Developing a data management framework for a clinical research institution

Ho Van Hien

Oxford University Clinical Research Unit

Introduction

In the rapidly evolving field of clinical research, the effective management of data is paramount. A robust data management framework is essential for ensuring the integrity, security, and accessibility of data throughout the research lifecycle. This framework not only supports compliance with regulatory requirements but also enhances the efficiency and reliability of research outcomes.

Developing a data management framework for a clinical research institution involves a comprehensive approach that integrates best practices in data collection, storage, processing, and analysis. It requires a thorough understanding of the unique challenges and needs of clinical research, including the handling of sensitive patient information, adherence to ethical standards, and the facilitation of collaborative research efforts.

This framework introduces sets the stage for exploring the key components and strategies involved in creating an effective data management framework, highlighting its critical role in advancing clinical research and improving quality of research data throughout comprehensive data management processes. This is the result of a long-journey of studying the SCDM's Good Clinical Data Management Practices (GCDMP©) (SCDM, 2013), ICH E6 (R2) Good clinical practice (ICH) and practical experiences of Data Managers of the Oxford University Clinical Research Unit.

Objectives

The Data Framework serves as a roadmap for managing data throughout the entire lifecycle of a research project, from initiation to closure.

It provides general guidelines to help research staff understand that data management is not solely the responsibility of data management personnel, but rather a shared responsibility among various team members, including the Chief Investigator, Coordinator, Study Doctors, and Data Entry clerks.

The framework

The framework includes all necessary components and activities that are taken into account during the research project lifecycle. It is divided into three stages:

- Planning, preparation and training in pre-research stage
- Research conduct and policy compliance monitoring in the implementation stage
- Archiving and usage monitoring after the project completed

Each stage consists of components and sub-elements assigned to specific individuals or teams. The details of each component are outlined in the relevant policies, procedures, or operational manuals.

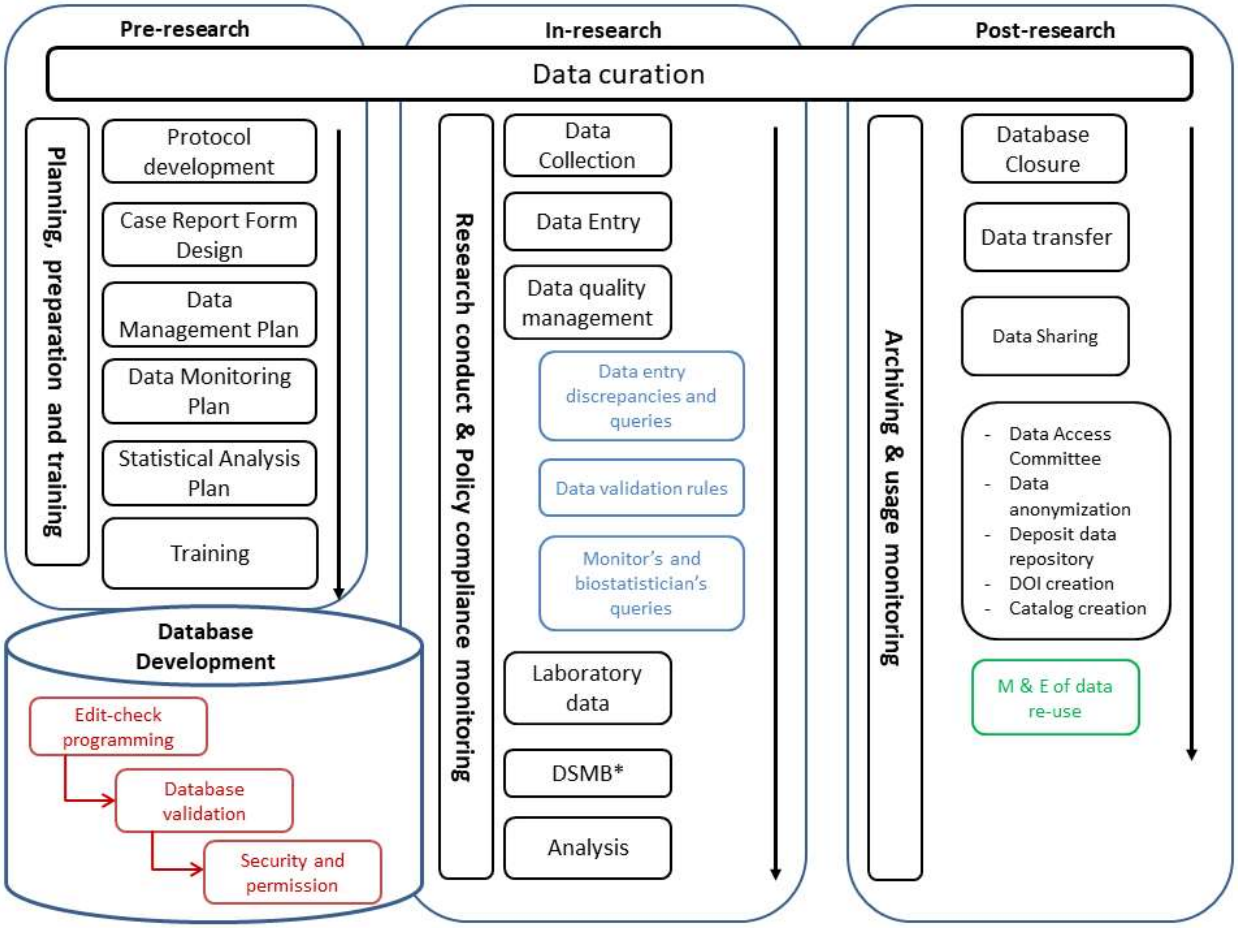

Figure 1. The Data Management Framework.

**Applicable to RCT. Blue boxes: data quality management activities; Red boxes: database validation and security; Green boxes: planned*

The framework is considered as a foundation for developing tailored data management processes for individual research projects within the institution. Based on the project's requirements, relevant components are incorporated into the data management plan. Responsibilities can also be adjusted to align with available resources and project needs

1. Protocol development

A research is conducted according to a plan (a protocol) or an action plan. The protocol demonstrates the guidelines for conducting the research. It illustrates what will be made in the study by explaining each essential part of it and how it is carried out. A protocol is directed by a chief investigators and contributed by various members such as co-investigators, study doctors, clinical research associates, research nurses, biostatisticians, data managers…

Data management should be involved and contributed during protocol development to determine appropriate methodology for data collection, project timelines and all data management related elements developed in the protocol.

2. Case Report Form design

Good Clinical Practice (ICH E6 R2) defines the term "case report form" as, "A printed, optical, or electronic document designed to record all of the protocol-required information to be reported to the sponsor on each trial subject."

Clinical data can be collected with a variety of tools, but case report forms are the most frequently used data collection tool. Great care must be given to ensuring each CRF accurately and consistently captures data specified in the study protocol.

To ensure the protocol specifies data collection strategies that are reasonable and achievable, CRF design should be taken into consideration before the protocol is finalized. The design and development of CRFs should be achieved through a multidisciplinary approach that includes inputs from CDM, statistical, clinical, safety monitoring and regulatory personnel.

Following aspects should be considered as best practices during the CRF design and Development:

- Establish and maintain a library of standard forms and associated edit checks (CRFs, CRF completion guidelines, subject diaries, etc.).
- Use a multidisciplinary team to provide input into the CRF design and review processes. Data entry personnel, biostatisticians, the internal study team, and clinical operations personnel may be able to provide valuable perspectives to help optimize CRFs.
- Design CRFs with safety and efficacy endpoints in mind. Consult the protocol, study biostatistician(s) or review the statistical analysis plan (SAP) (if available) to ensure all key endpoints are collected.
- Keep the CRF's questions, prompts, and instructions clear, concise and conformant to CDISC CDASH standards, where possible.
- Design the CRF to follow the data flow from the perspective of the person completing it, taking into account the flow of study procedures.
- Whenever possible, avoid referential and redundant data points within the CRF. If redundant data collection is used to assess data validity, the measurements should be obtained through independent means.
- Use carbonless copy paper (NCR) paper or other means to ensure exact replicas of paper collection tools.

