York and Scarborough Teaching Hospitals NHS Foundation Trust R&I Department
Standard Operating Procedure
R&D/S29

Data Management

IT IS THE RESPONSIBILITY OF ALL USERS OF THIS SOP TO ENSURE THAT THE CORRECT VERSION IS BEING USED

All staff should regularly check the R&I Department's website and/or Q-Pulse for information relating to the implementation of new or revised versions. Staff must ensure that they are adequately trained in the new procedure and must make sure that all copies of superseded versions are promptly withdrawn from use unless notified otherwise by the SOP Controller.

The definitive versions of all R&I Department SOPs appear online. If you are reading this in printed form check that the version number and date below is the most recent one as shown on the R&I Department website: www.research.yorkhospitals.nhs.uk/sops-and-guidance-/ and/or Q-Pulse

SOP Reference:	R&D/S29
Version Number:	5.0
Author:	Monica Haritakis
Implementation date of current version:	13th May 2026

Approved by:	Name/Position:	Deborah Phillips, Research Advisor
	Date:	21st April 2026
	Name/Position:	Sarah Sheath, SOP Controller
	Date:	15th April 2026

This SOP will normally be reviewed every 3 years unless changes to the legislation require otherwise

Version History Log

This area should detail the version history for this document. It should detail the key elements of the changes to the versions.

Version	Date Implemented	Reviewers	Details of significant changes
1.0	6th September 2010		
2.0	27th February 2012		Minor changes to incorporate University of York. Edited to apply to non-CTIMP studies. Change of SOP Controller.
3.0	22nd August 2017		Removal of references to the North and East Yorkshire R&D Alliance
4.0	2nd October 2019		Change of author. Change of link to R&D website. Addition of GDPR and updated to DPA 2018. Incorporation of addendum to ICH E6 updates
5.0	13th May 2026	Liz Johnson	Change of author. Minor formatting changes.

Contents

1 Introduction, Background and Purpose

Good Clinical Practice (GCP) states that all clinical study information shall be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification, irrespective of the type of media used. This Standard Operating Procedure (SOP) outlines procedures for managing study data on paper and in electronic systems. It covers the collection, entry, validation and management of clinical data according to the principles of GCP and applicable regulations in order to support statistical analysis and subsequent reporting.

A fundamental factor in conducting a research study is efficient data collection and management. Only data that is relevant for the purpose of the study should be recorded.

In general, the practices required for high quality conduct of research are similar, whatever type of study is involved. In York and Scarborough Teaching Hospitals NHS Foundation Trust, all studies, CTIMP and non-CTIMP, should be run to GCP-equivalent standards to ensure consistent practice and scientific quality.

2 Who Should Use This SOP

This SOP should be used by:

- Chief Investigators (CIs) and research co-ordinators of studies sponsored or co-sponsored by the Trust;
- Principal Investigators (PIs) and research staff at Sites where multi-site studies sponsored or co-sponsored by the Trust are being run;
- R&I Department personnel, who manage the sponsorship of research studies on behalf of the Trust;
- PIs for externally-sponsored studies “hosted” by the Trust where no Sponsor instructions exist.

3 When this SOP Should be Used

In accordance with Good Clinical Practice (GCP) the Sponsor should ensure appropriately qualified individuals are responsible for the overall conduct of the research study, handling the data, verifying the data, conducting the statistical analyses, and preparing the study reports.

The Sponsor should normally delegate data management within a research study to the Chief Investigator (CI). Where the CI further delegates data management to another member of the research team this should be clearly outlined on the Delegation Log.

4 Procedure(s)

4.1 Basic Terms

4.1.1 Source data

Source data are all information in *original* records and certified copies of original records of clinical findings, observations, or other activities in a study necessary for the reconstruction and evaluation of the research. Source data are contained in source documents (original records or certified copies). The investigator/institution should maintain adequate and accurate source documents and trial records that include all pertinent observations on each of the site's trial subjects. Source data should be attributable, legible, contemporaneous, original, accurate, and complete. Changes to source data should be traceable, should not obscure the original entry, and should be explained if necessary (e.g., via an audit trail)

4.1.2 Electronic Source (eSource) Data

Source data captured initially into a permanent electronic record. N.B. In this context 'permanent' means that all changes in the data are recorded in an audit trail (the minimum standard for this is a record of who made the change and when).

