CIRCLE P50 Center of Excellence

Data Quality Guidance

Sylvie Perkins, Rachel Steinberg, Lucy Baker, Timothy Werwie, Dr. Christopher Kemp, Dr. Suzanne Grieb, Dr. Jill Fish, Dr. Emily Haroz.

January 28, 2025

Data quality control is important to consider at all stages of a project. Starting early and discussing a plan with the entire research team can prevent errors and ensure that the process is efficient and rigorous. Oftentimes, this section of an IRB protocol is brief, but the more thought that goes into it before you begin a project, the better.

Data quality control may also be called continuous quality improvement (CQI), or data quality assurance (DQA). The goal is to ensure data is accurate and complete.

This document is by no means an all-inclusive guide for data quality control. Rather, this is a summary of approaches used by CIH and JHU faculty and staff. Data quality processes, alongside other regular checks like Data and Safety Monitoring Board (DSMB) meetings, are one key element of a study's internal monitoring systems.

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Quantitative Data

Before Data Collections Starts:

1) Identify key persons responsible for data quality control (e.g., Data Manager)
2) Create data collection forms
    - Standardize naming conventions for your variables. Naming convention best practices include:
        - Use lower case
        - Do not start variable names with question numbers (i.e., q1, q2, q3)
        - Do start variable names with overall thematic construct or scale name (i.e., phq9, audit, employment)
        - Use underscores to add sub-constructs and specify individual variables (i.e., phq9_1, phq9_2)
    - Consider adding data validation checks to data collection forms (depending on data collection software). Data validation check options can typically be found in data collection software trainings, but may include things like:
        - The user cannot enter a number into a field intended for names.
        - The user cannot leave a field blank.
        - The user must enter a certain number of decimals in a field.
        - The user cannot enter a number outside of a certain range.
3) Pilot Data Collection Instruments
    - Build in time to pilot all forms. Ask 3-4 members of the study team to complete the instruments independently and note any issues and average length of time from start to end. If additional piloting is required, recruit a small number (2-5) of eligible participants to complete the instrument and provide feedback. Issues may include:
        - Typos
        - Survey Logic (e.g., branching or skip logic malfunction). Examples include:
            (1) Someone who is determined ineligible is then prompted to complete data collection forms.
            (2) Someone with no children is then prompted to respond to questions about their child, or someone with children is not prompted to respond to questions about their child.
    - Test it to break it!
4) Write detailed data quality check protocols
    - This should be a team effort – those building the data collection instruments should be included in these conversations, as well as Data Collector(s) and Data Manager(s)
    - Create codebook(s) with variable names, questions, answer choices, measure sources/citations, and other key information
    - Determine which software(s) you are using for data management and analysis and how you plan to export and back up data
    - Describe your timeline for data quality control procedures

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

    - Some choose to review and fix mistakes in real time; others have set times when all data is assessed
    - If you are editing data in real time, determine threshold for editing values (e.g. within 48 hours) **see 3-4 During Data Collection*
  - Write up data quality control procedures (including which data collection instruments need to have data quality checks and what discrepancies to look for)
    - For example, review all fields to identify:
      (1) Missing values
      (2) Typos or unclear short answers
      (3) Accurate dates (current date, birth year, etc.)
      (4) Accurate and coherent numerical fields (e.g., number of decimal points, or if a 29-year-old's age is entered as 290)
      (5) Outliers
    - There may be fields where it is unclear whether there is a data entry error or an outlier. Work with the Principal Investigator (PI) and/or Data Analyst to decide protocols to explore outliers or incoherent values.