Process and People in charge

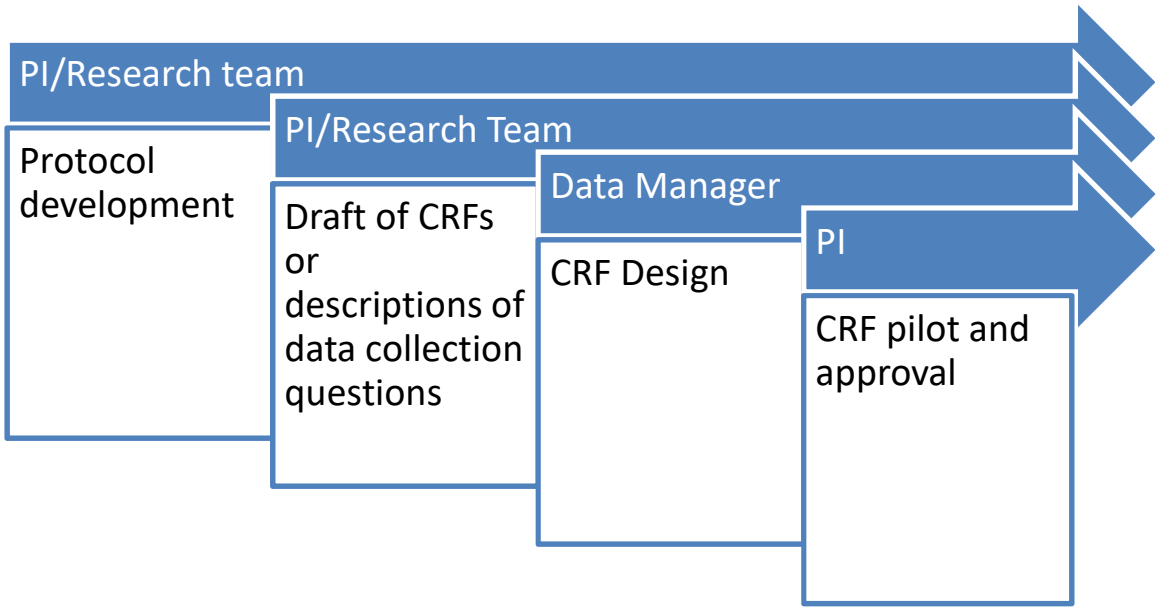

Figure 2. Process and people in charge for CRF design

The process of design the case report forms start from the beginning of the study protocol development. The Principle Investigator and the research team members draft the CRFs or describe data points to be collection for the protocol. CDM personnel (assigned Data Manager) will be responsible to develop the CRFs based on the draft CRFs, in consultation with protocol, PI and Biostatistician. The designed CRFs should be in piloting phase for collecting data on some subjects. This pilot is only for testing the feasibility of the CRF design, not for data entry and analysis so dummy data can be collected. The final Case Report Forms should be reviewed and agreed by participate research team members and approved by the PI before they are in used for conducting data collection.

SOPs, Forms and Templates

- CRF design and development process (SOP)
- CRF Approval Form (Form)

3. Data Management Plan

The Data management plan (DMP) documents the processes and procedures to promote consistent, efficient and effective data management practices for each individual study. The primary goal of the DMP is to communicate to all stakeholders the necessary knowledge to create and maintain a high-quality database ready for analysis. The DMP serves as the authoritative resource, documenting data management practices and decisions that are agreed at study initiation. The DMP should comply with all applicable regulatory guidelines (e.g., FDA, ICH, GCP) or local laws of the country; as well as the standard operating procedures (SOPs) of the institution. The DMP should also address any procedural or protocol updates that are made during conduct of the study.

The DMP should be created during the setup phase of each study and should contain information relating to all aspects of data management activities to be performed. The DMP should be considered a living document throughout the life cycle of a study, capturing any changes impacting data management made to the protocol or processes being used. The DMP must be uniquely identifiable, carry such identification on each page (e.g., study code/title) and be subject to version control. Each version should be documented and include date, author, reason for version change and an individual version identifier.

The organization, structure and order of components presented in a DMP may differ between the studies.

Process and People in charge

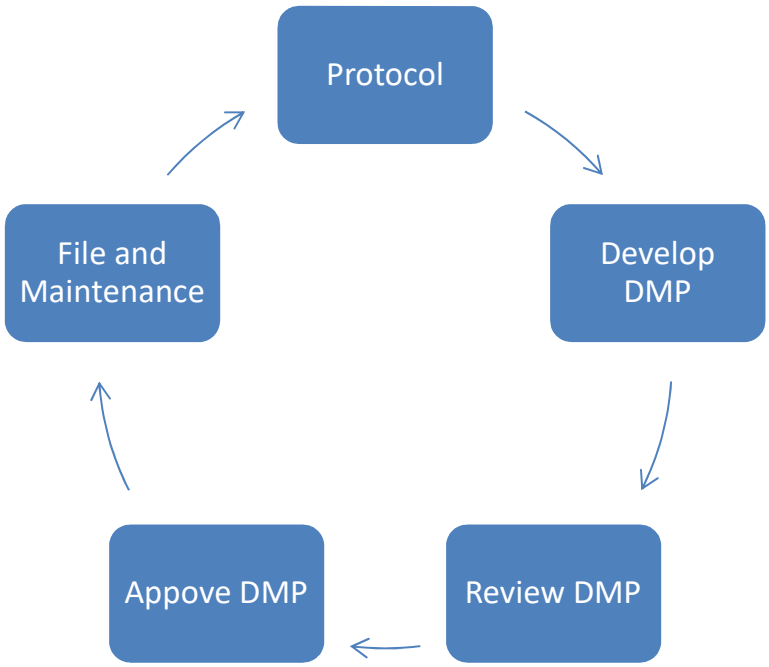

Figure 3. Data Management Plan (DMP) development process

It is required that all clinical trials conducted must have a validated protocol-specific DMP. Other observational or community-based studies may be optional and up to the PI's decision. However, any studies without a specific DMP must apply practices described in the general Data Management Plan that has been approved by the Research Committee.

For each new study that requires a specific DMP, clinical data management (CDM) personnel should compose a detailed DMP based on the protocol, work scope, contract, analysis plans, dataflow, case report forms (CRFs), other supporting documents, and data management standards and practices. The entire DMP should be drafted and reviewed by all responsible parties to ensure all components described in the DMP are feasible during the course of the study. The DMP must be reviewed by the Lead Data Manager and approved by the Study PI prior to commencement of the work it describes. The clinical data manager should ensure the DMP is kept current, including proper version control, and that all parties involved agree with the content. Upon conclusion of the study, the DMP should be archived with all other pertinent study documentation.

SOPs, Forms and Templates

- General Data Management Plan (SOP)
- Data Management Plan Development Process (SOP)
- Data Management Plan Template (Template)

4. Database development

Database development is a process to create a validated relational database to collect data and store in the CDMS. The database development process starts with a completed set of approved CRFs, then following with designing of the database, programming edit-check, validating the database, managing security and permission before the database is ready for production data collection.

