FHIR-PIT

A tool to smooth the join of FHIR records with Environmental and Exposures data. It assumes a patient has lived in the same address for the study period. It joins FHIR records with exposure data through **patient's address**, and environmental data through **patient's address, study period** and **date-of-visit**.

Technical Overview:

Built using sbt

It expects a particular folder structure.
Build.sbt:

- Name, version, scalaVersion
- Library Dependencies
- AssemblyMergeStrategies
- Resolvers

All scala files should go under src/main/scala. Build.sbt acts similar to requirements.txt, lists all dependencies and compiles all scala code that matches the folder structure.

Configured using the DHALL programming language

Example.dhall contains an example format configuration file for running FHIR-PIT. It calls pipeline.dhall.

The input parameters are:
Report:
Progress:
Configdir: Directory with extra configuration files (mainly icees_features.yaml)
Basedirinput: Data directory. The following structure is expected:

FHIR/: Contains FHIR records

EPR/ : EPR files
ICEESPCD/: Xwalk files
other/spatial/: Contains exposure data
other/env/: Contains environmental data.

Basedir: Directory to save intermediate results.
Basediroutput: Directory to save outputs.
Fhirconfig: JSON that configures the FHIR PIT run
SkipList: List configurations for the PreprocCSVTable transformation of each study period. (Poorly named)
Pyexec: Python exec.

The configuration file is YAML and defines a list of steps.
Each step has parameters: name, skip, arguments and function.
The function defines the object Step: It inherits the StepImpl trait and overrides a config type, a decoder and a step function. Afterwards, The main method parses the config.yaml file; creates a queue of steps; and executes each step in the queue. Results are saved in the directories specified by the arguments of each step.

Scala design-patterns/notions used:
Implicits: Passing the “wrong” type. “Unknown” method call. It performs type conversion implicitly when the wrong type is passed. The programmer defines the conversion. It can also define methods that extend functionality with conversion
Encoder/Decoder JSON: Defines a case class object with attributes that match the fields in a JSON file. It then uses io.circe.generic.semiauto.deriveDecoder to parse the json into the respective class. The classes can be nested to match the nested nature of the JSON.

Others: Type classes, Case classes, companion objects, Nullary functions for exec. time, Mapper, HashMap with Multimap, Match-case, foreach.

Files used in demo:

Basedirinput = /FHIR-PIT/data/input

[36M] FHIR data
{basedirinput}/FHIR/all
{basedirinput}/FHIR/all

[4M] Environmental data:
{basedirinput}/other/env/merged_cmaq_2010.csv

[1M] Environmental data:
{basedirinput}/other/env/cmaq2010

[589M] Census_data:
{basedirinput}/other/spatial/env/US_Census_Tracts_LCC/US_Census_Tracts_LCC.*

[1M] ACS_data:
{basedirinput}/other/spatial/acs/ACS_NC_2016_with_column_headers.csv

[36M] ACSUR_data:
{basedirinput}/other/spatial/acs/Appold_trans_geo_cross_02.10.10 - trans_geo_cross.csv

[45M] geoid_data:

{basedirinput}/other/spatial/acs/tl_2016_37_bg_lcc.shp

[489M] NearestRoadTL:
{basedirinput}/other/spatial/nearestRoadTL/tl_2015_allstates_prisecroads_lcc.shp

[2.6G] NearestRoadHPMS:
{basedirinput}/other/spatial/nearestRoadHPMS/hpms2016_major_roads_lcc.shp

[8.3M] CAFO + Landfill:
{basedirinput}/other/spatial/BDT_PointDatasets/

[2M] XWalkData:
{basedirinput}/ICEESPCD/RegistryPtsXWalkForHao.csv
{basedirinput}/ICEESPCD/8000PtsXWalkForHao.csv

[56K] EPR:
{basedirinput}/EPR/UNC_NIEHS_XWalk_for_Hao_shape_h3.csv
{basedirinput}/EPR/TLR4_AllData_NewHash_01292020 NO PII_no_new_line.csv

Step description:

Each step (bold and underlined) describes the input files, output files and highlights the contribution/operational-role of each input file in the output schema. Each input file is colored in the description. Each output file is the result of transformations/joins of multiple input files. The output schema is listed under each filename in a box. Data-elements in the output file's schema are highlighted with a paler color to match its corresponding input-file.

PreprocFHIR

Takes in: FHIR Data

Description:

Process encounters and resources on a per-patient basis.
It reorganizes data into the following file structure:
FHIR_processed/<Resource_type>/<Patient_num>/<Resource_num>@<iter_num>.json. Additionally it creates the frequently used geo.csv file.

Outputs:

FHIR_processed
"FHIR-PIT/data/output/FHIR_processed/Patient/d0f9bf93-f99f-4544-b095-1d3c5265b5bb"
geo.csv
"FHIR-PIT/data/output/FHIR_processed/geo.csv"

Patient_num, lat, lon

EnvCoordinates

Takes in: cmaq files, geo.csv, start date and end date

Description:

For each patient in **geo.csv** and for all years between start date and end date, save the contents of **cmaq<year>/<Row><Col>Daily.csv** into other_processed/env_coordinates/<Patient_num>.csv. Where Row and Col are estimated from the patient's lat lon coordinates in **geo.csv**.
The result is all the daily exposures recorded throughout the study period in the Row,Col address of a patient.

Outputs:

"FHIR-PIT/data/output/other_processed/env_coordinates/d0f9bf93-f99f-4544-b095-1d3c5265b5bb"

start_date,o3_avg,pm25_avg,o3_max,pm25_max,o3_min,pm25_mi

n,o3_stddev,pm25_stddev

Notes: It is assumed the patient lived in the same address (i.e. lat lon coordinates in geo.csv) during each year between start and end dates.

LatLongtoGEOID

Takes in: US_Census_Tracts_LCC.shp and geo.csv

Description:

For each patient in **geo.csv**, map their lat-lon coordinates to fips using **US_Census_Tracts_LCC.shp**. Expand geo.csv with the FIPS column and save the result to: other_processed/lat_lon_to_geoid/geoids.csv.

To obtain GEOID, it uses the function getGeoidForLatLon. It creates an LCC point with lat, lon coordidanates. Then it asks for the census block containing point. It does so by looking at the geometry and checking if it contains the point.

Outputs:

“FHIR-PIT/data/output/other_processed/lat_lon_to_geoid/geoids.csv”

patient_num,lat,lon,FIPS

PreprocEnvDataFIPS

Takes in: merged_cmaq_<year>.csv, geoids.csv, start_date, end_date.

Description:

Union all **merged_cmaq_<year>.csv** files across all years between start_date and end_date. Then inner join the result on FIPS with **geoids.csv**. Expand the data frame with column: **start_date** (i.e. yyyy/mm/dd) from Date. The result is all the daily exposures recorded throughout the study period in the FIPS address of a patient.

Outputs:

“FHIR-PIT/data/output/other_processed/env_FIPS/preagg”

FIPS,Date,Longitude,Latitude,CO_ppbv,NO_ppbv,NO2_ppbv,NOX_ppbv,SO2_ppbv,ALD2_ppbv,FORM_ppbv,pm25_daily_average,pm2

5_daily_average_stderr,ozone_daily_8hour_maximum,ozone_daily_8hour_maximum_stderr,BENZ_ppbv,start_date,patient_num

Notes: It is assumed the patient lived in the same address for every year between start and end dates. If columns in the env file are missing Mapper.envInputColumns2, the cols are aggregated as null (This causes an ERROR because writeDataframe does not accept null columns). start_date column is created to match daily join later.

PreprocSplit

Takes in: other_processed/env_FIPS/preagg.csv

Description:

Extract the rows from each patient in **preagg.csv** and save them to its own csv file.

Outputs:

"FHIR-PIT/data/output/other_processed/env_split_FIPS/d0f9bf93-f99f-4544-b095-1d3c5265b5bb.csv"

FIPS,Date,Longitude,Latitude,CO_ppbv,NO_ppbv,NO2_ppbv,NOX_ppbv,SO2_ppbv,ALD2_ppbv,FORM_ppbv,pm25_daily_average,pm25_daily_average_stderr,ozone_daily_8hour_maximum,ozone_daily_8hour_maximum_stderr,BENZ_ppbv,start_date,patient_num

PreprocEnvDataAggregate

Takes in: /env_coordinates, indices, statistics, study_periods, study_period_bounds

Description:

For each patient's environmental exposures throughout the study period (i.e. /**env_coordinates)**, group exposures by the intervals [study_period_bounds(i), study_period_bounds(i+1)]. Then **aggregate/compute** statistics for a set of indices. Expand the original data with the computed set, and append the values from the **previous day**. Save the result to other_processed/env_agg_coordinates/<patient_num>.csv

Outputs:

“FHIR-PIT/data/output/other_processed/env_agg_coordinates/d0f9bf93-f99f-4544-b095-1d3c5265b5bb.csv”

```
start_date,pm25_max,pm25_avg,o3_max,o3_avg,pm25_max_avg,pm25_avg_avg,o3_max_avg,o3_avg_avg,pm25_max_max,pm25_avg_max,o3_max_max,o3_avg_max,pm25_max_prev_date,pm25_avg_prev_date,o3_max_prev_date,o3_avg_prev_date
```

PreprocEnvDataAggregate

Takes in: /env_split_FIPS, indices, statistics, study_periods, study_period_bounds

Description:

Same as above but done on the **/env_split_FIPS** data.

For each patient’s environmental exposures throughout the study period (i.e. **/env_split_FIPS**), group exposures by the intervals [study_period_bounds(i), study_period_bounds(i+1)]. Then **aggregate/compute** statistics for a set of indices. Expand the original data with the computed set, and append the values from the **previous day**.

