Standard Operating Procedures for the LAKANA trial
SOP Data 01: Data storage, access, backup, and editing
Version 1.0 (2022-11-09)

1. Purpose and overview:

This SOP provides instructions for the setup of and management of systems to facilitate data storage, access, and backup for the LAKANA trial. In addition, it provides an overview of the data querying and editing process.

2. Applicability to and responsibilities of various staff members

Staff member	Responsibility
Study Data manager (TAU)	Sets up the data structure on the CommCare data collection application, troubleshoots problems in the application, and monitors the dataflow.
Lead data manager (CVD)	Oversees all data collection activities pertaining to the data collection tablets, including the application of updates, data transmission and synchronization.
Study statistician (TAU)	Reviews data descriptives and identifies data issues through the development and use of a statistical software script.
Study data reviewer (CVD)	Reviews queries identified by the statistical software script and provides answers to the queries.

3. Required materials

Item	Specification
CommCare HQ access	Data management user interface

4. Definitions and general instructions

4.1. Definitions

4.1.1. Data descriptives: a set of data visualizations and summary statistics, and associated R/STATA/ other software code, which can be used to view the distributions of each variable in the collected data and identify potential issues with the data collection process.

4.1.2. Synchronized database: a separate database (often residing on a different server in different location) that is synchronized with the main study database. The server running the database would ideally be hosted by Tampere University/CVD-Mali.

4.1.3. Data export: an on-demand, real-time set of comma-separated variable (CSV) files which represents the state of the data at the time of export. Data exports can be automated and scheduled using a CommCare-based application.

4.1.4. CommCare HQ: User interface for data management

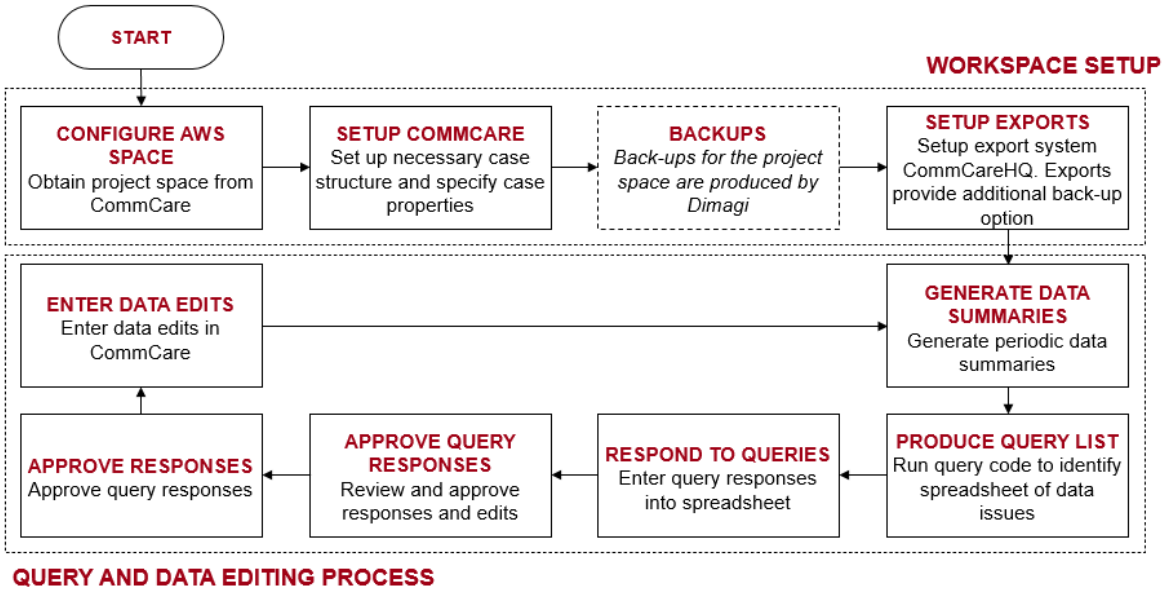
4.1.5. Amazon Web Services (AWS): Cloud platform, that provides storage space allocated for LAKANA trial.

4.1.6. Lead data manager: the CVD lead data manager, or designee, who oversees all data collection activities pertaining to the data collection tablets, including the application of updates, data transmission and synchronization.