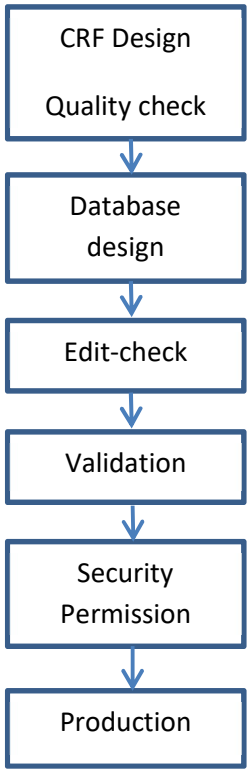

Figure 4. Database development workflow

The database development is responsible to CDM personnel and other relevant parties such as PI, Study Coordinators, Data Entry staff and Biostatistician, whose names are listed in the Responsibility of the Study Data Management Plan and Study Delegation Log.

Designing phase is various and depends on which CDMS/software platform is in use to develop the database. Each platform should have its own SOPs or manuals of operation.

SOPs, Forms and Templates

- Database Development Process and Guidelines (SOP)

4.1. Edit-check programming

The ultimate goal of clinical data management is to complete every study with a dataset that accurately represents data captured in the study. Data inconsistencies and errors can be alleviated with careful review and data-cleaning activities which may be performed by different personnel according to their knowledge and training. And edit checks are invaluable tools for increasing data quality and providing greater efficiency during data review and cleaning activities. Carefully designed edit checks can greatly increase efficiency and data quality by automating many data review processes within the clinical database or clinical data management system (CDMS). CDM personnel and members of the study team should collaborate to determine what edit checks should be in place to fulfill study requirements and reduce potential data errors and inconsistencies.

Although assignment of responsibilities varies between organizations, CDM would be involved with all phases of edit check specification and testing, with the possible exception of edit check programming.

SOPs, Forms and Templates

- Edit-Check Development and Validation Process (SOP)
- Real-time Edit-check specification Template (Templates)
- Post Entry Validation Rule specification Template (Templates)

4.2. Database validation

The following section is described in consumption that a CDMS has been validated and approved for use within an organization. The validation then only focuses on study- or protocol-specific database design and implementation. Validation at this phase can be addressed in three major categories: database design, data entry or capture, and other study specific programming.

When testing a study's data capture system, the most important considerations are to ensure that data entered through a data entry screen or captured via some other transfer process (e.g., electronic lab data transfers) map to the correct variables in the clinical study database and that the parameters for the variable correctly house the data provided. Useful validation measures include entering test or "dummy" data into the screens or loading test data transfer files so that output data listings and data extracted from the database can be reviewed to ensure that the variables were correctly added and saved within the database structure. Testing should be performed on all data, regardless of whether the data do or do not meet defined data structures. It is critical to Will all study data be accepted by the database? Are variable lengths sufficient to prevent truncating or rounding? Do character and numeric formats provide the necessary output for analysis files, query management software and other modules within the sponsor's overall CDMS? If the database is programmed to flag out-of-range data, are flags appropriately triggering at data entry or import?

Database entry or capture validation testing should help identify key records management issues. For example, the database should not accept entry of duplicate records, and primary key variables should be appropriately assigned and managed by the definition of the database's structure. When discrepancies between the first and second passes of data entry are resolved for double data entry systems, validation should ensure that one record with the correct data is permanently and correctly inserted into the study database and can be extracted. Most importantly, the audit trail for the study should be validated and protected so that all manipulations of the study database or external files are recorded by date, time, and user stamps in an unalterable audit trail that can be accessed throughout the life of the data.

Other examples of study-specific programming are data loading or transfer programming (e.g., loading adverse event coding variables or loading central lab data), and programming written to validate the data (e.g., edit checks, query rules, procedures). This programming includes any code written to check the data and can occur at the time of entry or later as a batch job. This programming must be validated if action is taken regarding clinical data intended for submission as a result of the programming. Examples include programming that identifies data discrepancies such that queries are sent to clinical investigators or in-house data-editing conventions followed for items identified by the programming.

Best practices include identifying all intended uses of study-specific programming and testing each logic condition in the programming based on a validation plan. Algorithms for variable derivations occurring within the database must be validated. Practical suggestions include utilizing organization standards to document as much of the programming specification and validation plans as possible and code libraries to reduce the amount of new code generated for a protocol. The entire validation plan can be a standard operating procedure containing testing methodology, scope, purpose, acceptance criterion, approvals and the format for test data and problem reporting.

Process and People in charge

The database validation is responsible to CDM personnel who follow instructions and guidelines to develop the Validation Plan, preparing testing scripts and conducting testing with other relevant parties such as PI, Study Coordinators and Data Entry staff. The User Acceptance Testing are performed by Data Entry Team

SOPs, Forms and Templates

- Database Validation process (SOP)
- Database UAT Report (Template)
- Database Validation Summary Report

4.3. Security and permission

The following section is described in consumption that physical and electronic security measures have been covered in security and data protection policies within the organization. The security and

permission then only focuses on study- or protocol-specific database security and permission in a way to control access permissions and protection of data within the desired study database.

Role-based access control would be a key element of the database security permission. There are profiles for available database roles within the system being used to support the study. Roles are assign privileges based upon the duties performed in the study.

In multicenter studies, site-based permission is also critical. Security should be implemented in the way users from one site cannot access to other sites' records, other than they are authorized to have. Eg. Site investigator in Vietnam can only have access to records of subjects enrolled in Vietnam and same as Site Investigators from other sites. However, Chief Investigator or Project Manager can be authorized to have access all records of the study.

Form-based (or CRF-based) permission also can be applied in such study database in which different forms / CRFs are responsible to different personnel or group of people. Eg. Clinicians are allowed to access, enter and modify data within certain clinical forms/CRFs of the database such as EXAMINATION, MEDICAL HISTORY. In contrast, lab technician is requested to have access the LABORATORY INVESTIGATIONS.

Process and People in charge

The Database security and permission is responsible to the Study Data Manager who follows instructions and guidelines to develop, maintain and file relevant documentation. Granting access permissions may be executed by Study Data Manager or Database Administrator after obtaining approval from Chief Investigator or Head of Data Management.

SOPs, Forms and Templates

- Database Security Management (SOP)
- Database Access Privilege Form (Form)

5. Data collection

Accurate completion of case report forms (CRFs) is paramount to the quality of data that are captured during a clinical study. Accurate completion of CRFs is not only from the research staff who completes the CRFs during the study conduct, but also from the design of the CRF, methods of data collection (paper-based CRFs or electronic CRFs), the aide of complete, concise and logical guidelines, training to the assignee staff and quality control during the project lifecycle.

Data collection is responsible to Site staff following appropriate training and data collection guidelines.

Paper-based CRFs

If paper-based CRFs are used as the data collection instrument for the study, data collection guidelines should be given prior to starting data collection.

Electronic CRFs

If eCRFs are developed for data collection, completion instructions should be programmed to provide adequate guidance to complete certain fields for individual questions for the eCRFs.

SOPs, Forms and Templates

- Development of the data collection guidelines

6. Data entry and Process

The purpose of data entry processes is to ensure data are reliable, complete, accurate, of high quality, and suitable for statistical analyses. Data entry processes encompass the efficient receipt, tracking, entering, cleaning, coding, reconciling and transferring of data. Data entry process may differ from various studies depending on a number of factors such as the skill level and training of personnel, and the amount of time allocated for data entry, attributes and design of the studies.

