Clinical Data Management Plan - Guidance

Version 1.0

About EDCTP

The European & Developing Countries Clinical Trials Partnership (EDCTP) is a public– public partnership between 14 European and 16 African countries, supported by the European Union.

EDCTP's vision is to reduce the individual, social and economic burden of poverty-related infectious diseases affecting sub-Saharan Africa.

EDCTP's mission is to accelerate the development of new or improved medicinal products for the identification, treatment and prevention of infectious diseases, including emerging and re-emerging diseases, through pre- and post-registration clinical studies, with emphasis on phase II and III clinical trials. Our approach integrates conduct of research with development of African clinical research capacity and networking.

About TBVI

The Tuberculosis Vaccine Initiative (TBVI) aims to support, integrate, translate, and prioritise R&D efforts to discover and develop new TB vaccines that are accessible and affordable for all. In an effort to optimise the discovery and development of new TB vaccines and biomarkers, TBVI facilitates and supports the generation of new knowledge and exchange among R&D partners. TBVI creates an enabling environment for consortium members to promote knowledge sharing through scientific meetings and workshops, publication in scientific and non-scientific journals, formal and informal networking.

This document was developed by TBVI, in collaboration with Patrick O'Meara and Richard Liwsky (C-Path), as one of the deliverables of the project 'Development of tools and documents to support coordination of EDCTP TB-vaccine funded research', which is part of the EDCTP programme support by the European Union. The document reflects the views of the authors. The European Union is not liable for any use that may be made of the information contained herein.

We gratefully acknowledge the contribution from all EDCTP-funded TB vaccine projects to the development of this document.

For more information about this document, please contact the EDCTP Secretariat at info@edctp.org.

Contents

1 Introduction

A core aspect in supporting the success of vaccine discovery and development is the use and accessibility of quality data. The collection of clinical data possesses the immense potential to drive progress in the tuberculosis (TB) healthcare arena by providing the means to measure outcomes, to develop and refine best practices, and to enable rapid discovery and innovation. With clinical data growing exponentially in volume, and with a substantial number of different stakeholders managing and processing data within the vaccine development arena, the risk of divergence in approach, data content and quality becomes increasingly probable. Such potential divergences also increase the risk of data being misinterpreted, resulting in inaccuracies with regards to data analysis and modelling efforts.

In an effort to minimise such risks and to establish a consistent approach, this document has been developed to cover the following areas:

- Recommendations on Data Management Plan content and approaches for clinical trial studies
- Recommendations regarding the sharing of data for secondary research purposes
- Considerations regarding FAIR (findability, accessibility, interoperability, and reusability) data principles alignment

For all recommendations outlined within this document, it is expected that all relevant regulatory requirements (e.g. ICH-GCP, General Data Protection Regulations (GDPR) etc.) and relevant organisational Standard Operating Procedures (SOPs) for a specific activity are adhered to where applicable. In addition, compliance with policies, terms and conditions of the funding agency (ies) is required. Importantly, any data collection should take place within an ethical framework, ensuring the protection, rights and freedoms of data subjects.

2 Clinical Data Management Plan

This section focuses on the recommended content expected within a Data Management Plan (DMP) for a clinical study. The core purpose of a DMP is to act as the central document to clearly present all key activities conducted by the relevant data management (DM) service group and to provide an overview to other relevant functional groups of the activities undertaken by the DM service group and their contribution to such tasks.

It is recommended that an initial DMP is available before data processing commences and should be updated regularly and periodically throughout the lifecycle of the study. The DMP should also address any procedural or protocol updates that are made during conduct of the study.

The DMP is an auditable document and is commonly used by auditors to verify the project team's adherence to the procedures described as well as a means of identifying perceived gaps in activity. Therefore, the DMP should clearly describe all aspects of the data management process and/or where applicable reference Standard Operating Procedures (SOPs) and/or other study specific documents that further describe the activity in question.

The following sections provide recommendations on specific topics or components of the DMP. If these components are fully and clearly described in other documents (e.g. SOPs and other study specific documents), it is recommended that references are made within the appropriate section of the DMP. The order of these topics or components do not necessarily need to be presented in the exact order as prescribed below, however it is recommended that all are appropriately broached and presented within the study DMP.

2.1 Study Protocol Overview

For the purpose of providing clarity and context, it is recommended that a broad overview of the protocol should be provided in a dedicated section, including a reference to the reader of the full study protocol and/or study synopsis, including the stored location where the document(s) can be accessed. A specific focus should be made on describing the objectives of the study as well as highlighting the primary, secondary and other objectives/endpoints for the study.

In the event of protocol amendments, the key changes between the relevant protocol versions can also be described within this section if the change has a significant impact or influence on the data management activities.

2.2 Scope and Responsibilities

To ensure there is clarity both within the DM service group and functions participating within a specific study, it is recommended that an overview or table listing out the various activities undertaken by the DM service group is clearly outlined.

This section should also highlight any vendors that the DM service group are working with (e.g. eCRF developed using a vendor platform), and any functions or third parties who conduct an activity that the DM service group collaborate directly with (e.g. external data reconciliation activities) or that the DM group has an upstream dependency on.

References to scope of work documentation can also be referenced here if deemed appropriate.

2.3 Data Management Personnel and Training

Further to **Section 2.2**, it may also be useful to describe the roles and responsibility of each type of personnel assigned to the study by the DM service group. Alternatively, an external document can be referenced which provides such information.

This section should also address expectations and requirements on study specific training for DM personnel, including descriptions on how such training is formally documented and baseline training requirements per role. Where applicable, SOPs and/or other study specific documents can be referenced if these provide sufficient information pertaining to DM specific training within the study.

2.4 Project Timelines

It is recommended that a list or table highlighting the key milestones for the DM service group are included in the DMP highlighting the expected completion targets for all key deliverables. Examples of milestones include the following:

- Protocol Finalisation
- Case Report Form (CRF) Approval
- Database Go – Live
- First Patient In (FPI)
- Last Patient Last Visit (LPLV)
- Data Entry Complete
- Last Query Out (LQO)
- Final/Interim Database Lock
- Archival Completion

In the event of more detailed project timelines and plans being maintained for the purpose of set up, maintenance and database lock activities, it is recommended that a reference of these plans is provided to the reader along with its available location.

2.5 Case Report Forms (CRFs)

When describing the process for developing the CRF for a specific study, it is recommended that the DMP describes or refers to the appropriate SOP which addresses the following topics below:

- **Process overview:** Provide an overview of the development process including describing the following aspects:
  - Describe any rate-limiters or requirements to proceed with the process (e.g. final protocol required)
  - Refer to any relevant CRF library standards used to develop the CRF (e.g. CDISC CDASH)
  - Describe how the CRF is developed, who reviews and approves the CRF content.
  - Describe how CRF review findings or comments are documented and addressed
  - Describe how CRF design changes are managed during the lifecycle of the study (e.g. changes required due to a protocol amendment)
- **CRF completion guidelines:** Provide a reference to the CRF Completion Guidelines document (or equivalent) including any other resource materials which provide all relevant personnel clarity on how to accurately complete the CRF. The guidance document(s) should ideally address the following:
  - Provide comprehensive guidance on the completion of each unique CRF form
  - Provide agreed conventions on reporting/flagging of: missing data, partial dates or unknown values
  - How to handle data corrections/updates
  - Stipulate expectations on data entry turnaround (most notably for Electronic Data Capture systems)
  - Process for requesting access to an eCRF (if applicable)
  - If an eCRF is utilised and form dynamics have been applied (i.e. certain forms or visits will only appear based a specific detail being reported), guidance's on these triggers should be provided
  - If the eCRF is utilised to alert a specific function of specific study event (e.g. Serious Adverse Event, Hy's Law Event etc.), detailed guidance should be provided to ensure such events are appropriately reported.

