# Data Retrieval Criteria

Please remember to acknowledge the Clinical and Translational Science Institute (CTSI) in all publications that received services through R3. The CTSI provides the following citation for your convenience: **The project described was supported by the National Institutes of Health through Grant Number UL1 TR001857**

Start date for population definition: 1/1/2011

End date for population definition: 1/31/2021

Minimum age of subjects at time of encounter: N/A

Maximum age of subjects at time of encounter: N/A

Encounters: ED

Any of the following Diagnoses:

* Influenza
  + ICD9 = ('487', '487.0', '487.1', '487.8')
  + ICD10 = ('J09', 'J09.X', 'J09.X1', 'J09.X2', 'J09.X3', 'J09.X9', 'J10', 'J10.0', 'J10.00', 'J10.01', 'J10.08', 'J10.1', 'J10.2', 'J10.8', 'J10.81', 'J10.82', 'J10.83', 'J10.89','J11', 'J11.0', 'J11.00', 'J11.08', 'J11.1', 'J11.2', 'J11.8', 'J11.81', 'J11.82', 'J11.83', 'J11.89')
* Parainfluenza
  + ICD9 = 480.2
  + ICD10 = J12.2
* RSV
  + ICD9 = ('466.11', '079.6')
  + ICD10 = ('B97.4, J21.0')
* hMPV
  + ICD10 = J12.3
* COV19
  + ICD10 = ('U07.1', 'Z11.52', 'Z20.822', 'J12.82')

**OR**

**Labs performed on same date as ED visit:** COMPONENT\_ID = ('12682', '13392', '13393', '14893', '16921', '6175', 'METAPN', '10131', '10132', '10448', '10566', '10719', '10720', '10721', '10722', '10931', '11081', '11082', '11083', '12883', '1326', '1327', '13395', '13396', '13399', '13400', '13402', '13403', '1354', '1355', '14361', '14362', '1438', '1517', '15696', '15697', '15893', '16345', '16453', '16599', '16609', '19691', '19711', '3820', '3821', '4094', '4095', '4096', '4108', '4109', '4937', '4938', '6803', '6804', '7419', '7420', '7947', '7948', '7957', '7958', '9435', '9436', '9437', '9552', '9553', '9563', '9626', '9627', 'CIAQ', 'CIBQ', 'FLBAGQ', 'FLUAQ', 'FLUAVQ', 'FLUBVQ', 'FLUCAQ', 'FLUCBQ', 'IAQ', 'IARPQ', 'IBQ', 'IBRPQ', 'INAA', 'INBA', 'INFA', 'INFAH1', 'INFAH3', 'INFB', 'INFH1Q', 'INFLUA', 'INFLUB', 'INRNAQ', 'ITAC', 'ITBC', 'SOR17Q', '14320', '17704', '20048', '3600', '3601', '3602', '7949', '7950', '7951', 'PAIF1Q', 'PAIF2Q', 'PAIF3Q', 'PARIN1', 'PARIN2', 'PARIN3', '10635', '12541', '12542', '13046', '13253', '13298', '13393', '13394', '13397', '13437', '13444', '13556', '14369', '17236', '17237', '17238', '19697', '19752', '7254', '7255', '7262', '7656', '7946', '9229', '9340', '9960', 'ROTA1Q', 'RSPCRQ', 'RSV15', 'RSV224', 'RSVABQ', 'RSVRA', 'RSVRB', '13051', '17235', '17236', '17237', '17238', '14515', 'FLUBQ', 'FLUBVQ', 'INFLUACONFIRM', '20126','202303','20240','20241','20242','20243','20244','20245','20264','20274','20275','20276','20277','20278','20279','20280','20283','20285','20286','20287','20288','20289','20290','20291','20292','20293','20298','20300','20301','20303','20308','20346','20347','20364','20365','20369','20379','20380','20381','20389','20390','20391','20392','20460','20463','20464','20465','20466','20469','20640','20647','20652','20655','20670')