4.1.3 Source documents

Source documents are original documents and records where study data are *first* recorded e.g. hospital records, clinical and office charts, laboratory notes, memoranda, subjects' diaries or evaluation checklists, pharmacy dispensing records, recorded data from automated instruments, copies or transcriptions certified after verification as being accurate copies, microfiches, photographic negatives, microfilm or magnetic media, x-rays, subject files, and records kept at the pharmacy, at the laboratories and at medico-technical departments involved in the study. In some circumstances the Case Report Form (CRF) may be considered a source document. This should be clarified with the Sponsor at the time of set up and documented.

Source documents are considered "Essential Documents" that allow evaluation of the study and ensure the quality of the data and serve to certify Sponsor and CI compliance with the relevant regulatory requirements. The process referred to as Source Data Verification (SDV) is an evaluation of the data recorded in the data collection tool against the source documents. Information in source document should mirror that entered in the CRF.

4.1.4 Case Report Form (CRF)

A printed, optical, or electronic document designed to record all of the protocol required information to be reported on each study participant. The sponsor should ensure that the investigator has control of and continuous access to the CRF data reported to the sponsor. The sponsor should not have exclusive control of those data.

4.1.5 Certified Copy

A copy (irrespective of the type of media used) of the original record that has been verified (i.e., by a dated signature or by generation through a validated process) to have the same information, including data that describe the context, content, and structure, as the original. When a copy is used to

replace an original document (e.g., source documents, CRF), the copy should fulfil the requirements for certified copies.

4.1.6 Validation (of Computerized Systems)

A process of establishing and documenting that the specified requirements of any computerized system can be consistently fulfilled from design until decommissioning of the system or transition to a new system. The approach to validation should be based on a risk assessment that takes into consideration the intended use of the system and the potential of the system to affect human subject protection and reliability of trial results.

4.1.7 Monitoring Plan

A document that describes the strategy, methods, responsibilities, and requirements for monitoring the trial. This is outlined in R&D/S08

4.1.8 Protected Personal Information (PPI)

Protected personal information is any patient identifiable data received from the NHS; or any information that links an identifiable individual with information that, if released, would put them at significant risk or harm or distress; or any source of information relating to 1,000 or more individuals not in the public domain, even if the information is not considered likely to cause harm or distress.

4.2 The Data Management Plan (DMP)

The DMP is a document that describes and defines all data management activities for the study and should ideally be in place prior to the start of recruitment and definitely before any data queries are raised. The extent of the data management activities described in the DMP will be dependent on the complexity of the trial and the associated risks. Relevant sections of the DMP must be reviewed by the R&I Department on behalf of the Sponsor and evidence of any input documented. Any subsequently amended versions of the DMP should also be reviewed and approved by the Sponsor prior to implementation

The DMP should include:

- All data collection tools to be used
- The data management system to be used
- Process for maintaining the blind if applicable
- Data Entry requirements
- Data verification requirements (may be a reference to the applicable section of the monitoring plan)
- Data validations included in the electronic CRFs (eCRFs) and other external data validations and consistency checks to be carried out and the timing of these validations
- The query process
- Content and frequency of data quality reports
- The process and timings for Serious Adverse Event (SAE) reconciliation
- Data coding guidelines where applicable

- The process for handling any additional data from other systems along with quality checks required to ensure that the data is consistently and accurately imported.
- Pre interim data release/database lock checks
- Process for database lock and release of data
- Process for data back-up and archiving

4.3 Data Management Process

The data management process involves evaluating data collected using data collection tools, normally referred to as CRFs, which are transferred into an electronic format to allow statistical analysis to be conducted.

4.4 Data Management Databases

Note: All computer systems, both hardware and software, being utilised for the collection and analysis of research data, are expected to have undergone full validation.

Once the CRF has been designed in accordance with the protocol and has been approved for use, the CI (or delegated individual/s) should develop an appropriate database to store the relevant information. The type and size of the database will depend on the size of the study and could vary between a standard Excel spreadsheet (minimum) to a more technical Data Management System (recommended). Any database should be designed in such a way that data can be exported to a data analysis packages (e.g. SPSS) with minimal effort.

The following instructions may therefore need to be amended according to the size and complexity of the study. The following points should be addressed when developing a database system:

(a) Ensure data integrity
- Raising data queries
- Traceability of data corrections

(b) Ensure the security of the database and study data
- Develop back up and disaster recovery plans

(c) Maintain the database according to user requirements
- Verify appropriate levels of access (unique username and password)
- Be aware of appropriate software upgrades relating to the database and action and document these upgrades

(d) Provide training for relevant staff on the use of the database

(e) Ensure database systems are checked/ maintained on a regular basis

(f) Define database lock-down procedures to ensure access to the final dataset is permanently restricted for final analysis and report, prior to archiving

(g) Ensure archiving of data is maintained correctly

(h) Ensure that the database will be complete and accessible throughout the retention period by looking at digital preservation issues.

