Project Identification: 2021-1-CZ01-KA220-HED-000023177

Investor/Program/Project type: European Union - Erasmus+ Key Action 2: Cooperation for innovation and the exchange of good practices; Strategic Partnerships in the field of education, training and youth

Curriculum Development of Human Clinical Trials for the Next Generation of PhD Students and Early Career Researchers in the Medical, Science, Pharmacy and Health Professions

CHAPTER 6

DATA MANAGEMENT (FAIR DATA, CRF, DATABASE) AND STATISTICAL ANALYSIS – STANDARDIZING, AND ANALYZING GENERATED DATA (STATISTICAL CONSIDERATIONS FOR THE CLINICAL TRIAL PLANNING)

Authors: Robert Herczeg
University of Pécs, Pécs, Hungary

Reviewers: Marta del Alamo, Gerd Felder, Salma Malik, Keiko Ueda
European Clinical Research Infrastructure Network (ECRIN)

Date first created: 11/10/2022

Last revision: 29/05/2024

Disclaimer: The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the Erasmus+ Programme of the European Union

Content

Time required to complete this chapter

Core content:	1h 00m
Additional/advanced content (yellow boxes):	30m
Activities/practical exercises (blue framed boxes):	00m
Total time:	**1h 30m**

1 Introduction to the chapter

This chapter highlights fundamentals of the research data management according to the FAIR Principles and standards which have been implemented worldwide. Data management (DM) is a set of steps that handle information during medical research. Its main aim is to ensure quality, integrity, and compliance with different protocols and regulations. It also helps the main stakeholders to stay on track: sponsors, research entities, and clinical centres where the data collection takes place. Effective data management is essential to ensuring high-quality data, reports, validation, and a well-designed clinical trial. Data handling and management are also well defined in the ICH-GCP: EMA/CHMP/ICH/135/1995, guideline for good clinical practice E6(R2) section 5.5 (Trial management, data handling, and record keeping). This document gives an overall view about the good clinical practice regarding of data/trial management too.

2 Introduction to data management

Effective data management means understanding where the data is, and the ability to get the data into some form where it can be appropriately managed. Besides this, it is also important to know to whom the data belongs. In a simplified data management model, there are at least the following steps/tasks (Figure 1) which are based on three main phases, namely planning, performing, and finishing. Planning involves reviewing the study protocol, managing data, developing and validating CRF and database, and training study personnel. During the performing phase, data is collected at patient sites and submitted to the database, while ensuring an audit trail, consistency checks, and query management as per GCP. The phase ends with a data lock in the database. The third phase, which includes analysis and interpretation, can only begin after the database is locked. Furthermore, the data needs to be stored for a long time for inspection and future use.

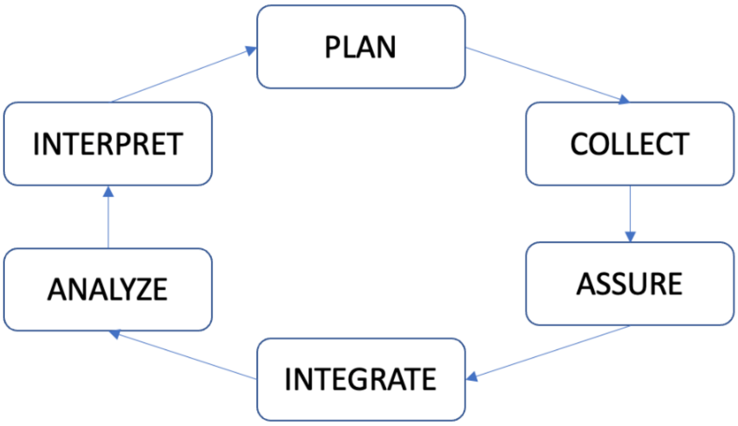


Figure 1: Simple data management model

Data management is also a multidisciplinary activity that includes, for example (Figure 2):

- Patients
- Data managers
- Nurses
- Investigators

- Biostatisticians
- Clinical managers
- etc.

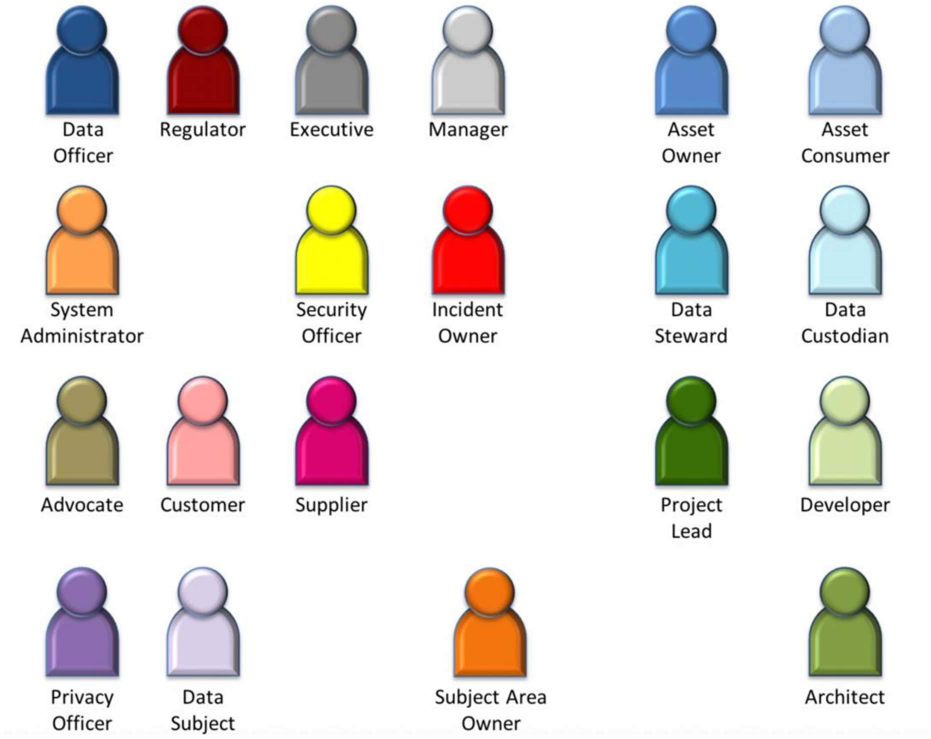


Figure 2: Data management model participants

Several types of roles involve different kinds of roles and activities as well.

2.1 Roles of data management

Several roles are relevant to data management. It is important to note that these roles are not positions, therefore context counts. Several roles can be defined based on the given trial, for example:

- data administrator,
- data owner,
- data user,
- data steward,
- data custodian,
- data analyst,
- medical coder,
- database developer,
- database designer,
- data manager.

Data access restrictions are assigned by clinical data management software depending on the users' designated roles and responsibilities. This coding ensures that there is an audit trail, and that each user may only access the necessary capabilities, without allowing them to make other changes. The protocol specifies several levels of access for each employee, including managers, programmers, administrators, medical coders, data coordinators, quality control workers, and data input personnel. These access levels may be restricted by the principal investigator using a clinical data management system (CDMS). CDMS is a tool used in clinical research to manage the data from a clinical trial. The data (from the given clinical trial) are collected from the participating sites in the case report format and are stored in the CDMS.

The exchange of data between patients and doctors is essential for the practice of medicine – and patient data are essential for medical research and progress. The following figure shows a data flow from the patient to the database and/or until a data analysis (Figure 3).