Deceased: Included

# Research Datasets

1. R3\_1971\_YE\_ENCOUNTER\_NOTES\_2021\_03\_01.zip

* R3\_1971\_YE\_DEMOGRAPHICS\_2021\_03\_01
* R3\_1971\_YE\_ENCOUNTERS\_2021\_03\_01
* R3\_1971\_YE\_DIAGNOSES\_2021\_03\_01
* R3\_1971\_YE\_PROBLEM\_LIST\_2021\_03\_01
* R3\_1971\_YE\_LAB\_RESULTS\_2021\_03\_01
* R3\_1971\_YE\_ED\_DATES\_2021\_03\_01
* R3\_1971\_YE\_PROCEDURES\_2021\_03\_01
* R3\_1971\_YE\_PROCEDURE\_NOTES\_2021\_03\_01
* R3\_1971\_YE\_ENCOUNTER\_NOTES\_2021\_03\_01

1. R3\_1971\_YE\_NOTES\_JANUARY\_2021\_03\_05.zip

* R3\_1971\_YE\_ENCOUNTER\_NOTES\_2021\_03\_05
* R3\_1971\_YE\_PROCEDURE\_NOTES\_2021\_03\_05

*Table 1: Patient Demographics*

File Name: R3\_1971\_YE\_DEMOGRAPHICS\_2021\_03\_01

Columns (one row per patient):

Patient\_Study\_ID

Age at admission

Gender

Death\_Date

Zip Code (3 digit)

*Table 2: Encounters (all encounters for seven (7) days following the initial ED encounter)*

File Name: R3\_1971\_YE\_ENCOUNTERS\_2021\_03\_01

Columns (one row per encounter):

Patient\_Study\_ID

Visit\_Start\_Date (date only)

Visit\_End\_Date (date only)

Enc\_Type

Location

Appointment\_Status

Admit\_Source

Hospital\_Service

Patient\_Type

Patient\_Class

Chief\_Complaint

Chief\_Complaint\_Onset\_Date

*Table 3: Diagnoses (all Dx identified in cohort criteria for seven (7) days following the initial ED encounter)*

File Name: R3\_1971\_YE\_DIAGNOSES\_2021\_03\_01

Columns (one row per diagnosis):

Patient\_Study\_ID

Diagnosis\_Type

Diagnosis\_Code

Diagnosis\_Name

Primary\_Diagnosis\_YN

Diagnosis\_From\_Date

Diagnosis\_To\_Date

*Table 4: Problem List Diagnoses (all Dx identified in cohort criteria for seven (7) days following the initial ED encounter)*

File Name: R3\_1971\_YE\_PROBLEM\_LIST\_2021\_03\_01

Columns (one row per diagnosis):

Patient\_Study\_ID

Diagnosis\_Type

Diagnosis\_Code

Diagnosis\_Name

Reported\_Date

Onset\_Date

Resolved\_Date

Resolved\_Reason

Problem\_Status

*Table 5: Lab Results (all Labs identified in cohort criteria for seven (7) days following the initial ED encounter)*

File Name: R3\_1971\_YE\_LAB\_RESULTS\_2021\_03\_01

Columns (one row per result):

Patient\_Study\_ID

Component\_ID

LOINC code

Order\_Procedure\_ID

Result\_Date

Component\_Name

Ord\_Value

Ord\_Num\_Value

Reference\_Unit

Reference\_Low

Reference\_High

Result\_Flag\_Title

Lab\_Result\_Status

Specimen\_Collected\_Date

Specimen\_Received\_Date

Specimen\_Type

Specimen\_Source

*Table 6: Notes (All Notes for the cohort within 7 days of ED encounter dates)*

Notes to be delivered in text file format, and must be able to be linked with patients who potentially have multiple ED encounters

1. R3\_1971\_YE\_ED\_DATES\_2021\_03\_01
2. R3\_1971\_YE\_PROCEDURES\_2021\_03\_01
3. R3\_1971\_YE\_PROCEDURE\_NOTES\_2021\_03\_01
4. R3\_1971\_YE\_ENCOUNTER\_NOTES\_2021\_03\_01
5. R3\_1971\_YE\_ENCOUNTER\_NOTES\_2021\_03\_05
6. R3\_1971\_YE\_PROCEDURE\_NOTES\_2021\_03\_05

# Important Folders

C:\Users\yey5\Dropbox\000KGrant\000.data\new-neptune-data-retrieval\ most\_important\

* + R31971\_Ye\_Specification\_2021\_03\_01.docx
  + COMPONENTS.csv

C:\Users\yey5\Documents\00.K99\_data\R3\_1971\_YE\_ENCOUNTER\_NOTES\_2021\_03\_01\

