Consortium on Vulnerability to Externalizing Disorders and Addictions

Standard Operating Procedure

Data Management

Table of Contents

1. Data Cleaning

Sources of error	Resolution
Using escape when task ends rather than continue resulting in task being recorded as incomplete.	Marked as complete at Delosis.
Data not returned for some technical reason.	Re-synchronization of all computers.
Some tasks are not administered, an informant may be unavailable, or participant refuses a task.	Reported in debrief.

1.1 Recruitment files

Recruitment files contains details of all participants included in the study. All information is anonymized, and each participant is represented by a randomly assigned PSC1 code.

Action: The recruitment files are checked for bogus PSC (incorrectly entered code) and corrected.

1.2 Discrepancy management

Identify and resolve the discrepancy. Examples are:

1. Codes, age and gender matched across instruments and across recruitment, genetic sampling and neuroimaging files.
2. Each participant completed timestamp matched across instruments to identify log in errors.

1.3 Identifying Duplicates

Identify duplication in data, which may come from different sources such as:

1. Participants entered under more than one age band under same PSC, will show as multiple users.
2. Login errors – Participants entered under multiple PSC codes

Action:

a) Examination of excess data and age band duplicates.
b) Comparison of site testing PSCs to utilized PSCs
c) Data is matched against DOB and gender before all transfers.

2. Generation of summary scores

R scripts are used to calculate summary scores for all questionnaires at Delosis server. Details available at https://github.com/delosis/psytools

3. Data QC

Data is analysed to obtain the following QC measures:

Total number of subjects
Distribution based on – site, age band, gender
Strange values in dataset – to be listed along with the PSC1 code
Refused values in dataset (R, NR) – to be listed along with the PSC1 code
Missing values in dataset – PSC1 codes to be listed with details of missing values

Summaries – Frequencies (%) for count data; Mean (SD) for score data
Outliers (outside 2SD) – to be listed

*Missing values –
- Missing values may be skipped in some cases (e.g. female questions for male)

*Outliers
- Outliers are values that are less or more than 2 SD of the mean value.

*Strange values

- Correction if year of date is wrong and it is clearly a typo
- If a query needs to be sent it is important that the question is not 'leading' the recruiter towards an answer. E.g. Do not write 'Is the value 5' but ask 'Please recheck this value and confirm'.
- Keep consistency between queries. If possible, make template questions that can be reused for similar cases.

4. Data release

All releases are notified on the cVEDA website: https://cveda.org/dataset/