 chapters, it covers the CDM process by highlighting the minimum standards and best practices.

Clinical Data Interchange Standards Consortium (CDISC), a multidisciplinary non-profit organization, has developed standards to support acquisition, exchange, submission, and archival of clinical research data and metadata. Metadata is the data of the data entered. This includes data about the individual who made the entry or a change in the clinical data, the date and time of entry/change and details of the changes that have been made. Among the standards, two important ones are the Study Data Tabulation Model Implementation Guide for Human Clinical Trials (SDTMIG) and the Clinical Data Acquisition Standards Harmonization (CDASH) standards, available free of cost from the CDISC website (www.cdisc.org). The SDTMIG standard [4] describes the details of model and standard terminologies for the data and serves as a guide to the organization. CDASH v 1.1[5] defines the basic standards for the collection of data in a clinical trial and enlists the basic data information needed from a clinical, regulatory, and scientific perspective.

IV. THE CDM PROCESS

The CDM process, like a clinical trial, begins with the end in mind. This means that the whole process is designed keeping the deliverable in view. As a clinical trial is designed to answer the research question, the CDM process is designed to deliver an error-free, valid, and statistically sound database. To meet

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.06.12 Published on 05 (8) 2016

this objective, the CDM process starts early, even before the finalization of the study protocol.

V. REVIEW AND FINALIZATION OF STUDY DOCUMENTS

The protocol is reviewed from a database designing perspective, for clarity and consistency. During this review, the CDM personnel will identify the data items to be collected and the frequency of collection with respect to the visit schedule. A Case Report Form (CRF) is designed by the CDM team, as this is the first step in translating the protocol-specific activities into data being generated. The data fields should be clearly defined and be consistent throughout. The type of data to be entered should be evident from the CRF. For example, if weight has to be captured in two decimal places, the data entry field should have two data boxes placed after the decimal as shown in Figure 1. Similarly, the units in which measurements have to be made should also be mentioned next to the data field. The CRF should be concise, self-explanatory, and user-friendly (unless you are the one entering data into the CRF). Along with the CRF, the filling instructions (called CRF Completion Guidelines) should also be provided to study investigators for error-free data acquisition. CRF annotation is done wherein the variable is named according to the SDTMIG or the conventions followed internally. Annotations are coded terms used in CDM tools to indicate the variables in the study. An example of an annotated CRF is provided in Figure 1. In questions with discrete value options (like the variable gender having values male and female as responses), all possible options will be coded appropriately.

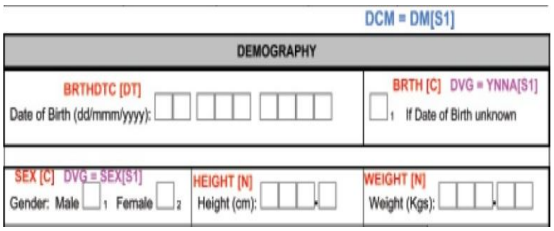

Fig1. Annotated sample of a Case Report Form (CRF).

Annotations are entered in coloured text in this figure to differentiate from the CRF questions. DCM = Data collection module, DVG = Discrete value group, YNNA [S1] = Yes, No = Not applicable [subset 1], C **...**

Based on these, a Data Management Plan (DMP) is developed. DMP document is a road map to handle the data under foreseeable circumstances and describes the CDM activities to be followed in the trial. A list of CDM activities is provided in Table 1. The DMP describes the database design, data entry and data tracking guidelines, quality control measures, SAE reconciliation guidelines, discrepancy management, data transfer/extraction, and database locking guidelines. Along with the DMP, a Data Validation Plan (DVP) containing all edit-checks to be performed and the calculations for derived variables are also prepared. The edit check programs in the DVP help in cleaning up the data by identifying the discrepancies.