C:\Users\yey5\Documents\00.K99\_data\R3\_1971\_YE\_NOTES\_JANUARY\_2021\_03\_05\

# Initial Analysis of Individual Tables

## Table 1. Patient Demographics

File Name: R3\_1971\_YE\_DEMOGRAPHICS\_2021\_03\_01

Data elements: Patient\_Study\_ID, Age at admission, Gender, Death\_Date, Zip Code (3 digit)

Comments on table 1:

* one row per patient

Code name and location (jupyter notebook/python files/sql query shared to Ye only through Microsoft overdrive; Keep all intermediate files):

Provide following information (with comments in the code for each question):

* + 1. Number of patients:there are 148369 records, and 117672 patients(117672 unique STUDY\_IDs)

For missing value:

In this Demographics data, only ***Death\_Data*** and **ZIP\_code** get missing values

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Number of NaN** | **proportion** |
| DEATH\_DATE | 130858 | 88% |
| ZIP\_CODE | 82 | 0.055% |

* + 1. Age distribution, percentage of missing data

For age distribution:

Age attribute does not get a missing value

The minimum of age: 0, maximum of age: 119

In this data set, newborn (age:0) gets the largest proportion (7.5%), and age 105 gets the smallest proportion(only one record in dataset)

Chart, histogram

Description automatically generated

Furthermore, exploring age distribution by different gender, in the early age (0-12), there are more male records than female’s, after then, (13-119), female recodes are more than male records.

Chart

Description automatically generated

* + 1. Gender distribution, percentage of missing data

There is no missing value in Gender.

There are three unique values in Gender attribute----MALE, FEMALE, UNKNOWN

|  |  |  |
| --- | --- | --- |
| Gender | number | proportion |
| Female | 79066 | 53.2% |
| Male | 69302 | 46.7% |
| Unknown | 1 | 0.0007% |

* + 1. Number of death:

For DEATH\_DATE, there are 130858 rows of missing value----missing rate 88%

17511 rows get death record----Death record rate is 11.8%

* + - 1. Number of death within 7 days of ED visit (DEATH\_DATE-ED\_DATE=7)

Here, I found that once a patient dies, all previous records on that patient are updated.

Thus, a person may have several death records.

In this dataset, to analyze deaths within 7 days, we only need to use the latest `ED\_date` of that patient.

Table

Description automatically generated

There are 2,094 patients died within 7 days

Chart, histogram

Description automatically generated

In 2020-12-09, there was 23 persons died within 7 days after his/her ED\_DATE, which is the maximal value from 2011 to 2021

* + - 1. Number of death within 30 days of ED visit (DEATH\_DATE-ED\_DATE=30)

There are 4,920 patients died within 30 days

Chart, histogram

Description automatically generated

Same as the situation in within\_7 case, In 2020-12-09, there was 41 persons died within 30 days after his/her ED\_DATE, which is the maximal value from 2011 to 2021

Comparing the age distribution of died within 7 and 30 days:

In general, the data distributions of died within\_7 days and within\_30 days are quite similar—the peak appear in the range of age 70-90.

However, in detail, more men than women aged 60-80 died within 30 days. Combined with the fact that more senior females are recorded at this age range, this graph does suggest that males are more vulnerable than females at this age range.

Evaluating the death case along time series, the death cases within 7 days and 30 days reach their peaks in the period of the outbreak of Covid-19

* + 1. zip code distribution

There are 82 rows missing values in ZIP\_CODE (0.055%)

There are 560 different ZIP\_CODEs

25% of data coming from ZIP\_CODE of 152

13% of data coming from ZIP\_CODE of 151

12% of data coming from ZIP\_CODE of 166

10% of data coming from ZIP\_CODE of 161

9% of data coming from ZIP\_CODE of 150

Chart, pie chart

Description automatically generated

Exploring the death situation of specific region(ZIP\_CODE), the proportion of deaths corresponds to the proportion recorded in each region (ZIP\_CODE)

More patient recodes in ZIP\_CODE 152, more mortality in this area.in this region

Chart, pie chart

Description automatically generated

## 1.2 Table 2: Encounters (all encounters for seven (7) days following the initial ED encounter)

File Name: R3\_1971\_YE\_ENCOUNTERS\_2021\_03\_01

Data elements:

