World Health Organization
African Region

EXPANDED SPECIAL PROJECT FOR ELIMINATION OF NEGLECTED TROPICAL DISEASES
ESPEN

SOP 7: Data Cleaning and Validation Procedures

1. Introduction

This standard operating procedure (SOP) is the seventh and last in the series. It provides comprehensive guidance on data cleaning procedures for surveys conducted using the ESPEN Collect platform. The data cleaning is essential to ensure the accuracy and reliability of survey data, which is essential for assessing the impact of neglected tropical diseases treated with preventive chemotherapy (PC-NTD).

Good data is essential for making informed decisions, such as whether to resume drug distribution programmes. To support this process, ESPEN Collect offers data monitoring tools through the Metabase dashboard, enabling real-time monitoring of data quality during ongoing surveys. In addition, the platform provides ongoing support for data management, both during the survey and data cleansing phases.

To speed up the data cleaning process and improve overall data quality, data managers from the Ministry of Health (MOH) and partner organisations are encouraged to actively participate. This guide will detail how they can effectively contribute to data cleansing efforts, highlighting the benefits of their involvement.

By following this SOP, users will be equipped with the necessary procedures and best practices to ensure high quality data, supporting the successful elimination of targeted diseases.

2. Data Cleaning Procedures

Data cleansing procedures are essential to guarantee the accuracy, consistency and reliability of the data collected. They involve identifying and correcting errors, managing missing data and standardizing data formats. As part of ESPEN Collect survey management, data cleansing includes error correction, data cleaning and data collection monitoring.

Metabase is an essential tool for the data cleaning process, and its use is detailed in “SOP 6: Guide to Metabase, Monitoring, and Supervising Results”.

We can consider two scenarios for data cleansing. In the first scenario, a data management team is present in the country. This team may be made up of data managers from the Ministry of Health (MOH), partners or the WHO. Their role is to focus on data management, which includes error correction, data cleaning and monitoring data collection. In this way, they

ensure good data quality. In the second scenario, there are no dedicated data managers in the country to work on data cleansing. This can pose additional challenges for ensuring data quality. In this case, the ESPEN Collect data managers will work with the survey supervisors to carry out the data cleaning.

2.1. Scenario with a Dedicated Data Management Team

When a team of data managers is present to work on data cleansing to ensure good data quality, the process should be as follows. The data manager should have at least a basic knowledge of Excel. They should also understand how the Metabase dashboard works and be able to identify inconsistencies from Metabase. There are three stages to monitoring and cleaning up data: identifying errors, cleaning up errors and applying error corrections in Metabase.

a. Error identification:

The identification of errors is done from Metabase. The data manager connects to the dashboard in the "Monitoring and Errors" section to see the details and any inconsistencies detected. Each time data is sent to the server, this section is updated automatically. The data manager can download details of any inconsistencies and review them. They can also download the entire database in XLS or CSV format for in-depth analysis of data quality.

b. Error clean-up:

The main data clean-up tools we need are Metabase and Microsoft Excel. The data manager downloads error details from Metabase. Then, thanks to the data registers and the hard copies, he makes corrections to the Excel file, and then sends the Excel file back so that the corrections can be applied. Data managers have the option of downloading the entire database, making corrections and then returning the database so that the ESPEN Collect data managers can check and apply the corrections in Metabase. Data cleaning should start from day one for optimisation reasons. The advantage is that data entry operators can easily remember the information they have entered and can quickly recognise errors and suggest solutions.

c. Applying corrections:

On completion of the data cleaning, the data managers send the data back to the ESPEN Collect data managers who will cross-check the data. If there are any inconsistencies, they will contact the country data managers to correct the remaining errors. This back-and-forth process continues until there are no more errors in the data. Once the data is error-free, the ESPEN Collect Data Managers will apply these corrections in Metabase so that everyone can access the corrected data. After this step, the ESPEN Collect data managers will generate the EPIRF, if it is a survey whose result can be included in an EPIRF, and then they will share this EPIRF with the country so that it can verify and complete the missing information. The EPIRF contains the information collected with ESPEN Collect and the information submitted by the country at the time of the request, but often information needs to be completed, such as morbidity. The complete EPIRF is then submitted by the country via the ESPEN portal using the JAP upload tool.

2.2 Scenario without a dedicated Data Management Team

In this scenario, there are no dedicated data managers in-country to work on data cleansing. ESPEN Collect data managers will monitor the data via Metabase and send identified errors to field supervisors. The supervisors will receive details of the errors, contact the data

collectors and check the hard copies for correct information. They will then send the corrected information back to the ESPEN Collect data managers. It is important to note that the ESPEN Collect data managers cannot make any corrections without the approval of the country data managers, as the data belongs to the country and not to WHO.

Once the data has been cleaned and validated, the ESPEN Collect data managers will apply the corrections in Metabase, making the corrected data accessible to all. Finally, if the survey allows, an EPIRF will be generated and shared with the country for verification and additional information before submission via the ESPEN portal.

3. Tools and Software

To ensure effective data cleansing, it is essential to use the right tools. The main tools recommended for ESPEN Collect surveys are Metabase and Microsoft Excel.

- **Metabase**: This open-source data visualisation tool allows data to be monitored and analysed in real time, identifying errors and inconsistencies. For more details, see SOP 6: Guide to Metabase, Monitoring, and Supervising Results.
- **Microsoft Excel**: Excel is widely used for data cleansing and analysis. Tutorials to improve your Excel skills are available on platforms such as YouTube

Using these tools, data management teams can ensure the quality and reliability of ESPEN Collect survey results.

4. Conclusion

This standard operating procedure (SOP) on data cleaning and validation procedures is the seventh and final in a series of documents to guide users of the ESPEN Collect platform. By following the steps outlined in this document, data management teams can ensure the accuracy, consistency and reliability of the data collected, which is essential for assessing the impact of neglected tropical diseases treated with preventive chemotherapy (PC-NTD).

The following SOPs provide a solid foundation for understanding and effectively using the ESPEN Collect Platform:

- SOP 1: Overview of the ESPEN Collect Platform
- SOP 2: Procedure for Requesting Access to the Platform
- SOP 3: ESPEN Collect Step-by-Step Guide
- SOP 4: Training Manual for Instructors
- SOP 5: Organizing Surveys
- SOP 6: Guide to Metabase, Monitoring, and Supervising Results

By incorporating best practices for data cleaning and validation, users can ensure that data are ready for in-depth analysis and informed decision-making. Collaboration between data managers, supervisors and data collectors is essential to maintain data quality throughout the process.