Introduction to Data Management For Clinical Research Studies

Contents

Course Objectives and Contents

Objectives

Upon completion of this course, you will have an understanding of:

- What data management is and the purpose of a data management plan
- The factors to be considered in the design and type of a case report form
- What should be taken into account when completing and data entering a case report form and closing the data set
- Considerations for data analysis
- What is important when deciding on a data management system

Contents

- Contents

- Introduction
- The Data Management Plan
- Case Report Form Design
- Paper or Electronic Case Report Forms?
- Case Report Form Completion
- Choosing a Data Management System
- Data Entry
- Data Errors and Resolution
- Quality Assurance, Data cleaning and Locking
- Data Analysis
- Data Archiving
- Staff
- Summary

- Key Points to Remember

- References & Resources

This section provides the references and resources used in this course which you may find useful for further reading.

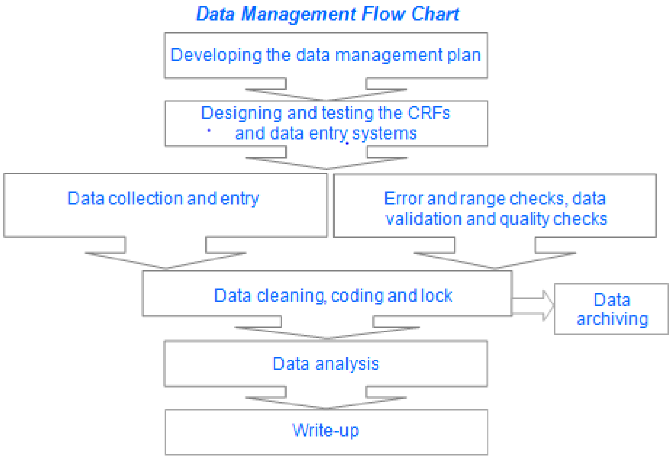
- Quiz

This section provides questions that will allow you to test what you have learned from the course.

Introduction

Data management in clinical research relates to the processes of gathering, capturing, monitoring, analysing and reporting on data. Data management begins with the development of the data management plan and design of the data capture instrument (e.g. the case report form), continues with data collection and regular quality control procedures, the database cleaning, locking and ends with the analysis, archiving and write-up.

Good data management requires proper planning and as McFadden (2007) states '*in parallel with the development of the protocol, the data to be collected to answer the study objectives should be defined'*. Friedman et al (1998) pointed out that '*no study is better than the quality of its* data'. This statement highlights the importance of capturing good quality data that is valid, auditable, and accurate, which can easily be replicated, and measures the intended variables in the research question. As high data quality is essential, recording the study data is the most crucial stage of the data management process. Therefore, developing the data management practices alongside the protocol ensures that all of the protocol-specified data are accurately captured on the case report form (CRF).


Data Management Flow Chart
Developing the data management plan
Designing and testing the CRFs and data entry systems
Data collection and entry
Error and range checks, data validation and quality checks
Data cleaning, coding and lock
Data archiving
Data analysis
Write-up

The objectives of good clinical data management are to ensure that the study database is:

- An accurate and true representation of what took place in the study
- sufficiently clean to support the statistical analysis and its interpretation

The European Clinical Research Infrastructures Network has written some helpful and straightforward guidance on good clinical practice (GCP) compliant data management in multinational clinical studies. Their website address is supplied in the 'Resources' section of this course.

The Data Management Plan

The Data Management Plan is a very important piece of study documentation. Depending on the study the plan may be made up of several documents. It should give a complete picture of how the data will be handled throughout the study by outlining all of the information relating to the study's data management procedures. These should include:

- database structure specifications
- a description of the database building and testing procedures
- a list of SOPs for the data management processes (e.g. data entry and data validation)

which will be used to ensure consistency
- a description of how the data will be reviewed (e.g. data queries and resolution processes) and information about how changes in data will be managed
- details of how the data will be coded, analysed and archived

The plan should also describe the roles and responsibilities of all staff involved in gathering or handling the study data. It is a valuable tool both for study planning and in case of auditing.