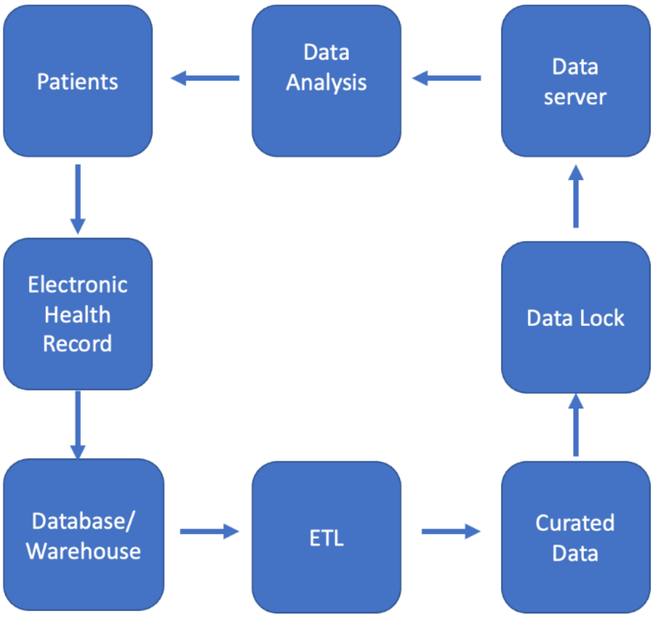


Figure 3: Simple data flow model

The first role in data management is that of the data administrator. This person is responsible for the overall management of the data, including its design, implementation, and maintenance. They are also responsible for ensuring that the data is secure and compliant with any relevant regulations.

Another important role of data management in healthcare is data security and privacy. The healthcare industry is subject to strict regulations and laws, such as the Health Insurance Portability and Accountability Act (HIPAA)[1], that govern the handling and storage of patient data. Data management professionals are responsible for ensuring that all data is protected from unauthorized access and breaches, and that patient privacy is always maintained. This includes implementing security measures such as encryption, firewalls, and regular backups to protect against data loss or corruption. Data analyst also plays a crucial part in the case of clinical data management, especially when the data should be evaluated. This person is responsible for analysing and interpreting data and using it to make informed decisions. They may also be responsible for creating reports and visualizations to help others understand the

[1] https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act

data. Data management professionals also play a critical role in data governance and quality assurance. They are responsible for ensuring that data is accurate, complete, and consistent across different systems and providers. This includes implementing data governance and quality assurance processes, such as data validation, data cleaning, and data standardization, to ensure that data is accurate, reliable, and consistent.

Data Controller and Data Processors are also important roles. The data controller is the organization or individual who determines the purposes and means of processing personal data. They are responsible for deciding what personal data will be collected, how it will be used, and who it will be shared with. In other words, the data controller is the "owner" of the data and has ultimate responsibility for ensuring that it is processed lawfully and in accordance with the relevant data protection regulations. The data controller can be an individual, a company, or any other legal entity. On the other hand, a data processor is a third-party organization or individual who processes personal data on behalf of the data controller. Data processors are responsible for carrying out the actual processing of the data and must only process the data in accordance with the instructions of the data controller. They can be companies or individuals who provide services such as hosting, data analysis, or customer support. Examples of data processors include cloud service providers, payroll companies, and IT support providers.

It is important to note that the data controller remains responsible for ensuring that personal data is processed in compliance with data protection regulations, even if the processing is carried out by a third-party data processor. Therefore, it is essential for the data controller to choose a data processor that is reliable, trustworthy, and can demonstrate compliance with relevant data protection regulations. Additionally, data controllers must have a written contract or other legal agreement in place with their data processors, outlining the data processor's obligations and responsibilities with regards to data protection.

2.2 Data quality

Problems with data quality tend to fall into two categories. The first category is related to inconsistency among data resources such as format, syntax, and semantic inconsistencies. The second category is related to poor ETL (Extract, Transform, Load) processes. Ensuring data quality is one of the most important parts of a clinical trial. Most data quality problems have been the result of poorly trained professionals, did not follow the trial protocol, did not record the required information, or poorly designed protocols. Also, worth mentioning missing and wrongly recorded data as well. These problems can be minimized with a well-defined protocol in accordance with a properly designed case report format and clear data roles during the clinical trial. Ensuring high-quality, reliable, and statistically relevant data is the main goal of clinical research. In clinical research, it is important to create and maintain organization-wide standards for data quality management to ensure consistency across/during the whole clinical trial. Data quality can be defined in different aspects therefore it is necessary to understand its dimensions (Figure 4).

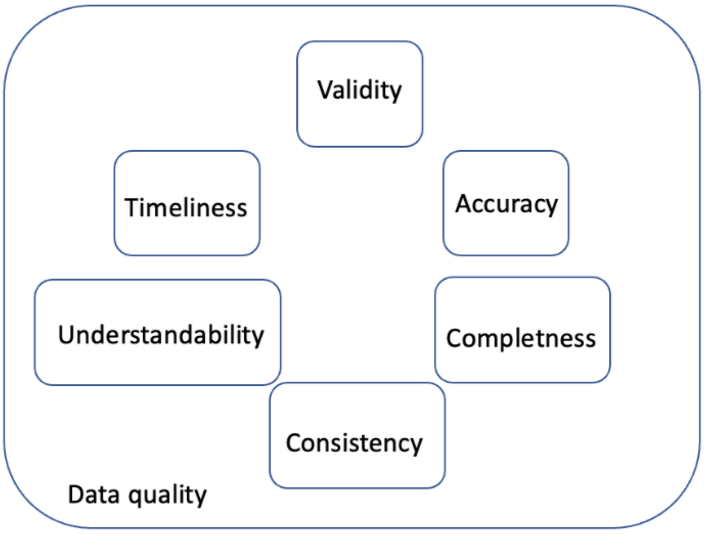


Figure 4: Data quality main dimensions

- **Validity**
  Does the data follow the relevant data guidelines? Is it follow the specific format what was defined in the data management plan?
- **Accuracy**
  Reflects how the object is described, how well the piece of data characterizes the given object.
- **Completeness**
  Does it fulfil the data's expectations? Is it in the right amount or missing?
- **Consistency**
  Does the data match with other data stored somewhere else? Are they connected?
- **Understandability**
  Is it easy to understand the raw data by the data user? Is the documentation contain the relevant explanation of the given data or data source?
- **Timeliness**
  Is it possible to up-to-date the data when it is necessary?

2.3 Risk assessment

Risk assessment is detailed in the Chapter 4.[2]

2.4 Data management standards

Data management standards play a vital role in ensuring the accuracy, completeness, and consistency of healthcare data. These standards are designed to provide a framework for the management and use of healthcare data (example, Figure 5), and to ensure that data is

[2] CONSCIOUS II Chapter 4: Quality and regulatory information; https://www.conscious2.eu/

handled and stored in a way that complies with legal and ethical standards. In this essay, we will explore the importance of data management standards in healthcare and the impact they have on patient care.