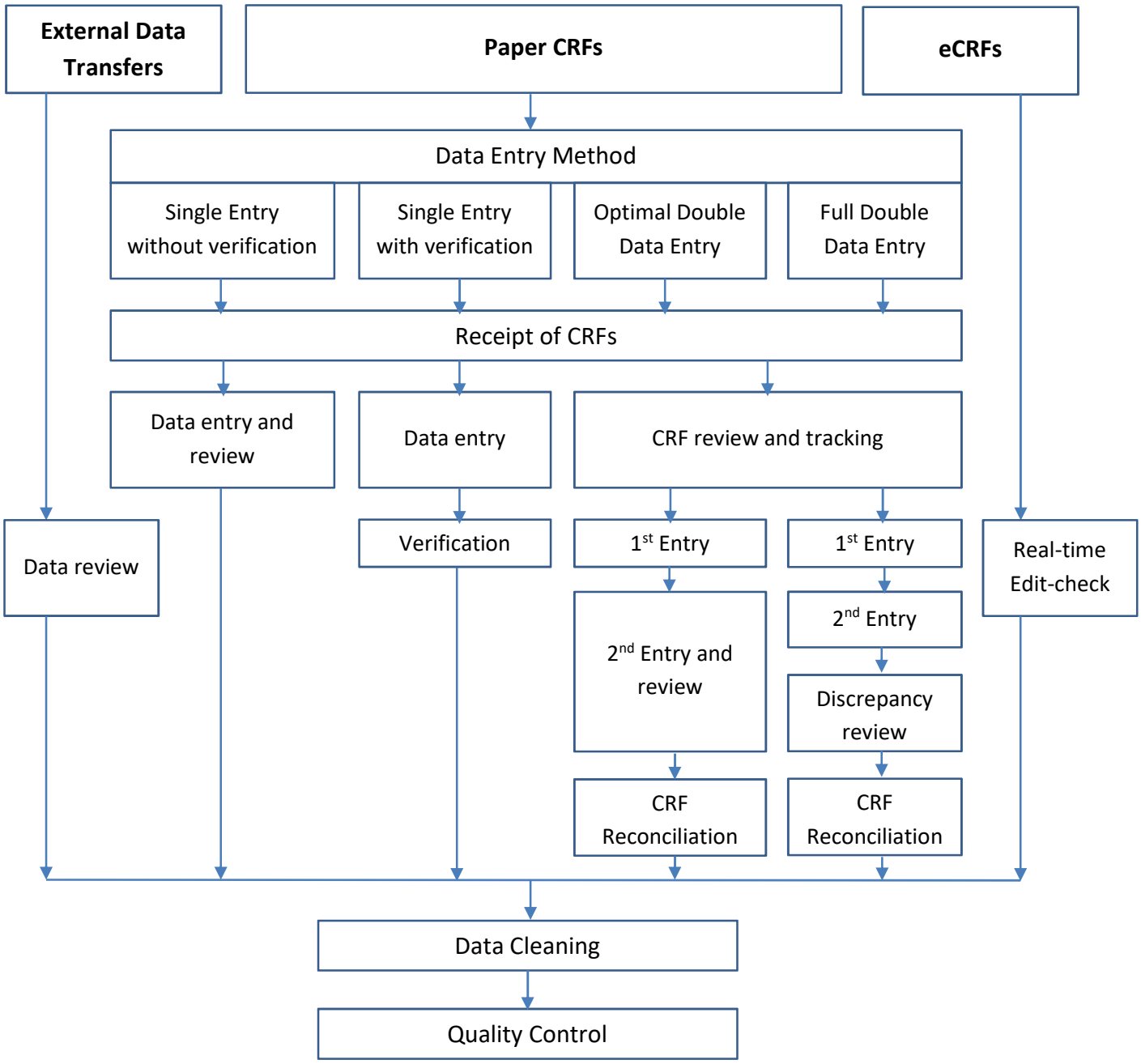

Figure 5. Data entry processes

There are three typical data will be entered into study databases and each of type of data will be processed differently: External data transfer, electronic CRFs and paper CRFs.

6.1. External Data Transfers:

This type of data will not be entered by data entry personnel, but imported from external data sources such as laboratory data, imaging, and third-party data. Such data should be reviewed in the way data and its structure is consistent with the designed data type and structure of the CDMS. Data cleaning and quality control should be considered to ensure the expected quality of data for the study.

6.2. eCRFs:

Data will not be entered by data entry personnel but Site staff who may transcribe data from source documents or collect source data from observation or interview the study subjects. eCRFs should be designed with proper real-time edit-checks to ensure data is collected precisely at time of data collection. Data cleaning and quality control should be considered to ensure the expected quality of data for the study.

6.3. Paper CRFs

Data entry processes should address data quality needs of the study. The following are some commonly used data entry strategies for studies using paper CRFs.

Single Entry without verification: Although not recommended, situations may occur where one person enters data and the data are not subsequently reviewed.

Single Entry with verification: One person enters the data and a second person reviews the data entered against the source data.

Full Double Data Entry: the process starts from CRF preparation by Site staff in batch before sending to data entry for double data entry. Assigned staff at Data Entry team will initially review the CRFs and records the receipt in the CRF tracking module. The batch of CRFs will be 1st entered by a data entry staff and 2nd entered by another data entry staff. The 2nd data entry staff will receive a notification on any field where data is entered differently with the 1st entry and have a chance to review the CRF and make any corrections if applicable. A third person (considered as Data Entry supervisor) independently reviews and resolves any discrepancies between 1st and 2nd entry. The process is complete once the batch of CRFs are reconciled to confirmed all received CRFs are double entered and discrepancies have been solved.

Optimal Double Data Entry: the process is similar with the Full Double Data Entry starts but the 2nd Data Entry staff will take the role as Data Entry Supervisor and perform data review and corrections during 2nd data entry step. This process can applied in such studies where double entry is highly required with a small and qualified data entry team.

Process and People in charge

Responsibilities for data entry vary depending on types of data within the study. In practice, type of data and selection of data entry methods is described in the study specific data management plan. Generally, external data transfers are responsible to Study Data Manager and/or Study Coordinator, eCRFs will be electronic data captured by research site staff. Data entry for paper CRFs is subject to data entry team.

SOPs, Forms and Templates

- Data Entry Process (SOP)
- Development of the data entry operational guidelines

7. Data quality management

High quality clinical research data provides the basis for conclusions regarding the safety and efficacy of a medical treatment. Data quality may be composed of numerous attributes. For clinical research data, attributes of quality may include accuracy, consistency, timeliness, consumability, currency, completeness, relevance, granularity, unambiguity, precision and attribution. Clinical research data quality may therefore refer to a dataset that accurately represents data points collected from subjects, has acceptable completeness, is defined sufficiently for use, is current, is attributable, and contains relevant data at the appropriate level of precision to answer the study's primary hypotheses.

A quality system includes quality assurance (QA) and quality control (QC). QA refers to a systematic process to determine whether the quality control system is working and effective. In the other hand, QC means periodic operational checks within each functional department to verify that clinical data are generated, collected, handled, analyzed, and reported according to protocol, SOPs, and GCPs.

This section focuses on the QC measurements after data has been entered / collected into database to improve data quality during the course of the study. QA is specifically described in the section of Data and safety monitoring plan and various QC measurements also considered during data collection and data entry processes.

7.1. Source data verification

The ICH Harmonised Tripartite Guidelines for Good Clinical Practice, the WHO Guidelines for Good Clinical Practice for Trials on Pharmaceutical Products, and the Code of Federal Regulations require that source data verification must occur for all clinical trials in phases I–IV. An evaluation of the conformity of data presented in CRFs with source data, SDV is conducted to ensure data collected are reliable and allow reconstruction and evaluation of the study. The SDV responsibilities of the principal investigator, sub-investigator, study coordinator, monitor, quality assurance auditor, and the clinical trial manager must be made clear at the outset of the clinical trial, and adequate training should be provided to all staff involved. So there are no misunderstandings or errors when SDV is undertaken, special emphasis should be placed on confidentiality and direct access to data. All staff involved must realize that SDV adds to the scientific and ethical integrity of a clinical trial. Records of what was done and found, including an evaluation of findings, must be made in the same way as for any other aspect of the trial.

