

STANDARD OPERATING PROCEDURE (SOP)

Title: Data Cleaning and Validation

Document ID: MCTC238

Version: 1.0

Applicability: MCRI-Sponsored Clinical Trials

Author

NAME and TITLE: Laura Galletta; Project Manager | Clinical Trial Manager

The author is signing to confirm the technical content of this document

Signature:  **Date:** 12 October 2026

Institution/Department name: Melbourne Children’s Trial Centre (MCTC)

Reviewed and Approval

This reviewer agrees with the technical content of the document and that this document is approved for implementation at the Melbourne Children’s.

NAME and TITLE: Andrew Davidson, MCTC Director, Melbourne Children’s Trial Centre (MCTC)

This document is effective from the date of the last approval signature and will be reviewed in three years, or earlier if required.

Document History

Revision	Modified by	Effective Date [DD/MMM/YYYY]	Description of Change
1.0	Laura Galletta	26/OCT/2026	New Issue

CONTENTS

1. PURPOSE	3
1.1. Quality Improvement.....	3
2. BACKGROUND	3
3. SCOPE	4
4. RESPONSIBILITY	4
5. RELATED DOCUMENTS.....	5
6. PROCEDURE	5
6.1. Workflow.....	5
6.2. Types of Data Checks	6
6.3. Data Cleaning and Validation Planning	7
6.4. Development and Testing of Post-Entry Checks.....	8
6.5. Data Review Process	8
6.6. Query Management.....	9
6.7. Query Resolution	9
6.8. Missing Data	9
6.9. Protocol Deviations	10
6.10. External Data Reconciliation.....	10
6.11. Medical Coding.....	10
6.12. Principal Investigator Data Sign-Off.....	11
6.12.1. PI Sign-Off in OpenClinica	11
6.12.2. PI Sign-Off in REDCap	11
6.13. Final Data Review Process	11
6.14. Deviations from this SOP	12
7. GLOSSARY	13
8. REFERENCES	15
9. COLLABORATORS	15

1. PURPOSE

This Standard Operating Procedure (SOP) describes the processes and responsibilities for the identification, review, correction, documentation, and resolution of data discrepancies within clinical trial datasets. While the objective of this process is not to specify a working method, it attempts to create a framework within which quality assurance is attained throughout the project.

The purpose of this SOP is to ensure that clinical trial data are complete, accurate, consistent, traceable and fit for analysis, and that data cleaning activities are conducted in a controlled, documented and reproducible manner.

1.1. Quality Improvement

This SOP improves the quality of clinical trial data by establishing a standardised, controlled and documented process for identifying, assessing, correcting and preventing data errors throughout the trial, and in turn, increasing the quality of data for statistical analysis.

2. BACKGROUND

The collection of complete, accurate and reliable data is fundamental to the conduct and credibility of a clinical trial. Clinical trial data are used to assess participant safety, evaluate the efficacy and effectiveness of interventions, support statistical analyses, and inform conclusions about the trial outcomes. Data quality therefore needs to be actively managed throughout trial conduct, rather than addressed only at the end of the trial.

Data cleaning and validation provide a systematic process for identifying, investigating and resolving errors, omissions, inconsistencies and discrepancies in a timely manner. These activities help identify issues such as missing or implausible data, inconsistent dates, duplicate or conflicting information, protocol deviations, and discrepancies between source records and the clinical trial database.

Ongoing data cleaning enables issues to be identified and corrected closer to the time of data collection, when source information and relevant staff are more readily available. It can also identify recurring issues with data collection processes, Case Report Forms (CRFs), database configuration or site practices, enabling corrective actions to be implemented promptly.

Data cleaning and validation should be proportionate to the risks and criticality of the data, and should be predefined, documented and traceable. All discrepancies should be appropriately reviewed and resolved, with database modifications supported by an auditable record.

Effective ongoing data cleaning and validation reduces unresolved issues at database lock, supports efficient trial close-out, and provides confidence that the final dataset is accurate, complete, consistent and fit for purpose.

3. SCOPE

This guidance applies to all clinical trials sponsored by the Murdoch Children's Research Institute (MCRI) and is applicable to all MCRI researchers involved in clinical trials, including Sponsor-Investigators, Statisticians, Database Programmers/Developers, Data Managers, Clinical Trial Managers/Trial Coordinators, Clinical Trial Assistants, Data Assistants and those holding Honorary positions within MCRI.

This SOP applies to all clinical trials where data is entered from paper forms into a database or directly into a database using an Electronic Data Capture (EDC) process (e.g. via OpenClinica or REDCap). It also applies where mobile data collection equipment is used to capture data on/offline and imported into a database.