2.6 Database Set-up and Maintenance

2.6.1 General

The DMP should refer to any relevant SOP which describes the database set up activities for a specific platform. Either within the DMP or associated SOP the following items should ideally be described:

- System name and version (e.g. Medidata RAVE EDC, Version 3.6.2)
- Rate-limiting factors for the initiation of database development (e.g. requirement of a final annotated CRF)
- System configuration processes and documents that are required to be in place prior to initiating database design activities
- Data Integration process set up (e.g. web service integration with an IVRS platform) including references to data integration specification documents
- Reference to database specification documents
- Testing plan(s) and procedures for testing
- Tracking and resolution of test findings
- Process on managing database changes during the maintenance period (i.e. post-production updates)
- User roles and privileges.

Details addressing the topic of edit check testing will be covered in **Section 2.7.2**.

2.6.2 Paper/Hybrid Studies

2.6.2.1 Paper CRF Scanning and Tracking

In the event of the study having a paper (pCRF) component, the set-up of the paper scanning and tracking platform should be described, or an appropriate SOP referenced.
Subject to the capabilities of the platform used, the scope of testing should also be able to demonstrate that:

- The assigned barcodes per pCRF scan automatically and correctly to the relevant CRF type/page (e.g. Visit 2 – Vital Signs, page 37)
- Where a CRF workflow is employed (e.g. CRF moves to Pass 1 Entry to Pass 2 Entry buckets), testing should demonstrate that the CRFs move through the workflow correctly.

2.6.2.2 Data Entry

Where a pCRF component is utilised and data entry is conducted by the DM service group, the process to describe the entry testing into the database should be elaborated upon within the DMP or referencing the relevant SOP.

In terms of entry testing, it is recommended that the following activities are conducted before study start using a test dataset.

- Entry testing should include test/dummy CRFs which encompass all subject types per study protocol (e.g. Screen Failure, Early Withdrawal, Study Completion etc.)
- Should demonstrate Pass 1 and Pass 2 entry procedures (where double-data entry is applied) for all unique CRFs
- The data extracted after entry completion matches the expected data content as per test/dummy CRFs to ensure there are no data extraction issues.

2.6.3 Electronic Data Capture (EDC)

Where an eCRF component is utilised on the study, the basic recommendations for testing align with the recommendations described in **Section 2.6.2.2**. In addition, the following additional elements should also be considered (where applicable) and described in the DMP or relevant SOP:

- Describe how each individual user role is being tested (e.g. Data Manager, CRA, Principal Investigator)
- Testing of form dynamics, field calculations, entry and view restrictions.

- Verifying field/form status workflow as expected (e.g. if data point is updated field/form status changes to requiring SDV etc.)
- Verify that the correct user roles can see, raise, answer and close the relevant query types (e.g. system generated, DM manual queries, CRA manual queries etc.)
- Verify that the correct user roles can see the relevant status reports applicable to their role and that the information presented therein reflects the information reported in the test and/or production environment.

2.7 Edit Checks

2.7.1 General

In addition to CRF and database design, the finalised database that reflects the approved eCRF or pCRF will include data validation checks (or edit checks) to ensure that data issues are detected and subsequently addressed accordingly with the relevant study site. Subject to the Clinical Data Management System (CDMS) platform utilised, other types of edit checks may include programmatic triggers to generate derivation values (e.g. BMI measurements), trigger the appearance/disappearance of specific forms/fields (e.g. Pregnancy Test form will only appear if subject is reported as female on the Demographics) or the sending of e mail alerts (e.g. SAE alert e-mails to the Pharmacovigilance group).

A Data Validation Specification (DVS) or equivalent should be created to document all data validation checks applied within the CDMS platform or ran external from the CDMS to detect and address data issues. If necessary, separate documents can exist to document system checks and manual edit checks (e.g. SAS listings). It is recommended that each defined check should at least contain the following details:

- **Edit Check ID:** Unique edit check identifier (e.g. DM001)
- **Version:** In the event of post-production changes, a means to version track each edit is recommended
- **Form/Module:** Defined unique CRF(s) or dataset(s) that this check applies to.
- **Field Name/Variable:** Field(s) or variable(s) involved in the program
- **System/Manual:** Indicate if the data validation check is programmed within the CDMS or is conducted externally via programmed listings
- **Check Type:** Indicate of check is a query, derivation, e mail alert, etc.
- **Description:** Describe the intent or scenario of the edit check (e.g. Date of Randomisation should not be prior to Informed Consent Date)
- **Logic:** Provide the general logic that should cause a query to fire or populate in a listing. For example, Date of Randomisation should not be prior to Informed Consent Date, the logic could be represented as follows:
  (DM.RANDAT is not AND ICF.CONSDAT is not null) AND (DaysBetween (DM.RANDAT, ICF.CONSDAT)<0)
- **Query Text:** Standard text to be utilised for querying purposes.
- **Listing Output:** Where manual listings are generated, the expected output (e.g. columns/headers) and sort order of output should ideally be defined.
- **Additional Instructions:** In the event of very specific scenarios, it may be useful to provide additional guidance to the data manager and/or programmer on how to approach such scenarios.

Within the DMP itself a description of the DVS development activity should be provided, or relevant SOP being reference. In addition, the actual DVS documents created should be reference including where the reader can access the referenced documents.

2.7.2 Edit Check Testing:

The DMP should describe the approach to testing system and manual checks or reference the appropriate SOPs. For the testing of each iteration/version of the DVS the following is recommended:

- A testing plan(s) and procedures for testing (e.g. test scripts)
- Use of test data to ensure all expected scenarios are robustly tested for all edit check types

- A means to track and resolve test findings including use of a pass/fail table
- Use of audit trail reports, screenshots, and other reference files (e.g. e mail alerts, test datasets, listing outputs, etc.) to demonstrate the successful outcome of a specific check.