Outputs:

“FHIR-PIT/data/output/other_processed/env_agg_FIPS/d0f9bf93-f99f-4544-b095-1d3c5265b5bb”

```
start_date,pm25_daily_average,ozone_daily_8hour_maximum,CO_ppbv,NO_ppbv,NO2_ppbv,NOX_ppbv,SO2_ppbv,ALD2_ppbv,FORM_ppbv,BENZ_ppbv,pm25_daily_average_avg,ozone_daily_8hour_maximum_avg,CO_ppbv_avg,NO_ppbv_avg,NO2_ppbv_avg,NOX_ppbv_avg,SO2_ppbv_avg,ALD2_ppbv_avg,FORM_ppbv_avg,BENZ_ppbv_avg,pm25_daily_average_max,ozone_daily_8hour_maximum_max,CO_ppbv_max,NO_ppbv_max,NO2_ppbv_max,NOX_ppbv_max,SO2_ppbv_max,ALD2_ppbv_max,FORM_ppbv_max,BENZ_ppbv_max,pm25_daily_average_prev_date,ozone_daily_8hour_maximum_prev_date,CO_ppbv_prev_date,NO_ppbv_prev_date,NO2_ppbv_prev_date,NOX_ppbv_prev_date,SO2_ppbv_prev_date,ALD2_ppbv
```

```
_prev_date,FORM_ppbv_prev_date,BENZ_ppbv_prev_date
```

PreprocPerPatSeriesACS

Takes in: geo.csv, tl_2016_37_bg_lcc.shp, ACS_NC_2016_with_column_headers.csv, icees_features.yaml/acs

Description:

Map the lat, lon columns of **geo.csv** to GEOID using tl_2016_37_bg_lcc.shp. Then inner join the result with **ACS_NC_2016_with_column_headers.csv** on GEOID.

Outputs:

"FHIR-PIT/data/output/other_processed/acs.csv"

```
patient_num,EstResidentialDensity,EstProbabilityHighSchoolMaxEd
ucation,EstProbabilityNoHealthIns,EstProbabilityHouseholdNonHisp
White,EstProbabilityESL,EstProbabilityNonHispWhite,EstHouseholdI
ncome,EstResidentialDensity25Plus,EstProbabilityNoAuto
```

PreprocPerPatSeriesACSUR

Takes in: geo.csv, tl_2016_37_bg_lcc.shp, Appold_trans_geo_cross_02.10.10, trans_geo_cross.csv, icees_features.yaml/acsUR

Description:

Map the lat, lon columns of **geo.csv** to GEOID using tl_2016_37_bg_lcc.shp. Then inner join the result with **Appold_trans_geo_cross_02.10.10 - trans_geo_cross.csv** on GEOID.

Outputs:

"FHIR-PIT/data/output/other_processed/acsUR.csv"

```
patient_num,ur
```

PreprocPerPatSeriesNearestRoad

Takes in: geo.csv, tl_2015_allstates_prisecroads_lcc.shp, icees_features.yaml/nearestRoadTL

Description:

For each patient in **geo.csv**, compute distance from patient's address to nearest road using **tl_2015_allstates_prisecroads_lcc.shp**.

Outputs:

"FHIR-PIT/data/output/other_processed/nearestRoadTL.csv"

```
patient_num,MajorRoadwayHighwayExposure
```

PreprocPerPatSeriesNearestRoadHPMS

Takes in: geo.csv, hpms2016_major_roads_lcc.shp, icees_features.yaml/nearestRoadHPMS

Description:

For each patient in **geo.csv**, compute distance from patient's address to nearest road distance and features using **hpms2016_major_roads_lcc.shp**.

Outputs:

"FHIR-PIT/data/output/other_processed/nearestRoadHPMS.csv"

```
patient_num,RoadwayDistanceExposure,RoadwayType,RoadwayAADT,RoadwaySpeedLimit,RoadwayLanes
```

PreprocPerPatSeriesNearestPointCafo

Takes in: geo.csv, Permitted_Animal_Facilities-4-1-2020.shp, icees_features.yaml/cafo

Description:

For each patient in **geo.csv**, find distance in meters between the patient's address and **Permitted_Animal_Facilities-4-1-2020.shp**.

Outputs:

"FHIR-PIT/data/output/other_processed/cafo.csv"

```
patient_num,CAFO_Exposure
```

PreprocPerPatSeriesNearestPointLandfill

Takes in: geo.csv, Active_Permitted_Landfills_geo.shp, icees_features.yaml/landfill

Description:

For each patient in **geo.csv**, find distance in meters between the patient's address and **Active_Permitted_Landfills_geo.shp**.

Outputs:

"FHIR-PIT/data/output/other_processed/landfill.csv"

patient_num,Landfill_Exposure

PreprocPerPatSeriesToVector

Takes in: FHIR_processed/, start_date, end_date, icees_features.yaml/

Description:

Load a patient json into a Patient object, which contains a list of Medication objects, Address objects, Encounter objects, Condition objects, Lab objects, Procedure objects, bmi objects.

An Encounter object also contains a list of: Condition, Lab, Medication, Procedure objects; in addition to an id, start_date, end_date. Encounters is a data structure to aid operations between all these pieces of information.

For each patient, group their encounters, medications (converted to Encounter), conditions (converted to Encounter), labs (converted to Encounter), procedures (converted to Encounter) by **day** using a HashMultiMap (i.e. A dictionary with day as key and set of encounters as value). Accordingly, encounters are grouped-by on a daily basis.

Then aggregate features from the encounters by: Count, First, Last.

Finally, collect all the vectors, union all the feature names collected each day, and populate a csv with missing values replaced by “”.

Outputs:

“FHIR-PIT/data/output/FHIR_vector/patient/d0f9bf93-f99f-4544-b095-1d3c5265b5bb.csv”, which contains the demographic information of a patient.

```
patient_num,Ethnicity,Sex,birth_date,Race
```

“FHIR-PIT/data/output/FHIR_vector/visit/d0f9bf93-f99f-4544-b095-1d3c5265b5bb.csv”, which contains the patient’s vectorized/aggregated encounters.

```
patient_num,start_date,Ethnicity,AgeVisit,ObesityBMIVisit,encounter_num,VisitType,Sex,birth_date,Race
```

PreprocPerPatSeriesToCSVTable

Takes in: /FHIR_vector/visit/<patient_num>, landfill.csv, cafe.csv, nearestRoadHPMS.csv, nearestRoadTL.csv, acsur.csv, acs.csv, /other_processed/env_agg_FIPS/<patient_num>, /other_processed/env_agg_coordinates/<patient_num>, study_periods, study_period_bounds.

Description:

For each patient in /FHIR_vector/visit/<patient_num>.csv: Left join the **visits vector** with **environmental data (i.e.** /other_processed/env_agg_coordinates/<patient_num>**)** on start_date. Thus attaching the aggregated environmental exposures that occurred the day of the encounter. Similarly, associate the **exposure** information through the patient_num (Note exposures were associated to the patient_num with the patient's address). Lastly, **expand** the data with **bucket**, **study_period** and **year**. Extra columns that all carry the same information (i.e. Which study_period_bound is associated with a given row). Save the result to icees/<year>/<patient>.

Outputs:

"FHIR-PIT/data/output/icees/2010/d0f9bf93-f99f-4544-b095-1d3c5265b5bb"

```
bucket,patient_num,start_date,Ethnicity,AgeVisit,ObesityBMIVisit,encounter_num,VisitType,Sex,birth_date,Race,pm25_daily_average,ozone_daily_8hour_maximum,CO_ppbv,NO_ppbv,NO2_ppbv,NOX_ppbv,SO2_ppbv,ALD2_ppbv,FORM_ppbv,BENZ_ppbv,pm25_daily_average_avg,ozone_daily_8hour_maximum_avg,CO_ppbv_avg,NO_ppbv_avg,NO2_ppbv_avg,NOX_ppbv_avg,SO2_ppbv_avg,ALD2_ppbv_avg,FORM_ppbv_avg,BENZ_ppbv_avg,pm25_daily_average_max,ozone_daily_8hour_maximum_max,CO_ppbv_max,NO_ppbv_max,NO2_ppbv_max,NOX_ppbv_max,SO2_ppbv_max,ALD2_ppbv_max,FORM_ppbv_max,BENZ_ppbv_max,pm25_daily_average_prev_date,ozone_daily_8hour_maximum_prev_date,CO_ppbv_prev_date,N
```

O_ppbv_prev_date,NO2_ppbv_prev_date,NOX_ppbv_prev_date,SO2_ppbv_prev_date,ALD2_ppbv_prev_date,FORM_ppbv_prev_date,BENZ_ppbv_prev_date,EstResidentialDensity,EstProbabilityHighSchoolMaxEducation,EstProbabilityNoHealthIns,EstProbabilityHouseholdNonHispWhite,EstProbabilityESL,EstProbabilityNonHispWhite,EstHouseholdIncome,EstResidentialDensity25Plus,EstProbabilityNoAutour,MajorRoadwayHighwayExposure,RoadwayDistanceExposure,RoadwayType,RoadwayAADT,RoadwaySpeedLimit,RoadwayLanes,CAFO_Exposure,Landfill_Exposure,study_period,year

Note: Bucket, Study_period, year information is redundant when study_period_bounds are yearly.

PreprocCSVTable

Takes in:

"FHIR-PIT/data/output/icees/2010/d0f9bf93-f99f-4544-b095-1d3c5265b5bb.csv",
"FHIR-PIT/data/output/FHIR_Vector/patient/d0f9bf93-f99f-4544-b095-1d3c5265b5bb.csv"

Description:

Per year, the visits of all patients are "union" into a single table, grouped by (patient_num, study_period) and aggregated with the following sql functions:

first(pm25_avg, false) AS `AvgDailyPM2.5Exposure`
first(pm25_avg_avg, false) AS `AvgDailyPM2.5Exposure_StudyAvg`
first(pm25_avg_max, false) AS `AvgDailyPM2.5Exposure_StudyMax`
first(pm25_max, false) AS `MaxDailyPM2.5Exposure`
first(pm25_max_avg, false) AS `MaxDailyPM2.5Exposure_StudyAvg`
first(pm25_max_max, false) AS `MaxDailyPM2.5Exposure_StudyMax`
first(o3_avg, false) AS `AvgDailyOzoneExposure`
first(o3_avg_avg, false) AS `AvgDailyOzoneExposure_StudyAvg`
first(o3_avg_max, false) AS `AvgDailyOzoneExposure_StudyMax`
first(o3_max, false) AS `MaxDailyOzoneExposure`
first(o3_max_avg, false) AS `MaxDailyOzoneExposure_StudyAvg`
first(o3_max_max, false) AS `MaxDailyOzoneExposure_StudyMax`
first(pm25_daily_average_avg, false) AS `AvgDailyPM2.5Exposure_2`
first(ozone_daily_8hour_maximum_avg, false) AS `MaxDailyOzoneExposure_2`
first(CO_ppbv_avg, false) AS `AvgDailyCOExposure_2`
first(NO_ppbv_avg, false) AS `AvgDailyNOExposure_2`
first(NO2_ppbv_avg, false) AS `AvgDailyNO2Exposure_2`
first(NOX_ppbv_avg, false) AS `AvgDailyNOxExposure_2`
first(SO2_ppbv_avg, false) AS `AvgDailySO2Exposure_2`
first(ALD2_ppbv_avg, false) AS `AvgDailyAcetaldehydeExposure_2`
first(FORM_ppbv_avg, false) AS `AvgDailyFormaldehydeExposure_2`
first(BENZ_ppbv_avg, false) AS `AvgDailyBenzeneExposure_2`
max(ObesityBMIVisit) AS `ObesityBMI`
totaltypevisits(VisitType, RespiratoryDx) AS `TotalEDVisits`
totaltypevisits(VisitType, RespiratoryDx) AS `TotalInpatientVisits`
(totaltypevisits(VisitType, RespiratoryDx) + totaltypevisits(VisitType, RespiratoryDx)) AS `TotalEDInpatientVisits`
first(Sex2, false) AS `Sex2`
first(birth_date, false) AS `birth_date`
first(Sex, false) AS `Sex`
first(Race, false) AS `Race`
first(Ethnicity, false) AS `Ethnicity`
first(MajorRoadwayHighwayExposure, false) AS `MajorRoadwayHighwayExposure`
first(RoadwayDistanceExposure, false) AS `RoadwayDistanceExposure`
first(RoadwayType, false) AS `RoadwayType`
first(RoadwayAADT, false) AS `RoadwayAADT`
first(RoadwaySpeedLimit, false) AS `RoadwaySpeedLimit`
first(RoadwayLanes, false) AS `RoadwayLanes`

first(CAFO_Exposure, false) AS `CAFO_Exposure`
first(Landfill_Exposure, false) AS `Landfill_Exposure`
first(EstResidentialDensity, false) AS `EstResidentialDensity`
first(EstProbabilityHighSchoolMaxEducation, false) AS `EstProbabilityHighSchoolMaxEducation`
first(EstProbabilityNoHealthIns, false) AS `EstProbabilityNoHealthIns`
first(EstProbabilityHouseholdNonHispWhite, false) AS `EstProbabilityHouseholdNonHispWhite`
first(EstProbabilityESL, false) AS `EstProbabilityESL`
first(EstProbabilityNonHispWhite, false) AS `EstProbabilityNonHispWhite`
first(EstHouseholdIncome, false) AS `EstHouseholdIncome`
first(EstResidentialDensity25Plus, false) AS `EstResidentialDensity25Plus`
first(EstProbabilityNoAuto, false) AS `EstProbabilityNoAuto`
first(ur, false) AS `ur`
CAST(sum(Propranolol) AS INT) AS `Propranolol`
CAST(sum(Cetirizine) AS INT) AS `Cetirizine`
CAST(sum(PregnancyDx) AS INT) AS `PregnancyDx`
CAST(sum(Fluoxetine) AS INT) AS `Fluoxetine`
CAST(sum(ObesityDx) AS INT) AS `ObesityDx`
CAST(sum(Fluticasone) AS INT) AS `Fluticasone`
CAST(sum(Mometasone) AS INT) AS `Mometasone`
CAST(sum(Leuprolide) AS INT) AS `Leuprolide`
CAST(sum(Albuterol) AS INT) AS `Albuterol`
CAST(sum(CroupDx) AS INT) AS `CroupDx`
CAST(sum(ReactiveAirwayDx) AS INT) AS `ReactiveAirwayDx`
CAST(sum(Ciclesonide) AS INT) AS `Ciclesonide`
CAST(sum(Ipratropium) AS INT) AS `Ipratropium`
CAST(sum(KidneyCancerDx) AS INT) AS `KidneyCancerDx`
CAST(sum(Diphenhydramine) AS INT) AS `Diphenhydramine`
CAST(sum(Escitalopram) AS INT) AS `Escitalopram`
CAST(sum(Fexofenadine) AS INT) AS `Fexofenadine`
CAST(sum(EndometriosisDx) AS INT) AS `EndometriosisDx`
CAST(sum(OvarianDysfunctionDx) AS INT) AS `OvarianDysfunctionDx`
CAST(sum(Metaproterenol) AS INT) AS `Metaproterenol`
CAST(sum(TesticularCancerDx) AS INT) AS `TesticularCancerDx`
CAST(sum(Tamoxifen) AS INT) AS `Tamoxifen`
CAST(sum(Goserelin) AS INT) AS `Goserelin`
CAST(sum(DepressionDx) AS INT) AS `DepressionDx`
CAST(sum(AutismDx) AS INT) AS `AutismDx`
CAST(sum(Progesterone) AS INT) AS `Progesterone`
CAST(sum(CoughDx) AS INT) AS `CoughDx`
CAST(sum(Omalizumab) AS INT) AS `Omalizumab`
CAST(sum(AnxietyDx) AS INT) AS `AnxietyDx`
CAST(sum(Citalopram) AS INT) AS `Citalopram`
CAST(sum(Beclomethasone) AS INT) AS `Beclomethasone`
CAST(sum(Theophylline) AS INT) AS `Theophylline`
CAST(sum(FibromyalgiaDx) AS INT) AS `FibromyalgiaDx`
CAST(sum(Sertraline) AS INT) AS `Sertraline`
CAST(sum(Venlafaxine) AS INT) AS `Venlafaxine`
CAST(sum(DrugDependenceDx) AS INT) AS `DrugDependenceDx`
CAST(sum(Formoterol) AS INT) AS `Formoterol`
CAST(sum(AlcoholDependenceDx) AS INT) AS `AlcoholDependenceDx`
CAST(sum(AlopeciaDx) AS INT) AS `AlopeciaDx`
CAST(sum(MenopauseDx) AS INT) AS `MenopauseDx`
CAST(sum(CervicalCancerDx) AS INT) AS `CervicalCancerDx`
CAST(sum(Mepolizumab) AS INT) AS `Mepolizumab`
CAST(sum(TesticularDysfunctionDx) AS INT) AS `TesticularDysfunctionDx`
CAST(sum(Estropipate) AS INT) AS `Estropipate`
CAST(sum(Histrelin) AS INT) AS `Histrelin`
CAST(sum(Triptorelin) AS INT) AS `Triptorelin`
CAST(sum(Salmeterol) AS INT) AS `Salmeterol`
CAST(sum(Arformoterol) AS INT) AS `Arformoterol`
CAST(sum(Paroxetine) AS INT) AS `Paroxetine`
CAST(sum(Flunisolide) AS INT) AS `Flunisolide`
CAST(sum(Testosterone) AS INT) AS `Testosterone`
CAST(sum(Budesonide) AS INT) AS `Budesonide`
CAST(sum(DiabetesDx) AS INT) AS `DiabetesDx`
CAST(sum(Metformin) AS INT) AS `Metformin`
CAST(sum(Nandrolone) AS INT) AS `Nandrolone`
CAST(sum(Prasterone) AS INT) AS `Prasterone`
CAST(sum(AsthmaDx) AS INT) AS `AsthmaDx`
CAST(sum(Indacaterol) AS INT) AS `Indacaterol`
CAST(sum(Androstenedione) AS INT) AS `Androstenedione`
CAST(sum(Duloxetine) AS INT) AS `Duloxetine`
CAST(sum(Prednisone) AS INT) AS `Prednisone`
CAST(sum(PneumoniaDx) AS INT) AS `PneumoniaDx`
CAST(sum(UterineCancerDx) AS INT) AS `UterineCancerDx`
CAST(sum(Medroxyprogresterone) AS INT) AS `Medroxyprogresterone`
CAST(sum(Hydroxyzine) AS INT) AS `Hydroxyzine`
CAST(sum(ProstateCancerDx) AS INT) AS `ProstateCancerDx`
CAST(sum(Estrogen) AS INT) AS `Estrogen`
CAST(sum(Trazodone) AS INT) AS `Trazodone`
CAST(sum(Estradiol) AS INT) AS `Estradiol`
CAST(sum(OvarianCancerDx) AS INT) AS `OvarianCancerDx`

The result of the aggregation is in /icees2/2010/all_patient. Intermediate files are created /icees2/2010/visit_combined and /icees2/2010/all_visit which contain the union of all patient's visit info. The only difference between the two are renames of columns and default missing values.

Outputs:

"FHIR-PIT/data/output/icees2/2010/visit_combined"

It contains all the icees/<year> information for all patients combined. The schema is the union of all columns from all files in icees/<year> with 0 as a default for missing values.

```
AgeVisit,ALD2_ppbv,ALD2_ppbv_avg,ALD2_ppbv_max,ALD2_ppbv
_prev_date,BENZ_ppbv,BENZ_ppbv_avg,BENZ_ppbv_max,BENZ_
ppbv_prev_date,birth_date,bucket,CAFO_Exposure,CO_ppbv,CO_p
pbv_avg,CO_ppbv_max,CO_ppbv_prev_date,encounter_num,EstHo
useholdIncome,EstProbabilityESL,EstProbabilityHighSchoolMaxEdu
cation,EstProbabilityHouseholdNonHispWhite,EstProbabilityNoAuto,
EstProbabilityNoHealthIns,EstProbabilityNonHispWhite,EstResidenti
alDensity,EstResidentialDensity25Plus,Ethnicity,FORM_ppbv,FORM
_ppbv_avg,FORM_ppbv_max,FORM_ppbv_prev_date,Landfill_Exp
osure,MajorRoadwayHighwayExposure,NO2_ppbv,NO2_ppbv_avg,
NO2_ppbv_max,NO2_ppbv_prev_date,NO_ppbv,NO_ppbv_avg,NO
_ppbv_max,NO_ppbv_prev_date,NOX_ppbv,NOX_ppbv_avg,NOX_
ppbv_max,NOX_ppbv_prev_date,ObesityBMIVisit,ozone_daily_8hou
r_maximum,ozone_daily_8hour_maximum_avg,ozone_daily_8hour_
maximum_max,ozone_daily_8hour_maximum_prev_date,patient_nu
m,pm25_daily_average,pm25_daily_average_avg,pm25_daily_aver
age_max,pm25_daily_average_prev_date,Race,RoadwayAADT,Roa
dwayDistanceExposure,RoadwayLanes,RoadwaySpeedLimit,Roadw
ayType,Sex,SO2_ppbv,SO2_ppbv_avg,SO2_ppbv_max,SO2_ppbv_
prev_date,start_date,study_period,ur,VisitType,year
```

