Data Management

SOP Reference: RGIT_SOP_020	
Version Number: 11.0	
Effective Date: 09 Jan 2024	Review by: 09 Jan 2027
Author: Thomas Lewis, Ethics and Research Governance Co-ordinator	
Approved by: Ruth Nicholson Head of Research and Integrity	Date: Ruth Nicholson Digitally signed by Ruth Nicholson Date: 2024.01.09 12:51:31 Z

Version	Date	Reason for Change
Version 1.0	31 May 2007	Change of office name from CRO to CRGO
Version 2.0	01 Sep 2007	Annual review
Version 3.0	20 Jun 2008	Annual review
Version 4.0	08 Feb 2010	Formation of Joint Research Office
Version 5.0	14 Jul 2011	Annual Review
Version 6.0	30 Nov 2012	Annual Review
Version 7.0	18 Feb 2015	Scheduled Review
Version 8.0	25 Oct 2017	Scheduled Review
Version 9.0	21 Jan 2020	Updated Data Protection Act from 1998 to 2018
Version 10.0	19 Oct 2020	Scheduled Review Templates removed and administrative changes to SOP. JRCO name change to RGIT.

Version 11.0	09 Jan 2024	Scheduled review

Table of Contents

1. PURPOSE

This SOP describes the process for data management for Imperial College Academic Health Science Centre (AHSC) sponsored clinical trials. Specifically, is involved with collecting, validating, analysing and archiving such data.

2. INTRODUCTION

An essential element of conducting a clinical trial is an efficient data collection and management. Only data that is essential for the purposes of the study should be collected as stated in the clinical trial protocol. It is advisable to seek advice from a trial statistician as early as possible in the trial design process to facilitate this. This SOP describes the full data management process including data entry; data cleaning; and resolving data queries. This SOP also describes the use of data monitoring committees (DMCs) for assessing data during interim analyses, and how such a committee should operate.

ICH GCP Guidelines E6 R2 section 5.5.1 states "The sponsor should utilize appropriately qualified individuals to supervise the overall conduct of the trial, to handle the data, to verify the data, to conduct the statistical analyses, and to prepare the trial reports." This SOP will not describe procedures for developing the case report form (CRF)/ data collection tool, as this is covered in the RGIT_SOP_007 (Imperial College AHSC Case Report Form SOP).

3. PROCEDURE

3.1. Data Management Process

The process of data management involves converting the data collected using data collection tools, most commonly case report forms (CRFs), into electronic data that can then be statistically analysed. This SOP will need to be appliedaccording to the size and complexity of the trial being conducted, as smaller trials may not require a full use of the processes described.

3.1.1 Data Management Software

Once the CRF has been designed in accordance with the protocol; the database to store the information collected should be designed. Depending on the size and type of study this database could be a standard spreadsheet, or a more technical data management system (DMS) may be required (for example REDcap). When developing a database please refer to the Data Management Guidance (*cited on 22 Mar 2023*).

For clinical trials involving Investigation Medicinal Products (IMPs), all AHSC sponsored trials must use the OpenClinica electronic data capture system as their CRF unless data management and monitoring have been formally contracted to an external organisation with robust data management systems. In this case, a risk assessment for the use of the external CRF will be performed during the sponsorship review process. Information about OpenClinica can be found at:

OpenClinica - Clinical Trials Unit (*cited on 13 Mar 2023*)

Under ICH GCP E6 R2 there should be a specific SOP for managing the study database in place. The database should allow changes to be made to the data in a documented manner, and it should not delete data entry to ensure an audit trail for the data is maintained (ICH GCP

E6 R2 5.5.3). The database should be secure, an appropriate password-protected access to prevent unauthorised access to the data, with a list identifying those individuals permitted to make changes to the data. TheICH GCP E6 R2 also requires that there is adequate backup for the data, and that, if there is blinding involved in the study, the data entry and processing systems allow this to be maintained (ICH GCP E6 R2 5.5.3).

3.2. Coding CRF Responses

Before any data entry to the database, the responses from the paper CRFs must be coded, using either analphabetical or numerical code that can then be used for analysis, which is different to the MEDDRA coding. These codes should be determined before data entry begins (e.g., codes 0 and 1 for Yes and No respectively). Codes should also be in place for answers such as 'not known' or 'not applicable' (e.g., 999 to show missing data). It is important to make sure that whatever value is chosen to represent missing data, that value would be unfeasible as an actual response. Any data entry should be collected via coded fields, which should minimise the use of text fields. Clinical data also needs to be coded for the recording of ~~all~~ adverse events (AEs, SAEs). It can be used for medical history, but this is not mandatory.
The coding can be done at various stages of the trial such as: during the initial data collection from the participant by the investigator or research nurse; after the data collection, but prior to entering the data on the database; or when the data are entered onto the database.