In the SDV process, information reported by an investigator is compared with the original records to ensure that it is complete, accurate, and valid. Strictly speaking, every item of data that appears in a CRF should be documented somewhere else to allow verification, audit, and reconstruction. The main objective of SDV is to confirm that the information collected during a clinical study is complete, accurate, reliable, and verifiable so as to give confidence to the sponsor and the regulatory authorities in the data being used to support a marketing application. SDV is also required to provide confidence in any data reported, for example, in published manuscripts and at scientific conferences. Without SDV or stringently controlled electronic source data collection methods, no scientist can have confidence in the data presented and in the conclusions derived.

All information in original records of clinical findings and in certified copies of original records are necessary for the reconstruction and evaluation of the trial. These records may include hospital records, clinical and office charts, laboratory notes, memoranda, subjects' diaries or evaluation checklists, pharmacy dispensing records, recorded data from automated instruments, microfiches, photographic negatives, microfilm or magnetic media, X-rays, subject files, records at the laboratories and at medico-technical departments involved in the clinical trial, observations, and documentation recording activities in the clinical trial.

Process and People in charge

The monitor team is responsible for conducting the SDV following the monitoring plan they developed. Any findings related to SDV must be reported to the study team and appropriate corrections to the CRFs and database must be performed.

7.2. Data entry queries

The purpose of data entry processes is to ensure data are reliable, complete, accurate, of high quality, and suitable for statistical analyses. Queries generated during data entry process can be considered as an efficient tool to prevent data error from being entered into data and reduce time for validating and cleaning data.

Data entry queries can be generated at the time CRFs received for data entry and must be corrected before CRFs are entered into system. These queries would be focused on such key variables as subject identifiers, date of enrolment, dates of visit…

During data entry, data entry staff may experience various errored and ambiguous data collected without recognizing at the time of receiving of CRFs. Collected data includes, but not limited to different types that prevent it from being entered appropriately:

- Is carelessly written and cannot be read by data entry staff
- Has different data type than the data type designed in the EDC/database. Eg. Letters are written on a numeric field.
- Is out of value range or length defined in the database/EDC. Eg. 45^{o}C was collected in the CRF but the field is set from 36 – 44 for the temperature; a very long sentence is written on a 10-letter free-text field.
- A clearly error data is written on CRFs and should not be entered. Eg. Data is written as 30/02/3020.

Instead of stopping data entry and send the CRFs back for correction, data entry staff can make data entry queries on those questions and continue entering data for the rest of the CRFs. Data entry staff then generate a list of all queries and send back to site staff for correction before those data can be updated into the database appropriately. Some CDMS or EDC (CliRes DMS) can support this process by building a function that data entry staff can make an electronic query attached to the field and print/export queries to send to site staff for response.

Process and People in charge

Data entry staff or Data Entry supervisor who receives completed CRFs are responsible for verifying and issuing data entry queries before the CRFs being sent for data entry.

While performing data entry, Data entry staff is responsible for creating queries on such data they would not enter into the database.

SOPs, Forms and Templates

- Data Entry Process (SOP)

7.3. Data entry verification

Data entry verification is an important process to ensure data entered into the database are reliable, complete, accurate, of high quality, and suitable for statistical analyses. Data entry verification provide a cross check on data single entered by a data entry staff and this verification includes correcting data if any discrepancies occur between the entered data and the CRFs.

If the CDMS include a function of data entry verification, it should be used for all single data entry studies. Otherwise, this verification should be confirmed by a stamp on the CRFs.

All data change during data entry verification must be recorded in the audit trails report for later audit and inspection purpose during the study.

Process and People in charge

A member of data entry team or the Data entry team leader can be responsible for data entry verification. If applicable, data monitor from monitoring team can be best position to perform the data entry verification.

SOPs, Forms and Templates

- Data Entry Process (SOP)

7.4. Double data entry discrepancies

Double data entry process is considered as one of the efficient QC procedures to ensure data are reliable, complete, accurate, of high quality, and suitable for statistical analyses. The major factor leads double data entry has more advantage than single data entry is that it can be triggered the data entry errors during the data entry process rather than data correction after it is entered into the database. However, this advantage is only valuable if the discrepancies between 1st pass entry and 2nd pass entry is reviewed and processed precisely.

In the double data entry process, there will be many discrepancies between 1st pass entry and 2nd pass entry. Such discrepancies should be automatically recorded and generated for an independent review and correction.

Each CDMS may have different mechanism to handle double data entry process in its system. With some systems (CliRes, Clintrial™), 2nd data entry staff will receive a notification for any single discrepancy on the field he/she is entering and have a great opportunity to review with the CRF and correct for any error typing or confirm with their entry. Whatever the 2nd data entry staff decides to correct or confirm the entry, the CDMS will generate a discrepancy record for the 3rd review or data monitor.

Person in charge for the review of double data entry discrepancies may differ from the data entry model of the study. If the study applies the optimal double data entry process, the 2nd data entry staff is responsible for receiving notification and deciding the final decision. This requires that 2nd data entry staff have good skill and experience to handle the process and be granted to additional data modification permission. If the study applies the full double data entry process, the 3rd person (considered as data entry supervisor) is responsible to review the discrepancies reports and make decisions appropriately.

Process and People in charge

The data entry supervisor, or in the mode of Optimal Double Data Entry, the 2nd pass data entry staff is responsible for reviewing and resolving any issues on double data entry discrepancies.

SOPs, Forms and Templates

- Data Entry Process (SOP)

7.5. Data validation queries

Data validation is one of the most important processes for data cleaning after data has been entered into database or collected in the eCRFs. Data validation queries are triggered by characteristics of related or aggregate data, and are more likely to notify CDM personnel of potential data errors after data entry has occurred. The potential data errors identified by triggered Data validation queries may prompt CDM personnel to perform data-cleaning activities such as performing self-evident corrections or generating queries to a site.

Data validation rule specifications (DVS) are crucial to identify invalid data, missing data, inconsistent data, and out-of-range values. DVS planning requires information from a number of sources and should be performed with a comprehensive strategy for specification development in place prior to creating the initial draft.

Process and People in charge

Data manager, in coordination with PI, Study Coordinator and Biostatistician, is responsible for developing the Post-entry Data Validation Rule specification (DVS) and implement it into the study database.

Study Coordinator is responsible for seeking responses and corrections for the DVS queries following schedule described in the study specific DMP. The data responses or corrections will be updated into the study database by either Study Coordinator or Data entry team.

SOPs, Forms and Templates

- Edit-Check Development and Validation Process (SOP)
- Post-Entry Data Validation Rule specification Template (Templates)

7.6. Monitor's and biostatistician's queries

Queries from monitors during monitoring visits and biostatisticians from their analysis also are considered an important aspect for improving quality of research data.

All queries from monitors or biostatistician would be reviewed and responded by Study PIs or designated research staff and data must be corrected and updated appropriately into the database before it is freeze and exported for the analysis datasets.

8. Managing laboratory data

In clinical research, laboratory data is collected to provide information on the efficacy and safety of the medication and may also use it in the screening of patients. It includes blood chemistry, hematology, lipids, urinalysis values, cultures, microbiology, virology, and assays specific to the protocol. Nearly all studies have at least some lab data.

Sometimes the values are printed in a form of paper lab report. In other cases, the values are obtained from a central laboratory that supplies the results in the form of an Electronic file.

8.1. Laboratory data written in the paper forms

If the laboratory data is in a paper forms. Result of the laboratory then can be transcribed into case report forms (CRFs) either paper-based CRFs or electronic data capture.

8.2. Laboratory data is in the form of an Electronic file.

If laboratory data is in the form of electronic file and includes adequate variables including test name, results, and key elements required for central database integration such as Subject identifiers, Laboratory does not need to be written in the paper CRFs or electronic data captured into the eCRFs, but need to be loaded into the central database. Loading the lab data into a central database is done either through programs written specifically for each study or by configuring an application. Whether users write a program or configure an application that affects clinical data, it should be subject to a validation process. Given the importance of lab data, this should be especially true for applications that load lab data across studies.

