CITF Data Sharing: Data transfer procedures

version 1.1 June 2022

Data inventory and transfer

Once the data sharing agreement is complete and your data from one or more study visits (or waves) are prepared, you and the CITF Data Management (CITF-DM) staff can initiate data upload.

These steps could be done with or without a simultaneous online call with a CITF-DM staff member.

File upload

- CITF-DM staff create a password-protected folder on the secure databank server that only the study team and CITF-DM staff can access. (Each study gets their own folder.)
- You will receive an email with
  - the web link to the folder,
  - the login and password,
  - a data upload inventory form,
  - and instructions.
- You or a designated member of your study team uploads the study data files and the inventory form. It is necessary to refresh the page after each file is uploaded.

Data upload inventory confirmation

CITF-DM staff run basic data checks to confirm that the number of data tables, variables, rows, and unique participant IDs match the information provided in the inventory form. CITF-DM staff will notify the study team if any discrepancies or questions arise.

Preparing your data

Variable list

CITF-DM staff will send you the list of your study variables that should be transferred to the CITF, the variables that potentially map to the CITF core data elements. We generate the list from the data dictionary that you shared with Maelstrom (Karla Ordonez); therefore, the variables names are those you used in that data dictionary.

You can transfer your data in almost any standard format such csv, R files (.Rdata, .Rds), SPSS (.sav), SAS (.sas7bdat), Stata, Excel. Please use the same variable names, category codes, and measurement units as in the data dictionary that you shared with Maelstrom, or describe any changes, in detail, in the Excel data dictionary we send you.

If using Excel, please be sure that each sheet containing data is a simple rectangular data set with variables as columns and records as rows; move any notes or variable definitions or higher-order headers to a Notes sheet. Variable names should be in row 1.

Basic data checks

Please do these **basic data checks**:

1. Do the number of columns, rows and participant IDs in each table make sense?
2. Check for suspicious values: e.g. run univariate descriptive statistics and check the min-max for continuous variables, and categories/levels of nominal and ordinal variables.
3. IMPORTANT - **Participant ID instructions**:
    - You must not send participant names or full address (only forward sortation area [FSA], if available).
    - If your unique identifiers (IDs) are constructed from name, MRN, etc, please generate a random ID.
    - Keep the lookup key for ID number and contact information in a secure place.
    - Participant IDs should remain constant over study visits. Use a separate variable/column to indicate study visit.
    - If a participant drops out, do not recycle his or her ID number.
    - A single unchanging ID for each participant is needed to communicate with you about data updates and in case a participant requests that their data be removed from the database.
    - **Cluster ID**: Create a coded unique identifier for cluster along the same principles as participant ID (does not directly identify the institution, consistent over study visits, not recycled, backed up look-up key).
        - Clusters: If you recruited participants using a clustered design (such as recruit schools, then teachers within those schools) or if participants are from a limited number of institutions because of geographic scope of your study, for example, it may be good statistical analysis practice to account for clusters. Cluster ID for geographic clusters may not be necessary if you are already sharing FSA. If FSA cannot be shared because it may uniquely identify participants (such as studies with participants living in LTCH or corrections facilities and the study includes multiple facilities), then coded cluster ID will be especially useful.