Table 1
List of clinical data management activities

Data collection
CRF tracking
CRF annotation
Database design
Data entry
Medical coding
Data validation
Discrepancy management
Database lock

5.1 Database Designing

Databases are the clinical software applications, which are built to facilitate the CDM tasks to carry out multiple studies.[6] Generally, these tools have built-in compliance with regulatory requirements and are easy to use. “System validation” is conducted to ensure data security, during which system specifications, [7] user requirements, and regulatory compliance are evaluated before implementation. Study details like objectives, intervals, visits, investigators, sites, and patients are defined in the database and CRF layouts are designed for data entry. These entry screens are tested with dummy data before moving them to the real data capture.

5.2 Data Collection

Data collection is done using the CRF that may exist in the form of a paper or an electronic version. The traditional method is to employ paper CRFs to collect the data responses, which are translated to the database by means of data entry done in-house. These paper CRFs are filled up by the investigator according to the completion guidelines. In the e-CRF-based CDM, the investigator or a designee will be logging into the CDM system and entering the data directly at the site. In e-CRF method, chances of errors are less, and the resolution of discrepancies happens faster. Since pharmaceutical companies try to reduce the time taken for drug development processes by enhancing the speed of processes involved, many pharmaceutical companies are opting for e-CRF options (also called remote data entry).

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.06.12 Published on 05 (8) 2016

5.3 CRF Tracking

The entries made in the CRF will be monitored by the Clinical Research Associate (CRA) for completeness and filled up CRFs are retrieved and handed over to the CDM team. The CDM team will track the retrieved CRFs and maintain their record. CRFs are tracked for missing pages and illegible data manually to assure that the data are not lost. In case of missing or illegible data, a clarification is obtained from the investigator and the issue is resolved.

5.4 Data Entry

Data entry takes place according to the guidelines prepared along with the DMP. This is applicable only in the case of paper CRF retrieved from the sites. Usually, double data entry is performed wherein the data is entered by two operators separately.[8] The second pass entry (entry made by the second person) helps in verification and reconciliation by identifying the transcription errors and discrepancies caused by illegible data. Moreover, double data entry helps in getting a cleaner database compared to a single data entry. Earlier studies have shown that double data entry ensures better consistency with paper CRF as denoted by a lesser error rate.[9]

5.5 Data Validation

Data validation is the process of testing the validity of data in accordance with the protocol specifications. Edit check programs are written to identify the discrepancies in the entered data, which are embedded in the database, to ensure data validity. These programs are written according to the logic condition mentioned in the DVP. These edit check programs are initially tested with dummy data containing discrepancies. Discrepancy is defined as a data point that fails to pass a validation check. Discrepancy may be due to inconsistent data, missing data, range checks, and deviations from the protocol. In e-CRF based studies, data validation process will be run frequently for identifying discrepancies. These discrepancies will be resolved by investigators after logging into the system. Ongoing quality control of data processing is undertaken at regular intervals during the course of CDM. For example, if the inclusion criteria specify that the age of the patient should be between 18 and 65 years (both inclusive), an edit program will be written for two conditions *viz.* age <18 and >65. If for any patient, the condition becomes TRUE, a discrepancy will be generated. These discrepancies will be highlighted in the system and Data Clarification Forms (DCFs) can be generated. DCFs are documents containing queries pertaining to the discrepancies identified.

5.6 Discrepancy Management

This is also called query resolution. Discrepancy management includes reviewing discrepancies, investigating the reason, and resolving them with documentary proof or declaring them as irresolvable. Discrepancy management helps in cleaning the data and gathers enough evidence for the deviations observed in data. Almost all CDMS have a discrepancy database where all discrepancies will be recorded and stored with audit trail.

Based on the types identified, discrepancies are either flagged to the investigator for clarification or closed in-house by Self-Evident Corrections (SEC) without sending DCF to the site. The most common SECs are obvious spelling errors. For discrepancies that require clarifications from the investigator, DCFs will be sent to the site. The CDM tools help in the creation and printing of DCFs. Investigators will write the resolution or explain the circumstances that led to the discrepancy in data. When a resolution is provided by the investigator, the same will be updated in the database. In case of e-CRFs, the investigator can access the discrepancies flagged to him and will be able to provide the resolutions online. Fig 2 illustrates the flow of discrepancy management.