2.8 CDMS Workflow

Within the DMP, it is recommended that there is a dedicated section which describes the overall workflow of data processing activities within the CDMS. It is recommended that the following topics are addressed within the DMP, references to other relevant documents (e.g. SOPs) can be made where applicable:

- **pCRF Scanning and Tracking:** In the event of a pCRF being utilised, the process of retrieving, tracking, and scanning of such forms should be described.
- **Data Entry:** Describe the responsible party or parties involved in the entry of data within the CDMS platform. In the event of Double-Data Entry (DDE) being conducted, the process for adjudicating differences in first and second round entries should also be described. References to the relevant CRF completion guidelines should also be made.
- **Self-Evident Correction (SECs):** If applicable to the study, the process for applying and quality control of self-evident corrections should be described. The DMP should also include the current list of SEC scenarios that can be applied on the study, and how approvals to apply such actions are tracked and obtained from study sites.
- **Query Generation / Data Clarification Forms (DCFs):**
  - **System Queries:** Describe the types of system queries being generated, who are responsible in responding to such queries, and who are responsible in closing each query type. Frequency of edit checks being triggered should be referenced if not generated automatically upon data entry.
  - **Manual Queries:** Define the user types who can raise manual queries with the CDMS and also describe which roles can respond to and close each type of manual query.
  - **Edit Check Listings:** Define the frequency edit check listings review. Describe also how old/ongoing issues are tracked/documented across multiple runs of edit check listings.
  - **Paper DCFs:** If a paper query is sent to site, describe the process for delivering a paper query to study site, how such queries are tracked and subsequently processed.
- **CDMS Review Tracking:** If the CDMS utilised is facilitating the ability to tracking of specific review activities. The DMP should describe this activity including any rate limiters which may affect the user's ability to conduct a specific activity. Examples of the types of reviews include:
  - DM Review
  - Source Data Verification (SDV)
  - Safety Monitoring

2.9 Access Management

Where internal or external access is managed for a specific platform, the process for requesting, granting and revoking access should be described or the relevant document referenced (e.g. SOPs), including any rate limiters (e.g. training requirements) that may affect a user in gaining access to the production environment of a specific platform.

In addition to describing the access procedures, it is recommended that the DMP describes or references quality procedures to mitigate the following risks:

- Users being assigned the incorrect user role
- Users being assigned to an incorrect study or site
- Identify users who have access but no longer are assigned to the study or no longer work for a specific organisation.
- Unauthorized user access

2.10 SAE / AESI Reconciliation

The DMP should describe or refer to documents that describe the process for reconciling Serious Adverse Events (SAE) or Adverse Event of Special Interest (AESI), the descriptions for this activity should include the following:

- **Safety Reports:** Describe the safety reports being received from the relevant Safety Group, including frequency of reception, content and formatting of reports being received to support the reconciliation process
- **Data Elements Reconciled:** Define the variables being reconciled between the CDMS and Safety Reports, indicating if an exact match is required (e.g. MedDRA Preferred Term requires an exact match) for each field being reconciled
- **Issue Tracking and Resolution:** Describe how discrepancies are being tracked and communicated to relevant parties.
- **Approval Process:** Describe the required approval process ahead of a significant deliverable or milestone (e.g. Database Lock)

The programs utilised for the reconciliation of safety data should be defined within the DVS or in a dedicated specification document. The recommendations for these reconciliation checks in terms of content and testing are described in **Section 2.7**.

2.11 Non-CRF Data Handling and Reconciliation

2.11.1 General

The DMP should ideally refer to any source of data that is being received by the DM Service group that is not part of the CDMS, but is required to support data reconciliation or other downstream processes that is ultimately incorporated into final data delivery for analysis and submission. Examples of Non-CRF (or external data) include:

- Central Imaging, Laboratory, ECG Data
- Data from Clinical Trial Management Systems (e.g. site information, protocol deviation reports)
- Interactive Voice Response Systems (IVRS) Data
- ePRO Data
- Local Laboratory Worksheets
- Safety Reports

Within the DMP, it is recommended that each data source is described or referenced and addresses the following topics:

- **Milestones / Frequency:** Define the expected frequency or milestones on receiving data from a specific source
- **Data Structure:** Describe the content and formatting of the data files being received, references to data transfers specifications can be made if applicable.
- **Storage:** Define the locations where each acquired data source is being stored
- **Data Pre-Processing:** Describe or reference any programs applied to the source data prior to further downstream use including any structural checks done as part of the data acquisition process (e.g. SAS PROC compare, conversion of JSON files to SAS, etc.)

2.11.2 Data Reconciliation Process

Where the acquired data is being reconciled against the CRF data, the DMP should address or reference the following:

- **Data Elements Reconciled:** Define the variables being reconciled between the CDMS and the external data source in question.
- **Issue Tracking and Resolution:** Describe how discrepancies are being tracked and communicated to relevant parties.

The programs utilised for the reconciliation of the external data source should be defined within the DVS or in a dedicated specification document. The recommendations for these reconciliation checks in terms of content and testing are as per **Section 2.7**.

2.12 Local Laboratory Data

Local laboratories are typically not able to perform electronic data transfers to a DM Service group, sites are usually required to report reference range information via CRFs or an approved worksheet template.

The DMP should ideally describe or refer to how the reference ranges are delivered to the DM Service group, addressing the following topics:

- How/where are results and reference ranges reported by sites
- How are changes in references ranges throughout the course of the study being managed
- Describe how reference ranges are linked to local results (if reported in separate locations)
- What consistency checks are in place to ensure that the reference ranges per analyte are accurate including references to programming specification documents where applicable.
- How inconsistencies are addressed with a study site or local lab.

2.13 Medical Dictionary Coding

The DMP should indicate which medical coding dictionaries (e.g., MedDRA, WHO DD) and the version of the dictionary that will be used for the study. In addition, the period of dictionary version updates should be specified if known. It is also recommended that the DMP should also address the following points:

- A full list of all datasets/CRFs in which coding is being applied for each dictionary type/version.
- There should be descriptions or references on how dictionary updates are being handled including the validation steps to ensure that the update has been successfully applied.
- Descriptions for auto-encoding or study-specific conventions used including the maintenance of any synonym lists supporting the coding process should be referenced within the DMP
- Process for review and approval of coded terms should be described, including the personnel responsible for conducting such review. The timepoints and/or frequency of such reviews should also be specified.

Where such conventions are not study specific, references to the relevant SOP will be sufficient. In addition, specification documents pertaining to the set up and implementation of the coding procedures should be referenced where applicable.

2.14 Protocol Deviations

A protocol deviation is described as an alteration to an approved study protocol. The detection and reporting of such non-compliances are important in addressing any site performance or training issues and may influence the number of valid subjects per analysis population.

Subject to project scope for the assigned DM Service group and organisational SOPs, the DMP should describe or refer to the following topics:

- **Protocol Deviation Reporting:** Describe where site and/or CRA personnel report deviations detected during the course of the trial, and how such details are delivered to the DM service group. If the reported protocol deviations are a non-CRF data source, then refer also to **Section 2.11.1**.
- **Protocol Deviation Detection:** Describe any programmed reports used to support the detection of protocol non-compliance. The responsible parties reviewing the reports and specification documents defining the programmed reports should also be referenced where applicable. It is recommended that reports which capture protocol deviations should be programmed at the beginning of the study and ran frequently to monitor compliance.

- **Subject Classification:** Describe the process on how protocol deviations are classified, including how this process feeds into determining the analysis population(s) of a specific subject. In the event of periodic classification meetings, such activities should also be described in the DMP.

In the event of deviations beyond the control of the study (e.g. as a result of natural disasters, regional conflicts etc.), such cases should ideally be centrally reported as part of the Protocol Deviation process and classified in a manner that distinguishes these cases from more typical deviations (e.g. site errors/negligence). If an alternative route or method of reporting is required/expected, then such an approach should be described and/or referred to within the DMP.

