Royal Papworth Hospital
NHS Foundation Trust

GD012

Data Cleaning

One major role of data management in clinical trials is to ensure that the data is valid, auditable and accurate. It is essential to provide reliable data for statistical analysis, if this is not done the data integrity can be compromised and the outcome of study can change. Much of this is done by designing data collection tools that make gathering data as simple and accurate as possible, however errors will always occur. The data cleaning process needs to efficiently and accurately find the errors and correct them, while making sure all changes are documented. This guideline describes some of the processes that could be involved.

1. Pre-Entry Validation – Pre-Entry Validation applies to users entering data into the electronic data capture (EDC) system from a paper source. The user entering the data into the EDC should review the source document prior to entry. The following should be checked:
   a. Missing pages
   b. Participant’s study ID and initials are present on each page
   c. Illegible data
   d. Changes have been initialled and dated
   e. No data to identify subject

2. On Entry Validations – On Entry Validations are made as the data is being entered into the EDC.
   a. System-based validation should be used as far as possible. Data repositories designed for clinical studies will have these built in, but even if using Microsoft Excel these can be designed into the spreadsheets. These will include:
      i. Warnings when values are entered outside of the expected range.
      ii. Warnings if the type of value entered is incorrect (e.g. a numeric value entered rather than text)
      iii. Warnings for the omission of mandatory fields
      iv. Consistency across visits
      v. Logic checks

   Care must be taken when implementing these, as how different data repositories handle the warning must be taken into account. This is because the on-entry validation is there to improve data quality, not hinder data entry.

   Hard – In systems like Excel the built in validations are hard, which means there is no way round a validation. So if a validation range for height was from 1.2 to 2.2m and the subject was 2.3m tall it would not allow the value to be entered; if a field is mandatory the field must contain a value before moving to another field or saving the file. So for the height example, the better range would then be set from 0.5 to 3.5m.

   Soft – In other systems the validations are soft, which means that it is possible to override the warning. In better data repositories this would lead to an automatic data query being raised, which is the best of the options. If this is not an option, then post entry validation is essential.

3. Post Entry Validation – These are validations that are run, usually via an automated program after the data has been entered, they will often generate a report listing the possible validation issues. Post entry validation often will include:
    a. Complex validations that might check data across multiple visits, multiple CRFs, or the capacity of the data repository for on-entry validation.
    b. Those that could have been included for the on-entry, but might have hindered data entry.
    c. A review of any soft on-entry validations that have been overridden.

    How often to run these will depend on how much data is being entered and the study type. For a single visit study, it might be best for the user to run the report directly after entering the data; for others it might mean weekly or monthly. Regularly running the reports will improve data quality. One consideration in using post entry is how to handle unresolvable validations, either missing data that cannot be populated, or question validations where the data is correct.

4. Validation of Imported Data – If the study requires data to be imported into the EDC, the imported data should be validated. When data has been imported it should be exported from the EDC then compared to the data to be imported, any discrepancies should be reported (FRM071 Importing Data into Openclinica should document any imports and validation).

5. Data Queries (OpenClinica is the current PTUC preferred data repository, but if this is changed in future, any comments in this Guidance Document regarding OpenClinica will apply to the new repository.) - The OpenClinica Notes and Discrepancies feature provides a means for users to document, communicate, and manage issues about data in a clinical trial. There are various situations where you use Discrepancy Notes, for example:
    a. You can create a Discrepancy Note when capturing or validating data in order to flag an item as incomplete or as having a value that is not expected.
    b. You can leave a required field in a CRF empty if you provide a Discrepancy Note that provides an explanation.
    c. OpenClinica can automatically create a Discrepancy Note when you save a CRF that contains errors in the data, as determined by OpenClinica's edit checking.
    d. OpenClinica can also automatically generate Discrepancy Notes when Rules run.

    Please see GD011 General OpenClinica User Guide for information on using Notes and Discrepancies in OpenClinica.

    If the study is not using OpenClinica data queries can be managed in a form, see FRM026 Data Query Form for a template.

6. Source Data Verification (SDV) – SDV is a check that the data collected on the EDC can be verified by looking at a primary source (e.g. medical records). This is completed by the study monitor who should be able to:
    a. Identify who recorded the information
    b. Access and read it
    c. Identify where the information was first recorded or reported

    d. Verify the information is correct with no errors

For studies using OpenClinica, the built in Source Data Verification (SDV) feature can be used. Please see GD011 General OpenClinica User Guide for information on using SDV in OpenClinica. If a post-live design change is required, a completed SDV status of any participants will be undone. If this is the case, the following process should be followed:

1. Design is approved knowing that SDV will be undone.
2. Data manager creates a list of SDV’d CRFs that will be affected by the change.
3. Design change is implemented.
4. Monitor uses the list provided to review the Audit log for each patient CRF to review changes made to the CRF since previous SDV. If no changes then the SDV can be remarked. If there are changes, then the SDV would need to redone as normal.

7. Data Completeness –This could be a manual or automatic process that reviews each patient, making sure all the expected data is entered, and marked complete. The level and method of review should be specified in the Data Management Plan (DMP).

8. Other methods of validation include:
    a. Read and Verify
    b. Double data entry – A process where the data for a CRF is entered twice to ensure the integrity of the captured data. It is typically used when CRF data is first captured on paper forms, then entered into an EDC system. OpenClinica includes this as an inbuilt function which flags any differences between the values in the initial data entry and the second data entry, and provides options for resolving the differences.

For larger studies it may be appropriate to keep a log of any data issues. The log also serves as the single place that the study team, data team and the statisticians can review if there are questions about the data that may be raised repeatedly during cleaning. Keeping a log of the issues makes understanding data issues quicker. Please see FRM057 Data Issues Log for guidance and a template.