DATA CLEANING AND QUALITY SOP

VERSION 1.0

APPROVALS

Author	**Position**	**Signature**
Lindsey Masters	Clinical Project Manager	Lindsey Masters Digitally signed by Lindsey Masters Date: 2019.09.03 19:57:28 +01'00'
Reviewer(s)	**Position**	**Signature**
Nafisah Atako	Clinical Project Manager	Nafisah Atako Digitally signed by Nafisah Atako Date: 2019.09.04 17:07:01 +01'00'
Ben Sydes	Data Manager	Ben Digitally signed by Ben Date: 2019.09.05 15:10:56 +01'00'
Claire Cook	Data Scientist	Claire Cook Digitally signed by Claire Cook Date: 2019.09.03 16:19:28 +01'00'
Marion Hall	Database Programmer	Marion Hall Digitally signed by Marion Hall Date: 2019.09.04 17:23:25 +01'00'
Approver	**Position**	**Signature**
Fleur Hudson	Head of Trial & Study Management	Fleur Hudson Digitally signed by Fleur Hudson Date: 2019.09.03 17:29:24 +01'00'

All appropriate approvals must have been completed prior to uploading to SOPbox.

UPLOAD TO SOPBOX

Name	**Position**	**Signature**	**Date uploaded to SOPbox**
Nancy Tappenden	SOPbox Administrator	Nancy Tappenden Digitally signed by Nancy Tappenden Date: 2019.09.05 20:00:18 +01'00'	06-Sep-2019

The effective date of this SOP is the day on which it is uploaded to SOPbox and is available to use. This is the date associated with the signature of the SOPbox Administrator.

For the Revision History please see the Version History Summary in SOPbox.

DATA CLEANING AND QUALITY SOP

TABLE OF CONTENTS

The following symbols may be used in this SOP:

➡ Indicates a link to a related document

✏ Indicates instructions to document trial-specific processes elsewhere

Throughout this document the MRC Clinical Trials Unit at UCL, will either be referred to as 'MRC CTU at UCL' or 'the unit'. In instances where neither read well in the sentence, 'the CTU' may be used.

1 PURPOSE

The purpose of this SOP is:

- To outline the procedures related to ensuring queries are raised, managed and resolved appropriately
- To detail methods of data cleaning and monitoring throughout the trial

This SOP applies to all CTU studies where the MRC or UCL is the sponsor, or where the sponsor has delegated data handling activities to the CTU, unless it has been agreed with the sponsor that other SOPs will be used.

1.1 DATA MANAGEMENT MODELS

The CTU employs a number of data management models:

1. Use of case report forms (CRFs) which are received by and entered centrally at the CTU. These may be received on paper or electronically, e.g. via secure email. These will be referred to as Paper CRFs throughout the SOP.
2. Electronic data capture (eDC), where sites enter data directly onto the study database into electronic CRFs (eCRFs).
3. Remote data capture (rDC) where data are collected on a paper CRF and then entered on to the study database at a location offsite (i.e. not at the CTU).

Some studies may use a mixture of these models.

2 RESPONSIBILITY AND ROLES

The following table lists the roles relevant to this SOP and a brief description of their responsibilities.

This SOP will be circulated to be Read and Understood to all appropriate roles identified in the training matrix.

ROLE	RESPONSIBILITIES
Trial Management Team	• Identify required database validations and missing data checks • Specify data checks and the frequency of those checks • Define the critical queries that require expedited resolution with sites • Undertake query reviews where appropriate • Define and undertake ongoing data entry QC as specified in the Data Management Plan (DMP) • Ensure data is ready for IDMC, interim and final analyses • Ensure an escalation policy is in place to deal with any issues relating to eligibility, safety or the integrity of the study • Ensure participant related documents are filed appropriately
Data Manager	• List database validations and missing data check for programming • Maintain a list of standard manual queries • Ensure sites are regularly notified of data queries and missing CRFs/eCRFs • Manage data queries to resolution
Clinician	• Provide specialist input to database validations and missing data checks
Statistician	• Provide specialist input to database validations and missing data checks • Document acceptable data quality levels prior to database lock
Database Programmer/ Data Scientist	• Provide specialist input to database validations and missing data checks

3 PROCEDURES

There are a number of methods of data cleaning which are used within the unit. These include, raising and resolving data queries, and identifying and chasing missing CRFs. All of these activities contribute toward ensuring data quality.

There are two main methods of resolving data queries:

1. Queries/issues which need to be addressed immediately, and will require a detailed escalation policy
2. Queries which are saved in the database (either automatically through validations/missing data checks or added manually at data entry/checking) and resolved through routine querying procedures such as query reports.