In relation to the medical coding, The World Health Organisation Adverse Reaction Terminology (WHO-ART) and Medical Dictionary for Regulation Activities (MedDRA) both have a system of coding to assist with this categorised by System Organ class, Preferred Term and Lower Level Term. A code is assigned for each term and adverse reaction. You can access WHO-ART and MedDRA through WHODrug Portfolio and MedDRA respectively (*cited on 27 Mar 2023).* Please note that WHO-ART is no longer maintained and is referenced here for legacy purposes.

Medical coding would be done after data has been entered on the database and performed by the trial manager or a designated person of the trial management team on the study.

3.3. Data Entry

The study specific delegation log should specify which individuals are authorised to make data changes to study CRFs and the database respectively.
For clinical trials involving Investigation Medicinal Products (IMPs), all AHSC sponsored trials must use the OpenClinica electronic data capture system as their CRF, unless data management and monitoring has been formally contracted out to an external organisation with robust data management systems. For non-CTIMPs, it is not mandatory to use eCRFs and the use of paper CRFs is acceptable.

In the case of paper CRFs, on initial receipt of CRF, the form should be date stamped and checked for responses that are incomplete or missing. If any inconsistencies are found, these should be inquired with the investigator and a record should be kept of all the queries. Instructions for sites to respond to data query responses should include no use of Tippex/correction fluid, not to obscure the original data entry and to initial and date any amendments made.

The data must be entered onto the database, by trained data entry staff, once the paper CRFs are completed.For multicentre studies where the CRFs are being sent to a coordinating centre

for data entry, a copy of the CRF should be retained by the investigator, with the originals (usually 2 copies from No Carbon Required Paper CRFs) going to the coordinating centre. The data manager for the study at the coordinating centre must keep and maintain a log of all CRFs received.

All stored CRFs should be kept in a secure environment such as a locked filing cabinet in a locked room. Secure also means protection against environmental damage such as damp or fire.. This should include checking that there are no water sprinklers above the cabinets.

The guidance for the design of paper CRFs can be found in the RGIT_SOP_007 (Case Report Forms).

3.3.1. Double and Single Data Entry with Control Checks

During data entry by trained staff, an average of 5% of errors is expected to occur. Two methods can be used to reduce the risk of errors: Double Data Entry or Single Data Entry with Control Checks. It should be noted that most EDCs (Electronic Data Capture) to date don't support this approach.

i. Double Data Entry

Double data entry involves two people entering the same CRF data onto the database independently of each other. The data may be entered twice onto the database on two separate files, depending on the software used, which are then compared by the system for accuracy. If the two entries do not correspond this would be flagged up by the database. Alternatively, when the second data entry person enters the data, if it differs from that entered by the first person, a message immediately appears on screen and the original data can be verified. This method depends on the availability of a technically capable database, and is usually applicable to older EDCs and paper CRF processes

ii. Single Data Entry with Control Checks

This method may be more suitable for smaller single centre studies with fewer staff available for data entry and/or less sophisticated database software. Once the data has been entered, a visual check can be done between what is recorded on the paper CRF, and what was entered on screen. This should be particularly highlighted for critical data and dates, where an easy typo could be made during the data entry.

3.4. Data Cleaning and Validation

An integral part of the data management process is data validation; to ensure the most accurate 'clean' set of data is provided for the statistical analysis. In OpenClinica and other EDCs, the validation checks are implemented and will run at the entry point. Data validation can be carried out at three stages during the trial:

a. When CRFs are completed by the investigator

To improve accuracy at this stage all staff completing CRFs should be sufficiently trained in their completion. A CRF completion manual would assist with this (see RGIT_SOP_007). Validation should also be carried out as part of the on-going monitoring of the study; either by members of the research team or by independent monitors (see RGIT_SOP_015). Validation via monitoring is done through Source Data Verification (SDV). SDV involves checking the data entered into the CRFs against that in the original source records (e.g., patient's hospital files for accuracy).

b. When data are entered in the database by data entry staff

If the study database has software enabled for automatic data entry checks, where data entry checks are being used, the clinicians/statisticians/data staff involved with the study should put together an Edit Check Specification (ECS) document. This document should provide full details of the data entry checks that have been set up, and all checks should be tested before the trial begins. This can also be documented on a Study User Requirements Template.

Depending on the database software, it is also advisable to set up warnings to alert data entry staff when values are entered outside of the expected range, or if the type of value entered is incorrect (e.g., a numeric value entered rather than text). It is also useful to set up alerts for missing values where possible.

c. When data have been entered and are available for the data manager

At this stage it is advisable to carry out systematic post-entry data checks. Data listings should then be created (either through an automatic database software system, or manually) of the following data queries:

- All missing values will be listed
- All values outside of pre-defined range

Logical checks should also be performed to ensure consistent reporting between relevant fields, withall values outside of pre-defined range listed. This is to ensure there are no implausible differences between fields.

All checks should be defined before the study starts and should be described in the Edit Check Specification/Study User Requirements Template document described previously. Data validation should continue until all missing values and inconsistencies are corrected or clarified.