Researchers are advised to use commercially produced and validated software wherever possible. Specialist software that has been produced 'in-house' or as a one-off application by a commercial company must be subjected to strict validation processes (R&D/T10) that must be agreed in advance by the Sponsor. The sponsor should base their approach to validation of such systems on a risk assessment that takes into consideration the intended use of the system and the potential of the system to affect human subject protection and reliability of trial results. Results of validation testing should be filed in the Trial master File (TMF).

It should be noted, in addition, that computerised Laboratory Information Management Systems which capture analytical results of tests conducted during a research study are considered part of the data management and investigators should seek assurances at the study planning stage that the accreditation status of the laboratory computerised system is suitable.

Stored data should be held on an appropriate NHS Trust or University server with security/access requirements.

4.5 Data entry, data processing and data validation

Quality control should be applied to each stage of data handling to ensure that all data are reliable and have been processed correctly. If data are transformed during processing, it should always be possible to compare the original data and observations with the processed data.

4.5.1 Data entry

When the CRF has been received by the delegated individual responsible for data management (e.g. Data Manager, Trial Manager), the form should be reviewed for any missing data, incomplete fields or data outside normal ranges. If any discrepancies are raised at this stage, these must be clarified with the PI (or delegated individual) at the originating centre and any queries recorded.

Where paper CRFs are in use correction Fluid must not be used for any data query responses and original data entry must not be obscured. Any amendments made on the CRF should be initialled and dated by the PI or delegated individual/s. A record of all amendments should also be recorded.

The data entry process should be defined for the specific study. Users should always check their work for accuracy and completeness and may request a second person checks the data entry if appropriate.

Overall the Principal Investigator (PI) should ensure the accuracy, completeness, legibility, and timeliness of the data reported to the CI/Data Manager in the CRFs and in all required reports. In addition data reported on the CRF, that are derived from source documents, should be consistent with the source documents or the discrepancies should be explained.

If required for a study based on risk assessment, software should allow printing of individual records that can be printed and counter-signed by the PI when checked against the CRF.

4.5.2 Data processing

The following factors should be followed when processing research study data:

- All transactions to the database (insert, update, delete) must have a clear and complete audit trail. For some software (e.g. Excel) this may necessitate the printing of data and the certification and dating of the data as an accurate record of the previous and current versions of the database.
- Data should only be accessible to authorised personnel
- The Data Handler must comply with GCP and is responsible for keeping data secure and confidential at all times
- Coding should be performed using appropriate dictionaries (e.g. MedDRA,)
- Where autocoding is not possible, manual coding may be performed.

4.5.3 Source Data Verification (SDV)

Source Data Verification (SDV) for all required data should be carried out during monitoring visits in accordance with the agreed study specific monitoring plan. In addition, study site personnel may also perform SDV during self-monitoring. Other data checks and central data monitoring may be performed and will be detailed in the DMP. In some instances direct data verification can be performed on studies where the paper CRFs or other source data e.g. Patient Reported Outcome Measures (PROMs), are held and the data entered by R&I Department study staff.

4.5.4 Data validation

An essential aspect of data management is the process of data validation. It is important that all data validation/cleaning is carried out in accordance with the documented plan for a study. The frequency and methodology for the validation checks to be performed will be decided at the start of the study and documented in the DMP. This process aims to ensure the most accurate validated set of data is provided for statistical analysis. Data validation can be undertaken at three stages during the study:

1. **When CRFs are completed by the PI or delegated staff member**

   To ensure accuracy all staff completing CRFs should be sufficiently trained to do so. It is advisable to validate some data early in the study so that any issues of lack of accuracy can be addressed at an early stage. SDV should therefore be conducted as part of the on-going monitoring of the study either by members of the research team or by independent monitors.

2. **When data are entered in the database by data Entry staff**

   Where data entry checks are used, if the data management system has software enabled for automatic data entry checks, this should be validated before the study begins. Depending on the database software used, some may have built in checks that can be viewed and exported directly.

   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.

Logical checks should also be performed to ensure consistent reporting between relevant fields and that there are no implausible differences between fields e.g. male and pregnant. All checks should be defined before the study starts. Data validation should continue until all missing values and inconsistencies are corrected or clarified. These checks can also be performed manually on an ad-hoc basis to check for any other inconsistencies which require clarification.