The Clinical Data Acquisition Standards Harmonization (CDASH) collated input from a spectrum of clinical trialists, including data managers, statisticians, investigators, monitors and study coordinators, and developed a set of consensus-based data collection standards. CDASH created conventions for individual data collection fields and sets out standards for element name, definition, and metadata. The web address for the CDASH guidance can be found in the 'Resources' section of this course. CDASH will become the standard for regulatory submissions in the USA and possibly Europe. Researchers setting up new data management systems may wish to consider following these standards as this will help their studies to be fairly considered alongside those projects already complying.

The Association for Clinical Data Management has prepared a very comprehensive data management plan template. Though it is primarily aimed at commercial study operations it could be adapted for use in non-commercial studies. The web address for the template can be found in the 'Resources' section of this course.

Case Report Form Design

A CRF (also known as a data collection form) turns the protocol into the data capture system. The CRF is essential in obtaining accurate and complete study data. The1996 International Conference on Harmonization Principles of Good Clinical Practice (ICH GCP) guidelines define a CRF as a printed, optical, or electronic document that *'is designed to record all of the protocol required information to be reported to the sponsor on each study subject*'.

The CRF design should run parallel to protocol development. In section 3 of CDASH's Draft Guidance V1 there is some useful guidance on 'Best Practices for Creating Data Collection Instruments'. Some of the main points of this document are:

- Necessary data only: collect only data that will be used for analysis and avoid collecting redundant data. The protocol team should draft a statistical analysis plan to define what data is essential. .
- CRFs should record sufficient identifiers to ensure that data can unambiguously be assigned to the correct participant but this needs to be balanced against data protection and anonymity requirements
- Control and document the process of CRF design, printing and distribution. Create standard operating procedures for CRF design, development, quality assurance, approvals, version control and site training.
- Ensure that all members of the study team have adequately reviewed the CRFs before they are finalised.
- Pilot the CRFs if at all possible.
- Keep the end-user in mind and consider the workflow at the study site so that CRF is quick and easy for site personnel to complete. Also, consider the source data: will there be reliable medical charts at the site or will site staff require study-specific worksheets in which to record study observations or measurements; what will the lab reports look like; is it possible to use a central laboratory to ensure consistent results?
- Employ standards for data collection and use CDASH standards wherever possible.
- Use standardised and validated tools for the collection of qualitative data wherever possible (e.g. the Euroqol group's EQ-5D which is a standardised instrument for use as a measure of health outcome.)
- Keep the CRF questions clear and unambiguous and ensure that they are not 'leading'
- Wherever possible, avoid collecting 'free text' as it requires coding before it can be

analysed. It is preferable to use ‘yes/no’ checkboxes or to provide a pre-defined list of possible responses.
- Ensure that translated CRFs are prepared using the same development process as the originals and are reviewed to ensure that the questions have a consistent meaning in all languages.
- Prepare CRF completion guidelines to assist site personnel in completing the forms.

Paper or Electronic Case Report Forms?

The choice to use paper or electronic CRFs will be largely influenced by the data management system in use and the local infrastructure. Pavlovic et al. (2009) showed that electronic CRFs are considerably less expensive to process because it can be faster (no data entry required as there is with paper data capture), save on staffing, space, transportation and cost because of the lack of paper documentation. However, factoring in the additional cost of computers or mobile devices and internet or GPRS access may mean that paper is cheaper for use in remote sites. Other drawbacks may include:

- difficulty in checking the source data (so extra monitoring may be needed)
- no records of the visits or measurements from which the data were taken
- the need for training of staff in computer skills and the relevant programmes, which can take additional time and add additional expense and may be less appropriate for resource-poor settings.

As McFadden (2007) points out ‘*if using paper data capture forms, there is a need for systems for (a) distributing blank paper forms to the participating sties and (b) for returning completed forms in a timely way to the coordinating centre. If data is captured electronically, hardware and software must be developed and fully tested and validated’.* When using paper CRFs, it is important to think ahead regarding the flow of data and to account for the extra time that will be required for data entry. For multi-site study’s, it is also important to consider the time and expense required for the CRFs to be shipped to and from the coordinating centres. The US Food and Drug Administration (FDA) 2010 draft guidance to ‘Electronic Source Documentation in Clinical Investigations’ can be found in the ‘Resources’ section of this course.

