

c-VEDA Dataset Version 1.0

Release date: 2018-11-21

DOI: 10.25720/veda-c10e

This is the first release of c-VEDA Psytools baseline (BL) data obtained from the automated pipe-line that runs on the c-VEDA server:

1. A Python script downloads raw CSV files from the [Delosis server](#) into directory `/cveda/databank/RAW/PSC1/psytools/`:
[cveda psytools download.py](#)
2. Another Python script further de-identifies data. It converts subject identifiers from PSC1 to PSC2 and removes dates or converts them to age in days. The output goes into directory `/cveda/databank/RAW/PSC2/psytools/`:
[cveda psytools deidentify.py](#)
3. Finally an R script based on an R library provided by Delosis derives the CSV files, switching from long format to wide format, eliminating multiple sessions for a single subject. The output goes into directory `/cveda/databank/processed/psytools/`:
[cveda psytools derive.R](#)
[psytools task derivations.R](#)

At this point most subject identifier misassignments seem to have been addressed, at least for baseline (BL) data.

Psytools files can be downloaded via SFTP:
`sftp://cveda.nimhans.ac.in/data/1.0/psytools/`

In addition to Psytools data, we provide an excerpt of the recruitment files maintained by each recruitment centre. These recruitment files provide reference values for both the date of birth and sex of each subject. They have been checked and re-checked multiple times. The reference date of birth is not published, rather we use it internally to calculate the age in days associated to dates during de-identification. The reference value of sex is published and should be used in preference to values found in Psytools files.

The de-identified excerpt of the recruitment files can be downloaded via SFTP:
`sftp://cveda.nimhans.ac.in/data/1.0/recruitment_files/`

Errors and caveats

Among other issues, we have found issues in the pseudonymization of a dozen participants. We recommend you wait for release 1.1.