4. RESPONSIBILITY

Reviewing, cleaning and validating data is the responsibility of the research project team. The following is the responsibility matrix for data validation activities:

Function	Activities and Responsibilities	Scope
Data Manager Data Management Team	• Defines and manages the process for data cleaning and validation • Leads the development of the data validation plan (DVP) in conjunction with the Sponsor-Investigator • Programs all items specified in the data validation plan into the database • Undertakes all data cleaning activities • Generates and manages data queries and data clarification requests • Oversight of query resolution and management of data corrections • Documents data cleaning activities • Coordinates database reconciliation activities • Develops database validation reports • Escalates data discrepancies/issues to the Sponsor-Investigator, as required • Prepares dataset for database lock	Trial Coordinating Centre (Sponsor)
Statistician	• Develops and reviews data validation specifications • Reviews of data listings and programmed checks, where applicable • Identification of statistical or cross-form inconsistencies • Develops database validation reports in liaison with the Data Management Team • Final review of analysis datasets prior to database lock, where applicable.	Trial Coordinating Centre (Sponsor)

Function	Activities and Responsibilities	Scope
Sponsor-Investigator	- Defines data management responsibilities - Leads the development of the data validation plan - Reviews and approves the data validation plan - Addresses data discrepancies/issues, as required	Trial Coordinating Centre (Sponsor)
Participating Trial Site Team Site Data Manager Data Entry Staff	Addresses data clarification requests/queries in relation to their sites data.	Participating Trial Site Team

5. RELATED DOCUMENTS

- [MCTC195 Template | Data Management Plan](#)
- [MCTC239 Template | Data Validation Plan](#)
- [MCTC240 Template | Edit Check Specifications document](#)
- [MCTC241 Form | Completion of Data Cleaning Activities](#)
- [MCTC242 SOP | Query Management](#)
- [MCTC216 Form | eCRF Data Review and Sign Off by PI](#)
- [MCTC231 SOP | CRF Design and Development](#)
- [MCTC232 Guidance | CRF Design and Development](#)

6. PROCEDURE

6.1.Workflow



6.2. Types of Data Checks

Data checking, review and validation should be performed regularly throughout the study to maintain data quality. Validation may occur at multiple stages of the data lifecycle and involve various processes for identifying, cleaning, reviewing and validating clinical trial data.

- **On-entry Data Checks:** Pre-defined automatic rules or automated validation checks that are developed and embedded as part of the database. These may be used to identify missing, inconsistent, implausible or otherwise invalid data. Automated edit checks can generate warnings, prevent data entry, or trigger queries for further review and resolution.

Examples include range checks, required-field checks, date checks, cross-form consistency checks, and checks for discrepancies between related data fields.

- **Post-Entry Data Checks:** A planned and structured review of clinical trial data after data entry has been completed and used to identify errors, inconsistencies, omissions, discrepancies or potential data quality issues that may not be detected by automated edit checks. These checks may include manual or programmed reviews of individual data fields, forms, visits, participants or datasets, and may result in data corrections, queries or further investigation.

Automated systematic post entry checks may be executed automatically through an inbuilt module of the Electronic Data Capture (EDC) system being used e.g. REDCap's Data Quality Module or OpenClinica's Rule Designer.

- **Manual Data Reviews:** The systematic, human-led examination of clinical trial data to identify errors, inconsistencies, omissions, implausible values, discrepancies or other data quality issues that may not be detected through automated validation checks.

Manual data review may include:

- Reviewing participant records and CRFs for completeness and consistency.
- Assessing data against predefined clinical or study-specific requirements.
- Reviewing critical endpoint and eligibility data.
- Identifying inconsistencies between related forms or visits.
- Reviewing unexpected or clinically implausible values.
- Assessing missing or incomplete data.
- Reviewing adverse events, protocol deviations and other important clinical data.
- Identifying issues requiring a data query.
- Confirming that data corrections appropriately resolve previously identified issues.

- **Source Data Verification:** either members of the research team or independent monitors perform validation as part of the ongoing monitoring of the study. Such validation is done via Source Data Verification (SDV). SDV involves checking the data entered into the CRFs is consistent with that in the original source records e.g. patient's hospital files.
- **Data Review Reports and Listings:** Data Managers may generate reports, listings or extracts to identify potential data issues. Reports can be used to review and identify:
 - missing data
 - out-of-range values
 - inconsistent dates
 - inconsistent participant information
 - unexpected visit patterns
 - duplicate records
 - incomplete CRFs
 - unresolved queries
 - discrepancies between related CRFs; and
 - other predefined data quality issues.