2.15 Status Reporting

The DMP should ideally provide an overview of the reports that the DM Service groups are generating for both internal and external use, it is recommended that the following information is provided per formal report generated:

- Report Name
- Summary of report contents
- Intended audience
- Frequency of report generation

In the event of a report requiring customised programming or is generated externally from a specific platform (e.g. CDMS), the following information is recommended to be described or referenced within the DMP:

- Reference to data specification documents
- Testing plan(s) and procedures for testing
- How issues are tracked and resolved during the testing process
- How changes in programming (e.g. due to a protocol amendment) are managed.

2.16 Unblinded Data Handling

2.16.1 General

Where the study has a blinded component, the DMP should describe or refer to the responsible parties who own or generate the information which has the potential to unblind, whom within the study will have access to this information, and the associated process to ensure that the blind is maintained during the expected period of a study.

The DMP should ideally also describe the conditions or circumstances which permit unblinding per source, and the general process steps to be followed in the event that a member of the study team is inadvertently unblinded.

2.16.2 External Data Reconciliation

In situations where external data results or files have the potential to unblind a study team member (e.g. PK, antibody results), there may be a requirement to reconcile these data sources during the course of the study to minimise the potential bolus of data issues towards the end of an important deliverable (e.g. database) and/or to ensure critical samples are available in time for analysis.

Where data reconciliation is required, factors to consider include:

- Whether this source of data requires a dedicated unblinded team to conduct the reconciliation process and address issues directly with site and/or data provider?
- Are there alternative reports (e.g. sample inventories) which can be shared with the blinded team to facilitate reconciliation that will not compromise blinding?

Within the DMP, the approach to reconciliation should be described as per the recommendations outlined in **Section 2.11,** this should also include descriptions on who is responsible for conducting these activities including any access controls to this data in the event of a dedicated unblinded team being needed.

2.17 Outbound Data Transfers

In addition to the acquisition of data from various sources, the DMP should also describe or refer to activities pertaining to the preparation, delivery and documentation of data transfers delivered by the DM Service to other internal functions (e.g. statistics group) or to an external party.
Within the DMP it is recommended that the following topics are covered per data delivery pathway and type:

- Deliverable Name
- Purpose of the deliverable
- Data Structure or Format (e.g. raw or SDTM version x)
- Internal location where outputs are stored
- Frequency of transfer
- Data Cleaning Level (e.g. dirty transfer, 100% clean)
- Method of delivery (e.g. sFTP)
- Data recipient

Where an outbound transfer delivery requires programmatic transformation activities, the following information is recommended to be described or referenced within the DMP:

- Reference to data specification documents describing the outbound transfer in question
- Testing plan(s) and procedures for testing
- How issues are tracked and resolved during the testing process
- How changes or updates to data transfer outputs are managed during the course of the study.

2.18 Database Lock

2.18.1 Clean Data Definitions

For the purpose of transparency and clarity, it is recommended that there is a clear definition agreed upon by all parties on what constitutes a clean data transfer per lock milestone. In the event that there are multiple locks during the course of the study (i.e. interim locks) it is recommended that the criteria for each lock (if different) is clearly specified within the DMP.

In addition to the agreed definitions, the references used (e.g. status reports) to measure and confirm that the specific database lock criterion for clean data has been achieved should ideally be described in the DMP.

2.18.2 Database Lock Procedures

In addition to the clean data criterions, it is recommended that the DMP does describe or refer to the process steps and responsible parties in preparing for and documenting the necessary approvals for database lock and data release. This should also include any upstream processes which require formal approvals prior to proceeding with the final steps for database lock (e.g. medical/safety review approvals, medical coding approval etc).

2.18.3 Database Unlock

In the event of a database unlock being required for a specific reason (e.g. a significant entry error detected during statistical analysis), the DMP should describe or refer to the steps and approvals required to correct the data source and to redeliver the affected datasets to the relevant recipient.

2.19 Data Quality Assurance Procedures

In addition to the data review and reconciliation processes being conducted by the DM service group, it is recommended that additional measures or safeguards are in place to ensure the overall quality and

compliance in data processing meets the needs and objectives of the project, and to potentially address any personnel training issues.

Examples of quality assurance measures conducted during the course of a study includes:

- Independent review of completed data listings and closed queries for a random subset of subjects to identify the degree of error in correct resolution of data issues
- An independent review of all closed queries pertaining to datapoints considered critical (e.g. primary endpoints) to ensure appropriate handling and resolution of a data specific discrepancies
- Programmatic quality checks on critical data to ensure data quality is appropriate for study analyses as outlined in the Statistical Analysis Plan (SAP) or equivalent.

Dependant on the organisations standard approach or mandated study specific requirements, the quality process should ideally be described or referred to in the DMP, including addressing the following topics:

- A list and description of quality checks being conducted
- Frequency of running and reviewing these quality checks
- Definition of thresholds or quality gates to determine that the overall quality of data is considered acceptable
- Description of how remedial actions for cases which fail the accepted threshold or quality gate are handled and documented.

2.20 Data Storage and Backup

Where the DM service group stores data within their internal servers and/or is supported by a third-party vendor (e.g. an EDC vendor), the DMP should describe or refer to the measures in place to mitigate data loss and ensure a robust disaster recovery process is in place.

2.21 Data Archival

Per ICH-GCP guidelines, a copy of clinical data must be retained at the investigator site throughout the defined records retention period.

The DMP should ideally describe or refer to the process in ensuring that all relevant sites have been given the relevant clinical data records.

For studies using pCRFs, this is normally be achieved by site keeping a copy of the paper records at the site and verified by onsite monitoring activities. For EDC studies the following topics should ideally be addressed within the DMP:

- Define the rate limiting factors in initiating the data archival process
- Reference specification documents that define the configuration or design of the outputs being generated (if applicable)
- Indicate responsible parties in generating and conducting QC of the outputs generated
- Describe the method (e.g. CD) and responsible parties in disseminating the relevant data to a specific study site.
- Describe how site confirmation of receiving the files and verifying opening of files is tracked and documented
- Where a database is being decommissioned (e.g. by an EDC vendor), the criteria for initiating decommissioning should be described.

2.22 Study Documentation

If not specified in a study File Management Plan (FMP) or SOP, the DMP should ideally provide an overview of the study file maintenance process conducted for the DM Service group portion of the Trial Master File (TMF). Topics that ideally should be described or referenced in the DMP include:

- Information whether a paper TMF (pTMF) or an electronic TMF (eTMF) is being utilised for the storage of key DM documentation
- Details on the eTMF platform that is being utilised for file storage (if applicable)
- In the event of a pTMF being maintained, details regarding where scanned copies of the originals are maintained as back-ups.
- Provide or refer to the filing structure being adhered to by the DM Service group
- Described the QC process to ensure that all relevant sections of the TMF have been appropriately fulfilled with correct documentation and that any gaps found have been addressed.
- Where the pTMF is being transferred periodically or at study end to an agreed recipient (e.g. study drug asset owner), the process to document the delivery of the pTMF should also be described.