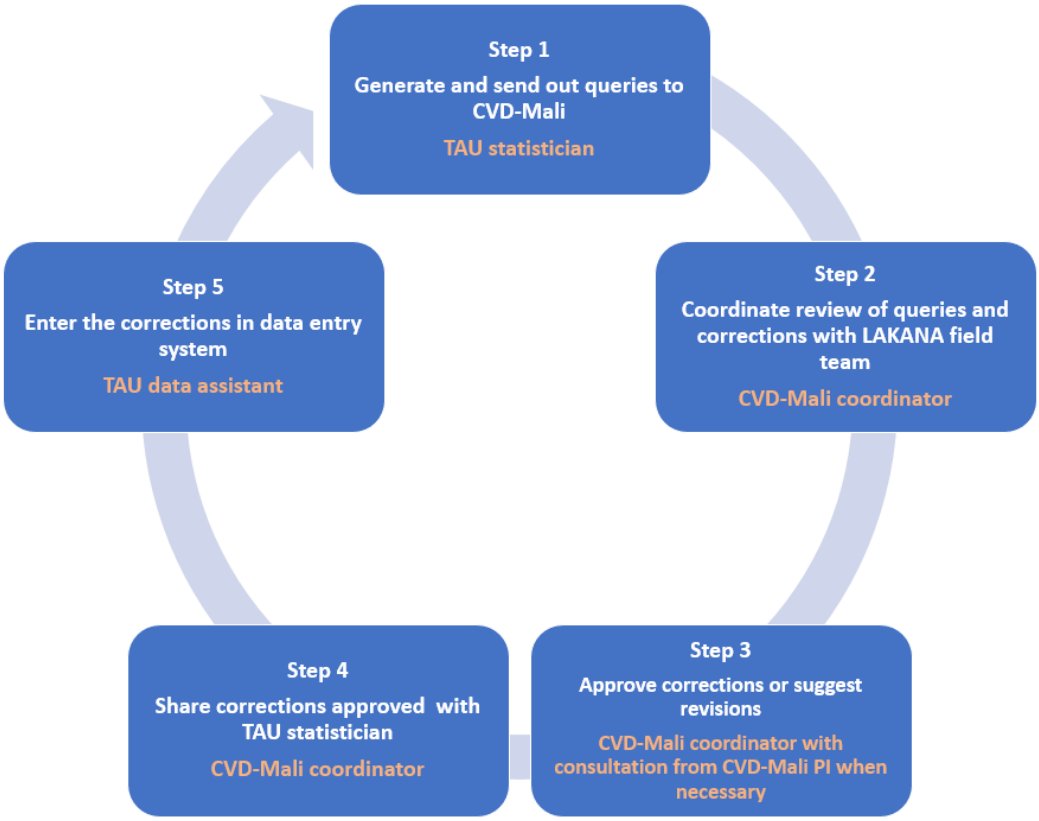
4.1.7. Study statistician: a TAU or CVD-Mali based statistician who reviews data descriptives and identifies data issues through the development and use of a software script.

4.2. General instructions

4.2.1. The data editing process requires the setup of weekly or monthly exports, generation of weekly or monthly summaries and queries, entry and approval of query responses, and finally edits to the database. These processes and their interaction are illustrated below



4.2.2. The main processes for query resolution are illustrated in the figure below and detailed instructions are provided in Section 5 Step-by-step procedures.



5. Step-by-step procedures

5.1. Data synchronization from tablets to database.

5.1.1. Data will be synchronized from tablets to AWS cloud and will be governed from CommCareHQ user interface.

5.1.2. Data will be synchronized to a MySQL database at a server at Tampere University and/or CVD Mali via CommCare's Data Export Tool. This synchronization can happen in approximately real-time, subject to internet connectivity and provides two main functions: (1) it provides a local instance of the data that can be accessed, if needed, without internet access, and (2) it provides a geographically diverse set of backups. This synchronization configuration will not be a substitute for a full backup policy, which is described below.

5.2. Data storage and security.

5.2.1. There is great amount of infrastructure around backups in CommCare's production environment, including multiple levels of redundancy in data storage as well as offsite snapshot backups for all data.

5.2.2. In addition to that, TAU team will export comprehensive data set regularly (approximately bi-weekly) from the cloud to a local computer, operated by statistician or a study researcher, so that in the case of extreme catastrophic incidence, the data can be imported back to the project environment and the study continued.

5.2.3. Exports to local devices will also act as a substitute for MySQL database, should the MySQL database be unavailable or under prolonged maintenance.