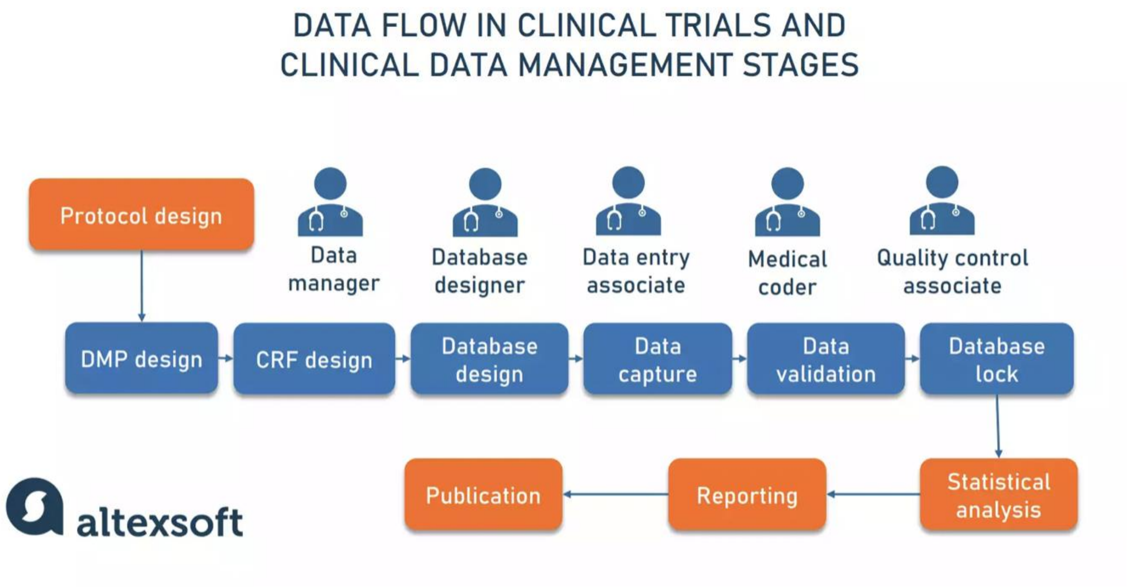


Figure 5: Data flow[3]

One of the key benefits of data management standards in clinical trials is the ability to improve the quality of data. These standards provide a framework for data collection, analysis, and storage, and ensure that data is accurate, complete, and consistent across different systems and providers. This improves the ability of healthcare providers to make informed decisions and improves patient outcomes. Another benefit of data management standards in healthcare is the ability to improve data security and privacy. Standards such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR)[4] provide guidelines for the protection of healthcare data from unauthorized access and breaches. Compliance with these standards is essential for healthcare organizations to protect patient data and to ensure that it is handled and stored in a way that complies with legal and ethical standards.

Data management standards also have an important part in data integration and interoperability. Standards such as the Health Level Seven International (HL7) and the Integrating the Healthcare Enterprise (IHE)[5] provide guidelines for the integration and interoperability of healthcare systems. Compliance with these standards is essential for healthcare organizations to ensure that data can be easily shared and exchanged between different systems and providers, which improves the overall quality of care. In addition, data management standards have a crucial part in data governance and quality assurance as well. Standards such as the NIST Cybersecurity Framework[6] provide guidelines for data governance and quality assurance, including recommendations for data validation, data cleaning, and data standardization. Compliance with these standards is essential for

[3] https://content.altexsoft.com/media/2022/03/word-image.png.webp
[4] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[5] https://www.hl7.org; https://www.ihe.net
[6] https://www.nist.gov/cyberframework

healthcare organizations to ensure that data is accurate, reliable, and consistent, which improves the overall quality of care.

In summary, data management standards in healthcare play a key role in ensuring the accuracy, completeness, and consistency of clinical trial data, as well as to collect data about adverse events and concomitant medications. They provide a framework for the management and use of the relevant data and ensure that data is handled and stored in a way that complies with legal and ethical standards. By following these standards, healthcare organizations can improve the quality of data, improve data security and privacy, improve data integration and interoperability, and improve data governance and quality assurance. With the growing importance of data in healthcare, compliance with data management standards is becoming increasingly essential to the healthcare industry.

2.5 Tools for data management

Data management (DM), especially in healthcare/clinical trials (CDM) is an important part of clinical research, which leads to having high-quality and reliable data. In addition, the following four main components are the most important: integrity, data quality, security, and privacy (Figure 6).

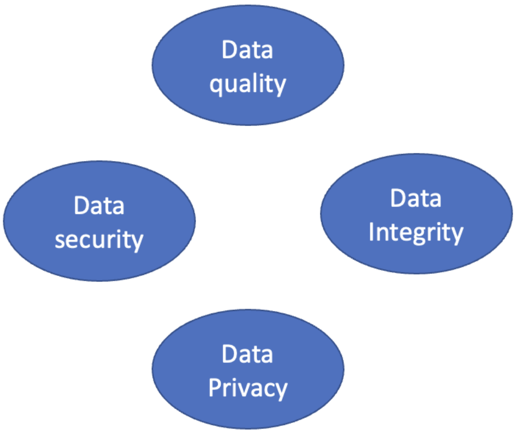
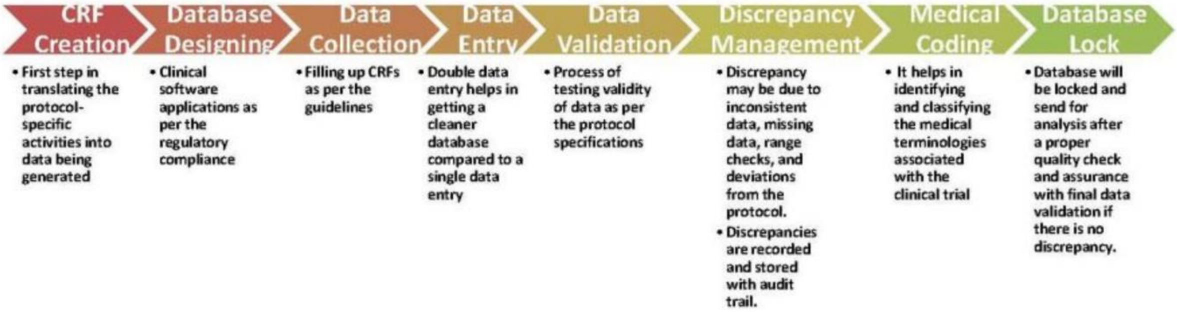