2.23 Other Data Processes

The details provided throughout **Section 2** cover the more common areas of clinical data management activities, however subject to study or organisational specific requirements there may be additional activities which are required to support DM processing, and ultimately support the data deliveries for statistical analysis purposes.

As a general rule of principle, any inbound or outbound process that the DM Service group is directly involved in which supports the overall data delivery should be documented within a dedicated section or subsection of the DMP. These sections or subsections should describe or refer to the processes to define, set up and validate the activity in question, including describing the responsible parties and overall workflow of this activity.

3 Data Sharing

The data generated during a clinical trial is key in determining the safety and efficacy of an investigated compound. Previously, the data generated for a clinical trial has remained confidential with only results being made more openly accessible.

With the emergence of various initiatives encouraging data openness to support innovation and progress within the vaccine discovery and development arena, grant-holder obligations for data sharing, as well as the requirement for a data-sharing statement for medical journals there is an increased shift towards having data (not just results) openly accessible.

Prior to making data available within either an open data sharing platform or via direct communications with authorised external parties (e.g. academic researchers), the following key elements should be addressed accordingly:

- Data Ownership
- Informed Consent
- Data Anonymization
- Data Collaboration and Harmonisation
- Data Sharing Agreement(s)
- Data User Agreement(s)
- Data Repository considerations
- Data Sustainability
- Data Use and Publishing

The points above, will be further described in Sections **3.1 – 3.10.**

3.1 Data Ownership

Prior to engaging in data sharing activities, the clinical data contributor should verify that they are the owner of all data generated and are authorized to share such data for secondary research purposes. It is recommended that ahead, or at the early onset of a clinical study, the legal underpinnings of data ownership and sharing should be clearly established by the relevant stakeholders, this is especially important in situations where a prospective compound is being developed or co-owned by more than one entity.

In addition to the legal ownership and authority to share data, the relevant stakeholders should also identify if any data generated within the clinical trial intended for sharing contains any propriety information. Upon identifying propriety information, the process for redacting or deidentifying such data should be defined within the Data Sharing Agreement (DSA).

3.2 Informed Consent

3.2.1 General Considerations

The basic principle behind informed consent is to protect the autonomy of clinical trial subject and, in particular, to ensure that the welfare and interest of the subject is the priority. Strict procedures are in place to review all clinical trial proposals, and to ensuring the adequacy of the informed consent procedure. These ethical principles which govern informed consent do have an impact on the rights to collect, process and share data. In light of the potential future use of the clinical trial data, it is recommended that there is adequate wording within the informed consent form (ICF) to describe to the subject; (1) the potential of secondary research, (2) the conditions, or reasons where secondary research may arise, and (3) providing assurances on the measures to protect personal data, such as anonymisation/pseudonymisation, (4) how data subjects can exercise their rights.

3.2.2 Genomic Data

The challenge in sharing genomic data is that the information can be so rich that it is theoretically possible to identify a specific research participant by looking at the single dataset alone and without referring to or requiring any further sources of information.

Subject to relevant regulatory requirements, and to facilitate the sharing of this data for secondary research purposes, it is likely that the ICF requires explicit and specific information, therefore the consent should refer to a clearly defined research project(s) and cannot be overly generalised. In the event of future research projects being identified, and subject to the relevant regulatory requirements, the need for reconsent is by the research participant is highly likely.

3.2.3 Research Participant Data

Anonymisation/pseudonymisation is seen as a means by which personal data can be rendered so that it can be processed further without harming the privacy of a research participant. The anonymization is fair if the possibility of reidentifying a research participant is nullified.

In accordance with ICH GCP 4.8.10(o), the subjects participating within a clinical trial will be assigned a study specific identifier for data reporting purposes in order to not be directly identified by name, however, special care should nonetheless be taken in ensuring that no inadvertent release of data occurs that could directly or indirectly identify an individual.

In general, the risk of re-identification increases with small studies or studies with only a small number of sites. Similarly, rare patient populations may also increase the risk of reidentification. Special care is needed when assessing the level of anonymization chosen for these studies due to the high level of personally identifiable information.

3.3 Personal Information

In order to further protect the subject's identity, the table below provides a list of types of personal information which should be redacted or removed. It is worth noting that the information in the table below also includes data which is not conventionally reported in a clinical trial setting, however, in the event that the open data sharing platform may also include other sources of personal health information such items are nonetheless worth highlighting. In all cases below, data redaction or removal is recommended.

Types of Personal Information	
Geographic subdivisions smaller than a state, county, or province.	Medical record numbers.
Postal codes.	Health plan beneficiary numbers.
All elements of dates directly related to an individual, including birth or death or dates of health care services or health care claims.	Account numbers.
Names and/or initials.	Certificate/license numbers.
Postal address information, other than town or city, state, and postal/ZIP Code.	Vehicle identifiers and serial numbers, including license plate numbers.
Telephone/Fax numbers.	Device identifiers and serial numbers.
Electronic mail addresses.	Web universal resource locators (URLs).
Social security numbers or other government identification credentials.	Internet protocol (IP) address numbers.
Socioeconomic details.	Biometric identifiers, including fingerprints and voiceprints.
Social identity.	Cultural Identify.

Pregnancy history.	Full-face photographic images and any comparable images.

3.3.1 Additional Privacy Safeguards

To quote GDPR Article 4(5) pseudonymisation is the,
"processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person"

To further mitigate the risk of subject identification, the following safeguards are recommended to be applied prior to the data in question being made available for secondary research activities:

a) Site/Subject Identifiers:
To minimise the risk of identifying a specific geographical location where a research participant may reside, the following measures are recommended:

- Removal of information which may identify the Principale Investigator.
- Removal of Study Site Name/Number. If for technical reasons a site number needs to be populated within the relevant datasets, then it is recommended that a pseudo-site number is assigned.
- If the subject identifier number contains the site number, at a minimum the original site number should be replaced by a pseudo-site number within the subject identifier.
- Additionally, to "break the link" between the source records/data and the data openly shared, in general it may be advisable to generate a unique pseudo-subject number which can include also incorporate the pseudo-site numbering convention described above.

NB: If there are sites with a very small number of subjects, it is recommended that a group of sites are merged together and assigned the same pseudo-site number, in such cases the subject number would need to be fully replaced.

b) Central / Local Service Providers
If a study is collecting data from central or local service (e.g. Imaging, ECG, PK, haematology etc.) the following sets of information should be removed, or a pseudo-value assigned if a variable is considered mandatory:

- Removal of local/central service provider name/details

c) Removal of reference identifiers (e.g. accession number for a laboratory sample, device number, etc.)
d) Dates

- All actual dates reported within the clinical trial should be removed
- Start and End Day variables should be utilised to identify the timepoint when an event or observation had occurred (e.g. vital signs taken at day 26)
- If there is a need for the date field not being empty, the recommendation would be that a false date is applied based on an offset value which is randomly generated per subject (see example below).