“FHIR-PIT/data/output/icees2/2010/all_visit”
Renamed and less columns than /icees2/2010/visit_combined

```
AgeVisit,Avg24hAcetaldehydeExposure_2,Avg24hBenzeneExposur
e_2,birth_date,bucket,CAFO_Exposure,Avg24hCOExposure_2,enco
unter_num,EstHouseholdIncome,EstProbabilityESL,EstProbabilityHi
ghSchoolMaxEducation,EstProbabilityHouseholdNonHispWhite,EstP
robabilityNoAuto,EstProbabilityNoHealthIns,EstProbabilityNonHispW
hite,EstResidentialDensity,EstResidentialDensity25Plus,Ethnicity,Av
g24hFormaldehydeExposure_2,Landfill_Exposure,MajorRoadwayHi
ghwayExposure,Avg24hNO2Exposure_2,Avg24hNOExposure_2,Av
g24hNOxExposure_2,ObesityBMIVisit,Max24hOzoneExposure_2,pa
```

tient_num,Avg24hPM2.5Exposure_2,Race,RoadwayAADT,RoadwayDistanceExposure,RoadwayLanes,RoadwaySpeedLimit,RoadwayType,Sex,Avg24hSO2Exposure_2,start_date,study_period,ur,VisitType,year,BeclomethasoneVisit,EstradiolVisit,ProstateCancerDxVisit,AlopeciaDxVisit,UterineCancerDxVisit,AutismDxVisit,DepressionDxVisit,TheophyllineVisit,FormoterolVisit,OvarianDysfunctionDxVisit,AlcoholDependenceDxVisit,MetforminVisit,ObesityDxVisit,ParoxetineVisit,DiphenhydramineVisit,PregnancyDxVisit,MepolizumabVisit,MedroxyprogesteroneVisit,LeuprolideVisit,CiclesonideVisit,ReactiveAirwayDxVisit,EndometriosisDxVisit,AnxietyDxVisit,FibromyalgiaDxVisit,FluticasoneVisit,IndacaterolVisit,CetirizineVisit,HydroxyzineVisit,FexofenadineVisit,PrednisoneVisit,TamoxifenVisit,MometasoneVisit,TrazodoneVisit,PropranololVisit,CoughDxVisit,NandroloneVisit,DrugDependenceDxVisit,ArformoterolVisit,BudesonideVisit,OvarianCancerDxVisit,CroupDxVisit,VenlafaxineVisit,MetaproterenolVisit,HistrelinVisit,OmalizumabVisit,MenopauseDxVisit,PneumoniaDxVisit,DiabetesDxVisit,DuloxetineVisit,EstropipateVisit,GoserelinVisit,CitalopramVisit,CervicalCancerDxVisit,AndrostenedioneVisit,ProgesteroneVisit,TestosteroneVisit,SertralineVisit,EscitalopramVisit,EstrogenVisit,TriptorelinVisit,SalmeterolVisit,IpratropiumVisit,KidneyCancerDxVisit,TesticularCancerDxVisit,AsthmaDxVisit,FlunisolideVisit,PrasteroneVisit,TesticularDysfunctionDxVisit,FluoxetineVisit,AlbuterolVisit,Sex2,Avg24hPM2.5Exposure,Max24hPM2.5Exposure,Avg24hOzoneExposure,Max24hOzoneExposure

"FHIR-PIT/data/output/icees2/2010/all_patient"
This file is the result of the aggregation.

Note: The last portion of the icees_features.yaml file lists the Dx whose presence indicates a RespiratoryDx. That is RespiratoryDx is True if the visit had any of the Dx listed. Then, for each given (patient, study_period),
TotalEDVisits/TotalInpatientVisits/TotalEDInpatientVisits count the total number of visits that indicate a RespiratoryDx and satisfy a given visitType. For instance, TotalInpatientVisits is the total count of "IMP"

type visits registered with “RespiratoryDx” flag for a given (patient, year); TotalEDVisits is the total count of “AMB”,“EMER” type visits registered with “RespiratoryDx” flag for a given (patient, year)

patient_num,study_period,AvgDailyPM2.5Exposure,AvgDailyPM2.5Exposure_StudyAvg,AvgDailyPM2.5Exposure_StudyMax,MaxDailyPM2.5Exposure,MaxDailyPM2.5Exposure_StudyAvg,MaxDailyPM2.5Exposure_StudyMax,AvgDailyOzoneExposure,AvgDailyOzoneExposure_StudyAvg,AvgDailyOzoneExposure_StudyMax,MaxDailyOzoneExposure,MaxDailyOzoneExposure_StudyAvg,MaxDailyOzoneExposure_StudyMax,AvgDailyPM2.5Exposure_2,MaxDailyOzoneExposure_2,AvgDailyCOExposure_2,AvgDailyNOExposure_2,AvgDailyNO2Exposure_2,AvgDailyNOxExposure_2,AvgDailySO2Exposure_2,AvgDailyAcetaldehydeExposure_2,AvgDailyFormaldehydeExposure_2,AvgDailyBenzeneExposure_2,ObesityBMI,TotalEDVisits,TotalInpatientVisits,TotalEDInpatientVisits,Sex2,birth_date,Sex,Race,Ethnicity,MajorRoadwayHighwayExposure,RoadwayDistanceExposure,RoadwayType,RoadwayAADT,RoadwaySpeedLimit,RoadwayLanes,CAFO_Exposure,Landfill_Exposure,EstResidentialDensity,EstProbabilityHighSchoolMaxEducation,EstProbabilityNoHealthIns,EstProbabilityHouseholdNonHispWhite,EstProbabilityESL,EstProbabilityNonHispWhite,EstHouseholdIncome,EstResidentialDensity25Plus,EstProbabilityNoAuto,ur,Prednisone,OvarianDysfunctionDx,Hydroxyzine,Albuterol,Diphenhydramine,TesticularDysfunctionDx,Estradiol,PneumoniaDx,Formoterol,DiabetesDx,Metformin,Beclomethasone,Omalizumab,OvarianCancerDx,Venlafaxine,PregnancyDx,Fexofenadine,Arformoterol,Mometasone,Medroxyprogresterone,DepressionDx,Propranolol,Salmeterol,TesticularCancerDx,Budesonide,Prasterone,Trazodone,AlcoholDependenceDx,Androstenedione,Ipratropium,Tamoxifen,Nandrolone,Duloxetine,EndometriosisDx,CroupDx,Escitalopram,Metaproterenol,ObesityDx,CoughDx,Histrelin,AutismDx,Estrogen,Ciclesonide,UterineCancerDx,Sertraline,Testosterone,Estropipate,Indacaterol,AsthmaDx,Triptorelin,KidneyCancerDx,Progesterone,ReactiveAirwayDx,Goserelin,ProstateCancerDx,AnxietyDx,Theophylline,DrugDependenceDx,Flunisolide,CervicalCancerDx,FibromyalgiaDx,Leuprolide,Fluoxetine,Citalopram,MenopauseDx,Cetirizine,Fluticasone,Paroxetine,Alope

ciaDx,Mepolizumab,AgeStudyStart,Active_In_Study_Period

BinICEES

Takes in:

Config_file: "FHIR-PIT/spark/config/icees_features.yaml",
Input_dir: "FHIR-PIT/data/output/icees2",
Output_dir: "FHIR-PIT/data/output/icees2_bins",
Study_periods: List of years

Description:

Note: icees2/year has three files {all_visit, all_patient, visit_combined}.

Add index information to dataframe.

The goal is to bin the features of all_visit and all_patient FHIR PIT outputs. Binning consists in categorizing the features based on bins. The bins are created to spread the data uniformly across each bin (i.e. qcut) or to bin the domain into evenly sized bins (i.e. cut). Note the difference between qcut and cut. Qcut may not produced evenly sized bins, and cut may not produced bins with the same amount of data in each.

The process for all_patient and all_visit is the same except for minor details. For each year get file {input_dir}/{year}/{all_patient, all_visit} and:

1. PreProcAge bins age: Bins age into two groups, ['<5', '5-17', '18-44', '45-64', '65-89'] and ['0-2', '3-17', '18-34', '35-50', '51-69', '70-89'].
2. PreProcEnv bins environmental variables.
3. PreProcSocial bins exposure variables.
4. Cut_col bins columns in the intersection between all_patient/all_visit cols and icees_features.FHIR.keys()

The data is then deidentified by dropping: patient_num and birthdate for all_patient. And dropping a larger list (cols_to_drop) for all_visit.

Save the result to: “/output/EPR_binned/EPR_binned”

Outputs:
"FHIR-PIT/data/output/EPR_binned/EPR_binned”

BinEPR
Takes in:
"/home/jjgarcia/projects/fhir_sample_data/TLR4_AllData_NewHash_01292020 NO PII_no_new_line.csv",
"/home/jjgarcia/projects/fhir_sample_data/UNC_NIEHS_XWalk_for_Hao_shape_h3.csv",
"FHIR-PIT/data/output/icees2_bins/",
"FHIR-PIT/data/output/FHIR_processed/geo.csv",
"FHIR-PIT/data/output/EPR_binned/EPR_binned",
study_periods
Description:
[Important] Cw has two columns [patient_num, HASH_VALUE]
Df has multiple columns, mostly on EPR_binned.
Df_pat_geo is loaded from **geo.csv**
Preprocess df and write it to **EPR_binned**.
JOIN LEFT df with cw on HASH_VALUE; creates column In_EPR; writes it to **EPR_binned_pat.** Save to df_pat
JOIN LEFT df with cw on hash, then OUTER JOIN result with df_pat_geo[patient_num]. Save to df_pat_ord.
The difference between df_pat and df_pat_ord are the geo.csv columns.
For each year in study_periods:

- Load Dfp from **icees2_bins/{year}patient**.