Once study enrolment begins every effort should be made to enter the participant information onto the study database as quickly as possible to enable accurate tracking of study progress and monitoring of safety data. It is also of utmost importance to track the flow of paper CRFs from the study sites to the coordinating centre and to have a system for chasing any that are not returned within a specified time period.

Case Report Form Completion

The CRFs are completed by the investigator, a clinician, nurse, laboratory staff, etc. using the relevant source documents. Defining the source data used to complete the CRFs is important, often this can be recorded on the CRF itself. Source data can include:

- medical notes
- scans
- X-rays
- bedside test forms
- nursing notes
- prescription charts
- fluid charts
- randomisation forms
- laboratory tests

When completing CRFs it is important to remember to write legibly in black, to complete all the relevant data fields (ensuring this includes the participant’s study ID) and, if providing necessary additional information, that comments are clear and concise. Only authorised site staff should complete and make corrections to the CRF. Each CRF should be signed and dated by the person that completed it.

Choosing a Data Management System

The data management system is typically a computer system used for data entry, editing, storage, and transmission into an analysis package. There is a wide variety of software that can be used for processing clinical study data. The choice of which software package to use is usually based on the complexity of data handling required and the costs of the different systems.

A minimum requirement for handling study data is that ‘clinical study’ specific software is used as it is important both to have an audit trail (where the source data is easily identifiable, errors and omissions are explained, any changes to the data is justified, etc.) and for the data to be secure (compliance with standards such GCP principles helps ensure that basic data privacy issues are adhered to). Generic spread sheet and database software such as Excel and Access do not meet these requirements, for example Excel cannot handle longitudinal data as successfully as the other programmes, it does not allow for an audit trial, it requires programming for most tasks which can be time consuming, and it does not allow for importing or merging data files, defining data types, or selecting subsets of data.

OpenClinica, is a free, open source, web-based electronic data capture data management system that is fully compliant with ‘US Food and Drug Administration’ (US FDA) 21 CFR Part 11 and can export data in Clinical Data Interchange Standards Consortium (CDISC) format. The CDISC is a non-profit organization, whose 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*’. CDSIC has developed the ‘Study Data Tabulation Model’ which is intended for standardisation of data submission to the US FDA. The CDSIC web address can be found in the ‘Resources’ section of this course.

Choosing a Data Management System

Security can be implemented to electronic data management systems (such as OpenClinica) by creating ‘password only’ access which limits the users to only those with the authority to work on the data. With OpenClinica, different access level accounts can be created to allow different users different access rights depending on the tasks they need to perform on the database.

Fegan and Lang (2008) suggest that OpenClinica might be a useful platform for clinical study data management in resource limited settings. A useful feature of this program is that it allows translation of CRFs into multiple languages. The OpenClinica web address is provided in the ‘Resources’ section of this course.

There is development being carried out on software to enable remote data collection in OpenClinica using mobile devices. The EpiHandy software could be an excellent solution for electronic data collection in areas that do not have access to broadband internet but do have mobile phone coverage. The Epihandy web address is provided in the ‘Resources’ section of this course.

Many study sponsors and funding agencies request that the date management system is compliant with the US FDA Title 21 Code of Federal Regulations Part 11 requirements which are:

- System validation
- Robust audit trail
- Security access controls
- Specification for system design and edit checks
- Archiving procedures
- Electronic signatures

Though these requirements may be exacting the result is that the data validity is assured, which is a strong incentive for having a robust data management system.