Original Date	Amended Date	Additional Information

Reference Date	19-Jul-16	28-Sep-16	Reference date is normally date of randomisation or date of initial study drug exposure Random offset = 71
Date of Assessment	14-Aug-16	24-Oct-16	Random offset = 71
Study Day	26	26	Unchanged/Unaffected

e) Geographic Information

As per Section 5.3.2.2.2 of the external implementation guidance on EMA Policy 0070 (https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-1.pdf) the following is stated;

"It might be necessary to aggregate or generalise from country to region or continent unless this information is critical to the analysis. The need to aggregate or generalise should be considered when performing the risk assessment. The link between individual patient data and the identity of the site should be removed since a frequency analysis can most likely reveal the most recruiting site in a country, which will in turn include many of the trial participants. However, it may not be the case where the recruitment is uniform across all sites".
Information should only ideally be reported at a minimum on a country level only. If this is too specific for certain study cases (e.g. rare diseases), then a larger geographical area should be used (e.g. region or continent).

f) Demographic Information

- Ethnicity: To safeguard cultural identity, details reported regarding the research participant ethnicity should not be available within the data output (unless needed for the purpose of the study and adequately justified).
- Ages 90 and above: Following the principles within the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, it is recommended that any subject aged 90 or above is assigned a general category value of **">=90"** as opposed to their actual age.

3.4 Data Structure

To optimise the value of the data, and alignment with FAIR data principles (specifically interoperability) the data presented for secondary research purposes should be displayed in a standard data structure including the use of standard terminologies.

Due to its prevalence of use within the drug development industry, including supporting regulatory submissions (e.g. EMA, FDA and PMDA), the recommendation is for the open sharing of data to be presented in an agreed version of CDISC SDTM for tabulated data and CDISC ADAM for analysis datasets.
In the event of another data structure being utilised, the basic principles of all projects presenting the data using the same datasets, variables and terminology conventions is key in ensuring interoperability for all data sources made openly available.

3.4.1 Data Collaborations

If the collaboration for open data sharing involves more than one entity who owns or contributes data, upon initiation of such a collaboration, the following items should be considered by all parties:

- What types of data can be made available?
- Data standard (including version) being utilised per study
- Terminologies/ dictionary standards applied per study

Upon the establishing the details for each point above, the collaborative group in question will then need to address the following topics:

- Identify the difference in how data is currently being presented by each group
- Based upon the gap analysis conducted, establish what common standard (including terminologies) that should be utilised by the group for open data sharing.
- Establish a plan or roadmap in applying the changes per study to align with the agreed standard for open data sharing.

3.4.2 Data Harmonisation

In the event of a data collaboration for open data sharing requiring a specific subset of data to be transformed into the agreed data standard as per **Section 3.4.1**, the following steps are recommended:

- Establish the responsible party for transformation (e.g. data owner or a single dedicated group)
- If transformation is delegated across multiple groups, a single standard template for data mapping specifications should be established by the group.
- Establish a dedicated Data Management Plan (DMP) outlining the process to harmonise the data for open sharing use. At a minimum, the DMP in question should address the topics as outlined by the European Commission at (https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm).

Aside from assigning the responsible parties and establishing a common template for documentation purposes, the actual transformation process itself should be traceable for auditing purposes and be able to demonstrate the following:

- Demonstrate data provenance from source
- All programs used to support the activity are version controlled
- All programming steps are traceable via a log or report.
- Demonstrate the transformation has been fully validated and aligned with agreed data structure and terminology standards (including any documentation tracking issues to its successful conclusion)
- Demonstrate that the defined data privacy safeguards have been correctly applied.

The steps describing the data harmonisation process (and associated templates) should be described in a Standards Operating Procedure (SOP) or Project Specific Procedure (PSP) document.

3.4.3 Metadata

Where SDTM and/or ADAM datasets are utilised for open data sharing, it is recommended that Define-XML document or equivalent is made available to represent the metadata for data artefacts such as case report forms (CRFs) and datasets created for use in clinical research. The Define-XML for example also contains detailed metadata describing data elements, controlled terminology, and the methods used for derivations and transformations of the data. Any proprietary data within the Define-XML or equivalent file should be redacted or removed in line with the data outputs for data sharing.
In the event of sharing data in a non-CDISC format or a Define-XML document not being sharable, the metadata should be presented in a consistent structure within the relevant group or consortium. The metadata should be able to inform the end user of the following key elements at a minimum:

- Datasets, variable and value-level metadata
- Analysis results metadata (where applicable)
- Study metadata
- Study code lists metadata
- Links to supportive documents (where applicable)

The availability and transparency of such metadata will in turn minimise the risk of researchers misunderstanding the data presented, and in turn ensures that data is accurately analysed for secondary research purposes.

3.5 Data Sharing Agreement (DSA)

To ensure accountability and to assign clear responsibilities to all parties involved, a relevant data sharing agreement should be established.

At its core, the DSA is an agreement between the data owner (or controller) and the data processor (or sharer) which not only addresses the scope between both parties, but also defines the terms of use, disclosure and the defined mitigations to nullify data privacy concerns including research participant information. The DSA should also reference any relevant regulatory requirements or principles which the terms of the agreement are expected to adhere or align to.

The table below provides recommendations on the key topics that the DSA should contain within the context of Clinical Trial Data.

Topic	Content
Objectives	Provide background information and define the core objectives of the project. Describe the intended use of the data being provided.
Terms of Use	Describe and define how the data can be used by approved users.
Data Summary	Provide non-confidential details of the types of data being shared, including descriptions on formatting, data content and ownership of data. Any supportive documents to provide clarity on the data content should also be referenced within this section.
Data Transfer and Storage	Define the methods on how data will be securely transmitted and subsequently stored, including describing the responsible party (or parties).
Grant of Licenses	Define the terms for which the data owner grants the sharing of data.
Scope of Disclosure	Describe how the data will be disclosed for downstream use (i.e. research). Factors to consider include if certain milestones are to be met prior to the data being allowed to be more openly accessible (e.g. CSR report for project x to be submitted to prior to data being available).
Personal Data Protection	Provide details on how research participant data will be protected including any methods/requirements to further anonymise data prior to the data being openly available. The responsible parties in applying these safeguards should also be defined.
Data Updates/Removal	Describe the process and responsible parties in handling data corrections or removal. **NB: The topic for "Right of Erasure" in accordance to GDPR should also be addressed in this section.**
Approved Users	Define the user profile(s), including affiliates and other relevant initiatives which will have the authority to access the data to support secondary research objectives.
Data Access Responsibilities	Define the responsible parties for managing data access, including describing how access authorisation will be managed.
Data Retention	Describe who is responsible for retaining the data and the duration of data retention. Factors for ensuring data sustainability should also be highlighted.
Provisions for Termination	Describe how breaches in agreement are to be managed and the provisions for terminating the agreement in question.

3.6 Data Use Agreement

The terms for data sharing by the relevant stakeholders who are facilitating data accessibility to the relevant end user (i.e. researcher) should be clearly defined. In addition, an agreement should be in place, outlining provisions for data use for the purpose of secondary research, to be agreed by the prospective user in question. Such terms should be formally acknowledged by the prospective user prior to gaining access to the data.

Whilst creating the relevant Data Use Agreement (DUA), the factors described in **Sections 3.6.1 – 3.6.3** should discussed and addressed by the relevant stakeholders involved in the data sharing initiative.