3.1 DATA ISSUES REQUIRING IMMEDIATE ESCALATION

Some data issues may require immediate escalation, e.g. issues relating to the safety and eligibility of the participants or the integrity of the study. These may be identified

- At paper CRF receipt
- At paper CRF data entry
- During the regular checking process for data that have been entered for eDC/rDC CRFs
- During statistical analysis
- During a query review

The nature and frequency of the checks performed by CTU staff should be risk based and studies should have an escalation policy in place for these checks. It should detail:

- what requires escalation
- who the issue should be escalated to
- a time scale for escalation
- any subsequent documentation that is required e.g. adding the issue to a tracker

Any resulting documentation such as emails should be filed in the participant's file and/or the site file if the issue affects multiple participants at a particular site

✎ Study teams should document which data items may require urgent queries and the processes for dealing with these in the DMP or associated documents.

➡ See the Data Receipt and Entry SOP for more information.

3.1.1 CHECKING ELIGIBILITY DATA

Data to verify eligibility criteria should be collected and checked prior to registration/randomisation occurring, for example if an eligibility criterion is that blood pressure must be under a certain value, then the systolic and diastolic values should be collected to verify this before trial entry. This will usually be performed through checks built into the system that is used to register/randomise the participant.

Validations should also be included on the study database to flag up any inconsistencies between eligibility data provided at registration/randomisation and data received on later CRFs.

3.1.2 CHECKING SAFETY DATA POST-RANDOMISATION

While a patient is in a study, there may be data received that flags up a safety concern for the participant, such as specific toxicities which would necessitate stopping trial treatment immediately or would indicate that a Serious Adverse Event should be reported. There should be a process in place to flag up this data within 1 working day of receipt at the unit.

As well as any manual processes, validations should also be included on the study database to flag up any safety issues.

3.2 DATABASE QUERIES – SPECIFYING DATABASE VALIDATIONS AND MISSING DATA

This section provides practical advice and guidance for specifying database queries. For full information on when and how these are programmed and tested in the database, please see the Database Development SOP and Database Change Control SOPs.

Where possible the study databases should be programmed to check for:

- Missing data - expected data that has not been entered onto a CRF
- Inconsistencies - data that is inconsistent with the expected values e.g. outside of an expected date range or inconsistent with previously provided data These are flagged using database validations

When these checks are triggered they result in database queries, which are stored in the database for resolution.

These checks should be specified during study set up, although further checks may be added, amended or removed as the study progresses, for example following a protocol amendment, CRF update or a query review

3.2.1 SPECIFYING MISSING DATA CHECKS

If a piece of data is being collected on a CRF, the assumption is that it is required for the running of the study or to answer the research question, therefore it should be chased as missing data if not provided. However, there are instances where this does not always apply. Examples include:

- Patient reported outcome data where it is not possible to query retrospectively e.g. diary card or quality of life data
- Non CRF data e.g. date of receipt of CRF at the CTU
- Where it is dependent on another piece of data e.g. if a test has not been done, there would not be a test result expected

This should be agreed within the study team based on the needs of the study and the method of data collection and the decision documented.

3.2.2 SPECIFYING VALIDATIONS

Validations allow us to check that data seems sensible within the context of the study and for the participant . Examples of the types of validations which might be specified are:

- Data is out of an expected range e.g. date ranges or lab ranges
- To check consistency of data provided across different questions and CRFs e.g. if chemotherapy has been indicated as being given, but number of chemotherapy cycles has been recorded as zero

When specifying range checks be aware of the following:

- Range checks should be suitable to the participant group e.g. where the participant group is unwell do not use lab ranges for healthy participants
- Ensure the range is not too restrictive – queries should only fire when the data seems implausible, in order to flag up errors in CRF completion or data entry

There should be input from the whole TMT when specifying the validations, however the following roles are responsible for providing input into this process based on their specialism, and for reviewing lists of any validations that will be programmed into the study database. This includes any changes throughout the lifecycle of the study:

- Trial physician/Clinical Research fellow
- Statistician
- Programmer/Data Scientist

Documentation of the review should be placed in the Trial Master File. This may be in the form of emails or meeting minutes.

The Data Manager is usually responsible for documenting the validations in the metadata for programming into the study database.

3.3 MANUAL QUERYING

Some study databases also have the ability to add manual queries, which the person entering or reviewing the data may use to raise further issues that have been identified. These may arise due to:

- Problems with the clarity of the data on the CRF e.g. unable to read the handwriting
- Data is provided in the wrong units
- Data has not been corrected appropriately on the CRF i.e. without crossing through and initialling and dating the correction
- Queries identified by the statistician during IDMC, interim or final analyses

The above is not an exhaustive list.