Data Entry

There is much debate on whether studies should use single or double data entry. In a double entry system two separate users independently enter the data from each paper CRF, the system then compares the two records and will only accept them if they are identical. A system must be in place for resolving discrepancies. The purpose of double data entry, as Shirima et al. (2007) explain is *'to ensure that the data available in the database are a true reflection of the information on the forms, but the original forms are usually archived in case of subsequent query'.* However Buchele et al. (2005) found in their research that whilst '*double data entry would only marginally enhance data quality……. expressed in monetary terms, double data entry increases cost by a factor of about 2.5 in comparison with single data entry'.* An alternative to double data entry is data validation, where another user confirms that what has been entered is consistent with what is recorded on the paper CRF. This might be limited to key data fields related to study endpoints.

The use of an electronic database system is recommended as it is more secure, provides an audit trail and allows for easier manipulation of the data for quality checks, reporting, etc. Using a system such as OpenClinica allows for data verification, consistency and range checks to be undertaken at the data entry stage. If data is double entered, comparisons and reconciliation should be regularly undertaken by the study coordinator or data manager to ensure that subjective data entry errors are reduced. Using the data management system to make fields mandatory allows for additional checks for missing data. Additional range and consistency checks should be performed on a regular basis once the datasets for analysis are constructed. All variables should be examined for unusual, outlying, unlabelled or inconsistent values.

Data Errors and Resolution

Errors can occur at several stages of the data collection and data entry process, and these errors must be captured, checked and amended through an audit-trailed system. Data queries are raised when there is a discrepancy in the data e.g. when a male is listed as being pregnant. The study coordinator or data manager will check the CRFs as they are received and chase any omissions, errors or inconsistencies using a 'variance' form which will record reasons for any corrections made. Variance forms must be kept for audit purposes.

Machin and Fayers (2010) stress that it is essential to chase missing data as soon as possible because '*a major cause for concern is that the lack of these data may result in bias, and that the apparent results of a clinical trial will then not reflect the true situation. We will not know if the difference we observe (or lack thereof) between treatments is a truly reliable estimate of the real difference'.*

Any discrepancies and/or problems with study data should be addressed immediately by the study coordinator, monitor, data manager, or statistician, as appropriate. To the best possible extent, data queries should be resolved. It is generally accepted that due to administrative reasons and data availability a small number of problems will continue to exist however, these must be noted on the relevant discrepancy notes within the data management system.

Quality Assurance, Data Cleaning and Locking

A range of quality assurance steps should be in place throughout the entire data management process to catch errors prior to data analysis, for example performing monthly error checks following data entry and monitoring data during review. Additionally, samples of CRFs should be reviewed for accuracy and completeness against the source.

Once the study is closed, a final data cleaning exercise should be undertaken. This involves chasing any essential data missing from the CRFs (e.g. in the case of a birthing survey this could be whether the birth outcome was a live or still birth) and resolving out of range or logical issues and inconsistencies. Hackshaw (2009) explains that once as much data as possible has been received and data entered, any data which remains missing or data issues which remain unresolved will be ‘coded’ to indicate, for the purpose of analysis, that that data item was missing, invalid, etc. After all the data has been entered, the database will be locked. Once a database is locked no data can be added to, or edited within, the system.

Data Analysis

The study protocol must state the project’s statistical objectives (both primary and secondary). The statistical analysis plan should explain how the study data will be analysed. The analysis plan outline can be altered during the course of the study, for example after requests for interim analyses, but it should be finalised before the database is ready for the full data analysis as this avoids the risk of the data being manipulated. A statistical analysis plan has the advantage of preventing unplanned primary analyses at the end of the study but it should not stop further examination if unusual features of the data are identified during the analyses.

Data analysis is carried out so that *‘under some underlying probability distribution assumption, a statistical inference can be derived based on clinical data collected from a representative sample of the targeted patient populations* (Chow and Liu 1994). They go on to explain that ‘*to provide a valid statistical assessment of the uncertainty with a desired accuracy and reliability, statistical and/or estimation procedures should possess the properties of unbiasedness and least variability whenever possible.*

It is important that if the study does not have a dedicated statistician that, at least, the advice and input of one is sought during the drafting of the statistical analysis plan and during the data analysis stages.