- OUTER JOIN df_pat with dfp on patient_num. LEFT JOIN the result with df_pat_ord on “patient_num” and “hash_value”. This will keep all the null values from the initial outer join and attach the geo information to non-null matches.
- Save result to dfpe.
- Preprocess na values, sort, cast floats to ints from dfpe, Save to **EPR_binned{year}patient**
- Drop [“patient_num”, “hash_value”] from dfpe. Save to **EPR_binned{year}patient_deidentified.**

Repeat the same process as above with two modifications: dfp is loaded from **icees2_bins/{year}visit**; dfpe is the OUTER JOIN df_pat with dfp on patient_num. It essentially contains the visit information and the hash information from df_pat (it contains patient_num, hash_value and all the original df information).
The results are saved to: **EPR_binned{year}visit, EPR_binned{year}visit_deidentified.**

Lastly, drop “hash_value” from df and save it to **EPR_binned_deidentified.**

Outputs:
"FHIR-PIT/data/output/EPR_binned/EPR_binned”

HASH_VALUE,IN_FINAL_SAMPLE,ORIGINAL_ANALYTIC_SAMPLE,TLR4_AGE,GENDER,RACE,ETHNICITY,CURRENT_AGE,RESPONDER_STATUS,QXAGE,BMI_CATEGORY,D28_ASTHMA,D28A_ASTHMA_AD_TEXT,D28B_STILL_HAVE_ASTHMA,D28C_ASTHMA_EPISODE_12M,D28D_ASTHMA_ER_VISIT_12M,D28E_ASTHMA_MED_TAKE_12M,D28F_ASTHMA_14D_NUM_NIGHTS_TEXT,D28G_ASTHMA_14D_LIMIT_DAYS_TEXT,D28H_ASTHMA_14D_NUM_WHEEZE_TEXT,S177_SMOKE_100_CIG_LIFETIME,SMOKE_CAT,SNP1,SNP2,SNP3,SNP4,ESTTOTALPOP,DISTANCE,AADT,RO

ADTYPE,SPEED,THROUGH_LANES,O3_N_OBS,PM25_N_OBS,TLR4_AGE2,CURRENT_AGE2,QXAGE2,D28A_ASTHMA_AD_TEXT2,TLR4_DIST_1X_qcut,TLR4_DIST_1X_cut,TLR4_DIST_2X_qcut,TLR4_DIST_2X_cut,TLR4_DIST_3X_qcut,TLR4_DIST_3X_cut,O3_ANNUAL_AVERAGE_qcut,O3_ANNUAL_AVERAGE_cut,PM25_ANNUAL_AVERAGE_qcut,PM25_ANNUAL_AVERAGE_cut,ESTTOTALPOP25PLUS_qcut,ESTTOTALPOP25PLUS_cut,ESTPROPPERSONSNONHISPWHITE_qcut,ESTPROPPERSONSNONHISPWHITE_cut,ESTPROPPERSONS25PLUSHSMAX_qcut,ESTPROPPERSONS25PLUSHSMAX_cut,ESTPROPHOUSEHOLDSNOAUTO_qcut,ESTPROPHOUSEHOLDSNOAUTO_cut,ESTPROPPERSONSNOHEALTHINS_qcut,ESTPROPPERSONSNOHEALTHINS_cut,ESTPROPPERSONS5PLUSNOENGLISH_qcut,ESTPROPPERSONS5PLUSNOENGLISH_cut,MEDIANHOUSEHOLDINCOME_qcut,MEDIANHOUSEHOLDINCOME_cut,DISTANCE2

"FHIR-PIT/data/output/EPR_binned/EPR_binned_pat”

HASH_VALUE,IN_FINAL_SAMPLE,ORIGINAL_ANALYTIC_SAMPLE,TLR4_AGE,GENDER,RACE,ETHNICITY,CURRENT_AGE,RESPONDER_STATUS,QXAGE,BMI_CATEGORY,D28_ASTHMA,D28A_ASTHMA_AD_TEXT,D28B_STILL_HAVE_ASTHMA,D28C_ASTHMA_EPISODE_12M,D28D_ASTHMA_ER_VISIT_12M,D28E_ASTHMA_MED_TAKE_12M,D28F_ASTHMA_14D_NUM_NIGHTS_TEXT,D28G_ASTHMA_14D_LIMIT_DAYS_TEXT,D28H_ASTHMA_14D_NUM_WHEEZE_TEXT,S177_SMOKE_100_CIG_LIFETIME,SMOKE_CAT,SNP1,SNP2,SNP3,SNP4,ESTTOTALPOP,DISTANCE,AADT,ROADTYPE,SPEED,THROUGH_LANES,O3_N_OBS,PM25_N_OBS,TLR4_AGE2,CURRENT_AGE2,QXAGE2,D28A_ASTHMA_AD_TEXT2,TLR4_DIST_1X_qcut,TLR4_DIST_1X_cut,TLR4_DIST_2X_qcut,TLR4_DIST_2X_cut,TLR4_DIST_3X_qcut,TLR4_DIST_3X_cut,O3_ANNUAL_AVERAGE_qcut,O3_ANNUAL_AVERAGE_cut,PM25_ANNUAL_AVERAGE_qcut,PM25_ANNUAL_AVERAGE_cut,ESTTOTALPOP25PLUS_qcut,ESTTOTALPOP25PLUS_cut,ESTPROPPERSONSNONHISPWHITE_qcut,ESTPROPPERSONSNONHISPWHITE_cu

t,ESTPROPPERSONS25PLUSHSMAX_qcut,ESTPROPPERSONS25PLUSHSMAX_cut,ESTPROPHOUSEHOLDSNOAUTO_qcut,ESTPROPHOUSEHOLDSNOAUTO_cut,ESTPROPPERSONSNOHEALTHINS_qcut,ESTPROPPERSONSNOHEALTHINS_cut,ESTPROPPERSONS5PLUSNOENGLISH_qcut,ESTPROPPERSONS5PLUSNOENGLISH_cut,MEDIANHOUSEHOLDINCOME_qcut,MEDIANHOUSEHOLDINCOME_cut,DISTANCE2,**patient_num**,**IN_EPR**

"FHIR-PIT/data/output/EPR_binned/EPR_binned2010patient”

HASH_VALUE,patient_num,index,IN_FINAL_SAMPLE,ORIGINAL_ANALYTIC_SAMPLE,TLR4_AGE,GENDER,RACE,ETHNICITY,CURRENT_AGE,RESPONDER_STATUS,QXAGE,BMI_CATEGORY,D28_ASTHMA,D28A_ASTHMA_AD_TEXT,D28B_STILL_HAVE_ASTHMA,D28C_ASTHMA_EPISODE_12M,D28D_ASTHMA_ER_VISIT_12M,D28E_ASTHMA_MED_TAKE_12M,D28F_ASTHMA_14D_NUM_NIGHTS_TEXT,D28G_ASTHMA_14D_LIMIT_DAYS_TEXT,D28H_ASTHMA_14D_NUM_WHEEZE_TEXT,S177_SMOKE_100_CIG_LIFETIME,SMOKE_CAT,SNP1,SNP2,SNP3,SNP4,ESTTOTALPOP,DISTANCE,AADT,ROADTYPE,SPEED,THROUGH_LANES,O3_N_OBS,PM25_N_OBS,TLR4_AGE2,CURRENT_AGE2,QXAGE2,D28A_ASTHMA_AD_TEXT2,TLR4_DIST_1X_qcut,TLR4_DIST_1X_cut,TLR4_DIST_2X_qcut,TLR4_DIST_2X_cut,TLR4_DIST_3X_qcut,TLR4_DIST_3X_cut,O3_ANNUAL_AVERAGE_qcut,O3_ANNUAL_AVERAGE_cut,PM25_ANNUAL_AVERAGE_qcut,PM25_ANNUAL_AVERAGE_cut,ESTTOTALPOP25PLUS_qcut,ESTTOTALPOP25PLUS_cut,ESTPROPPERSONSNONHISPWHITE_qcut,ESTPROPPERSONSNONHISPWHITE_cut,ESTPROPPERSONS25PLUSHSMAX_qcut,ESTPROPPERSONS25PLUSHSMAX_cut,ESTPROPHOUSEHOLDSNOAUTO_qcut,ESTPROPHOUSEHOLDSNOAUTO_cut,ESTPROPPERSONSNOHEALTHINS_qcut,ESTPROPPERSONSNOHEALTHINS_cut,ESTPROPPERSONS5PLUSNOENGLISH_qcut,ESTPROPPERSONS5PLUSNOENGLISH_cut,MEDIANHOUSEHOLDINCOME_qcut,MEDIANHOUSEHOLDINCOME_cut,DISTANCE2,IN_EPR,study_period,AvgDailyPM2.5Exposure,AvgDailyPM2.5Exposure_StudyAvg,

AvgDailyPM2.5Exposure_StudyMax,MaxDailyPM2.5Exposure,MaxDailyPM2.5Exposure_StudyAvg,MaxDailyPM2.5Exposure_StudyMax,AvgDailyOzoneExposure,AvgDailyOzoneExposure_StudyAvg,AvgDailyOzoneExposure_StudyMax,MaxDailyOzoneExposure,MaxDailyOzoneExposure_StudyAvg,MaxDailyOzoneExposure_StudyMax,AvgDailyPM2.5Exposure_2,MaxDailyOzoneExposure_2,AvgDailyCOExposure_2,AvgDailyNOExposure_2,AvgDailyNO2Exposure_2,AvgDailyNOxExposure_2,AvgDailySO2Exposure_2,AvgDailyAcetaldehydeExposure_2,AvgDailyFormaldehydeExposure_2,AvgDailyBenzeneExposure_2,ObesityBMI,TotalEDVisits,TotalInpatientVisits,TotalEDInpatientVisits,Sex2,Sex,Race,Ethnicity,MajorRoadwayHighwayExposure,RoadwayDistanceExposure,RoadwayType,RoadwayAADT,RoadwaySpeedLimit,RoadwayLanes,CAFO_Exposure,Landfill_Exposure,EstResidentialDensity,EstProbabilityHighSchoolMaxEducation,EstProbabilityNoHealthIns,EstProbabilityHouseholdNonHispWhite,EstProbabilityESL,EstProbabilityNonHispWhite,EstHouseholdIncome,EstResidentialDensity25Plus,EstProbabilityNoAuto,ur,PneumoniaDx,Ciclesonide,DrugDependenceDx,Leuprolide,CervicalCancerDx,AutismDx,TesticularCancerDx,Flunisolide,Escitalopram,AlopeciaDx,Propranolol,Ipratropium,Nandrolone,Testosterone,Prasterone,Trazodone,Venlafaxine,Estrogen,OvarianCancerDx,Sertraline,EndometriosisDx,CoughDx,Fluticasone,AnxietyDx,PregnancyDx,FibromyalgiaDx,Metformin,Citalopram,Paroxetine,Cetirizine,Androstenedione,Fluoxetine,Duloxetine,ObesityDx,AlcoholDependenceDx,Formoterol,Estropipate,OvarianDysfunctionDx,Budesonide,Omalizumab,Prednisone,Hydroxyzine,Albuterol,Metaproterenol,Theophylline,AsthmaDx,TesticularDysfunctionDx,Salmeterol,Goserelin,Tamoxifen,Indacaterol,KidneyCancerDx,CroupDx,Progesterone,MenopauseDx,Estradiol,Histrelin,Mometasone,Diphenhydramine,ReactiveAirwayDx,ProstateCancerDx,DiabetesDx,Medroxyprogresterone,Mepolizumab,DepressionDx,Fexofenadine,Triptorelin,Beclomethasone,UterineCancerDx,Arformoterol,AgeStudyStart,Active_In_Study_Period,AgeStudyStart2,AvgDailyPM2.5Exposure_StudyAvg_qcut,AvgDailyPM2.5Exposure_StudyMax_qcut,AvgDailyPM2.5Exposure_qcut,MaxDailyPM2.5Exposure_StudyAvg_qcut,MaxDailyPM2.5Exposure_StudyMax_qcut,MaxDailyPM2.5Exposure_qcut,AvgDailyOzoneExposure_StudyAvg_qcut,AvgDailyOzoneExposur