It is recommended that studies maintain a list of common manual queries with standard text to ensure consistency. Studies may review the manual queries raised in the database periodically, in order to check for consistency and to identify any further validations which could be programmed in to the database. See section 3.7.1 for more information on query reviews.

Please note: if data issues are identified that require immediate escalation, the processes in section 3.1 should be followed.

3.4 NOTIFYING SITES OF QUERIES

3.4.1 STUDIES USING PAPER CRFS WHICH ARE RECEIVED BY AND HANDLED CENTRALLY AT THE CTU

Queries which are generated by the study database following data entry at the CTU, should be compiled into reports which are then sent to site. These are usually known as Data Clarification Forms (DCFs), but may also be known as query reports or error reports.

At a minimum the queries sent to site should provide:

- Trial/Study Number
- Date of visit and/or CRF
- Name and/or number of CRF
- The question number and/or text
- Brief description of the error
- Space for authorised delegate at site to respond to the query and then sign and date this response (either signing off each response or per page of responses).

It is recognised that there are alternative methods of querying used by some studies, such as requesting updated CRFs instead of a completed DCF report. In all circumstances the site staff should be reminded that they must correct the site copy of the CRF if the data originally provided was incomplete or incorrect. Corrections should be initialled and dated with an explanation given if necessary.

➡ If sites are notified of queries electronically see the Management of Personal Data SOP to ensure they are sent securely.

3.4.2 STUDIES USING EDC AND RDC

In studies using eDC or rDC site staff/data centre staff have responsibility for the day to day management of data queries. They should be encouraged to review and resolve any queries that may arise on a regular basis. However CTU study staff should still have oversight of this process.

To facilitate the resolution of queries it may be helpful for studies to produce regular query listings to send to sites/data centres. These listings should contain:

- Trial/Study Number
- Date of visit and/or CRF
- Name and/or number of CRF
- The question number and/or text
- Brief description of the error

✎ The responsibilities of the CTU study team, the data centres and the sites in terms of the query process should be clearly defined in the DMP and any relevant site guidance documents.

rDC studies will also likely need to utilise the procedures described in sections 3.5.1 to ensure queries are addressed and resolved appropriately, depending on the set up of the sites and data centres.

3.5 RECORDING WHEN QUERIES ARE SENT

A record should be kept of when query reports/listings are sent out for resolution to each site e.g. a query report tracker. It is also recommended that studies keep a copy of each query report sent. Studies will set a regular timescale for the sending of queries, which should be documented in the DMP.

3.6 QUERY RESPONSES

Query response data should be treated in the same way as CRFs, for example:

- Query report should be signed off by someone listed on the Site Delegation of Responsibilities log appropriately
- Query reports should be tracked as received and entered in the same way as the study CRFs e.g. date stamped and initialled
- Query reports should be filed in the participants file after entry

3.7 MONITORING RESOLUTION OF QUERIES

In studies where paper CRFs are received at the CTU, the person entering the query responses at CTU should monitor whether the site is giving adequate reasons for resolution, and raise manual queries if they think more information is needed.

In eDC and rDC studies where site staff have the ability to close queries, study teams should check the reasons given for closure to ensure they are adequate. In these situations it is preferable for queries to be closed off centrally at the CTU, rather than allowing sites permission to do so. However, if it is not possible to limit access to query closure in this way, all queries closed by sites should be reviewed by the study team, and queries re-raised if the reason given for closure is not deemed adequate. The Data Management Systems (DMS) Project Manager for your study can provide guidance and tools to facilitate this.

➡ See the Closed Overruled Queries Tracker template in SOPbox.

Studies may also have a list of data items that should never be unavailable, as defined in the study DMP. If the site responds to a missing data query for one of these specific data items and says that it is unavailable, the Data Manager (or whoever is dealing with the queries) should follow the appropriate process as defined in the DMP.

3.7.1 QUERY REVIEWS

It is recommended that studies undertake regular query reviews. Reviewing the queries that have fired in the database can highlight whether:

- There are any misfiring queries (i.e. queries that have been programmed/specified incorrectly and not picked up during testing e.g. queries that are too stringent)
- There are queries that are occurring as a result of CRF completion errors or CRF wording errors

- There are issues with critical data e.g. whether there is missing data for items that should not be unavailable
- Queries are being dealt with by assessing the number of open queries compared to closed queries
- Sites are taking a long time to resolve queries
- There are any site-specific query issues that may require re-training

➡ For more information on conducting a query review see the Query Review Working Instruction in SOPbox

3.8 MISSING CRFS

Studies should have a process in place for identifying any missing CRFs or missed visits, based on the visit schedule for a participant as outlined in the study protocol.