4.5.5 Electronic Data Handling

When using electronic trial data handling and/or remote electronic trial data systems, the sponsor should:

a) Ensure and document that the electronic data processing system(s) conforms to the sponsor's established requirements for completeness, accuracy, reliability, and consistent intended performance (i.e. validation). The sponsor should base their approach to validation of such systems on a risk assessment that takes into consideration the intended use of the system and the potential of the system to affect human subject protection and reliability of trial results.

b) Maintains SOPs for using these systems.
The SOPs should cover system setup, installation, and use. The SOPs should describe system validation and functionality testing, data collection and handling, system maintenance, system security measures, change control, data backup, recovery, contingency planning, and decommissioning. The responsibilities of the sponsor, investigator, and other parties with respect to the use of these computerized systems should be clear, and the users should be provided with training in their use.

c) Ensure that the systems are designed to permit data changes in such a way that the data changes are documented and that there is no deletion of entered data (i.e. maintain an audit trail, data trail, edit trail).

d) Maintain a security system that prevents unauthorized access to the data.

e) Maintain a list of the individuals who are authorized to make data changes

f) Maintain adequate backup of the data.

g) Safeguard the blinding, if any (e.g. maintain the blinding during data entry and processing).

h) Ensure the integrity of the data including any data that describe the context, content, and structure. This is particularly important when making changes to the computerized systems, such as software upgrades or migration of data.

4.6 Data Discrepancies and Queries

In addition ongoing external data validation/data cleaning checks should be performed on a periodic basis to facilitate central monitoring of the data and to provide additional information in reports for Data Monitoring Committee (DMC) and Trial Steering Committee (TSC) meetings.

- Data entry should be monitored by review of the database against numbers randomised and reaching each time point
- Data entry should be monitored for missing data across sites
- Data should be reviewed for potential discrepancies in primary and secondary outcomes information
- SAE reconciliation should be carried out periodically in accordance with the study specific reconciliation process documented in the DMP.
- Reconciliation should be carried out prior to generation of Development Safety Update Report (DSUR) or DMC reports and finally prior to database lock.

Discrepancies, identified during the review of the automated validation output or following manual validation checks, will be sent to the appropriate member of the research team for resolution.

The automatically raised validations will be checked for each participant, so that any data that is not within the expected range for the study will be flagged.

(a) Discrepancies will be flagged to sites
(b) External validations identifying any discrepancies (e.g. missing or inconsistent data) will be raised as queries in the eCRF or notified to sites in an agreed and systematic manner
(c) The method for tracking and resolving discrepancy queries will be decided at the start of the study and documented in the DMP.
(d) Changes made to the data following resolution of a discrepancy, will be documented as part of the audit trail.
(e) Following any data changes, the data validations will be re-run until no further discrepancies are identified.
(f) A summary of the results of the validation checks will be documented and considered at regular intervals
(g) In circumstances where discrepancies are not able to be resolved, the Data Manger (or other individual with this responsibility) will document this prior to release of any interim or final datasets.

4.6.1 Data Changes

Amendments to data recorded on CRFs and eCRFs should always be handled at the local site by research staff. Data Managers/Trial Management personnel must not amend data themselves.

- Corrections to data on eCRFs should be made directly within the eCRF by site personnel to ensure there is an appropriate audit trail.
- Corrections to paper CRFs should be made by drawing a single line through the incorrect item and dating and initialing all corrections in pen.
- If paper CRFs are entered centrally, when completing a query, an amended copy of the CRF must be attached to the original CRF
- For electronic eCRFs these should be set up and auditing/data logging modules enabled so all corrections and changes made can be viewed.

4.7 Quality Monitoring

Data quality including errors identified during SDV, query rates, missing data, and safety data reconciliation issues will be monitored and reported to the Sponsor and the relevant study committees (e.g. TSC, DMC). The frequency and content of data quality reports should be detailed in the DMP.

4.8 Data Transfer

The following sections provide a description of the processes to be followed when conducting data transfer.

4.8.1 Paper CRFs

Trial Management Personnel are responsible for maintaining the following:

- Paper CRFs should be stored securely at all times and only be accessible by authorised personnel e.g. in a locked filing cabinet in a locked office.
- If paper CRFs are transferred for data entry, they should be sent by courier or registered post to ensure safe delivery.
- A log should always be maintained of documents sent and received at each site involved.