e_StudyMax_qcut,AvgDailyOzoneExposure_qcut,MaxDailyOzoneExposure_StudyAvg_qcut,MaxDailyOzoneExposure_StudyMax_qcut,MaxDailyOzoneExposure_qcut,AvgDailyPM2.5Exposure_2_qcut,MaxDailyOzoneExposure_2_qcut,AvgDailyCOExposure_2_qcut,AvgDailyNOExposure_2_qcut,AvgDailyNO2Exposure_2_qcut,AvgDailyNOxExposure_2_qcut,AvgDailySO2Exposure_2_qcut,AvgDailyAcetaldehydeExposure_2_qcut,AvgDailyFormaldehydeExposure_2_qcut,AvgDailyBenzeneExposure_2_qcut,MajorRoadwayHighwayExposure2,RoadwayDistanceExposure2,IN_ICEES

"FHIR-PIT/data/output/EPR_binned/EPR_binned2010patient_deidentified”

index,IN_FINAL_SAMPLE,ORIGINAL_ANALYTIC_SAMPLE,TLR4_AGE,GENDER,RACE,ETHNICITY,CURRENT_AGE,RESPONDER_STATUS,QXAGE,BMI_CATEGORY,D28_ASTHMA,D28A_ASTHMA_AD_TEXT,D28B_STILL_HAVE_ASTHMA,D28C_ASTHMA_EPISODE_12M,D28D_ASTHMA_ER_VISIT_12M,D28E_ASTHMA_MED_TAKE_12M,D28F_ASTHMA_14D_NUM_NIGHTS_TEXT,D28G_ASTHMA_14D_LIMIT_DAYS_TEXT,D28H_ASTHMA_14D_NUM_WHEEZE_TEXT,S177_SMOKE_100_CIG_LIFETIME,SMOKE_CAT,SNP1,SNP2,SNP3,SNP4,ESTTOTALPOP,DISTANCE,AADT,ROADTYPE,SPEED,THROUGH_LANES,O3_N_OBS,PM25_N_OBS,TLR4_AGE2,CURRENT_AGE2,QXAGE2,D28A_ASTHMA_AD_TEXT2,TLR4_DIST_1X_qcut,TLR4_DIST_1X_cut,TLR4_DIST_2X_qcut,TLR4_DIST_2X_cut,TLR4_DIST_3X_qcut,TLR4_DIST_3X_cut,O3_ANNUAL_AVERAGE_qcut,O3_ANNUAL_AVERAGE_cut,PM25_ANNUAL_AVERAGE_qcut,PM25_ANNUAL_AVERAGE_cut,ESTTOTALPOP25PLUS_qcut,ESTTOTALPOP25PLUS_cut,ESTPROPPERSONSNONHISPWHITE_qcut,ESTPROPPERSONSNONHISPWHITE_cut,ESTPROPPERSONS25PLUSHSMAX_qcut,ESTPROPPERSONS25PLUSHSMAX_cut,ESTPROPHOUSEHOLDSNOAUTO_qcut,ESTPROPHOUSEHOLDSNOAUTO_cut,ESTPROPPERSONSNOHEALTHINS_qcut,ESTPROPPERSONSNOHEALTHINS_cut,ESTPROPPERSONS5

PLUSNOENGLISH_qcut,ESTPROPPERSONS5PLUSNOENGLISH_cut,MEDIANHOUSEHOLDINCOME_qcut,MEDIANHOUSEHOLDINCOME_cut,DISTANCE2,IN_EPR,study_period,AvgDailyPM2.5Exposure,AvgDailyPM2.5Exposure_StudyAvg,AvgDailyPM2.5Exposure_StudyMax,MaxDailyPM2.5Exposure,MaxDailyPM2.5Exposure_StudyAvg,MaxDailyPM2.5Exposure_StudyMax,AvgDailyOzoneExposure,AvgDailyOzoneExposure_StudyAvg,AvgDailyOzoneExposure_StudyMax,MaxDailyOzoneExposure,MaxDailyOzoneExposure_StudyAvg,MaxDailyOzoneExposure_StudyMax,AvgDailyPM2.5Exposure_2,MaxDailyOzoneExposure_2,AvgDailyCOExposure_2,AvgDailyNOExposure_2,AvgDailyNO2Exposure_2,AvgDailyNOxExposure_2,AvgDailySO2Exposure_2,AvgDailyAcetaldehydeExposure_2,AvgDailyFormaldehydeExposure_2,AvgDailyBenzeneExposure_2,ObesityBMI,TotalEDVisits,TotalInpatientVisits,TotalEDInpatientVisits,Sex2,Sex,Race,Ethnicity,MajorRoadwayHighwayExposure,RoadwayDistanceExposure,RoadwayType,RoadwayAADT,RoadwaySpeedLimit,RoadwayLanes,CAFO_Exposure,Landfill_Exposure,EstResidentialDensity,EstProbabilityHighSchoolMaxEducation,EstProbabilityNoHealthIns,EstProbabilityHouseholdNonHispWhite,EstProbabilityESL,EstProbabilityNonHispWhite,EstHouseholdIncome,EstResidentialDensity25Plus,EstProbabilityNoAuto,ur,PneumoniaDx,Ciclesonide,DrugDependenceDx,Leuprolide,CervicalCancerDx,AutismDx,TesticularCancerDx,Flunisolide,Escitalopram,AlopeciaDx,Propranolol,Ipratropium,Nandrolone,Testosterone,Prasterone,Trazodone,Venlafaxine,Estrogen,OvarianCancerDx,Sertraline,EndometriosisDx,CoughDx,Fluticasone,AnxietyDx,PregnancyDx,FibromyalgiaDx,Metformin,Citalopram,Paroxetine,Cetirizine,Androstenedione,Fluoxetine,Duloxetine,ObesityDx,AlcoholDependenceDx,Formoterol,Estropipate,OvarianDysfunctionDx,Budesonide,Omalizumab,Prednisone,Hydroxyzine,Albuterol,Metaproterenol,Theophylline,AsthmaDx,TesticularDysfunctionDx,Salmeterol,Goserelin,Tamoxifen,Indacaterol,KidneyCancerDx,CroupDx,Progesterone,MenopauseDx,Estradiol,Histrelin,Mometasone,Diphenhydramine,ReactiveAirwayDx,ProstateCancerDx,DiabetesDx,Medroxyprogresterone,Mepolizumab,DepressionDx,Fexofenadine,Triptorelin,Beclomethasone,UterineCancerDx,Arformoterol,AgeStudyStart,Active_In_Study_Period,AgeStudyStart2,AvgDailyPM2.5Exposure_StudyAvg_qcut,AvgDai

lyPM2.5Exposure_StudyMax_qcut,AvgDailyPM2.5Exposure_qcut,MaxDailyPM2.5Exposure_StudyAvg_qcut,MaxDailyPM2.5Exposure_StudyMax_qcut,MaxDailyPM2.5Exposure_qcut,AvgDailyOzoneExposure_StudyAvg_qcut,AvgDailyOzoneExposure_StudyMax_qcut,AvgDailyOzoneExposure_qcut,MaxDailyOzoneExposure_StudyAvg_qcut,MaxDailyOzoneExposure_StudyMax_qcut,MaxDailyOzoneExposure_qcut,AvgDailyPM2.5Exposure_2_qcut,MaxDailyOzoneExposure_2_qcut,AvgDailyCOExposure_2_qcut,AvgDailyNOExposure_2_qcut,AvgDailyNO2Exposure_2_qcut,AvgDailyNOxExposure_2_qcut,AvgDailySO2Exposure_2_qcut,AvgDailyAcetaldehydeExposure_2_qcut,AvgDailyFormaldehydeExposure_2_qcut,AvgDailyBenzeneExposure_2_qcut,MajorRoadwayHighwayExposure2,RoadwayDistanceExposure2,IN_ICEES

"FHIR-PIT/data/output/EPR_binned/EPR_binned2010visit”

