**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 |
| Aug 22, 2019 | v2.0 | Characterization of study, improved list of covariates |

# 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.

This is an informatics study focused on data elements in a defined cohort of patients. There is no clinical hypothesis question being studied in this study.

# 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

## Site data extracted and study goals

* What Data Elements (DEs) are present in EHR data (subset of codes enumerated by the package; aggregated statistics for site, no individual patient level data is extracted)
* Including Rx history data
* What Rx are HIV patients taking (over time) (based on enumerated list of ingredient concept\_id’s)
* Including Dx history data
* Comorbidity data for HIV patients (based on enumerated list of concept\_id’s)
* Diagnostic evolution of HIV via a set of related cohort defintions (similar to treatment pathway analytical approach)
* Ever progression into HIV viral load > 200 (for LOINC:20447-9)
* Measure data density (with the goal to demonstrate higher data density of EHR data compared with clinical trial (e.g., longer observation periods)
* Describe cause of death (of HIV patients that are deceased)

# Centralized processing

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