5) Train Data Collectors, on topics such as:
   - Correct use of data collection software and/or devices
   - Techniques for effective probing (to minimize missing data)
   - Accurate clinical measurement-taking, measurement units, rounding rules, etc.
6) Ask Data Collectors to practice correct use of data collection software, techniques for effective probing, and accurate clinical measurement-taking, as applicable
   - Solidifies data collector skills and knowledge
   - Familiarizes data collectors with the instruments and questions
   - Provides opportunity to give feedback to data collectors prior to collecting real data
7) Certify Data Collectors
   - Ensures all data collectors meet baseline requirements for successful data collection
   - Allows for individualized refresher training, as needed

During Data Collection:

1) Update guidance and timelines regularly as new information becomes available
2) Review data using manual or automated data quality checks
   - Consider notifications of data upload
   - Automated data quality check options depend on data collection software and can typically be found in data collection software trainings. Examples include:
     - Set limits on age variables such that any value that falls outside the range is flagged.
     - Include automated notifications to confirm whether a respondent intended to skip certain questions, as a prompt to offer an opportunity for completeness.

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

- Create a pop-up box for the data collector that indicates the missing variables so they can verify with the participant whether that was intentional.

3) Investigate blank fields, errors, outliers
   - Option 1: Contact the Data Collector
     - Ask the Data Collector to correct their mistake or explain why the data looks the way it does
     - If due to Data Collector error, provide feedback and/or reminder(s) for future data collection
     - This is best if you are doing timely checks – it can be hard for someone to accurately remember what happened if a long time has passed

> Processes from one NIH funded clinical trial, Together Overcoming Diabetes (Award R01DK091250), included as an illustrative example:
> 1) Data is uploaded to REDCap online from the REDCap app by community-based data collectors. Data manager receives daily notification of upload
> 2) 24-hour data quality check completed by Data Manager and 48-hour data quality check completed by second study team member
> 3) Data Manager to update missing values or errors based on feedback from Data Collector. Value changed only if the correct entry is clearly identified by the Data Collector and provided within 48 hours of data collection.

   - Option 2: Contact the participant
     - If allowable per IRB protocol, and the participant can be found or easily contacted, see if they can provide clarification
   - Option 3: Discuss additional data error reconciliation options with the PI, Data Analyst, etc.
   - If you have an outlier that is NOT a data entry error:
     - Your team must decide whether to exclude it from analysis, transform the data, or use more robust forms of analysis that are outlier resistant
4) Clean up data
   - This can be done by:
     - Editing the value in the data collection form after reviewing with study team
       - For example, a participant said they have 2 children, but it was entered as 22. After reviewing with the study team, delete 22 and enter 2 in the data collection form.
     - Leaving the original entry and adding a note to a Data Quality Tracker (see below) or data analysis code
     - Different method that is well documented
5) Track errors and resolutions
   - Record all errors in a Data Quality Tracker or data analysis code, including any follow-up actions.

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

- Data Quality Trackers can be made in Excel and are meant to compile useful information for data analysis, as well as track frequency of user or technology error for follow-up (e.g., the same REDCap error has happened twice; the same data collector has made the same mistake three times). They are a tool for the Data Manager and Data Analyst and can be made to meet project needs.

Sample Data Quality Tracker

Date	Record ID	Category	Data Collector	Variable	Error	Resolution
		Data entry error	Initials	Variable name	Description	Action taken and name responsible
		Missing data				
		REDCap error				

6) Back up data regularly
    - Consider backing up to multiple locations per IRB protocol (e.g., Microsoft Teams, password-protected external hard drive, etc.)

After Data Collection Ends:

1. If data was not edited in real time during data collection, make and track edits to address duplicates, errors, outliers, etc.
2. Add notes about analytic methods used to address errors in specific fields (e.g., exclusions, transformations, etc.) to the codebook(s)
3. Maintain or delete data per IRB protocol

Additional Resources:

1. FDA Good Clinical Practices Guidance
2. FDA GCP Guidance specific to Data Monitoring

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Qualitative Data

Before Data Collections Starts:

1. Draft qualitative guide(s) (e.g., interview guides, narrative inquiry guides, focus group guides)
    - Guides must be developed according to the methodology you are using and provide step-by-step guidance for collecting qualitative data
2. Train interviewers as needed
3. Discuss positionality and reflect on potential personal biases of team members on an ongoing basis
4. Conduct mock interviews (e.g., interviewers practicing on each other, with a person who understands the issue(s) of investigation but would not be an actual participant)
5. Identify transcription company or software (I.e., NVivo, Otter.ai)
    - Confirm or set up any contracts necessary for future payments and/or identify and complete institutional agreements
6. Organize data quality control files
    - Activity log
    - Transcription and transcript cleaning log
    - Recruitment/participant log (as applicable)
    - Participant payment log (as applicable)

Activity Log with Examples

Activity	Date	Description/Notes	Who Participated
Interview Process Discussions	Mm.dd.yy	Ongoing discussions of interview process. Edited interview guide to create Interview Guide v# mm.dd.yy: description of changes and reasons for change	Name, name, name, name
Interview	Mm.dd.yy	Completed interview # with [participant type]	Interviewer name
Transcript Review and Memoing	Mm.dd.yy-Mm.dd.yy	Independent review of transcripts #-#	Name, name, name

Activities include, but are not limited to, interview planning, interviewing, interviewer feedback, interview process discussions, analysis planning, transcript review and memoing, analysis process discussions, codebook development, coding, coding checks and/or discussions, analysis/post-coding analysis

Transcription and Cleaning Log Example

ID	Time (min)	Interview Date	Audio Uploaded to OneDrive	Audio Submitted for Transcription	Transcription Confirmation	Transcription Received	Transcript Cleaned
	Length of audio				*Confirmation code from company*		*Date and staff initials*

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

7. Develop a project-specific standard operation procedure (SOP) that includes details regarding:
    - Recruitment (i.e., how recruitment is initiated and by whom, number of attempts to reach potential participant, how and where recruitment attempts are documented)
    - Audio recording (i.e., how interviews will be audio recorded, how to process audio files immediately after an interview, when to delete audio files from recording device)
    - How to document interview completion and payment (or, if applicable, need for designated staff to process payment)
    - Expectation for interview notes and reflection on potential biases
    - Location of various files (i.e., consent forms, interview guides, quality control files)
8. Identify key person(s) responsible for various data quality control considerations (e.g., Program Manager, designated interviewer(s))
9. Create a plan for cleaning transcripts (i.e., what identifying information should be removed, and what, if any, must remain for purposes of analysis)
10. If applicable, identify how saturation will be determined
11. Identify at what stages of data collection and analysis team members not directly involved in data collection and analysis, including Community Research Council members, should be brought in for updates and feedback (e.g., after a certain number of interviews, to review a draft codebook)
12. If applicable, identify plan for member checking
13. Identify plan for iterative process of data collection and analysis (e.g., independent review of transcripts, weekly/biweekly team meetings to discuss completed interviews and identify any necessary changes to interview guides and/or recruitment)

During Data Collection:

1. Update project-specific SOP as needed as new information becomes available
2. Ensure data quality control documents are updated with each interview
3. Keep audit trail of all documents – do not simply revise documents (i.e., interview guides, codebooks) in existing files
    - As an example, keep original file versions by labeling files: Interview Guide v1 mm.dd.yy; Interview Guide v2 mm.dd.yy; Interview Guide v3 mm.dd.yy
4. Implement iterative process of data collection and analysis as planned, adjusting if necessary based on process
    - Keep notes of discussions and any decisions made (e.g., reasons for editing an interview guide); these can be organized as meeting notes and summarized in the Activity Log OR the details of these meetings can be documented in the Activity Log

After Data Collection Ends:

1. Continue with analysis as planned and as appropriate for methodology

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

- Keep notes of discussions; these will usually be organized as meeting notes and summarized in the Activity Log

2. Check transcripts and/or coding on an ongoing basis to ensure high quality data
3. Destroy audio when appropriate for project

Additional Resources:

1. Practical guidance to qualitative research. Part 4: Trustworthiness and publishing
2. COREQ (COnsolidated criteria for REporting Qualitative research) Checklist

CIRCLE is funded by the National Institute on Drug Abuse of the National Institutes of Health under award number P50DA058619.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.