4.8.2 Electronic data transfer

Electronic data will be risk assessed before the transfer of the data and the risk assessment results documented in the DMP. The data transfer will be done either by email attachment, secure file transfer systems (e.g. Egress) or other removable media. Transferred data should be encrypted and passwords should be employed when transferring data. The following must be taken into consideration:

- Strong passwords of at least 15 characters should be used.
- Passwords for encrypted data must always be transmitted to the recipient of the data by a separate route, e.g. by telephone.

4.9 Data Backup and User Access Control

All data stored on computer systems must have adequate backup procedures in place. Satisfactory disaster recovery capability must be in place and documented.

Users should have password limited access to any database systems, which restrict access to their own particular role and site. Access should only be granted following confirmation that the user is on the appropriate delegation log(s). Access to the database must be removed from a user once they are no longer part of the study.

4.10 Data Management document control

All data management documentation must be version controlled and superseded in a controlled manner. All data management documentation forms part of the TMF and must be archived with all other TMF documents at study close.

4.11 General Data Protection Responsibilities

(a) Throughout the data management process it is vital that all study data are kept within a secure location and in accordance with the terms of the Data Protection Act 2018 and the EU General Data Protection Regulation as detailed in the regulatory applications.

(b) The investigator/institution should have control of all essential documents and records generated by the investigator/institution before, during, and after the trial.

(c) Any PPI should be kept within secure premises and secure systems. PPI should not be accessed via remote access unless the machine/device used to access the data is fully encrypted or the machine/device used to access the data is kept within the research site's premises at all times and held securely (i.e. locked away when not in use, not left unattended whilst in use and not used in a public or general access area). The downloaded material on the machine/device used to access must be wiped as soon as no longer required.

(d) Paper CRFs should be kept in locked filing cabinets in locked rooms only accessible by authorised personnel. Investigator Site Files (ISFs) and other files making up the TMF should be stored in lockable filing cabinets. Metal cabinets are advised.

(e) If paper CRFs must be transferred to another site for data entry, they should be photocopied and a copy retained at Site. Original CRFs should be sent either by courier or registered post to minimise loss of data. A log of documents sent should be maintained at the Site. It is good practice for the research team to initial and date the copy to evidence the original CRF has been sent to the Sponsor. N.B If the CRF is designated as source data the original should be retained at site and a certified copy sent to the Sponsor

(f) Copies of faxed CRFs must be retained in the ISF with the date of faxing logged.

(g) If electronic data transfer is used, this should be via a secure system, password protected and encrypted. If transferred via email, the password should never be transferred in the accompanying email.

(h) If electronic CRFs or any other data defined as PPI are transferred or stored off site using removable media (including laptops, memory sticks, smart phones, CDs, and portable hard drives) they must be encrypted. The transferred material on the removable media should be deleted as soon as transfer is successfully effected. Best practice would be to avoid the need to transfer any PPI outside of the originating organisation.

(i) All documents must be fully anonymised prior to postage. This includes images such as x-rays. The responsibility to ensure that this is the case resides with the PI (or delegated individual). Packages must be sealed securely before leaving the care of the investigator team to minimise document loss. If this is not possible due to the requirements of the collecting courier company then a delegated individual must be responsible for securely sealing the parcel once the contents have been verified.

(j) The database management system should be password protected, with each member of the research team responsible for data entry having their own password.

4.12 Database Lock and Study Close Down

The CI is required to provide the Sponsor with a copy of the locked database (where applicable) following completion of participant recruitment/follow up. The final database must be "locked" to ensure access to the final dataset is permanently restricted for final analysis and report, prior to archiving. For each study the specific data lock procedures should be documented in the DMP for that study.

A copy of the locked database should be provided to the Sponsor and a study close down visit undertaken before analysis can commence.

For randomised trials sponsored by the Trust, the final locked database should be confirmed by the Sponsor before code breaks are released.

The CI should consider the following points to ensure that all activities are completed prior to database lock:

- All research study subjects must complete their final visit and any follow-up visit activities
- All coding of clinical events must be completed
- All outstanding queries are resolved and database updated
- Any Serious Adverse Events (SAE) queries have been resolved and database updated
- Notification to all members of the study team of the proposed date of lock
- Research staff should make every effort to complete any outstanding queries by the data lock deadline date.

5 Related SOPs and Documents

International Conference on Harmonisation (ICH) of Good Clinical Practice

Data Protection Act 2018

EU General Data Protection Regulation

Acknowledgement

We thank Cardiff University for their kind permission to adapt and reproduce their SOP.