The validation process starts with a specification. A mapping of the layout of the electronic file to the database storage structures provides the basis of the specification, which also should include details on how to handle specific problems (examples below). The process continues with the program being written according to good practices or the application being configured according to guidelines and manuals. Documented testing and specific user instructions round out the requirements. Satisfying Validation needs adds significantly to the development effort for these loading applications. The more consistent the file structures, the less overall effort required.

Electronic laboratory data handling is a subject of decision before the study starts and should be stated in the study-specific data management plan (DMP). Data management personnel will be responsible for working with study team to assess the feasibility of the laboratory electronic files and implement the laboratory data loading process.

9. DSMB

A Data Safety Monitoring Board (DSMB) is a group of independent professionals with pertinent expertise that reviews, on a regular basis, data accumulated from one or more ongoing clinical trials. The DSMB continuously advises the sponsor on the safety of trial subjects and those yet to be recruited to the trial as well as the validity and scientific merit of the trial.

The function of a Data Safety Monitoring Board (DSMB) is to act as a monitor during the trial and execute a planned follow-up period to evaluate trial effectiveness, participant safety, study conduct and external data relevant to a trial. In addition, the Data Safety Monitoring Board (DSMB) provides recommendations to the sponsor regarding the continuation, modification, suspension or termination of a trial. The members also ensure written records of outcomes are maintained and the confidentiality of data is preserved.

It is crucial that the Data Safety Monitoring Board (DSMB) is provided with complete and accurate data because the outcomes of all DSMB meeting involve ensuring that members obtain sufficient information in order to:

- Adequately review a study
- Profit from the discussions with the sponsor and investigator representative
- Identify issues
- Develop approaches for addressing the issues
- Reach a consensus concerning trial recommendations

Data is the foundation for Data Safety Monitoring Board (DSMB) recommendations with respect to trial safety and efficacy, thus it is important that the information they receive is accurate, complete and as up-to-date as possible. The Data Manager is responsible for providing good quality data in order to respond to the DSMB's requirements and support the biostatistician in the preparation of the tables and statistical analyses that the Data Safety Monitoring Board (DSMB) will receive and use to make their recommendations.

The Data Manager will plan in advance during the CRF specification developing by defining: Which data is required; what level of cleaning is needed; and Timelines. Such definitions should be stated clearly in the study-specific data management of the study.

The Data Manager will discuss and agree on a monitoring plan to ensure the clinical data is captured in a timely manner:

- *For EDC studies: define target for when data has to be entered into CRF (e.g. within 2 days of clinic visit)*
- *For paper studies: define monitoring visits and shipment of CRFs (e.g. ship originals vs scanned copies, normal post vs courier, etc)*

- *Implement tracking reports/metrics to monitor what is in-house vs what is still outstanding and share updates with the study team to avoid backlogs*

The Data Manager also defines and tailors the cleaning strategy by clarifying if there is a cut-off date for subject enrollment, if there is a cut-off date for last data point entered into the clinical data management system, which data are required for the DSMB, if coding of medical terms is required, or whether medical terms should be reviewed before generation of listings for Data Safety Monitoring Board (DSMB).

It is important to define if there are any external data required for the DSMB. External data are usually on a critical path in terms of vendor management and data transmission. In this case, the Data Manager will:

- Define data transfer specifications and method of transfer
- Define timelines of data transfers with provider
- Define timelines for reconciliation to allow time for query issue and resolution
- Keep the study "blinded", for instance in the case of PK or Biomarkers, data could reveal what the subject is taking – such as active drug or placebo by viewing the data concentrations. In such cases, the Data Manager will perform a data reconciliation on sample nominal time with concentrations.

The Data Manager works closely with the study team having regular meetings to share the status of the clinical study. The Data Manager usually agrees with the study team if there is a need to run listings and tables in advance to check for potential issues or abnormal trends in clinical data entered into the database:

- *Data Manager can generate queries to address issues or get a justification/clarification prior to generating outputs for the DSMB*
- *Collection of protocol deviations and evaluation of possible impact on clinical data*
- *Define data checks in the clinical data management system: time window, inclusion/exclusion criteria, plausibility, ranges etc.*
- *Retrieve deviations collected through monitoring visits – liaise with CRAs*

10. Database closure

After the last subject's data has been collected from the sites, the race is on to close and "lock" the study database for analysis and archiving. Once a study has been locked, the final analysis can be made and conclusions drawn. Because there is usually high pressure to make those analyses (and related decisions) as soon as possible, companies frequently keep track of "time to database lock" and work constantly to minimize that time. The pressure to quickly lock a database for analysis comes up against a long list of time-consuming tasks that need to be performed first. The list involves many individual steps, including: collecting the final data, resolving outstanding queries, reconciling against other databases, and performing final quality control (QC).

Once these tasks are performed and someone signs off on them, the study is considered locked. (That is, no data value will be changed.) The data is now considered ready for analysis. Statisticians and upper management will begin to draw conclusions from the data under the assurance that it is unlikely to change: unlikely to change, perhaps, but it is not uncommon to detect problems or find missing data after a study lock. If there are enough changes or the changes are serious enough, the study will be unlocked. There are usually specific conditions that can lead to unlocking a database, restrictions on

what can be changed during the time it is unlocked, and requirements that must be met before it can be relocked

10.1. Data Entry completeness

It is important that all data must be entered into the database and all required external data to be integrated with the database before the database is locked and exported for analysis.

Data entry team must be responsible for checking the database records against the CRFs they have received for data entry.

Study Coordinator should be cross-checking the participant records against the master log or any other relevant documents to determine to completeness of data entry process.

10.2. Final data and queries

Before a study database can be locked, it must contain all data generated by the study which is original data from the subjects reported on case report forms (CRFs) or through electronic case report forms (eCRFs), plus such other data as corrections from the sites, calculated values, codes for reported terms, and data from labs. Any of this final data may generate discrepancies that will require resolution before study lock.

To account for all the original data, data management uses tracking information to ensure that all expected CRF pages have been received and data entry is completed. In addition, the lab data administrator will check that all laboratory data was loaded and that any other electronic loads are complete. Once in the central database, this data will go through the cleaning process, which may generate discrepancies. As the final data comes in, the final calculated values also must be derived. Discrepancies raised by calculated values are usually traced back to problems with the reported data and may have to go back to the site.

All reported terms (such as adverse events [AEs] and medications) must be coded and any changes to terms that come in as corrections must also be rerun. When a term cannot be coded, a query may have to be sent to the site. Just to be sure everything is in a final, coded state, many companies rerun coding at the end of the study to catch cases where the assigned code changed due to a change in the dictionary or synonyms table or cases were the term was changed but the code did not receive an update.

Resolutions for the new discrepancies, as well as those still outstanding from earlier in the study, are also required for completeness. Generally, all outstanding queries must have a resolution, even if the resolution is that a value is not available, before a study can be locked. Getting these last resolutions from the site can hold up the entire closure process, so CRAs frequently get involved in calling or visiting the sites to speed corrections. Because of the difficulties and time pressures at the end of the study, companies may choose not to pursue noncritical values at this stage of the data handling. Ideally, the list of critical values will have been identified at the start of the study in the data management plan and can be referred to when facing getting a resolution from an uncooperative site right before study lock.

10.3. Final QC

The quality of the data will affect the quality of the analyses performed on the data. At the close of the study, there is a particularly strong emphasis on checking on the quality of the data that is about to be handed over to a biostatistics group. Because there is, or should be, a high barrier to getting a study unlocked, it is worth making an effort to check the data thoroughly.