5.2.4. Tampere University hosted virtual machine(s) that maintains the synchronized database will have scheduled back-up procedure starting at midnight local time

5.2.4.1. The back-up procedure creates a virtual snapshot ("copy-in-time") of the virtual machine.

5.2.5. Local copies of the data in coherent, consistent state will be saved regularly in forms of data exports.

5.2.6. In summary, there are three layers of back-up for the data:

5.2.6.1. CommCare infrastructure in the AWS environment

5.2.6.2. Nightly snapshots of the virtual machines

5.2.6.3. Regular data exports

5.2.7. CommCareHQ access will be limited for LAKANA researchers, and the access requires a username and a password

5.2.8. In CommCareHQ, users have different levels of access, from read-only to admin. Access levels will be administered by TAU data manager.

5.2.9. Synchronized MySQL database will have even more restricted access, that is granted only for LAKANA researchers with required skillset to run maintenance on the database

5.2.9.1. Access is granted by TAU data manager

5.3. Data review and identification of suspicious data values.

5.3.1. The server administrator, study statistician, lead data manager and other staff approved by the LAKANA PSG will be able to perform data exports of the study database using CommCare's Data Export Tool.

5.3.2. The study statisticians will write a statistical software program (in R, STATA or other software) to generate data summaries. The study statisticians will generate these initially on a daily basis, but eventually on a weekly basis once initial issues with the data system have been resolved. The study statisticians or other authorized LAKANA researcher will use the data descriptives to identify

potential problems with the data and will issue a series of queries. Data descriptives will be generated by a statistical software program, which will be created and updated by the study statistician and shared with the Lead data manager and Study data reviewer at CVD. This program uses the study data exports as input and produces a series of summary statistics and visualizations

5.3.3. Any LAKANA researcher can identify and report suspicious values from the data summaries, or from other information received from the data collectors / LAKANA personnel.

5.3.4. TAU statistician will generate, using a statistical software script, a list of queries that compile issues identified from data received during a week.

5.3.5. In the beginning of the trial, TAU statistician will send the list of queries to the CVD-Mali study coordinator on a weekly basis.

5.3.6. The excel file with the list of queries will consist of 19 columns as described below:

5.3.6.1. Date_query_generated: indicates the date when the query was generated.

5.3.6.2. SITE: indicates the village of data collection.

5.3.6.3. DATA_COLLECTOR: indicates either the name or user ID of the data collector that collected the information.

5.3.6.4. DCF: stands for Data Collection Form, the number and name of the form is indicated in this column.

5.3.6.5. QUESTION_NUMBER: indicates the question number and text on a specific DCF.

5.3.6.6. DATA_ISSUE: provides a short description of the data issue.

5.3.6.7. CURRENT_VALUE: indicates the current problematic value in the database.

5.3.6.8. SUGGESTED_NEW_VALUE: the CVD-Mali coordinator will enter here the suggested value that will replace the current problematic value.

- If the current value should be replaced with a new value, the study coordinator will indicate the new value in this column.
- If the current value is correct or cannot be replaced for some reason, the study coordinator will repeat the current value in this column.

5.3.6.9. RESOLUTION_OR_COMMENT: the CVD-Mali coordinator will indicate here the reason for suggesting the new value or repeating the old value or removing the entry. This field will not be used in data management, its purpose is to work as a reminder and explanation to why a change was or was not made. The categories that can be entered in this column are listed below:

- "Duplicate": enter this comment if same compound/household/household member was entered more than once.
- "Entry error": enter this comment if the data collector has intended to enter another value but made a mistake when entering the value.

- "Suspicious value but no correction made": enter this comment if the suspicious value has been checked with the data collection team but it was not possible to identify an error or find a correct value.
- "Missing data": enter this comment if the value has been checked with the data collection team but the respective data was not collected, and the missing value cannot be replaced.
- "Other": enter this comment if none of the previous categories fit. If other is used, specify in the next column.

5.3.6.10. SPECIFY_OTHER: If in column RESOLUTION_OR_COMMENT, "Other" was used, describe in your own words the reason for data correction.