6.3. Data Cleaning and Validation Planning

- The level of cleaning should be proportionate to the risk, complexity and criticality of the trial data.
- Data cleaning and validation activities must be defined before or during trial database build and documented in a Data Validation Plan (DVP) and corresponding Edit Check Specifications (ECS) document. The DVP should define:
 - data review responsibilities
 - data cleaning methods
 - validation checks
 - percentage of source data verification
 - query management processes
 - review frequency
 - critical data elements
 - external data reconciliation
 - medical coding requirements e.g. coding of AEs and concomitant medications; and
 - medical review requirements, where applicable
- Review the CRF design, study forms and study protocols and draft the DVP (Refer to [MCTC239 Template | Data Validation Plan Template](#))
- Decide on the contents of the validation plan by:

- Deciding what needs checking/validating
 - Determining when the check should be used
 - Specifying how to resolve inconsistencies and issues identified by the checks
- For On-Entry data checks, draft an Edit Check Specification (ECS) (Refer to MCTC240 Template | Edit Check Specifications) providing full details of the data entry checks that need to be established and how each check will be tested. The ECS must be developed prior to or at the early stages of development of the database. Alternatively, it can be developed in parallel to when CRFs are being designed.
 - For Post-Entry data checks, develop a list of computer-programmed checks to identify missing data, perform context checks, logic checks, duplicate entries, etc, and record these also in the ECS document.
 - For Source Data Verification, agree on a percentage of records to carry out SDV on and define the data elements which will be checked and reviewed whilst performing SDV, and document these also in the ECS document.
 - Develop the Edit Check Specifications document ([MCTC240 Template | Edit Check Specifications](#)) based on the DVP requirements.
 - Obtain review of the DVP and ECS from all relevant stakeholders
 - Obtain approval and sign-off of the DVP by the Sponsor-Investigator
 - Ensure the approved DVP and ECS document is referenced in the Data Management Plan (DMP).

6.4. Development and Testing of Post-Entry Checks

- Some systems may not have the ability for all checks to be pre-programmed and embedded directly within the database. Therefore, alternative platforms external to the database may be required to check data e.g. SPSS, SAS, MS SQL Server, MS Access, R, etc.
 - Develop a query for each check using the platform of choice that is most suitable
 - Run a test on each check. The process involves using test data to test for violation, fixing any violations found and re-running the test until all checks returns non-violation results
 - Record all test checks in the ECS.

6.5. Data Review Process

- Data reviews should be undertaken periodically during the data collection phase. The frequency of reviews should be outlined in the study Data Validation Plan (DVP) and based on agreement with the research team. Reviews may occur after each batch of

data is collected, or at specific agreed intervals, e.g., quarterly, prior to any pre-defined interim analyses or at certain recruitment milestones are reached.

- Perform each check on the data as described on the DVP and ECS and record errors and/or discrepancies
- For each recorded error/discrepancy or clarification needed, a data query will be raised for investigation and correction by the participating trial study team. Details of this procedure are outlined in [MCTC242 SOP | Query Management](#)
- At any time during or at the end of the study and before database lock, additional checks may be added and the DVP and ECS document updated accordingly.

6.6. Query Management

- Data queries shall be raised where clarification or correction of data is required.
- Each query should:
 - Clearly identify the data requiring clarification
 - Describe the issue without directing the site toward an unsupported answer
 - Be assigned to an appropriate authorised user
 - Be tracked to resolution
 - Include the date raised and resolved; and
 - Maintain an auditable history of the query and response.
- Queries should be prioritised according to their impact on participant safety, trial integrity and analysis.
- Critical queries should be resolved as soon as practicable.

Refer to [MCTC242 SOP | Query Management](#), for further information on how to manage and execute data queries.

6.7. Query Resolution

- Participating Trial Site Team shall review each query against the source documentation and provide an appropriate response
- Where a data correction is required, the correction shall be made by an authorised user in accordance with OpenClinica or REDCap procedures.
- Corrections must not obscure the original information. The electronic audit trail must retain the history of the change, including the user, date/time and reason for the change, where supported by the system.
- Data should never be changed solely to make it conform to an expected result.
- Data must be changed/amended by the Trial Site Team and not by the Sponsor-Investigator or any member of their research team.

6.8. Missing Data

- Missing data shall be reviewed to determine whether:
 - the data were inadvertently omitted
 - the assessment was not performed

- the assessment was not applicable
- the participant withdrew from or otherwise discontinued study participation
- the visit did not occur; or
- another documented reason exists.
- Appropriate missing-data conventions should be predefined where possible.
- The Sponsor-Investigator or any member of their research team must not enter assumed or fabricated values to replace missing source data.