HASH_VALUE,IN_FINAL_SAMPLE,ORIGINAL_ANALYTIC_SAMPLE,TLR4_AGE,GENDER,RACE,ETHNICITY,CURRENT_AGE,RESPONDER_STATUS,QXAGE,BMI_CATEGORY,D28_ASTHMA,D28A_ASTHMA_AD_TEXT,D28B_STILL_HAVE_ASTHMA,D28C_ASTHMA_EPISODE_12M,D28D_ASTHMA_ER_VISIT_12M,D28E_ASTHMA_MED_TAKE_12M,D28F_ASTHMA_14D_NUM_NIGHTS_TEXT,D28G_ASTHMA_14D_LIMIT_DAYS_TEXT,D28H_ASTHMA_14D_NUM_WHEEZE_TEXT,S177_SMOKE_100_CIG_LIFETIME,SMOKE_CAT,SNP1,SNP2,SNP3,SNP4,ESTTOTALPOP,DISTANCE,AADT,ROADTYPE,SPEED,THROUGH_LANES,O3_N_OBS,PM25_N_OBS,TLR4_AGE2,CURRENT_AGE2,QXAGE2,D28A_ASTHMA_AD_TEXT2,TLR4_DIST_1X_qcut,TLR4_DIST_1X_cut,TLR4_DIST_2X_qcut,TLR4_DIST_2X_cut,TLR4_DIST_3X_qcut,TLR4_DIST_3X_cut,O3_ANNUAL_AVERAGE_qcut,O3_ANNUAL_AVERAGE_cut,PM25_ANNUAL_AVERAGE_qcut,PM25_ANNUAL_AVERAGE_cut,ESTTOTALPOP25PLUS_qcut,ESTTOTALPOP25PLUS_cut,ESTPROPPERSONSNONHISPWHITE_qcut,ESTPROPPERSONSNONHISPWHITE_cut,ESTPROPPERSONS25PLUSHSMAX_qcut,ESTPROPPERSONS25PLUSHSMAX_cut,ESTPROPHOUSEHOLDSNOAUTO_qcut,ESTP

ROPHOUSEHOLDSNOAUTO_cut,ESTPROPPERSONSNOHEALTHINS_qcut,ESTPROPPERSONSNOHEALTHINS_cut,ESTPROPPERSONS5PLUSNOENGLISH_qcut,ESTPROPPERSONS5PLUSNOENGLISH_cut,MEDIANHOUSEHOLDINCOME_qcut,MEDIANHOUSEHOLDINCOME_cut,DISTANCE2,patient_num,IN_EPR,AgeVisit,bucket,CAFO_Exposure,EstHouseholdIncome,EstProbabilityESL,EstProbabilityHighSchoolMaxEducation,EstProbabilityHouseholdNonHispWhite,EstProbabilityNoAuto,EstProbabilityNoHealthIns,EstProbabilityNonHispWhite,EstResidentialDensity,EstResidentialDensity25Plus,Ethnicity,Landfill_Exposure,MajorRoadwayHighwayExposure,ObesityBMIVisit,Race,RoadwayAADT,RoadwayDistanceExposure,RoadwayLanes,RoadwaySpeedLimit,RoadwayType,Sex,study_period,ur,VisitType,year,BeclomethasoneVisit,EstradiolVisit,ProstateCancerDxVisit,AlopeciaDxVisit,UterineCancerDxVisit,AutismDxVisit,DepressionDxVisit,TheophyllineVisit,FormoterolVisit,OvarianDysfunctionDxVisit,AlcoholDependenceDxVisit,MetforminVisit,ObesityDxVisit,ParoxetineVisit,DiphenhydramineVisit,PregnancyDxVisit,MepolizumabVisit,MedroxyprogresteroneVisit,LeuprolideVisit,CiclesonideVisit,ReactiveAirwayDxVisit,EndometriosisDxVisit,AnxietyDxVisit,FibromyalgiaDxVisit,FluticasoneVisit,IndacaterolVisit,CetirizineVisit,HydroxyzineVisit,FexofenadineVisit,PrednisoneVisit,TamoxifenVisit,MometasoneVisit,TrazodoneVisit,PropranololVisit,CoughDxVisit,NandroloneVisit,DrugDependenceDxVisit,ArformoterolVisit,BudesonideVisit,OvarianCancerDxVisit,CroupDxVisit,VenlafaxineVisit,MetaproterenolVisit,HistrelinVisit,OmalizumabVisit,MenopauseDxVisit,PneumoniaDxVisit,DiabetesDxVisit,DuloxetineVisit,EstropipateVisit,GoserelinVisit,CitalopramVisit,CervicalCancerDxVisit,AndrostenedioneVisit,ProgesteroneVisit,TestosteroneVisit,SertralineVisit,EscitalopramVisit,EstrogenVisit,TriptorelinVisit,SalmeterolVisit,IpratropiumVisit,KidneyCancerDxVisit,TesticularCancerDxVisit,AsthmaDxVisit,FlunisolideVisit,PrasteroneVisit,TesticularDysfunctionDxVisit,FluoxetineVisit,AlbuterolVisit,Sex2,AgeVisit2,MajorRoadwayHighwayExposure2,RoadwayDistanceExposure2,IN_ICEES,index

"FHIR-PIT/data/output/EPR_binned/EPR_binned2010visit_deidentified”

N_FINAL_SAMPLE,ORIGINAL_ANALYTIC_SAMPLE,TLR4_AGE,GENDER,RACE,ETHNICITY,CURRENT_AGE,RESPONDER_STATUS,QXAGE,BMI_CATEGORY,D28_ASTHMA,D28A_ASTHMA_AD_TEXT,D28B_STILL_HAVE_ASTHMA,D28C_ASTHMA_EPISODE_12M,D28D_ASTHMA_ER_VISIT_12M,D28E_ASTHMA_MED_TAKE_12M,D28F_ASTHMA_14D_NUM_NIGHTS_TEXT,D28G_ASTHMA_14D_LIMIT_DAYS_TEXT,D28H_ASTHMA_14D_NUM_WHEEZE_TEXT,S177_SMOKE_100_CIG_LIFETIME,SMOKE_CAT,SNP1,SNP2,SNP3,SNP4,ESTTOTALPOP,DISTANCE,AADT,ROADTYPE,SPEED,THROUGH_LANES,O3_N_OBS,PM25_N_OBS,TLR4_AGE2,CURRENT_AGE2,QXAGE2,D28A_ASTHMA_AD_TEXT2,TLR4_DIST_1X_qcut,TLR4_DIST_1X_cut,TLR4_DIST_2X_qcut,TLR4_DIST_2X_cut,TLR4_DIST_3X_qcut,TLR4_DIST_3X_cut,O3_ANNUAL_AVERAGE_qcut,O3_ANNUAL_AVERAGE_cut,PM25_ANNUAL_AVERAGE_qcut,PM25_ANNUAL_AVERAGE_cut,ESTTOTALPOP25PLUS_qcut,ESTTOTALPOP25PLUS_cut,ESTPROPPERSONSNONHISPWHITE_qcut,ESTPROPPERSONSNONHISPWHITE_cut,ESTPROPPERSONS25PLUSHSMAX_qcut,ESTPROPPERSONS25PLUSHSMAX_cut,ESTPROPHOUSEHOLDSNOAUTO_qcut,ESTPROPHOUSEHOLDSNOAUTO_cut,ESTPROPPERSONSNOHEALTHINS_qcut,ESTPROPPERSONSNOHEALTHINS_cut,ESTPROPPERSONS5PLUSNOENGLISH_qcut,ESTPROPPERSONS5PLUSNOENGLISH_cut,MEDIANHOUSEHOLDINCOME_qcut,MEDIANHOUSEHOLDINCOME_cut,DISTANCE2,IN_EPR,AgeVisit,bucket,CAFO_Exposure,EstHouseholdIncome,EstProbabilityESL,EstProbabilityHighSchoolMaxEducation,EstProbabilityHouseholdNonHispWhite,EstProbabilityNoAuto,EstProbabilityNoHealthIns,EstProbabilityNonHispWhite,EstResidentialDensity,EstResidentialDensity25Plus,Ethnicity,Landfill_Exposure,MajorRoadwayHighwayExposure,ObesityBMIVisit,Race,RoadwayAADT,RoadwayDistanceExposure,RoadwayLanes,RoadwaySpeedLimit,RoadwayType,Sex,study_period,ur,VisitType,year,BeclomethasoneVisit,EstradiolVisit,ProstateCancerDxVisit,AlopeciaDxVisit,UterineCancerDxVisit,AutismDxVisit,DepressionDxVisit,TheophyllineVisit,FormoterolVisit,OvarianDysfunctionDxVisit,AlcoholDependenceDxVisit,MetforminVisit,ObesityDxVisit,ParoxetineVisit,DiphenhydramineVisit,Pregn

ancyDxVisit,MepolizumabVisit,MedroxyprogresteroneVisit,LeuprolideVisit,CiclesonideVisit,ReactiveAirwayDxVisit,EndometriosisDxVisit,AnxietyDxVisit,FibromyalgiaDxVisit,FluticasoneVisit,IndacaterolVisit,CetirizineVisit,HydroxyzineVisit,FexofenadineVisit,PrednisoneVisit,TamoxifenVisit,MometasoneVisit,TrazodoneVisit,PropranololVisit,CoughDxVisit,NandroloneVisit,DrugDependenceDxVisit,ArformoterolVisit,BudesonideVisit,OvarianCancerDxVisit,CroupDxVisit,VenlafaxineVisit,MetaproterenolVisit,HistrelinVisit,OmalizumabVisit,MenopauseDxVisit,PneumoniaDxVisit,DiabetesDxVisit,DuloxetineVisit,EstropipateVisit,GoserelinVisit,CitalopramVisit,CervicalCancerDxVisit,AndrostenedioneVisit,ProgesteroneVisit,TestosteroneVisit,SertralineVisit,EscitalopramVisit,EstrogenVisit,TriptorelinVisit,SalmeterolVisit,IpratropiumVisit,KidneyCancerDxVisit,TesticularCancerDxVisit,AsthmaDxVisit,FlunisolideVisit,PrasteroneVisit,TesticularDysfunctionDxVisit,FluoxetineVisit,AlbuterolVisit,Sex2,AgeVisit2,MajorRoadwayHighwayExposure2,RoadwayDistanceExposure2,IN_ICEES,index

"FHIR-PIT/data/output/EPR_binned/EPR_binned_deidentified”