5.3.6.11. RESOLVED_BY: the name of the CVD-Mali coordinator recording the corrected value will be entered here.

5.3.6.12. DATE_RESOLVED: the CVD-Mali coordinator will indicate the date the suggested new value is recorded.

5.3.6.13. CORRECTION_TYPE: the CVD-Mali coordinator will indicate the category of correction here. This information will be needed by TAU for data processing. The corrections categories that can be entered in this column are listed below:

- "Corrected": enter this category if the current value has been checked and should be replaced with the new value entered in column SUGGESTED_NEW_VALUE. This category closes the case and it will no longer appear in the next list of queries.
- "Reviewed": enter this category if the current value has been checked but cannot be changed i.e. the old value remains. This category closes the case and it will no longer appear in the next list of queries.
- "Delete": enter this category if a record was created by mistake and the whole entry for the child/adult/household/compound should be deleted. This category closes the case and it will no longer appear in the next list of queries.
- "In process": enter this category if the issue has not been checked yet or is awaiting confirmation. The case will remain open and will appear in the next list of queries until resolution.
    - Once the value has been confirmed or if no verification is possible, enter one of the categories that closes the case (Corrected, Reviewed or Delete).

5.3.6.14. VARIABLE: indicates the variable name in the database. For data correction purposes. The value is generated automatically from the software script.

5.3.6.15. COMPOUND: indicates Compound ID. The value is generated automatically from the software script.

5.3.6.16. HOUSEHOLD: indicates Household ID. The value is generated automatically from the software script.

5.3.6.17. CHILD: indicates Child ID. The value is generated automatically from the software script.

5.3.6.18. DATEENTRY: indicates the date of data collection. Source of the date is case specific and can be obtained either directly from the DCF or alternatively from case/form metadata.

5.3.6.19. MDA_VISIT: indicates the MDA round.

5.4. Analysis of suspicious values (CVD-MALI)

5.4.1. The CVD-Mali study coordinator and field supervisor will receive the list of queries that compile issues identified from data send to LAKANA AWS server.

5.4.2. Through an internally defined process, CVD-Mali coordinator in collaboration with LAKANA data collection team will review the queries and provide corrections. The CVD-Mali coordinator will be responsible for filling in columns:

- SUGGESTED_NEW_VALUE,
- RESOLUTION_OR_COMMENT,
- RESOLVED_BY,
- DATE_RESOLVED,
- CORRECTION_TYPE.
- SPECIFY_OTHER

5.4.3. If a query is unclear or if more information is required, the CVD-Mali coordinator will contact TAU statistician for assistance.

5.4.3.1. The CVD-Mali coordinator will notify to TAU statistician any issues listed in the query file that should not be considered issues. The case will be discussed and if applicable will no longer be flagged as an issue.

5.4.3.2. Similarly, CVD-Mali team will report any issues that are not flagged by TAU. The case will be discussed and if applicable will be listed in the query file.

5.4.4. Once the CVD-Mali field supervisor has determined the resolution of the queries, s/he will share the list of queries, that now include corrective actions, with the Mali coordinator for review and approval.

5.5. Review and Approval

5.5.1. Provided that the data in question is related to the research data, as opposed to the metadata that the application requires for functioning, the CVD-Mali coordinator in collaboration with field supervisor will review the list of queries with corrections provided by the CVD-Mali field supervisor.

5.5.1.1. If changes are needed, the study coordinator will discuss the corrections with the Mali co-PI and update the file until approved.

5.5.2. If corrections are approved, the Mali co-PI or appointed person will share the approved query file with TAU statistician.

5.5.3. If the nature of the queries is related to the functioning of the app, the data manager (TAU) will review the queries and apply the necessary corrections in co-operation with the CVD-Mali coordinator.

5.6. Corrections review and entry to database (TAU)

5.6.1. TAU statistician will review the list of queries with corrections provided.

5.6.1.1. TAU statistician will check for unanswered queries, or corrections "In process" and keep these cases in the next list of queries generated.

5.6.1.2. If a correction is unclear or if more information is required, TAU statistician will contact the CVD-Mali coordinator for clarification.

5.6.2. TAU researcher will send the resolved queries to TAU data assistant who will enter the corrections in CommCare project space

5.6.3. CommCare maintains a log of every change done to the cases. The case history can be reviewed for each case, and the corrections can be reversed should there be a suspicion of a mistake.

6. Occupational Safety Issues

6.1. Maintaining a working data system, handling the accumulating information mass, and monitoring data integrity via statistical and/or computational methods will be time consuming and can cause stress and frustration. If prolonged, the impact on worker's well-being can be harmful.

7. Quality Assurance / Quality Control

7.1. Each step of the configuration will be reviewed by a second person to ensure that the server is configured correctly, data are loaded correctly, and data exports and descriptive programs are functioning correctly.

8. Appendices and other related documents

8.1. None.

9. Version history, authors and approvals

Version (date)	Edits to the SOP text (author)
1.0 (2022-11-09)	Authored by Juho Luoma. Approved by PSG on 2022-11-09.