Data Archiving

After data entry is completed, the CRFs should be filed according to the participant’s study ID number and by site (if it’s a multi-centred study) in the study coordinating centre. If there is more than one CRF to be completed for each participant these additional CRFs should be stored together with the previously received CRFs.

Copies of the source documents should be stored in each of the study sites (as they may have participant identifiers) for a minimum period of time which will be specified in the study protocol.

Data storage and CRF archiving should follow the principles laid out in the 1996 ICH GCP guidelines and any relevant regulatory or Sponsor requirements.

Staff

Data management staff can include a range of roles, including (but not limited to) filing clerks, data capturers, monitors, data managers, and statisticians.

To ensure high quality data and to comply with general GCP principles, these staff members should be suitably trained for their role, with up-to-date CVs.

An up-to-date list of data management staff should also be present in the delegation/responsibility log which details the level of data access/records to which each staff member is allowed.

Summary

Cox Gad (2009) warns that ‘t*he process of data management is complex, error prone and vital to reliable accurate data’*. Therefore good data management practices are vital to the success of a study as they help to ensure that the data collected is complete and reliable.

It is important to get the procedures for data collection correct from the start, as later on there is little you can do to improve the data quality.

Key Points to Remember

- Data management in clinical research relates to the processes of gathering, capturing, monitoring, analysing and reporting on data.
- It is essential to capture good quality data that is valid, auditable, and accurate, which can easily be replicated, and measures the intended variables in the research question.
- The data management plan and CRF design should evolve alongside the development of the protocol.
- The objectives of good clinical data management are to ensure that the study database is an accurate and true representation of what took place in the study and is sufficiently clean to support the statistical analysis and its interpretation.
- The data management plan describes the database structure, the procedures to be used for system testing, the validation, data entry, edit checks, data coding, data queries and query resolution, as well as outlines the roles and responsibilities of all staff.
- A CRF turns the protocol into the data capture system and is essential in obtaining accurate and complete study data.
- When choosing an electronic data management system, it is important that it allows for an audit trail and that the system is sufficiently secure.
- The choice to use paper or electronic CRFs will be largely influenced by the data management system in use and the local infrastructure
- A minimum requirement for handling study data is that ‘clinical study’ specific software is used as it is important both to have an audit trail and for the data to be secure.
- Errors can occur at several stages of the data collection and data entry process, and these errors must be captured, checked and amended through a robust audit system.
- Once recruitment begins the participant information should be data entered as quickly as possible to enable accurate tracking of study progress and monitoring of safety data.
- Only authorised site staff should complete and make corrections to the CRF.
- Any problems with study data should be addressed immediately by the study coordinator, monitor, data manager, or statistician, as appropriate.
- A range of quality assurance steps should be in place throughout the entire data management process to catch errors prior to data analysis.
- Once all data chasing, validation and cleaning is completed the database is locked so that nothing further can be added to, or amended within, it.
- It is important that if the study does not have a dedicated statistician that, at least, the advice and input of one is sought during the drafting of the statistical analysis plan and during the data analysis stages.
- The process of data management is complex, error prone and vital to reliable accurate data’. Therefore good data management practices are vital to the success of a study as they help to ensure that the data collected is complete and reliable.