All kinds of review of the data help provide assurance as to its quality and correctness, but study closure checklists frequently include these specific kinds of checks:

- Audits of the database
- Summary reviews of the data
- Reconciliation against other systems

10.4. Locking and unlocking

The exact checklist of procedures to follow before study can be locked comes from a data management standard operating procedure (SOP) or study specific requirements documented in the data management plan. Once all of the items in the study lock list have been completed, permission to lock the database is obtained. Typically, a manager reviews the closure procedures and associated results and signs off on the list. Permission is then obtained from the clinical group and biostatistics to show that everyone is in agreement that the data is as complete and accurate as possible at that point. (This dated lock form should always be filed in the study file.) The data management group then physically locks the study against changes. The idea is that the study stays locked, but unforeseen problems with the data or unexpected new data can force a temporary unlocking of the database.

Locking the database

The database lock involves removing permissions on the data so that only a privileged user (such as the database administrator) could make modifications. In many systems, this removal of permissions must be done manually, sometimes on a user-by-user basis.

Unlocking the database

Once the study database is locked and data analysis begins, it is not uncommon to find problems with the data that require corrections. (This is particularly true if final quality control does not include summary reviews of the data or draft runs of the analysis programs.) New information regarding AE data found during site close-out or site audits may also require edits to the data. A request to unlock the study usually requires review of detailed reasons by higher level management before the database administrator removes the locks. Unless the changes required are extensive, the database administrator will grant permissions to very few users. Appropriate quality control, review, and approval will again be required to relock the study. Many companies require that the lock checklist be used to relock a study.

The Study Data Manager will be responsible for coordinating with other stakeholders and obtaining Principal Investigator's approval for locking and unlocking database following the accompanied SOPs.

10.5. Study Archiving

Data is the most important property of any research projects. So the properly archiving research data and relevant documents are considered essential process.

The final datasets and their relevant elements such as audit trial reports, data dictionary, study metadata, annotations, discrepancy resolution and reports, analysis outputs and reports must be archived in the secured manner.

Head of Data Management will be responsible for coordinating with Study Data Manager and other stakeholders to perform archiving study data appropriately.

Process and People in charge

Head of Data Management will be responsible for coordinating with Study Data Manager and other stakeholders to perform database closure according to appropriate approved SOPs.

SOPs, Forms and Templates

- Study Database Closure (SOP)
- Study Close-Out Report (Template)

11. Creating Reports and transferring data

Data management is frequently responsible for producing reports or listings of study data for internal staff and management. Some of these reports are standard representations of the data, which are run over and over on current datasets. Other reports are ad hoc reports, that is, the format and content is requested by a user of the data for infrequent or one-time use. Users of standard or ad hoc reports may be data management staff, clinical research staff for medical review, management to monitor progress, auditors, and so on. The level of effort devoted to the creation of these reports depends on the users and the use to which they will put the report.

Transfers of data to internal or external groups may also fall to data management. For example, a small company may require the transfer of data to an external statistician for early or final analysis. In other cases, the transfer may be to a partner company or to a client (as in the case of a contract research organization [CRO]). Because transfers of data nearly always involve safety and efficacy data for storage or analysis, the level of effort devoted to a transfer must be much higher than that for most reports.

11.1. Exporting datasets

Some reports can be well defined and are used over and over either within the conduct of a study or across studies. Examples of this kind of report include patient accrual, lists of outstanding discrepancies, and data dumps of lab data. These reports can be considered Study datasets and usually are requested by Biostatistician and PIs for data analysis during the course of the study or at the end of the study.

Exported datasets are usually accompanied with a study data dictionary that includes all description of how tables and variables are designed and explanation for any coding terminologies used in the design. In addition, Annotated case report forms can be generated that is very useful for quickly interpret the meaning of the datasets in the way data is collected.

Data type of the exported datasets varies upon requested by users. It can be in excel spread sheets (xlsx), comma separated value (csv), Microsoft Access (mdb) or SPSS.

The study data manager is responsible for exporting datasets and data descriptions, keeping trach the records and transferring data to the appropriate requestors.

11.2. Creating ad hoc reports

Other reports are used infrequently or are created to deal with a particular data problem or need. These reports are frequently called ad hoc reports. Ad hoc reports are built to answer an immediate question or need and are typically not designed to work in other situations and so are quickly developed.

Ad hoc reports can be designed in advances and reports can be generated following the predesigned at any times of the study. In the other, a particular reports or subsets of the study datasets can be requested by study research staffs such as PIs, coordinators, monitors.

The study data manager is responsible for designing and creating reports keeping track the records and transferring reports to the appropriate requestors.

11.3. Transferring data

It's requires that datasets and study reports should be transferred securely either to internal members or external parties. Data transfer can be done with services that are designed to handle the specific challenges of securely sharing large amounts of data, especially in clinical research. Following are some key considerations for a data exchanging service.

- Encryption: to support encryption both at rest and in transit.
- Access Control: to offer robust access management, allowing you to define user roles and permissions.
- Compliance: to comply with relevant data protection laws (e.g., HIPAA, GDPR).
- Audit Trails: to provide audit logs for tracking access and modifications to data.
- Scalability: to handle large volumes of data as your needs grow.

In Oxford University Clinical Research Unit, exported Datasets or ad hoc reports will be transferred via the OUCRU Sharing File Service hosted at the https://sharefile.oucru.org. The service is in-house developed by the OUCRU IT team and the hosting server is on Microsoft Azure Cloud Service. This is a secured web services integrated with OUCRU Active Directory Service for authentication. Every member has registered OUCRU user name and password is able to access the service for secured sharing data to external collaborators.

When users transfer data, they access the OUCRU sharing file service to create a sharing folder and upload all data files into the created folder. They also need to provide recipient email address and determine a certain period of time the folders being available in the system.

The recipients are going through a 2-factor authentication for downloading the share folder data. They need to have the correct link sent from sharefile@oucru.org. Once the download link is activated, they need to obtain the correct PIN sent from another email to download the data.

All data transferred with the OUCRU sharing file service is encrypted with 256-bit SSL encryption.