Figure 6: Data management components

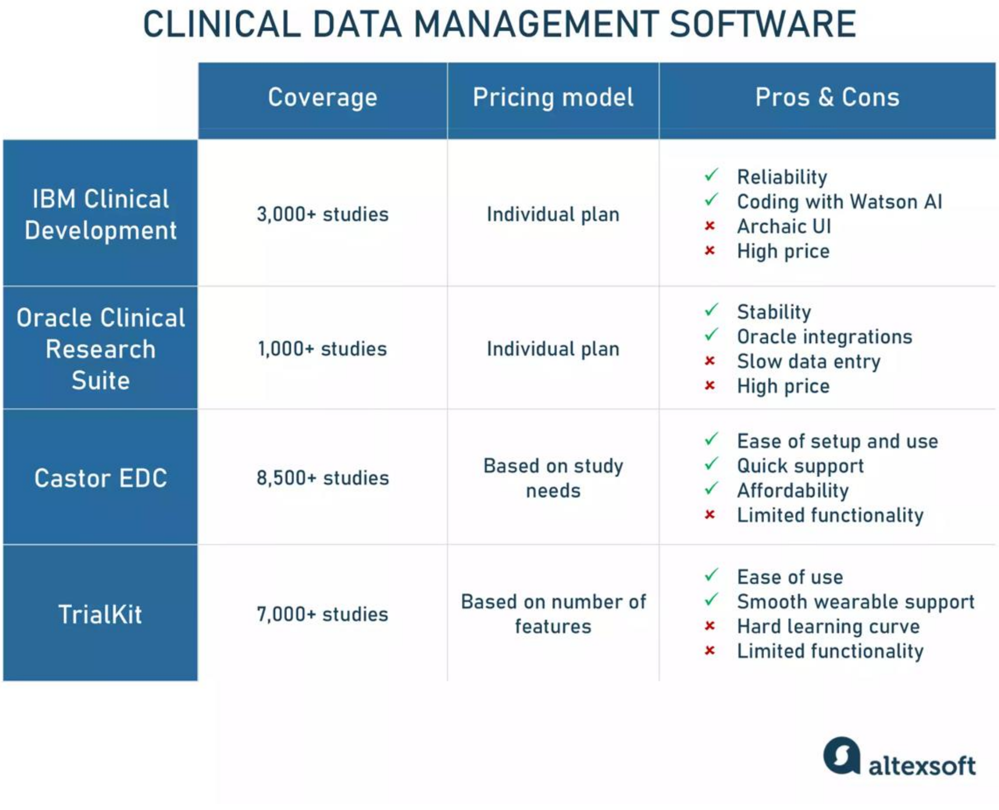
The main aim of CDM processes is to reduce errors and missing data as well as to ensure high-quality data. High-quality data should meet the defined protocol and be suitable for statistical analysis as per Statistical Analysis Plan (SAP). The CMD process begins with protocol development for the database and follows the different data processes (Figure 7).



- First step in translating the protocol-specific activities into data being generated
- Clinical software applications as per the regulatory compliance
- Filling up CRFs as per the guidelines
- Double data entry helps in getting a cleaner database compared to a single data entry
- Process of testing validity of data as per the protocol specifications
- Discrepancy may be due to inconsistent data, missing data, range checks, and deviations from the protocol.
- Discrepancies are recorded and stored with audit trail.
- It helps in identifying and classifying the medical terminologies associated with the clinical trial
- Database will be locked and send for analysis after a proper quality check and assurance with final data validation if there is no discrepancy.

Figure 7: Data management flow[7]

Spreadsheets and common office programs nowadays are obviously not enough to handle the data related to a clinical trial, and clinical data management. These programs/software are usually not able to comply with the requirements of clinical trial data management, for example: non-repudiation of origin or authentication. You need software capable of handling large amounts of customized documents and medical studies therefore, clinical data management systems (CDMS) and/or electronic data capture (EDC) systems should be applied. There are so many CDMS, EDC however, all of them have/should have common features covering basic data management operations. Below, we provide the most widely used CDMSs (Figure 8).



CLINICAL DATA MANAGEMENT SOFTWARE

	Coverage	Pricing model	Pros & Cons
IBM Clinical Development	3,000+ studies	Individual plan	✓ Reliability ✓ Coding with Watson AI × Archaic UI × High price
Oracle Clinical Research Suite	1,000+ studies	Individual plan	✓ Stability ✓ Oracle integrations × Slow data entry × High price
Castor EDC	8,500+ studies	Based on study needs	✓ Ease of setup and use ✓ Quick support ✓ Affordability × Limited functionality
TrialKit	7,000+ studies	Based on number of features	✓ Ease of use ✓ Smooth wearable support × Hard learning curve × Limited functionality

altexsoft

Figure 8: Data management software[8]

[7] https://pathbliss.com/wp-content/uploads/2018/09/14-Data-Management-Series-2-1024x363.jpg
[8] https://content.altexsoft.com/media/2022/03/word-image-2.png.webp

Besides this software, there are plenty of CDMS, which can be web-based or locally installed as well.

Clinical trials are way more complex projects, which can be documented in spreadsheets or on paper, and managing its various aspects is simply too complicated for a typical project management software. In addition, there is a need for a two steps validation in which

1) the CMDS should validated with different kinds of validation tests (IQ, OQ, Q),
2) validate the devices/instruments what are used during the clinical trial.

Individual software can follow budgets, data collection, compliance, research timelines, and more. Nonetheless, connecting these software together is most of the time not possible or it is hard to do. It is better to have this information all in one place to ensure the meta-data/data integration etc.

A good review about CDMS can be found at this link.[9] It reviews 20 tools, mostly online based, which are the top clinical trial data management top software nowadays. These systems are capable to collect different types of data depending on the therapeutic area of the clinical trial. For example, in oncology studies, the following data forms are typically used:

- Demographics,
- Medical history,
- Vital signs,
- Treatment data,
- Survival follow-up,
- Electrocardiogram data,
- Concomitant medications/ adverse events,
- Tumor assessments,
- Death information, etc.

CDM/EDC helps to record/collect various clinical trial data, such as:

- Patient's medical status and condition,
- Clinical trial safety data,
- Data from the lab reports and tests,
- Patient-related information (the quality of life),
- Other patient data that is monitored by medical equipment include blood pressure, oxygen saturation, and blood sugar levels.

3 FAIR Principles

FAIR is an acronym that stands for Findable, Accessible, Interoperable, and Reusable data (Figure 9).

[9] https://financesonline.com/clinical-trial-management/

Co-funded by the
Erasmus+ Programme
of the European Union

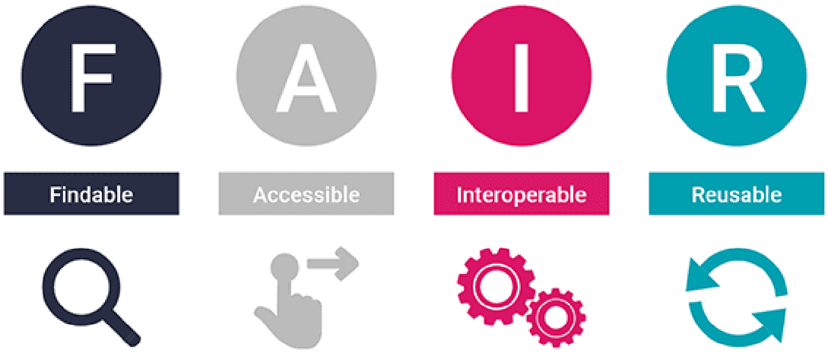


Figure 9: FAIR data

“The FAIR Data Principles[10] (Findable, Accessible, Interoperable, and Reusable), published in Scientific Data in 2016, are a set of guiding principles proposed by a consortium of scientists and organizations to support the reusability of digital assets. These principles have since been adopted by research institutions worldwide. The guidelines are timely as we see the unprecedented volume, complexity, and speed in creation of data.”

3.1 Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier.
- F2. Data are described with rich metadata (defined by R1 below).
- F3. Metadata clearly and explicitly include the identifier of the data they describe.
- F4. (Meta)data are registered or indexed in a searchable resource.

3.2 Accessible

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol.
  - A1.1 The protocol is open, free, and universally implementable.
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary.
- A2. Metadata are accessible, even when the data are no longer available.

3.3 Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

[10] https://www.nature.com/articles/sdata201618

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data.