3.6.1 User Access Criterion

At the early stages of a data sharing initiative, the relevant stakeholders need to agree and determine the criteria by which the end user will be given authorised access. Example criteria include:

- The prospective user is a member of an approved organisation or affiliate
- The prospective user meets the baseline academic qualifications
- The research proposal presented by a prospective user align with the provisions for data use
- Only specific defined user profiles will have access to specific types of datasets

Subject to a consensus being established on the required criterion, these details can be incorporated as part of the DUA itself or as a standalone application form which is to be completed by the prospective user. This consensus will also influence the approach in user access management as described in **Section 3.7**.

3.6.2 Data Sharing Method

Prior to determining the specifics of the Data Use Agreement (DUA), the relevant stakeholder(s) who are facilitating data sharing should at an early stage establish the mode/method by which the data will be made accessible to the end user, examples include:

- Data directly shared to data researcher by data owner or processor.
- Access provided to a data repository which allows the user to interrogate data and to download the datasets for secondary research use
- Access to a platform which only allows the researcher to analyse the data within the platform environment, allowing only the results generated being exportable.

Contingent to the mode/method of data sharing being utilised, the DUA may be one of the following:

- A standalone document which needs to be formally signed off by the researcher prior to the provision of access and/or data.
- Integrated into the data repository/platform requiring an eSignature by the prospective user. If utilising such an approach, the platform in question should be CFR 22 Part 11 compliant.

3.6.3 Terms of Data Use

The DUA should also explicitly define the terms by which the user is to adhere whilst accessing and utilising the data for secondary research purposes. The following terms are recommended as a minimum requirement:.

- End user will not attempt to reidentify subjects
- Data will not be disclosed beyond the defined scope of use
- The necessary safeguards for ensuring data is protected from misuse (applicable where user will be storing data locally)
- Will report incidence of data breach or misuse
- The results or intellectual property generated from the data will not be patented
- Citation of relevant stakeholder(s) that facilitated the sharing of data used in a publication

- Abide the defined terms and restrictions governing publications.
- The relevant stakeholder(s) that facilitated the sharing of data will be indemnified and held free of liability.

Subject to the project objectives and the nature of data being shared via a formalised process additional terms of use may need to be considered.

3.7 Data Repository Considerations

In the event of a data repository being utilised to facilitate data sharing the following factors should be considered in the selection process of an appropriate data repository platform (e.g. SAS Clinical Trial Data Transparency platform):

- The platform should facilitate the use of Persistent and Unique Identifiers (PIDs) to support data discovery, data searching and retrievals, and version controlling of data.
- Facilitates the input of metadata to enable finding of data including referencing related information (e.g. experimental protocols, publications, etc.)
- Enables the application of industry accepted data structures and metadata standards within the platform
- Facilitates user access management, including specifying conditions under which data can be viewed or consumed (e.g. specific user profiles may only view certain types of data or conduct certain actions within the platform)
- Possess adequate controls in ensuring confidentially rights of research participant data are upheld
- Adequate reporting capabilities including the generation of audit trails by the platform administrator(s).

3.8 Date Access Governance

Upon establishing the user profiles and criteria for gaining access to data the relevant stakeholders within the data sharing initiative should establish the procedures to formally manage and oversee the requesting, authorisation and where applicable the revoking of access. The established process steps should be documented within a dedicated plan for data governance.

3.8.1 Access Requests

Contingent to the technologies being utilised and the anticipated volume of users expected to gain access, it is recommended that the access request process follows one of the following approaches:

- The potential user can access a specific online portal and submit an access request.
- The potential users complete an access request form and submits a scanned copy to a specific e mail address.

Regardless of the agreed method, it is advisable that the application request form/portal should be designed in a manner which addresses the access criterion requirements as described in **Section 3.6.1**. For example, the form can be a combination of requester information and a set of questions that the potential user needs to answer prior to submitting the application for review.

3.8.2 Access Authorisation

In conjunction with the access request process, the relevant stakeholders should establish the application review and approval process. The data governance plan or equivalent should establish if a specific individual(s), entity, or committee will be given the responsibility to access the appropriateness of the potential user based on the details provided during the access request process.

If a significant volume of access requests is anticipated, and an electronic means for access requests is being utilised (i.e. requests submitted online) there is the possibility to further streamline the process by programmatically fast-tracking users to approval stage where the applicant meets the defined user criterion is met. Users that do not meet the necessary criterion via the automated process will require a manual review.

To enable this possibility, the online application form needs to be specific and descriptive enough in order to support such an automated process.

In the event that the defined user profiles for online users have very specific restrictions, it is recommended that there is a formal verification process to ensure that the requestor does not gain incorrect access (e.g. a researcher accidentally being assigned a power user role to the platform).

3.8.3 User Monitoring

Subject to the platform capabilities and the established user profiles (i.e. what a specific type of user can do on the platform), the user's online behaviour may need to be monitored. This for example may entail the system to flag specific actions or the activity is reviewed by a specific group via a system report (e.g. a user audit trail) to discern if any unauthorised activities have been observed.

It is recommended that a risk assessment is conducted to determine need of or extent of monitoring that may be required.

3.8.4 Revoking of Access

The platform being utilised for given users access to a specific set or subset of data should have the capabilities to easily revoke access. Exercising the need to revoke a person's access may come about where there is concrete evidence to demonstrate that the user has violated the terms of the DUA, for example the user may not have followed the agreed protocols for publications using this data.

In addition, the platform should allow for the facilitation of requests regarding "Right of erasure" in accordance with GDPR in cases where a user invokes their right to be forgotten.

3.9 Data Sustainability

As part of the coordinated effort to facilitate open data sharing, the collaborative group or initiative should determine how the data can be made sustainable both during and beyond the lifecycle of the current initiative to ensure continued use of the data to support further advances in drug development tools and approaches.

3.9.1 Data Interoperability

As per **section 3.4**, key decisions may need to be made by the relevant stakeholders in order to harmonise the data to ensure consistency of presentation and ease/interpretation of use by a prospective researcher. For such a data structure to have value both during and beyond the existence of an initiative the data structure should possess the following core facets:

- Data structure should align with a recognised industry standard (e.g. CDISC SDTM, HL7 etc.)
- A rich resource of metadata is made available to the researcher
- Terminologies and ontologies used are aligned with recognised industry standards (e.g. NCIT)

3.9.2 Data Hosting and Sustainability Considerations

If a platform is being utilised to host data for the purpose of facilitating open data sharing and/or analysis, the collaborative group or initiative should consider and weigh in on how the platform and/or data may be sustained beyond the current initiative.
Such strategies may include:

- Ongoing sourcing and identification of grants or funds to support ongoing platform maintenance and personnel resourcing.
- Prospective researchers or organisations to pay a subscription fee or provide a donation to support platform maintenance.
- Identify other open research initiatives where the data could be migrated to for continued use (**NB: A DSA should be created to define the terms of such an agreement**).

3.10 Data Use and Publication

In addition to ensuring that the data is being appropriately safeguarded prior to making it available to approved researchers and/or collaborative groups, the risk of the data being misinterpreted or misused is an aspect that does require a form of oversight or review.