References And Resources

References

1. Büchele G. Och G, Bolte G, Weiland S K. Single vs. double data entry. Epidemiology, 2005. 16(1): p. 130-131.
2. Cox Gad S (Editor). Clinical Trials Handbook 2009. Wiley & Sons Inc., New Jersey.
3. Chow SC and Liu JP. Design and Analysis of Clinic Trials: Concepts and Methodologies 2nd Ed. 2004. Wiley & Sons, New Jersey.
4. Fegan, G.W. and T.A. Lang, Could an open-source clinical study data-management system be what we have all been looking for? PLoS Med, 2008. 5(3): p. e6.
5. Friedman L, Furberg C and DeMets D. Fundamentals of Clinical Trials 4th Ed. 1998. Springer, New York.
6. Hackshaw A A. Concise Guide to Clinical Trials 2009. Wiley-Blackwell, Chicester.
7. International Conference on Harmonization Principles of Good Clinical Practice (ICH GCP) 1996 guidelines.
8. Machin D and Fayers PM. Randomized Clinical Trials: Design, Practice and Reporting 2010. Wiley-Blackwell, Chicester.
9. McFadden E. Management of Data in Clinical Trials 2nd ed. 2007. Wiley & Sons Inc., New Jersey.
10. Pavlovic, I., T. Kern, and D. Miklavcic, Comparison of Paper-based and Electronic Data Collection Process in Clinical Studies: Costs simulation study. Contemporary Clinical Studies, 2009. 30(4): p. 300-316.
11. Shirima K, Mukasa O, Armstrong Schellenberg J, Manzi F, John D, Mushi A, MrishoM, Tanner M, Mshinda H and Schellenberg D. The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania: Emerging Themes in Epidemiology 2007; 4(5): 1-8.

Resources

1. Association for Clinical Data Management data management plan template
2. CDSIC Study Data Tabulation Model
3. Clinical Data Acquisition Standards Harmonization Draft CDASH guidance V1.0
4. OpenXData Software for data collection
5. Euroqol group’s EQ-5D which is a standardised instrument for use as a measure of health outcome
6. European Clinical Research Infrastructures Network (ECRIN)
7. Food and Drug Administration (FDA). Guidance for Industry Part 11, Electronic Records; Electronic Signatures - Scope and Application 2003
8. Food and Drug Administration (FDA). Guidance for Industry Computerized Systems Used in Clinical Investigations 2007
9. International Conference on Harmonization Principles of Good Clinical Practice (ICH GCP) 1996 guidelines
10. OpenClinica web-based electronic data capture data management system
11. US Food and Drug Administration 2010 draft guidance to ‘Electronic Source Documentation in Clinical Investigations’

Quiz

Summary

1. Which of the following topics is not an aspect of data management:
    - ○ Designing the CRF
    - ○ Planning data collection

- ○ Developing quality control procedures
- ○ Designing the statistical analysis plan
- ○ Drafting data staff contracts
- ○ Planning database cleaning and locking

2. The data management plan and study CRFs should be designed alongside:
    - ○ The study handbook
    - ○ The investigators brochure
    - ○ The standard operating procedures
    - ○ The protocol
3. An objective of good clinical data management is to ensure the database reflects accurately the data collected in the study.
    - ○ True
    - ○ False
4. The data management plan does not cover the procedures used for:
    - ○ System testing
    - ○ Data validation
    - ○ Consenting participants
    - ○ Edit checks
    - ○ Data coding
    - ○ Chasing and query resolution
5. The CRF should be designed to collect additional information to allow for extra analysis at a later date.
    - ○ True
    - ○ False
6. A CRF can be (please select all that apply):
    - □ Open to the interpretation of those completing it
    - □ Longer than four pages
    - □ Time consuming to complete
    - □ Translated into other languages
7. Paper CRFs are always preferable over electronic CRFs.
    - ○ True
    - ○ False
8. Corrections can be made to the CRFs:
    - ○ As and when needed by whoever discovers the error
    - ○ Never, no matter what the reason
    - ○ Only by authorised staff, who must document change and the reason for it
9. Data queries should be:
    - ○ Should be batched and addressed at the end of the study
    - ○ Systematically coded as missing
    - ○ Addressed as and when they are discovered
    - ○ Completed by the person who discovers them with an estimated value
10. The USA FDA stresses the importance of a 21 CRF Part 11 compliant EDC system to ensure the following (please select all that apply):
    - □ System validation
    - □ Web interactivity
    - □ Robust audit trail
    - □ Security access controls
    - □ Archiving procedures
    - □ Electronic signatures
11. The benefits of double data entry far outweigh the benefits of single data entry.

- ○ True
- ○ False

12. Once a database is locked, it can only be amended in the following situation:
    - ○ Missing CRFs are returned
    - ○ Additional coding is applied
    - ○ Extra data for further analysis is available
    - ○ None of the above
    - ○ All of the above

Submit