IN_FINAL_SAMPLE,ORIGINAL_ANALYTIC_SAMPLE,TLR4_AGE,GENDER,RACE,ETHNICITY,CURRENT_AGE,RESPONDER_STATUS,QXAGE,BMI_CATEGORY,D28_ASTHMA,D28A_ASTHMA_AD_TEXT,D28B_STILL_HAVE_ASTHMA,D28C_ASTHMA_EPISODE_12M,D28D_ASTHMA_ER_VISIT_12M,D28E_ASTHMA_MED_TAKE_12M,D28F_ASTHMA_14D_NUM_NIGHTS_TEXT,D28G_ASTHMA_14D_LIMIT_DAYS_TEXT,D28H_ASTHMA_14D_NUM_WHEEZE_TEXT,S177_SMOKE_100_CIG_LIFETIME,SMOKE_CAT,SNP1,SNP2,SNP3,SNP4,ESTTOTALPOP,DISTANCE,AADT,ROADTYPE,SPEED,THROUGH_LANES,O3_N_OBS,PM25_N_OBS,TLR4_AGE2,CURRENT_AGE2,QXAGE2,D28A_ASTHMA_AD_TEXT2,TLR4_DIST_1X_qcut,TLR4_DIST_1X_cut,TLR4_DIST_2X_qcut,TLR4_DIST_2X_cut,TLR4_DIST_3X_qcut,TLR4_DIST_3X_cut,O3_ANNUAL_AVERAGE_qcut,O3_ANNUAL_AVERAGE_cut,PM25_ANNUAL_AVERAGE_qcut,PM25_ANNUAL_AVERAGE_cut,ESTTOTALPOP25PLUS_qcut,ESTTOTALPOP25PLUS_cut,ESTPROPPERSONSNONHISPWHI

TE_qcut,ESTPROPPERSONSNONHISPWHITE_cut,ESTPROPPERSONS25PLUSHSMAX_qcut,ESTPROPPERSONS25PLUSHSMAX_cut,ESTPROPHOUSEHOLDSNOAUTO_qcut,ESTPROPHOUSEHOLDSNOAUTO_cut,ESTPROPPERSONSNOHEALTHINS_qcut,ESTPROPPERSONSNOHEALTHINS_cut,ESTPROPPERSONS5PLUSNOENGLISH_qcut,ESTPROPPERSONS5PLUSNOENGLISH_cut,MEDIANHOUSEHOLDINCOME_qcut,MEDIANHOUSEHOLDINCOME_cut,DISTANCE2

YAML configuration file: icees_features.yaml

The Mapper object consumes this file for various purposes:

1. To label multiple FHIR concept codes with the same feature name. This is important for users to know code to feature correspondence of integrated tables.
2. To relabel data source features for the integrated feature tables: ACS, UR, HPMS, TL, Cafo, Landfill. This is important for users to have easier to understand feature names.
3. To aid a use case. For the asthma use case, it list the diagnosis of interest to count as “Respiratory” visits at ED and inpatient locations. This is important for asthma researchers.

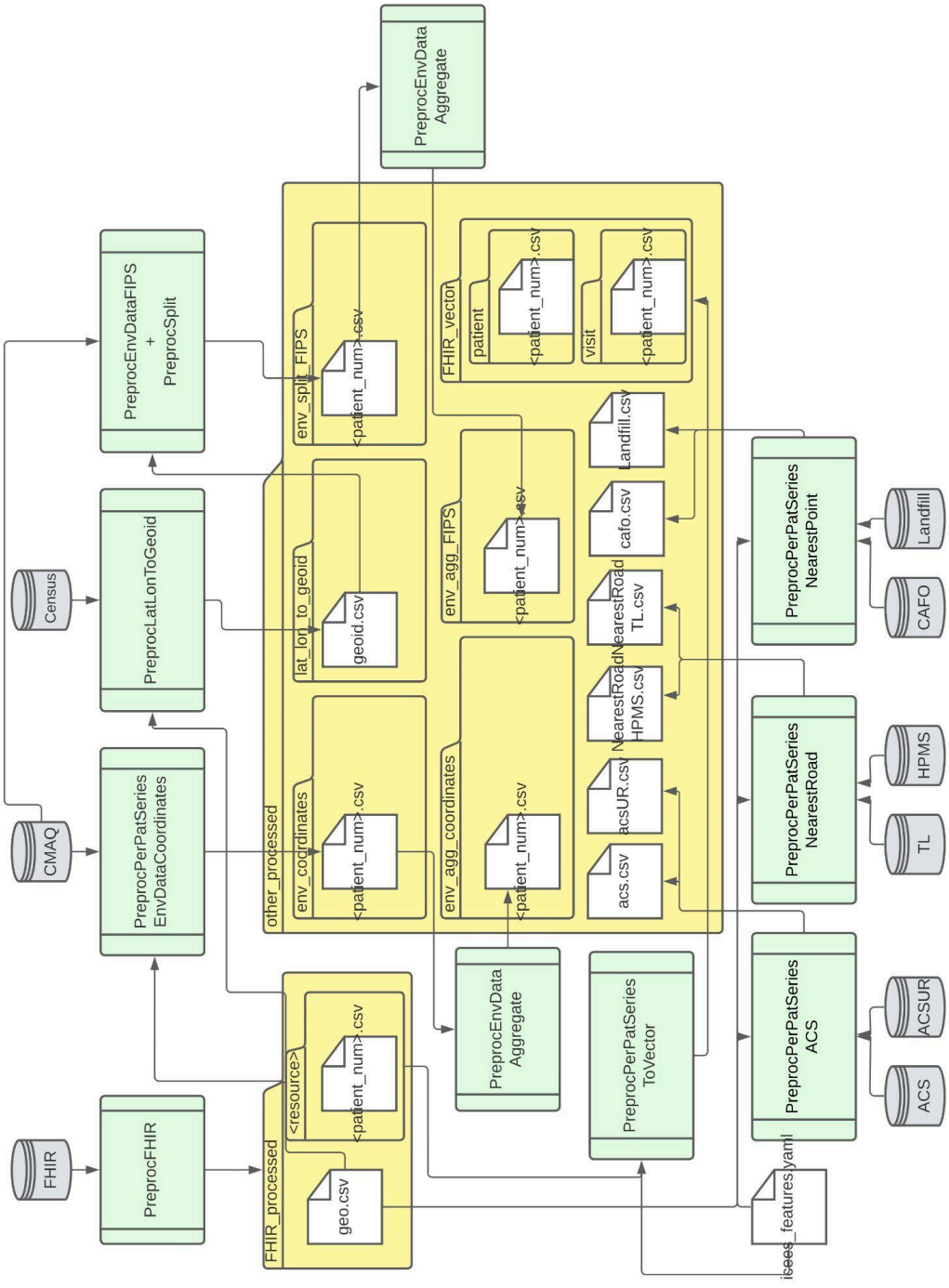

Figure 1: FHIR-PIT integrated feature tables pipeline. Arrows indicate dependency. Data sources are in gray, transformations in green, directories in yellow, and integrated feature tables in white. Please see the ‘Step description’ section for further details on the transformations.

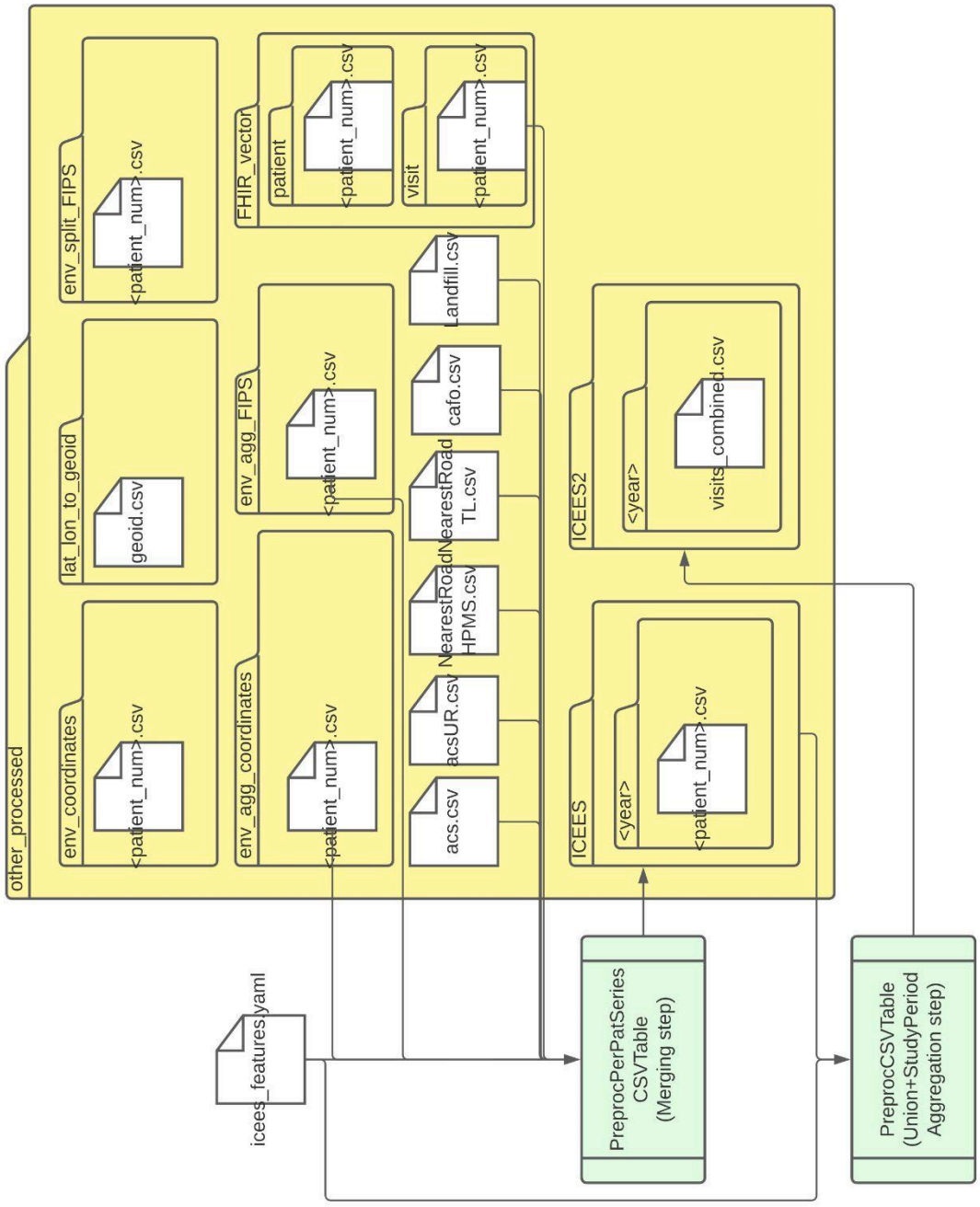

Figure 2: FHIR-PIT integrated feature tables merging and aggregation pipelines. Arrows indicate dependency. These transformations produce the final output to bin and deidentify. Data sources are in gray, transformations in green, directories in yellow, and integrated feature tables in white. Please see the ‘Step description’ section for further details on the transformations.