Patient\_Study\_ID, Visit\_Start\_Date (date only), Visit\_End\_Date (date only), Enc\_Type, Location

Appointment\_Status, Admit\_Source, Hospital\_Service, Patient\_Type, Patient\_Class, Chief\_Complaint, Chief\_Complaint\_Onset\_Date

Comments on table R3\_1971\_YE\_ENCOUNTERS\_2021\_03\_01:

* one row per encounter (some rows are not encounters, they are just lab results)
* Need to order by study\_id, visit\_start\_date, and ENC\_TYPE
* Number of ED encounters (ENC\_TYPE=EMERGENCY)
* Number of hospital encounters (ENC\_TYPE=HOSPITAL-ENCOUNTER):

Results:

* + 1. Summative analysis for each of elements
       1. number of patients in this table (unique study\_id):

there is no Nan in STUDY\_ID

the number of records: 1995731

the number of patients: 117672

how many records each patient has(on average): 16.960117954993542

* + - 1. START\_DATE: date range

There is no Nan in START\_DATE

The earliest day of START\_DATE ---- 2011, 1, 1

The latest day of START\_DATE ---- 2021, 1, 31

Duration ---- 3683 days

* + - 1. END\_DATE: date range

There are 1828933 Nan in END\_DATE----- 91.64% of the whole dataset

The earliest day of END\_DATE ---- 2011, 1, 1

The latest day of END\_DATE ---- 2021, 2, 16

Duration ---- 3699 days

Duration

41.87% of all the patients start and end their records in the same day (END\_DATE- START\_DATE=0)

92% of duration of records lasted less than 10 days

The largest duration is 292 days, and it takes 0.0006% of the whole records in this table

* + - 1. ENC\_TYPE: categories and frequency of each category, such as Number of ED encounters (ENC\_TYPE=EMERGENCY) and Number of hospital encounters (ENC\_TYPE=HOSPITAL-ENCOUNTER)

the number of Nan in ENC\_TYPE column: 3

the number of unique ENC\_TYPE: 150

the percentage of each value of ENC\_TYPE(descent):

LAB RESULTS 20.357433

IMAGING 10.544172

ER REPORT 7.157739

HOSPITAL-ENCOUNTER 5.684242

IP CONSULT 5.527557

...

PACT TELEPHONE 0.000050

MOTHER BABY LINK 0.000050

CRNP 0.000050

DIABETES EDUCATION 0.000050

EMAIL 0.000050

The percentage of EMERGENCY(%): 4.8826794032052465

The percentage of HOSPITAL-ENCOUNTER(%): 5.684241539929289

* + - 1. LOCATION

the number of NaN in location columns: 6675

the percentage of Nan in Location column: 0.334463913222774

the number of unique locations: 2959

the percentage of each location(%):

EXTERNAL DEPARTMENT 55.556455

CHILDREN'S HOSPITAL 3.769527

CHILDRENS HOSPITAL 1.806234

UPMC INPATIENT 1.184230

ALTOONA 1.163567

...

UPA NEPHROLOGY JHNSTWN 0.000050

UPA PULM PHONE CONSULT 0.000050

HEM/ONC SHADYSIDE INPT 0.000050

SIVA KEDARNATH BELLVUE 0.000050

CRS ERIE WEST 26 OT 0.000050

* + - 1. APPT\_STATUS

the number of Nan in APPT\_STATUS column: 1904268

the percentage of Nan in APPT\_STATUS column: 95.41706773107197

the number of unique APPT\_STATUS: 8

the percentage of each APPT\_STATUS(%):

COMPLETED 46.194636

CANCELED 39.919968

NO SHOW 11.125810

UNRESOLVED 2.326624

ADMITTED TO HOSPITAL 0.300668

LEFT WITHOUT SEEN 0.090747

ARRIVED 0.039360

SCHEDULED 0.002187

* + - 1. ADMIT\_SOURCE

the number of Nan in ADMIT\_SOURCE column: 1812929

the percentage of Nan in ADMIT\_SOURCE column: 90.84034872435213

the number of unique ADMIT\_SOURCE: 19

the percentage of each ADMIT\_SOURCE(%):