6.9. Protocol Deviations

- Potential protocol deviations identified during data cleaning shall be reviewed and managed in accordance with an applicable protocol deviation procedure.
- Data cleaning queries should not be used as a substitute for formal protocol deviation reporting.

6.10. External Data Reconciliation

- Where applicable, data held outside OpenClinica or REDCap must be reconciled against the clinical database. This may include:
 - laboratory data
 - central imaging data
 - randomisation systems
 - drug accountability data
 - safety databases
 - adjudication data
 - laboratory reference ranges; and
 - other external datasets.
- Reconciliation activities and outstanding discrepancies shall be documented before database lock.

6.11. Medical Coding

- Where applicable, coding of adverse events, medical history, medications or other clinical data shall be performed using the coding dictionary specified in the DMP. Coding should be reviewed for:
 - completeness
 - consistency
 - appropriate terminology
 - unexpected terms; and
 - appropriate coding hierarchy.
- Coding should be finalised before database lock unless otherwise justified and documented.

6.12. Principal Investigator Data Sign-Off

Site Principal Investigator (PI) sign-off on clinical trial data is a documented confirmation by the Site PI that the clinical trial data submitted by their site have been reviewed and, to the extent applicable, are accurate, complete and consistent with the source records, protocol and applicable trial requirements, and that any outstanding data issues have been appropriately addressed or documented.

6.12.1. PI Sign-Off in OpenClinica

In OpenClinica, Site PI sign-off can be facilitated via within the system. Following completion of data entry, data review, query resolution and applicable validation activities, the Site Principal Investigator or appropriately delegated investigator should review and electronically sign the applicable Study Event or Participant record in OpenClinica. The electronic signature confirms that the investigator has reviewed the data and considers the data entry and review requirements to have been completed in accordance with the protocol and applicable trial requirements.

6.12.2. PI Sign-Off in REDCap

In REDCap, where formal investigator sign-off is required, this should be implemented using the REDCap eSignature functionality and/or an appropriately validated electronic-signature process approved for the study. A simple Signature field embedded in a CRF is not automatically considered equivalent to a compliant electronic signature.

In REDCap, there is an important distinction between completing/signing a CRF and formally signing off clinical trial data. REDCap's standard functionality allows forms to be marked complete, forms and/or records to be locked, and data changes to be tracked through audit trails, but the exact electronic-signature workflow depends on how the REDCap instance and study have been configured (the ability to apply an electronic signature to forms must be explicitly enabled for each form, for example). REDCap itself is highly configurable, and compliance depends on the system environment and study-specific processes.

Alternatively, REDCap users can document PI sign-off on data outside of REDCap using the following form: [MCTC216 Form | eCRF Data Review and Sign Off by PI](#)

6.13. Final Data Review Process

- Once data collection is complete, ensure all final data review and validation activities are completed and that no unresolved or outstanding queries remain, except for documented/known issues that have been reviewed and agreed upon by the study team
- Perform a final run of all data checks and reviews as per the ECS document

- Ensure all external data sources have been reconciled, where applicable, e.g. central lab results, etc.
- Document any known issues or insoluble queries that have been agreed as insoluble by the data management team. Maintain comments explaining why the errors are deemed insoluble (Refer to [MCTC241 Form | Completion of Data Cleaning](#)).
- Ensure the Site Principal Investigator has provided attestation/sign-off of all data entered into the trial database (refer to Section 6.12 above)
- Sign and date the Completion of Data Cleaning Form ([MCTC241 Form](#)) and commence preparations for database lock. Completed Completion of Data Cleaning Forms must be filed in the TMF
- Should the database need to be unlocked to allow for changes to data, a final run of the checks will be carried out where necessary.

6.14. Deviations from this SOP

Any deviation from this SOP shall be documented, assessed for impact and managed in accordance with the applicable deviation/CAPA procedure.

Significant deviations affecting data integrity, participant safety or the validity of the trial data shall be escalated to the Sponsor.

7. GLOSSARY

Case Report Form (CRF)

A form that collects and contains Study Event information for a Study Participant.

A Case Report Form (CRF) is a crucial paper or electronic document used in clinical research to systematically record all protocol-required information for each participant in a study. CRFs serve as the primary data collection tool, ensuring consistent, accurate, and complete data for analysis, which helps evaluate the efficacy and safety of medical interventions and answer specific research questions. They are vital for regulatory compliance and the overall success of a clinical trial, capturing details such as demographics, vital signs, and treatment outcomes for each individual participant.