3.5. Data Protection

During the entire data management and validation process it is essential that all study data are kept in a secure location and in accordance with the terms of the Data Protection Act 2018. Participant confidentiality must be maintained at all times and all study records should be kept in pseudonymised form identifying participants by their study code rather than name, initials or hospital number. Identifiable data should not be captured unless used for ePROs (electronic Patient Reported Outcomes) and eConsent, and the system has correct security protocols in place to protect these data.

Any paper CRFs should be kept in locked filing cabinets in locked rooms only accessible by authorised personnel. The key to the participant code list should be kept separately to these documents, again in a locked, secure location. If paper CRFs must be transferred to a coordinating centre for data entry, they should be sent either by courier or registered post to minimise the risk of losing data. A log should always be maintained of documents sent and received at each centre involved. If electronic data transfer is used, this should be via a secure system, password protected and encrypted where possible.

The database itself should be password protected, with each data entry staff member having their own password. If data entry is performed at the investigator site it is essential that the investigator does not have access to the whole database, to protect against biases occurring due to investigators making decisions based on interim data. If a member of staff leaves the study team that information should be communicated so their access can be revoked.

Any data that is stored on Imperial College London networked computers must be stored in an anonymised form (fully anonymised or pseudo anonymised) with no identifiable information.
Any data processing activity that could result in high risk for data subjects are required under the Data Protection Act 2018 to conduct a data protection impact assessments (DPIAs) in order to identify and minimise data protection risk.

To facilitate compliance, the Faculty of Medicine has developed a Data Asset Registration Tool (DART)*(cited on 03 May 2023)*. The Faculty mandates the completion of this tool for research projects that handle health and social care data.
Where any data is stored on a database supported by a web application, please see the College Database Management Systems policy for further information on special Data Protection Act *(cited 08 June 2023)* requirements for such systems.

For further guidance on data protection please refer to the College Data Protection Policy available at: Imperial College London - Data Protection *(cited 08 June 2023).*

3.6. Data Management Plan

Before the study starts, it is essential that a data management plan for the recording of all data processing; management; and validation activities is created, and updated as necessary throughout the study. The data management plan should contain information on the following:

i. Key study contacts details
ii. Study Timepoints
iii. EDC Software Use
iv. Data Entry
v. Data Workflow
vi. Query Management
vii. Source Data Verification
viii. Data Security, Storage & Back up procedures

Although the above list is not exhaustive, it provides a basis for the data management plan that can be adapted and expanded as necessary.

3.7. Data Backup Systems

Whatever the format of the database software used to manage the study data, there should always be a back-up system in place to guard against the loss of data due to software or environmental disasters. The College ICT service has a data backup service that provides a reliable means of protecting data held on departmental and research groups file servers.
ICT does not backup files on local desktop machines. Owners of such machines are responsible for protecting local files.

The use of backup services for servers that are used to support research data exclusively is a charged serviceImperial - File recovery and backup link *(cited on 22 Mar 2023)*

3.8. Independent Data Monitoring Committees (IDMCs)

It is recommended for large, complex trials that an independent data monitoring committee (IDMC) is set up to carry out reviews of trial data at staged intervals during the study. The role

of the IDMC is to review interim results and determine whether or not there are any safety issues or any reason why the study should not continue, e.g., if interim results are showing strong evidence that the treatment/intervention is superior or inferior to the control.

The data reviewed by the monitoring committee should be as up to date as possible and should be validated up to the point of the interim analysis to ensure it is of sufficient quality. The membership of the committee should include experienced trial investigators, statisticians and clinicians; all of whom must be independent to the research team. A DMC charter must be established to monitor members of the committee. The results should be reviewed at regular intervals as sufficient data accumulates. Plans to establish a DMC must be described in the protocol.

If there is a trial steering committee (TSC) for the study, the IDMC would normally make their recommendations for action through them (see RGIT_SOP_015).

3.9. Archiving

Chief Investigators are required to make adequate arrangements for the archiving of all study data (including paper CRFs) and essential documents. This service is provided by Imperial College Corporate Records Unit. Please refer to RGIT_SOP_019 on archiving for further information regarding archiving; this SOP can be found on the SOP, Associated Documents & Templates page. Chief investigators must also make themselves aware of where the study database is to be archived and by whom; for Imperial College sponsored CTIMPs this will be arranged by the clinical data systems (CDS) team.

4. REFERENCES

Guideline on Data Monitoring Committees (*cited on 24 Apr 2023)*

DAMOCLES study group: A proposed charter for clinical trial data monitoring committees: helping them to do their job well (*cited on 24 Apr 2023)*

International Conference on Harmonisation (ICH) of Good Clinical Practice E6 R2 (*cited on 24 Apr 2023)*

Data Management Guidance *(cited on 03 May 2023)*

WHODrug Portfolio *(cited on 03 May 2023)*

MedDRA *(cited on 03 May 2023)*

Data Asset Registration Tool (DART) *(cited on 03 May 2023)*

RGIT_SOP_007_Case Report Forms

Data Protection Act 2018.

Imperial College Data Protection Policy (*cited on 22 Mar 2023)*)

RGIT_SOP_015_ Monitoring