MED STAFF REFERRAL 170669

HOSPITAL TRANSFER 5992

TRANS FRM SNF/ICF 3025

7 939

XFER OTH HLTH CR FAC 766

NON-STAFF REFERRAL 503

SPECIMEN 343

D 204

8 163

B 102

PGH CANCER INSTITUTE 45

CLINIC REFERRAL 36

BABYBORN OUTSIDE HOS 4

SICK BABY 2

U 2

H 2

EMERGENCY DEPARTMENT 2

COURT/LAW ENFORCE 2

P 1

* + - 1. HOSPITAL\_SERVICE

the number of Nan in HOSPITAL\_SERVICE column: 1812929

the percentage of Nan in HOSPITAL\_SERVICE column: 90.84034872435213

the number of unique HOSPITAL\_SERVICE: 58

the percentage of each HOSPITAL\_SERVICE (%):

EMERGENCY MEDICINE 41.160381

GENERAL MEDICINE 13.789236

INTERNAL MEDICINE 13.297448

FAMILY MEDICINE 8.830319

UNS 4.657498

PEDIATRICS 4.652575

PULMONARY DISEASE 2.259275

CRITICAL CARE MED 1.919016

HEMATOLOGY/ONCOLOGY 1.749434

CARDIOLOGY 0.977560

GERIATRIC MEDICINE 0.714434

PSY 0.676688

OBSTETRICS & GYN 0.637302

GENERAL SURGERY 0.605026

NEUROLOGY 0.464437

GASTROENTEROLOGY 0.462249

ORTHOPEDIC SURGERY 0.284461

TRANSPLANT 0.283914

NEUROSURGERY 0.272973

NEO 0.236868

CARDIOTHORACIC SURG 0.182711

ENDOCRINOLOGY 0.181617

OTORHINOLARYNGOLOGY 0.145513

PHYSICAL MED & REHAB 0.134572

NEPHROLOGY 0.124725

URO 0.115972

SURGICAL ONCOLOGY 0.096826

GYNECOLOGY 0.094638

INFECTIOUS DISEASE 0.091356

PATHOLOGY 0.076586

OPHTHALMOLOGY 0.074397

VAS 0.072209

PEDIATRIC SURGERY 0.067833

PSYCHIATRY 0.061268

RENAL ELECTROLYTE 0.050875

THORACIC SURGERY 0.050328

RHEUMATOLOGY 0.049781

ANESTHESIOLOGY 0.041028

PLASTIC SURGERY 0.035558

RADIOLOGY 0.033916

DENTAL MEDICINE 0.033369

TRAUMATIC SURGERY 0.028993

ALLERGY & IMMUNOLOGY 0.026258

COLON/RECTAL SURGERY 0.024617

TOXICOLOGY 0.024070

CARDIAC SURGERY 0.024070

PODIATRY 0.023523

RADIATION ONCOLOGY 0.021882

MIDWIFE 0.015317

FHC 0.014223

DERMATOLOGY 0.013129

GENETICS 0.012582

ALLERGY 0.012035

CLINICAL PSYCHOLOGY 0.003829

ORAL SURGERY 0.003829

OBT 0.002735

BREAST SURGERY 0.002188

NUCLEAR MEDICINE 0.000547

* + - 1. PATIENT\_TYPE

the number of Nan in PATIENT\_TYPE column: 1812978

the percentage of Nan in PATIENT\_TYPE column: 90.84280396506342

the number of unique PATIENT\_TYPE: 45

which Patient——type get the largest proportion(%): EMERGENCY DEPT : 53.30692245818126

which Patient——type get the smallest proportion(%): CR : 0.0005471866398910004

* + - 1. PATIENT\_CLASS

the number of Nan in PATIENT\_CLASS column: 1812929

the percentage of Nan in PATIENT\_CLASS column: 90.84034872435213

the number of unique PATIENT\_CLASS: 5

which PATIENT\_CLASS get the largest proportion(%): Emergency Department : 53.30630955897638

which PATIENT\_CLASS get the smallest proportion(%): I : 0.10174943381363442

Analyzing data distribution of PATIENT\_CLASS with duration. When duration is small(duration = 0 or 1), most of records coming from **Emergency Department,** and the records coming from **Inpatient** dominates the table whenduration>=2

* + - 1. CHIEF\_COMPLAINT

the number of Nan in CHIEF\_COMPLAINT column: 1814281

the percentage of Nan in CHIEF\_COMPLAINT column: 90.90809332520264

the number of unique CHIEF\_COMPLAINT: 74099

which CHIEF\_COMPLAINT get the largest proportion(%): FEVER : 5.425186001653348

which CHIEF\_COMPLAINT get the smallest proportion(%): DIFFICULTY SWALLOWING CONGESTION IN THROAT : 0.0005511160099200882

Analyzing the content of complaint within specific duration.

