About Raw Data

There are three CSV files which will be the input data. The data provided in events.csv are event sequences. Each line of this file consists of a tuple with the format (patient_id, event_id, event_description, timestamp, value).

- Patient id: De-identified patient identifiers.
- Event id: Clinical event identifiers. For example, DRUG19122121 means that a drug with RxNorm code as 19122121 was prescribed to the patient. DIAG319049 means the patient was diagnosed of disease with SNOMED code of 319049 and LAB3026361 means that the laboratory test with a LOINC code of 3026361 was conducted on the patient.
- Event description: Shows the description of the clinical event. For example, DIAG319049 is the code for acute respiratory failure and DRUG19122121 is the code for Insulin.
- Timestamp: the date at which the event happened. (Here the timestamp is not a real date.)
- Value: Contains the value associated to an event. See Table 1 for the detailed description.

event type	sample event_id	value meaning	example
diagnostic code	DIAG319049	diagnosed with a certain disease, value always be 1.0	1.0
drug consumption	DRUG19122121	prescribed a certain medication, value will always be 1.0	1.0
laboratory test	LAB3026361	test conducted on a patient and its value	3.690

Table 1: Event sequence value explanation

The data provided in mortality events.csv contains the patient ids of only the deceased people. They are in the form of a tuple with the format (patient id, timestamp, label).

The timestamp indicates the death date of a deceased person and a label of 1 indicates death. Patients that are not mentioned in this _le are considered alive.

The event feature map.csv is a map from an event id (SNOMED, LOINC and RxNorm) to an integer index. This file contains (idx, event id) pairs for all event ids.

Folder Structure:

```
|-- data
        |-- test
                |-- event_feature_map.csv
                |-- events.csv
        |-- train
                |-- event_feature_map.csv
                |-- events.csv
                |-- mortality_events.csv
        |-- features_svmlight.validate
|-- src
        |-- event_statistics.py
        |-- etl.py
        |-- models.py
        |-- model_main.py
        |-- cross.py
        |-- utils.py
|-- deliverables
        |-- etl_index_dates.csv
        |-- etl_filtered_events.csv
        |-- etl_aggregated_events.csv
        |-- features_svmlight.train
        |-- features.train
|-- environment.yml
|-- readme.pdf
```

All files have to be in the corresponding folder specify above. All code are in src folder and run with python 3.6.5 with environment specified in environment.yml which need to activate with docker.

event_statistics.py: compute raw data statistics
etl.py: preprocess data, construct features, and saved in SVMLight format
model.py: model implementation
model_main.py: run models to get performance metrics
cross.py: cross-validation with 5 folds
utils.py: helper functions