3.4 Reusable

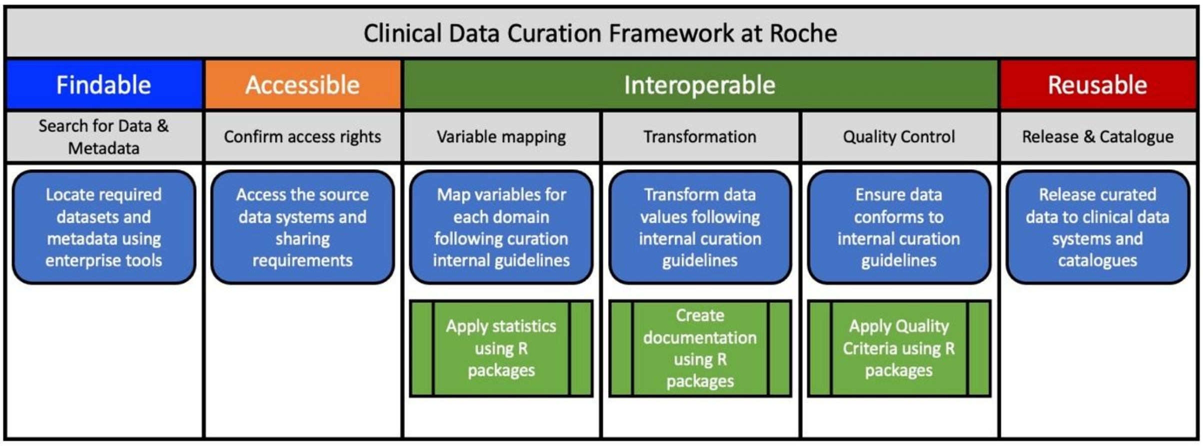
The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.
  - R1.1. (Meta)data are released with a clear and accessible data usage license.
  - R1.2. (Meta)data are associated with detailed provenance.
  - R1.3. (Meta)data meet domain-relevant community standards.

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

You can find more detail in the article: Wilkinson M, Dumontier M, Aalbersberg I et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).[11]

There is a good example of FAIR in the case of a clinical trial at Roche (Figure 10).



Clinical Data Curation Framework at Roche

Findable	Accessible	Interoperable			Reusable
Search for Data & Metadata	Confirm access rights	Variable mapping	Transformation	Quality Control	Release & Catalogue
Locate required datasets and metadata using enterprise tools	Access the source data systems and sharing requirements	Map variables for each domain following curation internal guidelines	Transform data values following internal curation guidelines	Ensure data conforms to internal curation guidelines	Release curated data to clinical data systems and catalogues
		Apply statistics using R packages	Create documentation using R packages	Apply Quality Criteria using R packages	

Figure 10: Example of FAIR[12]

As clinical trials increase in data and complexity, the FAIR principles have been widely adopted since publication. FAIR is a core element and driver for the digital transformation of drug development and life sciences research and development. FAIR is not a standard, and many different approaches exist regarding how to FAIRify data. To "become FAIR", it has to be sure that collected and associated data/metadata are both readable for humans (the clinical trial team or other researchers that might want to reuse the data) and machines (software that process and analyse the data). Thus, when incorporating the FAIR principles into a new trial,

[11] https://www.nature.com/articles/sdata201618
[12] https://fairtoolkit.pistoiaalliance.org/use-cases/fairification-of-clinical-trial-data-roche/

specific emphasis should be placed on amplifying the ability of machines to find and use data automatically and used by other researchers as well.

In a recently published paper, Perceptions and behaviour of clinical researchers and research support staff regarding data FAIRification,[13] the authors showed that 62.8% of researchers and 81.0% of support staff are currently putting forth some effort to achieve any aspect of FAIR. “Researchers are increasingly asked or obliged to make their data (more) FAIR, either by funders (e.g. Horizon Europe and The Dutch Research Council (NWO)) or their institutions (e.g., the Leiden University Medical Center and Radboud University). Research institutes often support individual researchers in the FAIRification of their data by offering help from trained research support staff (e.g., data stewards) that assist them during the FAIRification process. However, since FAIR is not a standard and the principles do not offer explicit guidance on the methodologies that should be used for data FAIRification, there are many different approaches to FAIRify data (i.e., make data FAIR). Various workflows have been published to guide researchers and researcher support staff during the FAIRification process. For the majority of the steps in these workflows, different types of expertise are required, and these steps should, therefore, be carried out in a multidisciplinary team guided by FAIR data steward(s), instead of by individual researchers or research support staff.”

To apply/use a FAIR Data ecosystem, data should be collected in concordance with metadata, and both the data and metadata should be findable in both a human- and machine-readable manner. In the cases where the data cannot be shared directly, which is often the case in medical fields due to the sensitivity of the health data, the sharing of metadata can be an alternative. Furthermore, anonymized datasets (due to GDPR) that aim to preserve the patient’s privacy can be used for sharing, or data-sharing agreements can be made to guarantee the data is handled safely. The largest open-access database for clinical trial data is clincialtrials.gov, which currently contains more than 300,000 studies across 208 countries. However, 88 % of these studies list only general data, such as the planned intervention, number of patients enrolled, and sponsors. They lack data related to the trial outcomes, which limits the usefulness of this repository. A good example of how to set up a platform to support FAIR data management (Figure 11).

[13] https://doi.org/10.1038/s41597-022-01325-2

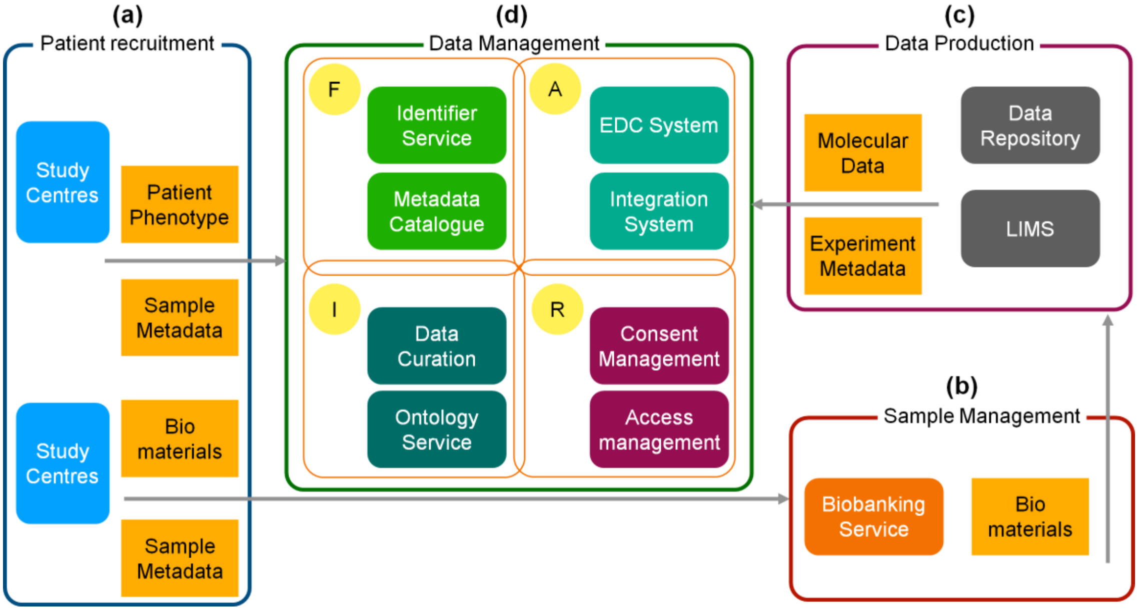


Figure 11: FAIR data management[14]

Such a data portal should also provide a secure, easy, and robust interface for the input and integration of new data from the ongoing recruitment of cohort studies. An API (Application Programming Interface) to enable programmatic access to the portal will provide structured and standardized data to bioinformaticians, statisticians, and data scientists working with clinically sized amounts of data.

