Verification Report: data_clean

Ellis Hughes

2021-05-12

Description

This is the verification report for the `data_clean` folder of the `correlates_reporting` project for CoVPN.

In this document, the output of `make_data_proc.R` is compared against the output of `process_data_raw.R`. The two scripts use the same base mock data. The two datasets generated by each of the two scripts will be compared with each other to confirm they contain the same values.

`process_data_raw.R` was independently double programmed based on the specifications found in `dat_clean_specifications.pdf`. This script outputs its processing to `data_clean/verification/verification_output/data_clean_verification_output.csv`

The file `data_clean/verification/verification_input/practice_data_w29.csv` was provided by the original programmer to the tester for verification purposes of `make_data_proc.R` when all time points are available.s Its md5 hash is e224a508c1fd6fda837d0e9b78fc85f2.

The file `data_clean/verification/verification_input/practice_data_wo29.csv` was provided by the original programmer to the tester for verification purposes of `make_data_proc.R` when only baseline and Day 57 were available. Its md5 hash is 83af631a06bb4a6efc2b3129022d21f1.

The file `data_clean/verification/verification_output/data_clean_verification_output.csv` was created by the tester for verification using the `process_data_Raw.R` script and using all time points available. Its md5 hash is 1dc6d4a45d4c8b3db1b614e6a66788c7.

The file `data_clean/verification/verification_output/data_clean_verification_output.csv` was created by the tester for verification using the `process_data_Raw.R` script and using only the time points Baseline and Day 57. Its md5 hash is cd644cc7731b630e4c0f957ee638a3aa.

Load Data

```
original_data <- read_csv(
  here("data_clean/verification/verification_input","practice_data_w29.csv"),
  guess_max = 30000)

original_data_no_d29 <- read_csv(
  here("data_clean/verification/verification_input","practice_data_wo29.csv"),
  guess_max = 30000)

verification_data <- read_csv(
  here("data_clean/verification/verification_output/data_clean_verification_output.csv"),
  guess_max = 30000)

verification_data_no_d29 <- read_csv(
  here("data_clean/verification/verification_output/data_clean_verification_output_no_D29.csv"),
  guess_max = 30000)
```

Verification

```
data_clean_comparison <- compare_datasets(
  cols =  colnames(original_data), index = "Ptid",
  ds1 = original_data, ds2 = verification_data
)
```

```
## There are 0 mismatched fields of 100.
```

```
data_clean_comparison_no_d29 <- compare_datasets(
  cols = colnames(original_data_no_d29), index = "Ptid",
  ds1 = original_data_no_d29, ds2 = verification_data_no_d29
)
```

```
## There are 0 mismatched fields of 87.
```

Output of `make_data_proc.R` is equivalent to the output of `process_data_raw.R` for cases when all time points are available. Output of `make_data_proc.R` is equivalent to the output of `process_data_raw.R` for cases when only Baseline and Day 57 are available. `make_data_proc.R` passes verification.

Signatures

Role	Name	Signature	Date
Tester	Ellis Hughes		