Whilst the effort of standardising data and complimenting it with a richness of metadata should in theory reduce the probability of misinterpretation during a research effort, this in itself does not necessarily fully guarantee that there would be no intentional or unintentional misuse of the data which could, in turn, introduce falsified information into the research community and beyond.

Subject to the range and complexity of data being openly shared for secondary research purposes, certain safeguards and conditions should be introduced in an effort to combat such risks. Means of combating such risks include:

- Ensure that the DSA and/or DUA associated with this data have specific legal terms and conditions regarding the appropriate use of this data, and liabilities pertaining to its misuse. Such terms should theoretically act as a deterrent to those who have a predisposed bias or intent to misrepresent the data.
- In relation to the data access application process (**Section 8.3**), it is recommended that each applicant or affiliated research group provides a synopsis of their research intent, including providing known details of their methodology or approach to analysing the data in question. Such details should then be reviewed by a central committee including requesting further information (where appliable) in order to support the application review process.
- The DSA and/or DUA should include terms and conditions regarding the use of the data in publications, including requiring data owner and/or central committee approval prior to being able to publish the results. These conditions may include a specified window where the results using the shared data are made available to the data owner and/or central committee to review and discern if the data has been fairly represented, in addition the analysis plan and other relevant supplemental information should also be made available as part of the review process.
- To address any serious doubts and/or concerns, the DSA and/or DUA could contain a condition allowing the data owner and/or central committee to be able to request and conduct an audit on the research activity conducted on the shared data to verify soundness in the methodologies used.

4 FAIR Data Considerations

Regarding the recommendations outlined in **Section 3**, an area of consideration when it comes to data sharing is to ensure continued value and usability of data that has been curated to support secondary research. Alignment to the FAIR data principles is a means in which such value and usability can be supported and achieved.

These principles precede any specific implementation choices (e.g. choice of technologies, data standards, etc.), and essentially act as a guide to achieve the continued accessibility of the data whilst presenting it in a manner that is interoperable and not prone to inadvertent misinterpretation. The general purpose of this section is to highlight the core aspects of data "FAIRness" to further inform the decision-making process for data sharing activities as described in **Section 3**.

Whenever a data sharing initiative is being considered, it is recommended that these principles are incorporated where possible. Furthermore, the DMP which is being developed to outline the data processing procedures of a specific data sharing project or initiative should also address how these principles are being implemented.

4.1 Making Data Findable

Data is considered findable when the data in question possesses a rich resource of metadata and is appropriately indexed in a searchable resource accessible to potential authorised users. The use of persistent unique identifiers (PUIDs) should be assigned in a manner such that the data can be accurately referenced and cited in research communications or publication, thus mitigating any ambiguities and /or perceived lack of credibility resources used (in this case clinical data).

The use of PUIDs allow for persistent linkages being established between the data, metadata, and associated materials to support data discovery and reuse. Examples of associated materials include:

- Information or code on how specific derivations or algorithms have been applied to a specific data value(s)
- Literature that provides further insights and context on of the study in question or on the results generated from a study.

To summarise, in order to achieve findability, the following items should be appropriately discussed and addressed as part of the data sharing initiative:

- The data and metadata that is intended to be shared can be easily searched within an online portal or platform (e.g. a data repository)
- The data intended to be shared have PUIDs assigned
- The metadata describing the data is rich in content and specifies the relevant PUID

4.2 Making Data Accessible

The agreed access procedure (including authentication and authorisation steps, where necessary), should allow end users and machines (if applicable) to gain access to the data with retrieval of specific data facilitated by the used of persistent unique identifiers.

In the event that the data is not or cannot be made directly accessible, the metadata associated with this data should be made openly accessible and presented in a manner that is broadly understood within the research community.

Therefore, the data being described should utilise recognised specifications and ontologies that determine the exact meaning of what the data in question represents. This implementation increases its "machine-readability" capabilities.

4.3 Making Data Interoperable

The implementation and adherence to standard specifications for representing both the data and metadata is a key component to achieve data interoperability and reusability. As outlined in **Section 3.4**, the topic on standards for a data sharing initiative should be broached by all relevant stakeholders at a relatively early stage to establish a harmonised approach to presenting the information in a manner that is recognisable within the research community.

4.4 Making Data Reusable

To enable data reusability, the need for rich metadata and documentation that aligns with recognised standards and provide information about data provenance are key in achieving such aspirations. This encompasses information on how data was created (e.g. laboratory processes), how the data has been refined or transformed to make the data usable.

To summarise the reusability can be achieved if:

- The data in question is accurately presented and well described in terms of metadata
- There is transparency on how, why and by whom the data have been created and processed by (i.e. data provenance)
- The data and metadata presented meet relevant and recognisable standards.

In addition to the above points, the appropriate data usage licenses should be in place to optimise use and reusability of the data where applicable.

Appendix I – Abbreviations/Acronyms

ADAM	CDISC Analysis Data Model (https://www.cdisc.org/standards/foundational/adam)
AESI	Adverse Event of Special Interest
CDASH	CDISC Clinical Data Acquisition Standards Harmonization (https://www.cdisc.org/standards/foundational/cdash)
CDISC	Clinical Data Interchange Standards Consortium (https://www.cdisc.org/)
CDMS	Clinical Data Management System
CFR	Code of Federal Regulations
CRF	Case Report Form
CRA	Clinical Research Associate
CRO	Contract Research Organisation
DBL	Database Lock
DCFs	Data Clarification Forms
DDF	Double-Data Entry
DM	Data Management
DMP	Data Management Plan
DSA	Data Sharing Agreement
DUA	Data User Agreement
DVS	Data Validation Specification
ECG	Electrocardiogram
eCRF	Electronic Case Report Form
EDC	Electronic Data Capture
EMA	European Medicines Agency
ePRO	Electronic Patient Reported Outcome
eTMF	Electronic Trial Master File
FAIR	Findable, Accessible, Interoperable, Reusable
FDA	Food and Drug Administration
FMP	File Management Plan
FPI	First Patient In
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation (https://gdpr.eu/)
HIPAA	Health Insurance Portability and Accountability Act

HL7	Health Level Seven
ICH	International Conference on Harmonisation (https://www.ich.org/)
ICF	Informed Consent Form
IVRS	Interactive Voice Response System
LQO	Last Query Out
LPLV	Last Patient Last Visit
MedDRA	Medical Dictionary for Regulatory Activities (https://www.meddra.org/)
NCIT	National Cancer Institute Thesaurus (https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/)
pCRF	Paper CRF
PID	Persistent and Unique Identifiers
PK	Pharmacokinetic
PMDA	Pharmaceuticals and Medical Devices Agency
pTMF	Paper Trial Master File
PUID	Persistent Unique Identifier
QA	Quality Assurance
QC	Quality Control
R&D	Research and Development
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SDTM	CDISC Standard Data Tabulation Model for Clinical Data (https://www.cdisc.org/standards/foundational/sdtm)
SDV	Source Data Verification
sFTP	Secure File Transfer Protocol
SOP	Standard Operating Procedure
TMF	Trial Master File
WHO DD	World Health Organisation Drug Dictionary