3.5 Role of data repositories and databases

In healthcare, data is one of the most valuable assets that medical professionals and organizations can have. Medical data includes patient records, diagnostic images, lab test results, and more. However, managing and storing this data is a complex task that requires specialized tools and techniques. That's where data repositories and databases come into play. In healthcare, data is one of the most valuable assets that medical professionals and organizations can have. Medical data includes patient records, diagnostic images, lab test results, and more. However, managing and storing this data is a complex task that requires specialized tools and techniques. That's where data repositories and databases come into play. The role of data repositories and databases in healthcare is critical. They provide a secure, efficient, and effective way to store and manage medical data. The metadata stored in the repository includes information such as the study title, protocol summary, study design, intervention details, inclusion and exclusion criteria, primary and secondary outcomes, and recruitment status. The repository also provides information about the trial sponsor, funding sources, investigators, and study sites.

ECRIN metadata repository: The Clinical Research Metadata Repository developed by the European Clinical Research Infrastructure Network (ECRIN), aims to establish a single, searchable database where metadata about all the data objects created by clinical research is

[14] https://academic.oup.com/cardiovascres/article/117/8/1823/6220322

gathered in one place. This means collecting metadata about data objects (not the data objects themselves) from a variety of sources – trial registries, bibliographic system, data repositories, etc. – and in a variety of forms, together with some basic data on the studies themselves. The harvested data is then transformed into a consistent metadata schema, in a single data store, and that store is then available for querying through search and filter mechanisms in a web portal.

3.6 Methods of future reuse and data sharing

Over the past few years, healthcare organisations have called for increased sharing of the data generated by publicly funded research. Although participant privacy concerns necessitate special preparation (such as de-identification), data from clinical research are not excluded from this request for increased data sharing. There are scientific, economic, and ethical justifications for sharing data from clinical research. From a scientific perspective, data sharing allows for comparisons and combinations of data from various studies and facilitates meta-analysis by aggregating the data. Ethically, data sharing is a better way to recognize the generosity of clinical trial participants by increasing the usefulness of their data and enhancing the value of their contribution.

There is now widespread acceptance of the need for greater sharing of individual participant data (IPD). Sharing individual participant data (IPD) and study documents should be considered a customary aspect of best practices in clinical research and be endorsed by all parties involved, including funders, patients' groups, researchers, academic institutions, professional associations, industry, editors, and regulatory and ethics authorities. IPD sharing model should be based on the concept of data stewardship rather than data ownership. The major steps/principles involved in data sharing and reuse in clinical studies and clinical trial, could be the following (Figure 12).

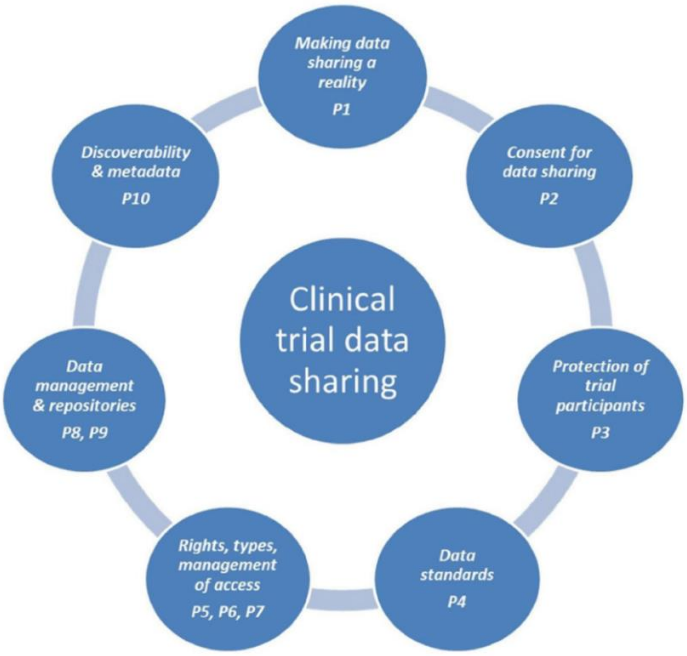

Clinical trial data sharing
Making data sharing a reality P1
Consent for data sharing P2
Protection of trial participants P3
Data standards P4
Rights, types, management of access P5, P6, P7
Data management & repositories P8, P9
Discoverability & metadata P10

Figure 12: Main principles of data sharing[15]

[15] https://bmjopen.bmj.com/content/7/12/e018647

The following list contains these 10 principles with more detailed description:

- **P1**: The provision of individual participant data should be promoted, incentivised and resourced so that it becomes the norm in clinical research. Plans for data sharing should be described prospectively and be part of study development from the earliest stages.
- **P2**: Individual participant data sharing should be based on explicit broad consent by trial participants (or if applicable by their legal representatives) to the sharing and reuse of their data for scientific purposes.
- **P3**: Individual participant data made available for sharing should be prepared for that purpose, with de-identification of data sets to minimise the risk of reidentification. The de-identification steps that are applied should be recorded.
- **P4**: To promote interoperability and retain meaning within interpretation and analysis, shared data should, as far as possible, be structured, described and formatted using widely recognised data and metadata standards.
- **P5**: Access to individual participant data and trial documents should be as open as possible and as closed as necessary, to protect participant privacy and reduce the risk of data misuse.
- **P6**: In the context of managed access, any citizen or group that has both a reasonable scientific question and the expertise to answer that question should be able to request access to individual participant data and trial documents.
- **P7**: The processing of data access requests should be explicit, reproducible and transparent, but so far as possible, should minimise the additional bureaucratic burden on all concerned.
- **P8**: Besides the individual participant data sets, other clinical trial data objects should be made available for sharing (e.g., protocols, clinical study reports, statistical analysis plans, blank consent forms) to allow a full understanding of any data set.
- **P9**: Data and trial documents made available for sharing should be transferred to a suitable data repository to help ensure that the data objects are properly prepared, are available in the longer term, are stored securely and are subject to rigorous governance.
- **P10**: Any data set or document made available for sharing should be associated with concise, publicly available and consistently structured discovery metadata, describing not just the data object itself but also how it can be accessed. This is to maximise its discoverability by both humans and machines.

The EU Clinical Trial Regulation 536/2014 recognizes the potential for data from clinical trials to be repurposed for future scientific research and emphasizes the significance of obtaining consent for using the data beyond the clinical trial's protocol, allowing individuals to withdraw their consent at any point, and establishing procedures to ensure the ethical and appropriate conduct of any secondary analyses.