Methods for producing a list of missing CRFs/visits for each participant include:

- A database report programmed by DMS e.g. Form Status Report
- Database validations to flag missing CRFs and/or visits
- A report run by the Statistician

The timescale for sending these lists to sites should be clearly defined in the DMP.

3.9 ONGOING DATA ENTRY QUALITY CONTROL

All new data entry staff will have their initial data entry checked for accuracy and understanding.

➡ See Data Management Documentation and Training SOP for more information.

Some studies may choose to implement ongoing data entry quality control. For example, this could be a regular check of a proportion of CRFs against the database, or a check of all primary endpoint data prior to analysis. The nature and frequency of the checking should be agreed based on the complexity and nature of the data being collected. For studies conducting ongoing data entry checking this process should be defined in the Data Management Plan.

3.10 CENTRAL DATA MONITORING

Data received at the CTU will be monitored centrally as part of the overall monitoring strategy for the study. Further details of what constitutes central monitoring can be found in the Monitoring SOP, and the type and frequency of central data monitoring to be conducted should be outlined in the study monitoring plan. Details of how to conduct central data monitoring, for example where to find and run reports should be detailed in the DMP.

➡ See Monitoring SOP for more information.

3.11 PREPARING FOR ANALYSES

The study statistician will perform analyses prior to an IDMC meeting, for a planned interim analysis and for the final analysis. This will generally require more intensive data entry and extra efforts to resolve outstanding queries prior to the analysis. This should be discussed within the TMT to identify timelines for this work.

When the TMT feel the data is ready for the statistician to prepare for the analysis, a dataset is extracted and the statistician will perform further consistency checks. Additional errors and inconsistencies discovered during this checking by the statistician should be fed back to the study team for queries to be raised and resolved and the data extracted again. Where possible these queries should be recorded in the study database using manual queries (if queries don't already exist for these issues in the database) and may be expedited for resolution. A record of the queries raised by the statistician and the action taken to resolve each of them should be maintained.

This process may be repeated in order to achieve the agreed levels of data quality required ahead of the analysis.

3.12 DOCUMENTING EXPECTED DATA QUALITY LEVELS

For analyses that will require a Database Lock, as defined in the Database Lock SOP, the trial statistician is responsible for defining data quality thresholds that should be met before analysis. This should be done in discussion with the TMT ensuring clinical input. In defining data quality thresholds, consideration should be given to the key information required for the analysis (eg. primary/secondary outcome and safety data), and ensuring that this is sufficiently accurate and complete, considering both the extent of missing data and the extent of unresolved queries.

The statistician should document the agreed quality levels in a *Database Lock Quality Checklist* which should be finalised to at least version 1.0 before the analysis specific cleaning activities begin as described in Section 3.12.

The checklist should contain a description of the specific data quality levels expected e.g.:

- Proportion of key CRFs to be present on the database
  - e.g. For >95% of expected visits, either follow-up CRF is present on the database or site has confirmed that visit was not attended
- Proportion of key variables to be present on the database
  - e.g. 100% of fatal SAEs have a death date recorded
- Priority queries resolved
  - e.g. >99.5% of queries resolved on date of progression

For studies where there are more than one analysis requiring database lock, there should be a Database Lock Quality Checklist produced prior to each one of these analyses.

The checklist should also be updated with detail of when each quality level was reached and if the quality level has not been reached, an explanation as to why the data can still be analysed. This should also be done in discussion with the TMT ensuring clinical input. The checklist should be up-versioned again (e.g. at least v2.0) once this information has been added and should then be submitted along with the Database Lock Request Form.

For more information see the Database Lock SOP.

See SOPbox for the Database Lock Quality Checklist template.

3.13 STORAGE OF PARTICIPANT-RELATED PAPER DOCUMENTS

Participant-related documents, which includes CRFs, printed emails, and completed DCFs/query reports, should all be kept in a secure area when not in use, and stored in such a way that unentered CRFs cannot become mixed up with those awaiting filing after data entry.

4 RELATED DOCUMENTS

For further information on this topic, see also:

- Management of Participant Personal Data SOP
- Query Review Working Instruction
- Data Management Documentation and Training SOP
- Monitoring SOP
- Database Lock SOP
- Dealing with data queries identified by statisticians WI
- Data Querying Checklist for Statisticians WI