Central Trial Coordinating Team

A group of MCRI researchers (at the Sponsor level) organised to coordinate the planning, development, operations and conduct of an MCRI-sponsored, Investigator-Initiated, clinical trial.

Clinical Trial

The World Health Organization (WHO) definition for a clinical trial is: 'any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes'.

Critical Data

Data considered essential to participant safety, eligibility, treatment assignment, primary endpoints, key secondary endpoints, regulatory requirements or the integrity of the trial analysis.

Data Cleaning

The systematic review and correction or resolution of incomplete, inconsistent, inaccurate or otherwise questionable clinical trial data.

Data Query

A documented request for clarification or correction of a data item or set of data items.

Database Freeze

A controlled point at which data entry and/or changes are restricted while final data review and reconciliation are completed before database lock.

Database Lock

The formal process by which a clinical trial database is placed into a controlled, read-only state after all required data cleaning, review, validation, query resolution and reconciliation activities have been completed and the dataset has been determined to be complete and fit for statistical analysis. Following database lock, changes to the clinical trial data are restricted and may only be made through a predefined, documented and authorised process.

Edit Checks

Predefined rules or validation checks applied to clinical trial data to identify missing, inconsistent, implausible or otherwise invalid data. Edit checks may be automated within the clinical trial data management system and can generate warnings, prevent data entry, or trigger queries for further review

and resolution. Examples include: range checks, required-field checks, date checks, cross-form consistency checks, and checks for discrepancies between related data fields.

Edit Check Specifications Template

A controlled template used to specify the business rules, validation logic, conditions, expected outcomes and actions for automated and/or programmed edit checks within a clinical trial data management system. The template provides a standardised and traceable method for documenting edit checks from specification and development through to testing, implementation and maintenance.

REDCap (Research Electronic Data Capture)

A secure, web-based application designed to support the collection, management and storage of research data, including clinical trial data. REDCap enables researchers to build electronic case report forms (eCRFs), implement data validation and edit checks, manage user access, track data changes through audit trails, and export data for analysis.

Rules

Customized methods used to trigger actions on records such as the automatic generation of a Discrepancy Note, email, etc. Rules are stored in XML files. They can be built directly in an XML file and uploaded to OpenClinica or built in OpenClinica using the Rule Designer feature.

OpenClinica

OpenClinica is a cloud-hosted, open-source software platform used for electronic data capture (EDC) and clinical data management (CDM) in clinical trials. It helps researchers design and manage clinical studies, capture patient and operational data electronically via eCRFs, and ensure regulatory compliance through features like audit trails and edit checks. OpenClinica supports various aspects of a trial, including patient-reported outcomes (ePRO), randomization, supply chain management, and reporting, with solutions designed to be flexible and user-friendly.

Validation Check

A programmed or manual check designed to identify data that may be incorrect, incomplete or inconsistent.

Site PI Data Sign-Off

A documented confirmation by the Site Principal Investigator that the clinical trial data submitted by their site have been reviewed and, to the extent applicable, are accurate, complete and consistent with the source records, protocol and applicable trial requirements, and that any outstanding data issues have been appropriately addressed or documented.

Source Data

Information contained in original records or certified copies of original records from which clinical trial data are derived.

Systematic Post-Entry Data Checks

A planned and structured review of clinical trial data after data entry to identify errors, inconsistencies, omissions, discrepancies or potential data quality issues that may not be detected by automated edit checks. These checks may include manual or programmed reviews of individual data fields, forms, visits, participants or datasets, and may result in data corrections, queries or further investigation. Examples include: review of missing or unusual data patterns, cross-form consistency checks, review of protocol-

required data, identification of duplicate or conflicting records, and comparison of related clinical or external data sources.

8. REFERENCES

- [TGA Integrated Addendum to ICH E6 \(R3\): Guideline for Good Clinical Practice 2025](#)
- [Australian Commission on Safety and Quality in Health Care. The National Clinical Trials Governance Framework and user guide for health service organisations conducting clinical trials. Sydney: ACSQHC; 2022](#)
- [NHMRC Australian Code for the Responsible Conduct of research \(2018\)](#)
- [NHMRC National Statement on Ethical Conduct in Human Research \(2025\)](#)

9. COLLABORATORS

Name / Role	Department / Group	Affiliation
<i>Luke Stevens</i>	<i>Clinical Epidemiology & Biostatistics Unit (CEBU)</i>	<i>MCRI</i>
<i>Richard Hall</i>	<i>Melbourne Children's Trial Centre (MCTC)</i>	<i>MCRI</i>
<i>Billy McMahon</i>	<i>Cancer Therapies</i>	<i>MCRI</i>

DOCUMENT END