Co-funded by the
Erasmus+ Programme
of the European Union

The following published articles provide deeper insight into data sharing and reuse:

- Data sharing is the future. Nat Methods 20, 471 (2023).[16]
- Khan N, Thelwall M, Kousha K. Data sharing and reuse practices: disciplinary differences and improvements needed. Online Information Review Vol. ahead-of-print, No. ahead-of-print (2023).[17]
- Ohmann C, Banzi R, Canham S et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open 7:e018647 (2017).[18]

4 Therapeutic Area User Guide (TAUG) – Data Management Plan (DMP) – Data Clean Plan (DCP)

Clinical Data Interchange Standards Consortium (CDISC) is a global not-for-profit organization that actively develops data standards with the collective knowledge and experience of volunteers within the pharmaceutical industry. CDISC data standards provide sponsors efficiency in structuring their raw data in alignment with globally accepted and required standards for successful data submissions. Therapeutic Area User Guides (TAUGs) extend the CDISC Foundational Standards to represent data that pertains to specific disease areas. TAUGs include disease-specific metadata, examples and guidance on implementing CDISC standards for a variety of uses, including global regulatory submissions.

You can find more information about TAUG's areas on the website of CDISC.[19]

4.1 DMP and DCP for clinical trials

Data management plan or DMP is a formal document that specifies how data will be handled throughout and after a research project/clinical studies/trials. It is important to highlight that DMP is a living document. It can be changed during the research due the nature of doing research. Therefore, any changes in the DMP should be recorded and review, validated by experts. DMP can consist of several elements, it is never always the same. In the following section you can see the most frequently appearing elements:

- **Types of data**: source, format, fixed or changing over time.
- **Contextual data** (metadata): document and describe the data.
- **Storage, backup, and security**: How to store and secure the data?
- **Provisions for protection/privacy**: What privacy and confidentiality issues must be addressed?
- **Policies for reuse**: How can the data reused?
- **Access and Sharing**: How can data shared, and with whom. How can data discovered?

[16] https://doi.org/10.1038/s41592-023-01865-4
[17] https://doi.org/10.1108/OIR-08-2021-0423
[18] https://doi.org/10.1136%2Fbmjopen-2017-018647
[19] https://www.cdisc.org/standards/therapeutic-areas

Co-funded by the Erasmus+ Programme of the European Union

- **Roles and plan oversight**: Who is responsible for the data management? What are the requirements for the resources?

More information and sample DMPs can be found here.[20]

Data clean plans (DCP)[21] are an essential component of effective data management in healthcare. Healthcare systems continue to adopt digital solutions, the sheer volume of data that is generated can be overwhelming. To make sense of this data, it is essential to implement a data clean plan, which is a systematic approach to ensuring data quality. A DCP involves several steps, starting with defining data quality standards. This step includes identifying the types of data that will be collected, the acceptable ranges and values, and any coding or formatting requirements. Once these standards are established, data collection can begin, and the next step is to perform data cleaning, which involves identifying and correcting errors, inconsistencies, and outliers in the data. The process of data cleaning involves a range of techniques, including visual inspection, statistical analysis, and automated tools. Visual inspection involves manually reviewing the data to identify any obvious errors or inconsistencies, such as missing values, incorrect formatting, or data that falls outside of the expected range. Statistical analysis can also be used to identify outliers and patterns in the data that may be indicative of errors or inconsistencies. Automated tools are becoming increasingly important in data cleaning, particularly as healthcare systems generate more and more data. These tools use algorithms to detect errors, flag inconsistencies, and suggest potential corrections. They can also be used to validate data, ensuring that it conforms to predefined standards and rules. Another important aspect of data clean plans is ensuring that data privacy and security requirements are met. This involves implementing appropriate access controls and data encryption measures, as well as ensuring that all data handling and storage practices comply with relevant regulations and standards. The benefits of implementing a data clean plan in healthcare are numerous. By ensuring that data is accurate and reliable, healthcare professionals can make better decisions and provide higher quality care.

4.2 Paper case report forms (pCRF) vs Electronic Case Report Forms (eCRFs)

Case Report Forms (CRFs) are a crucial tool in clinical research for collecting essential data that must be reported to the sponsor on each trial subject. These documents can be in the form of printed, optical, or electronic records and are designed to record all protocol-required information on each study subject. The CRF can be either a paper CRF (pCRF) or an electronic CRF (eCRF) that should be protocol-driven, robust in content, and comply with the regulations (for example GDPR in the EU). The design and completion of pCRF/eCRF have a significant impact on the quality of the data collected during a clinical study. A well-designed pCRF/eCRF can ensure that no essential data are missed, data queries are kept to a minimum, and aids good data management practice, statistical analysis, and reporting. Moreover, it should ensure the safety and eligibility of the participant, demonstrate compliance with study procedures, and adhere to Good Clinical Practice (GCP) and Principles of Good Data Management. Furthermore, pCRF/eCRFs should be optimized for collecting data in accordance with the

[20] https://dmptool.org/public_plans
[21] https://royalpapworth.nhs.uk/application/files/9915/5231/9593/GD012_-_Data_Cleaning_v1.1_in_developement.pdf

study protocol compliance, regulatory requirements, and enable the researcher to test the hypothesis or answer the study-related questions. In clinical research, individual patient data was usually collected on pCRFs by investigators in their offices summarizing medical charts on paper forms. However, nowadays electronic data capture has been increasingly used in both industry and academic research settings. While electronic CRFs have demonstrated their usefulness, concerns have been raised regarding the replacement of paper case report forms and questionnaires by their electronic counterparts. Nevertheless, eCRFs are preferred over pCRFs due to their efficiency, accuracy, and ability to streamline data collection and management processes.

When designing an electronic case report form (eCRF), the system automatically generates repetitive data, such as protocol ID, site code, subject ID, and patient initials, on all pages to avoid duplication. The eCRF also facilitates linking of data between related pages and includes built-in edit checks on each data field and the entire CRF, reducing the need for data cleaning by management personnel. This feature allows instant query resolution, saving time spent on clarification from the site/investigator and resulting in quicker acquisition of clean data. The eCRF streamlines the data cleaning process, leading to timely database lock, faster regulatory submission, and subsequent approval.

A good summary about case report form designing in a clinical research can be found in the article: Reddy S, Punjala SR, Allan P et al. First Report With Medium-term Follow-up of Intestinal Transplantation for Advanced and Recurrent Nonresectable Pseudomyxoma Peritonei. Annals of Surgery 277(5):p 835-840, May 2023.[22]

In conclusion, the development and implementation of a well-designed CRF, whether it be a pCRF or an eCRF, are essential for collecting reliable and accurate data in clinical research. This tool ensures that no vital information is missed, minimizes data queries, and adheres to Good Clinical Practice and Principles of Good Data Management. Although both pCRFs and eCRFs have their advantages and disadvantages, eCRFs are becoming increasingly popular due to their efficiency, accuracy, and ability to streamline data collection and management processes.

4.3 International quality guidelines for clinical trials

Good Clinical Practice (ICH GCP) [23]is an international ethical and scientific quality standard for the design, conduct, performance, monitoring, auditing, recording, analyses and reporting of clinical trials.

There are 13 core principles of ICH GCP and they are as follows:

1. Clinical trials should be conducted in accordance with ethical principles that have their origin in the Declaration of Helsinki, and that are consistent with ICH GCP and the applicable regulatory requirement(s).