Fig 2 : Discrepancy management (DCF = Data clarification form, CRA = Clinical Research Associate, SDV = Source document verification, SEC = Self-evident correction)

The CDM team reviews all discrepancies at regular intervals to ensure that they have been resolved. The resolved data discrepancies are recorded as 'closed'. This means that those validation failures are no longer considered to be active, and future data validation attempts on the same data will not create a discrepancy for same data point. But closure of discrepancies is not always possible. In some cases, the investigator will not be able to provide a resolution for the discrepancy. Such discrepancies will be considered as 'irresolvable' and will be updated in the discrepancy database.

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.06.12 Published on 05 (8) 2016

Discrepancy management is the most critical activity in the CDM process. Being the vital activity in cleaning up the data, utmost attention must be observed while handling the discrepancies.

5.7 Medical Coding

Medical coding helps in identifying and properly classifying the medical terminologies associated with the clinical trial. For classification of events, medical dictionaries available online are used. Technically, this activity needs the knowledge of medical terminology, understanding of disease entities, drugs used, and a basic knowledge of the pathological processes involved. Functionally, it also requires knowledge about the structure of electronic medical dictionaries and the hierarchy of classifications available in them. Adverse events occurring during the study, prior to and concomitantly administered medications and pre-or co-existing illnesses are coded using the available medical dictionaries. Commonly, Medical Dictionary for Regulatory Activities (MedDRA) is used for the coding of adverse events as well as other illnesses and World Health Organization–Drug Dictionary Enhanced (WHO-DDE) is used for coding the medications. These dictionaries contain the respective classifications of adverse events and drugs in proper classes. Other dictionaries are also available for use in data management (Eg, WHO-ART is a dictionary that deals with adverse reactions terminology). Some pharmaceutical companies utilize customized dictionaries to suit their needs and meet their standard operating procedures.

Medical coding helps in classifying reported medical terms on the CRF to standard dictionary terms in order to achieve data consistency and avoid unnecessary duplication. For example, the investigators may use different terms for the same adverse event, but it is important to code all of them to a single standard code and maintain uniformity in the process. The right coding and classification of adverse events and medication is crucial as an incorrect coding may lead to masking of safety issues or highlight the wrong safety concerns related to the drug.

5.8 Database Locking

After a proper quality check and assurance, the final data validation is run. If there are no discrepancies, the SAS datasets are finalized in consultation with the statistician. All data management activities should have been completed prior to database lock. To ensure this, a pre-lock checklist is used and completion of all activities is confirmed. This is done as the database cannot be changed in any manner after locking. Once the approval for locking is obtained from all stakeholders, the database is locked and clean data is extracted for statistical analysis. Generally, no modification in the database is possible. But in case of a critical issue or for other important operational reasons, privileged users can modify the data even after the database is locked. This, however, requires proper documentation and an audit trail has to be maintained with sufficient justification for updating the locked database. Data extraction is done from the final database after locking. This is followed by its archival.

5.9 Roles and Responsibilities in CDM

In a CDM team, different roles and responsibilities are attributed to the team members. The minimum educational requirement for a team member in CDM should be graduation in life science and knowledge of computer applications. Ideally, medical coders should be medical graduates. However, in the industry, paramedical graduates are also recruited as medical coders. Some key roles are essential to all CDM teams. The list of roles given below can be considered as minimum requirements for a CDM team:

- Data Manager
- Database Programmer/Designer
- Medical Coder
- Clinical Data Coordinator
- Quality Control Associate
- Data Entry Associate

