**HIV Descriptive study protocol**

**Version history**

|  |  |  |
| --- | --- | --- |
| Date | Protocol version | Changes |
| Nov 11, 2018 | v0.5 | initial draft |
| July10, 2019 | v1.1 | basic covariates collection for HIV cohort |
| July 22, 2019 | v1.3 | * Min cell count ([fixed #5](https://github.com/Sigfried/HivDescriptive/issues/5)) * Using FeatureExtraction from development branch ([fixed #3](https://github.com/Sigfried/HivDescriptive/issues/3)) **Fix if needed when changes are merged into FE:master** * Added sample output **MAKE SURE TO UPDATE IN NEW RELEASES** * Option in call to execute() to choose all possible covariates or a smaller list * Output includes table1 (from FE:createTable1) and new csv format that allows cross-site merging during central processing. * Using tidyverse now. (**Realize certain people may prefer that we don’t, and we can get rid of it if needed, but for now I can write code a lot faster with it – SG**) * Wrote code for including custom covariates. |
| Aug 2, 2019 | v6.0 | * Added mean\_visits\_per\_person\_by\_cohort. Does not use FeatureExtraction functions, just simple SQL and R code. |
| Aug 23 | v7.0 | * Making fake cohorts that match patients in two synthetic CDMs (1K SynPUF and Eunomia) |

# Study background

The study is part of a larger project titled Identification of Research Common Data Elements in HIV/AIDS using data science methods described at <https://github.com/lhncbc/CDE/blob/master/hiv/README.md>

Study is implemented as an R package.

# Study steps

The study consist of

1. Extracting data on HIV cohorts from multiple databases (Site Data Extraction)
2. Comparison of the data across sites (Centralized Processing)

# Site Data Extraction

After execution of the R study package, a set of files (in a .zip file) is transmitted to centralized processing (content of export folder). The extracted data is limited to aggregated data. No patient level information is included in the extract.

Study sites are encouraged to inspect the content of the export folder.

## Implementations

FeatureExtraction package is used to compute characteristics of the cohorts.

## Cohorts

cohorts.csv file contains a list of cohorts used by the package

# Centralized processing

Centralized processing consist of comparing datasets (from sites). The goal is to identify which data elements are common across datasets.