[22] https://doi.org/10.4103%2F2229-3485.140555
[23] https://www.ema.europa.eu/en/ich-e6-r2-good-clinical-practice-scientific-guideline

Co-funded by the
Erasmus+ Programme
of the European Union

2. Before a trial is initiated, foreseeable risks and inconveniences should be weighed against anticipated benefit for the individual trial subject and society. A trial should be initiated and continued only if the anticipated benefits justify the risks.
3. The rights, safety and well-being of the trial subjects are the most important considerations and should prevail over interest of science and society.
4. The available non-clinical and clinical information on an investigational product should be adequate to support the proposed clinical trial.
5. Clinical trials should be scientifically sound, and described in clear, detailed protocol.
6. A trial should be conducted in compliance with the protocol that has received prior institutional review board (IRB)/independent ethics committee (IEC) approval/favourable opinion.
7. The medical care given to, and medical decisions made on behalf of subjects should always be the responsibility of a qualified physician or, when appropriate, of a qualified dentist.
8. Each individual involved in conducting a trial should be qualified by education, training, and experience to perform his or her respective task(s).
9. Freely given informed consent should be obtained from every subject prior to clinical trial participation.
10. All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification.
11. The confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).
12. Investigational products should be manufactured, handled and stored in accordance with applicable Good Manufacturing Practice (GMP). They should be used in accordance with the approved protocol.
13. Systems with procedures that assure the quality of every aspect of the trial should be implemented.

The ICH GCP guidelines aim to protect the rights, safety, and welfare of trial participants and ensure the credibility and integrity of the data generated during the trial. ICH GCP guidelines are crucial in ensuring that clinical research is conducted ethically and that the data generated is scientifically valid and reliable. Compliance with ICH GCP guidelines is mandatory in many countries and is often a requirement for the approval of clinical trial protocols by regulatory authorities. The principles of ICH GCP also extend beyond clinical trials and are applicable to all research involving human participants. Adherence to ICH GCP guidelines is essential for the successful conduct of clinical trials, the protection of the rights and welfare of trial participants, and the generation of high-quality data that can be used to support the development of new drugs and treatments for a wide range of diseases and medical conditions.

More information can be found in:

- Guideline for good clinical practice E6(R2)[24]
- Gajic A, Herrmann R, Salzberg M. The international quality requirements for the conduct of clinical studies and the challenges for study centers to implement them. Annals of Oncology 15:1305–1309, 2004.[25]

5 Statistical considerations for the clinical trial planning

Clinical trials that are well-designed, well-conducted, and appropriately reported and published, regardless of the results, are considered successful. Physicians and statisticians play important roles in clinical trials, and successful trials are the result of collaboration between these two groups from the beginning of an idea to the publication of a manuscript.

Here are some additional details about each of these roles:

- **Physicians**: Physicians are responsible for the clinical aspects of clinical trials. They design the trials, recruit patients, and oversee the conduct of the trials. They also interpret the results of the trials and communicate the results to patients and the public.
- **Statisticians**: Statisticians are responsible for the statistical aspects of clinical trials. They design the trials, analyse the data, and interpret the results. They also help to ensure that the trials are conducted in a fair and unbiased manner.

Both physicians and statisticians are essential to the success of clinical trials. By working together, they can ensure that clinical trials are conducted in a rigorous and ethical manner, and that the results of the trials are accurate and reliable.

It has to be noted the collaborations between statisticians and clinical investigators by highlighting the essential role of statisticians as co-investigators. Statisticians work with clinical investigators from the early stages of trial development, through data collection, analysis, interpretation, and publication of results. Additionally, clinical investigators provide statisticians with valuable disease-specific knowledge that helps with trial design and analysis. Successful clinical trials require a true collaboration between statisticians and clinical investigators, and effective communication is vital.

During the statistical planning/analysis the following terms appear in almost every clinical trial:

- **Type of Trials**: Phase I, II, III.
- **Sample Size**: The determination of sample size for a clinical trial is a joint effort that necessitates substantial contribution from the principal investigator, while the calculation of potential sample size should be performed by the statistician.
- **Determination of Sample Size is a process**: Sample sizes for clinical trials are calculated by the statistician, who considers various parameters like effect size, error rates, accrual rate, and length of follow-up for time-to-event endpoints, among others. These potential sample sizes are then discussed with the trial's principal investigator.

[24] https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-good-clinical-practice-e6r2-step-5_en.pdf
[25] https://doi.org/10.1093/annonc/mdh363

- **Hypotheses**: H0, and HA no difference vs the new treatment is better than the standard treatment.
- **Type I and II errors**: There is always an underlying truth for any hypothesis test; however it isn't known, which is why clinical trials are conducted
- **Statistical power**: When reading the results of a prospectively designed clinical trial, it's uncommon not to have information on the statistical power. It is important to know the statistical power for retrospective analyses as well, especially when "negative" results are reported.
- **1-sided, 2-sided hypothesis tests**: A hypothesis test can be either one-sided or two-sided depending on the research question. If the objective is to test for a difference between two treatments A and B in any direction, a two-sided hypothesis test is suitable. On the other hand, if the goal is to test whether an experimental treatment (A) is better than the control arm treatment (B), a one-sided hypothesis test is appropriate.
- **Endpoints**: It is crucial to clearly define endpoints in a clinical trial protocol and in publications that follow, as the definition of an endpoint may vary across trials even if they share the same name.
- **Effect size**: Effect size is a statistical measure that quantifies the magnitude of a treatment or intervention's effect on an outcome of interest. It is usually expressed as a standardized difference between the means of two groups, such as the treatment group and control group in a clinical trial. A larger effect size indicates a stronger relationship between the treatment and the outcome, while a smaller effect size indicates a weaker relationship. Effect size is an important consideration in study design, sample size calculations, and the interpretation of study results.
- **Stratification and Blinding**: To ensure unbiased evaluation of treatment effects, blinding is implemented in clinical trials where patients and/or physicians are unaware of the treatment being administered. Stratification, on the other hand, is a technique that aims to balance the levels of a factor between treatment arms.
- **p-value, definition and interpreting**: The p-value represents the probability of obtaining a result that is as extreme or more extreme than the observed result, based on the sample data, if the null hypothesis (H0) is assumed to be true.

A more detailed summary about these terms can be found in the article: Winter K, Pugh S. An Investigator's Introduction to Statistical Considerations in Clinical Trials. Urol Oncol. 2019 May; 37(5): 305–312.[26]

Clinical trials are crucial for developing new therapies and medical interventions. However, to ensure that the results of a clinical trial are valid and reliable, it is essential to take into account several statistical considerations. Some of the key statistical considerations for clinical trials include defining clear and measurable endpoints, calculating the appropriate sample size, blinding and stratification, hypothesis testing, and reporting statistical results. These considerations are vital in ensuring that clinical trial outcomes are accurately measured and interpreted and that the treatments tested are safe and effective. Collaboration between physicians and statisticians is essential to ensure that all statistical considerations are taken into account from the design stage to the publication of the results.

[26] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6959125/pdf/nihms-1056270.pdf

6 Conclusion

Clinical Data Management is a critical phase in clinical research, which leads to generation of high-quality, reliable data from clinical trials thus making it reliable for the future sharing and reuse. As data management is fundamental part of an open research therefore, CONSCIOUS II is aspired to focus on this topic. We detailed the process of the data flow from source to a usable data set for analysis at the sponsor site as well as explore the advantages and disadvantages of using paper CRF (pCRF) and electronic CRF (eCRF) for data capture.