The data manager is responsible for supervising the entire CDM process. The data manager prepares the DMP, approves the CDM procedures and all internal documents related to CDM activities. Controlling and allocating the database access to team members is also the responsibility of the data manager. The database programmer/designer performs the CRF annotation, creates the study database, and programs the edit checks for data validation. He/she is also responsible for designing of data entry screens in the database and validating the edit checks with dummy data. The medical coder will do the coding for adverse events, medical history, co-illnesses, and concomitant medication administered during the study. The clinical data coordinator designs the CRF, prepares the CRF filling instructions, and is responsible for developing the DVP and discrepancy management. All other CDM-related

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.06.12 Published on 05 (8) 2016

documents, checklists, and guideline documents are prepared by the clinical data coordinator. The quality control associate checks the accuracy of data entry and conducts data audits.[10] Sometimes, there is a separate quality assurance person to conduct the audit on the data entered. Additionally, the quality control associate verifies the documentation pertaining to the procedures being followed. The data entry personnel will be tracking the receipt of CRF pages and performs the data entry into the database.

VI. CONCLUSION

CDM has evolved in response to the ever-increasing demand from pharmaceutical companies to fast-track the drug development process and from the regulatory authorities to put the quality systems in place to ensure generation of high-quality data for accurate drug evaluation. To meet the expectations, there is a gradual shift from the paper-based to the electronic systems of data management. Developments on the technological front have positively impacted the CDM process and systems, thereby leading to encouraging results on speed and quality of data being generated. At the same time, CDM professionals should ensure the standards for improving data quality. [11] CDM, being a specialty in itself, should be evaluated by means of the systems and processes being implemented and the standards being followed. The biggest challenge from the regulatory perspective would be the standardization of data management process across organizations, and development of regulations to define the procedures to be followed and the data standards. From the industry perspective, the biggest hurdle would be the planning and implementation of data management systems in a changing operational environment where the rapid pace of technology development outdates the existing infrastructure. In spite of these, CDM is evolving to become a standard-based clinical research entity, by striking a balance between the expectations from and constraints in the existing systems, driven by technological developments and business demands.

Reference

[1]. Gerritsen MG, Sartorius OE, vd Veen FM, Meester GT. Data management in multi-center clinical trials and the role of a nation-wide computer network. A 5 year evaluation. Proc Annu Symp Comput Appl Med Care. 1993:659–62.

[2]. Lu Z, Su J. Clinical data management: Current status, challenges, and future directions from industry perspectives. Open Access J Clin Trials. 2010;2:93–105.

[3]. CFR - Code of Federal Regulations Title 21 [Internet] Maryland: Food and Drug Administration.[Updated 2010 Apr 4; Cited 2011 Mar 1]. Available from:http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=11.10 .

[4]. Study Data Tabulation Model [Internet] Texas: linical Data Interchange Standards onsortium. c2011. [Updated 2007 Jul; Cited 2011 Mar 1]. Available from: http://www.cdisc.org/sdtm .

[5]. CDASH [Internet] Texas: Clinical Data nterchange Standards Consortium. c2011. [Updated 2011 Jan; Cited 2011 Mar 1]. Available from: http://www.cdisc.org/cdash .

[6]. Fegan GW, Lang TA. Could an open-source clinical trial data-management system be what we have all been looking for? PLoS Med. 2008;5:e6.

[7]. Kuchinke W, Ohmann C, Yang Q, Salas N, Lauritsen J, Gueyffier F, et al. Heterogeneity prevails: The state of clinical trial data management in Europe□-□results of a survey of ECRIN centres. Trials. 2010;11:79.

[8]. Cummings J, Masten J. Customized dual data entry for computerized data analysis. Qual Assur.1994;3:300–3.

[9]. Reynolds-Haertle RA, McBride R. Single vs. double data entry in CAST. Control Clin Trials.1992;13:487–94.

[10]. Ottevanger PB, Therasse P, van de Velde C, Bernier J, van Krieken H, Grol R, et al. Quality assurance in clinical trials. Crit Rev Oncol Hematol. 2003;47:213–35.

[11]. Haux R, Knaup P, Leiner F. On educating about medical data management - the other side of the electronic health record. Methods Inf Med. 2007;46:74–9.