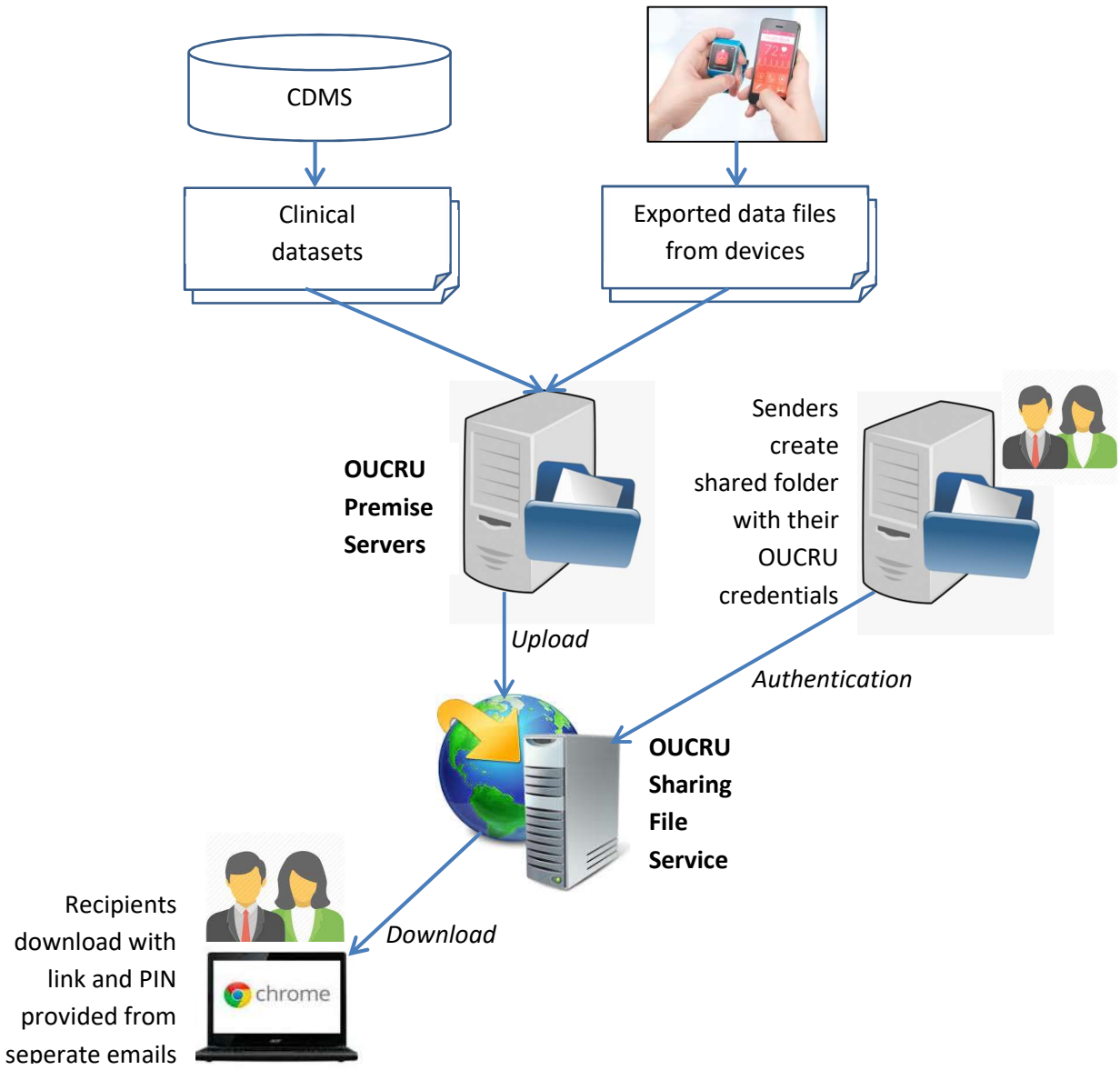

Figure 6. Data transfer workflow

SOPs, Forms and Templates

- Managing Data Transfer (SOP)

12. Data sharing

There has been a move towards greater transparency in design, conduct and reporting in both the public and private sectors over recent years. Better registration of trials has led to a clearer understanding of ongoing research. There is more reporting of results and clearer presentation of data within those reports, e.g. CONSORT. Many funders and sponsors now insist on open access publication whereby results are free to the reader, either immediately or within a set time. In parallel, a number of funders and sponsors have moved towards greater access to data from trials and other projects they have supported. They have done this by implementing policies that encourage or mandate access to

results or data, either via the host institution or through appropriate data repositories. There are a number of advantages for research transparency including the ability to:

(i) Place the results of the study in a larger context, e.g. as part of an individual patient data meta-analysis (IPD MA).

(ii) Make secondary use of the dataset, e.g. using the trial or cohort as a convenience sample of high quality, prospectively collected data to address a different question.

(iii) Collaborate directly with other researchers where data need to be transferred to an alternative location for planned analyses, e.g. as part of biological sub-studies.(iv)Provide supporting evidence to plan a new trial, e.g. estimating the expected event on their control arm, or estimating rates of recruitment

(v) Develop and validate new methodologies, e.g. new statistical methods or biomarker/prognostic indicators

(vi) Independently verify the analyses that the trial team has published or presented.

Every research institution should develop its own data-sharing strategy and policy that aligns with the open access requirements of sponsors and funders, while also ensuring compliance with local and global regulations, such as data privacy and protection laws. The strategy and policy should include but not limited to following aspects:

- Estabishing a Data Access Committee or relevant board to review and make decision on data sharing requests and collaboration.
- Develop a data anonymiztion or pseudonymization process for data sharing
- Deposit data in a data repository

13. Training

Just as standard operating procedures (SOPs) are required as infrastructure, so is training of staff, or perhaps more significantly, documentation of that training. Training is listed explicitly in the regulations, as we see in the ICH document “E6 Good Clinical Practice: Consolidated Guidance” which states: “2.8 — Each individual involved in conducting a trial should be qualified by education, training, and experience to perform his or her respective task(s).” 21 CFR11 echoes this in section 11.10, which states that the procedures and controls related to maintaining electronic records will include: “(i) Determination that persons who develop, maintain, or use electronic record/electronic signature systems have the education, training, and experience to perform their assigned tasks.”

The place to start with training is to decide who is trained on what. A training matrix where roles identify kinds of training that staff members must have in each of four different areas. The training table has columns not just for SOPs but also for CDM system training, pertinent guidelines, study specific training, and practice or tests. The SOP training applies if there is an SOP that covers the role or task in question. The system training refers to any training required for using the CDM system or other software. The guidelines column lists all the guidelines that apply. The “Study Specific” column indicates that there is study specific training required for each particular study. For example, someone joining a study already in progress to help with discrepancy management should be trained on the study’s edit checks and on any study specific discrepancy handling instructions before he or she begins work. Finally,

the test column indicates whether there is a test, practice, or work review required before the person can perform the task on production or "live" data.

ROLE	SOPs ON	CDMS	GUIDELINES	Study Specific	TEST or PRACTICE
Data Entry Staff					
Data Entry Supervisor					
Principle Investigator					
Study Coordinator					
Site Data Manager					
Biostatistician					
Research Nurse					
Study Doctor					
Study Monitor					

Data Managers will be responsible for coordinating with other stakeholders to perform training courses to appropriate staff when needed.

14. Summary

In summary, this framework provides the critical aspects of the data management for clinical research institutions to handle the vast and diverse datasets generated during research. Such a framework ensures data is collected, stored, analyzed, and shared in a secure, compliant, and efficient manner. It enables researchers to meet the stringent regulatory requirements, such as those related to data privacy and protection, while also fulfilling the open access mandates from sponsors and funders. Key components of the framework included in the three stages of the project lifecycle provices a comprehensive set of flows that enables clinical research institutions to improve data integrity, support collaboration, and enhance the reproducibility of their research outcomes.

However, as clinical research evolves, the framework must be continuously maintained and adapted to keep pace with the development of new Clinical Data Management Systems (CDMS) and the growing integration of AI in research practices.

References

ICH. (n.d.). *ICH E6 (R2) Good Clinical Practice.*

SCDM. (2013). *Good Clinical Data management Practices.* Society for Clinical Data Management.

About Author:

Hồ Văn Hiền
Head of IT and Data Management
Centre for Tropical Medicine
Oxford University Clinical Research Unit
764 Vo Van Kiet, Dist.5, Ho Chi Minh City, Vietnam
E-mail: hienhv@oucru.org; hien.ho@ndm.ox.ac.uk

Linkedin: https://www.linkedin.com/in/hien-ho-van/

Mr Ho Van Hien is currently the Head of IT and Data Management of the Oxford University Clinical Research Unit (OUCRU). OUCRU is a part of the Centre for Tropical Medicine and Global Health at the University of Oxford (UK). He has been working for more than 25 years in information technology and more than 17 years in clinical data management. In addition to qualifications in Information Technology, he got an FHNW's EMBA on Management Consulting, NUS' Chief Technology Officer Certification and SCDM's Certified Clinical Data Manager (CCDM®)

In 2006, He was a key collaborator in creating data management infrastructure in Viet Nam for the South East Asia Infectious Disease Clinical Research Network (SEAICRN) by holding the position of Regional Data Manager and sitting on the Data Management committee.

In addition, he designed and created a sophisticated data management system for clinical trials called CLIRES that has been used for most of OUCRU research projects. The system has also been widely used by the International Severe Acute Respiratory & Emerging Infections Consortium (ISARIC) for a number of studies, including SPRINT SARI and the WHO Clinical Characterisation Protocol, and by Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) as a core electronic data collection platform.

He currently focuses on enhancing the infrastructure for bioinformatics with Openstack cluster and promoting the development of an enterprise data warehouse (EDW) for the OUCRU Programme.