Data sets are separated by admission year, as ***ft\_zip\_$year.txt***, where **$*year*** ranges from 2010 to 2018

**Data Format:** the first two columns, *PATID, ENCOUNTERID* are primary keys identifying patient and encounter. Each row contains all information for one *encounter* (one patient may have multiple encounters) arranged in the following format:

<demo>|<vital>|<lab>|<DRG>|<comorbidity>|<icd>|<procedure>|<med>|<AKI\_label>

|  |  |  |
| --- | --- | --- |
| **Feature Group** | **Size** | **Format** |
| Demographics  <demo> | 4 | Age\_Hispanic\_Race\_Sex |
| Vital signs  <vital> | 7 | Height\_Weight\_BMI\_SBP\_DBP\_Smoking\_Tabacco\_TabaccoType  **for each vital with multiple records:**  val1,d1;val2,d2;val3,d3;…  d1,d2,… can be both positive or negative (>=-7) |
| Lab  <lab> | 812 | Lab1\_Lab2\_Lab3\_...  **for each lab with multiple records:**  labIndex1:val1,unit1,d1;labIndex2:val2,unit2,d2;…  d1,d2,… are expected to be positive |
| UHC DRG  <DRG> | 716 | DX1\_DX2\_DX3\_...  **Each encounter only has one admission DRG:**  dxIndex:0 |
| UHC Comorbidity  <comorbidity> | 706 | DX1\_DX2\_DX3\_...  **for each UHC historical diagnosis with multiple records:**  dxIndex1:d1,d2,…;dxIndex2:d1,d2,…;…  d1,d2,… are expected to be negative |
| ICD9 grouped with CCS  <icd> | 280 | CCS1\_CCS2\_CCS3\_...  **for each historical CCS diagnosis with multiple records:**  ccsIndex1:d1,d2,…;ccsIndex2:d1,d2,…;…  d1,d2,… are expected to be negative |
| Procedures  <procedure> | 12075 | PX1\_PX2\_PX3\_...  **for each procedure with multiple records:**  ccsIndex1:d1,d2,…;ccsIndex2:d1,d2,…;…  d1,d2,… can be both negative or positive |
| Medication  <med> | 11167 | RX1\_RX2\_RX3\_...  **for each medication with multiple records:**  rxIndex1:q1,d;q2,d2;q3,d3,…;rxIndex2:q1,d1;q2,d2;…;…  d1,d2,… are expected to be positive |
| AKI\_stage  <AKI\_label> | 3 | Label1\_Label2\_Label3  **Each label should only associated with one date and only valid discovered labels are reported:**  labelIndex1,d1; labelIndex2,d2; …  d1,d2,… are expected to be positive |

\*\*\*\*\*NOTE\*\*\*\*\*

For numerical features, such as lab and vitals, it is the original numerical values that are reported (discretization not applied). If a variable for <demo> or <vital> doesn’t exist, or no variables for the other feature groups exists, then a “0” is assigned.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Data Dictionary is saved as **feature.dict.csv**

|  |  |
| --- | --- |
| **Column Names** | **Description** |
| **VAR\_IDX** | **Variable Index** |
| VAR\_POS | Position of variable within each fact string |
| TABLE\_NAME | Original CDM table |
| **VAR\_CD** | **Encoded variable** |
| **VAR\_NAME** | **Human-readable variable label** |
| VAR\_TYPE | Variable type |

\*\*\*\*\*NOTE\*\*\*\*\*

VAR\_IDX are used in data table, which can be interpreted by mapping to VAR\_NAME. VAR\_CD are quoted in order to reserve leading “0”

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*