Since 92% of duration of records lasted less than 10 days , it is safe for me to peek the situation of complaint by viewing it within 20 days.

From duration = 1 to duration = 20, FEVER is the most frequent word appearing in the complaint. While as time pass, the proportion of Pneumonia and COVID become larger

* + - 1. CHIEF\_COMPLAINT\_ONSET\_DATE

the number of Nan in CHIEF\_COMPLAINT\_ONSET\_DATE column: 1814239

the percentage of Nan in CHIEF\_COMPLAINT\_ONSET\_DATE column: 90.9059888331

the number of unique CHIEF\_COMPLAINT\_ONSET\_DATE: 164729

which CHIEF\_COMPLAINT\_ONSET\_DATE get the largest proportion(%):

01-DEC-20 0800 : 0.013223723359707316

which CHIEF\_COMPLAINT\_ONSET\_DATE get the smallest proportion(%):

02-JAN-21 0803 : 0.0005509884733211381

​

1.3 Table 3: Diagnoses (all Dx identified in cohort criteria for seven (7) days following the initial ED encounter)

File Name: R3\_1971\_YE\_DIAGNOSES\_2021\_03\_01

Comments:

* one row per diagnosis

Data elements:

Patient\_Study\_ID, Diagnosis\_Type, Diagnosis\_Code, Diagnosis\_Name, Primary\_Diagnosis\_YN, Diagnosis\_From\_Date, Diagnosis\_To\_Date

* + 1. Summative analysis for each of elements
       1. Number of patients in this table

There are 1826776 records, and 117574 unique patients(117574 unique STUDY\_IDs)

* + - 1. Diagnosis\_type and frequency

There are NaN values in Diagnosis\_type

|  |  |  |
| --- | --- | --- |
| value | Occurrence time | Proportion in all recodes |
| NaN | 20750 | 1.135% |

There are 4 unique values of Diagnosis\_type:

Non\_ nan values’ distribution in Diagnosis\_type

|  |  |  |
| --- | --- | --- |
| DIAGNOSIS\_TYPE | Occurrence time | Proportion(among all of non-NaN values) |
| ICD10CM | 1658177 | 91.813573 % |
| ICD9CM | 145394 | 8.050493 % |
| ICD10 | 2439 | 0.135048 % |
| ICD9 | 16 | 0.000886 % |

* + - 1. DX\_CODE, DIAGNOSIS\_NAME, and frequency

**For DX\_CODE**: there are 1826776 records and 16794 unique values,

there is no NaN value in DX\_CODE

|  |  |  |
| --- | --- | --- |
| DX\_CODE (TOP 10) | Occurrence time | Proportion(among all of non-NaN values) |
| I10 | 38051 | 2.082959 % |
| K21.9 | 34811 | 1.905598 % |
| E78.5 | 27938 | 1.529361 % |
| Z87.891 | 26436 | 1.447140 % |
| Z79.82 | 25497 | 1.395738 % |
| F41.9 | 23335 | 1.277387 % |
| F32.9 | 21764 | 1.191389 % |
| I25.10 | 20049 | 1.097507 % |
| M19.90 | 17377 | 0.951239 % |
| N17.9 | 16578 | 0.907500 % |

**For DIAGNOSIS\_NAME**: There are 1806026 records and 20750 missing value .

|  |  |  |
| --- | --- | --- |
| value | Occurrence time | Proportion in all recodes (%) |
| NaN | 20750 | 1.1358809180764362 |

|  |  |  |
| --- | --- | --- |
| **DIAGNOSIS\_NAME** (TOP 10) | Occurrence time | percentage (among all of non-NaN values). (% ) |
| Essential (primary) hypertension | 38051 | 2.106891 |
| Gastro-esophageal reflux disease without esophagitis | 34811 | 1.927492 |
| Hyperlipidemia, unspecified | 27938 | 1.546932 |
| Personal history of nicotine dependence | 26436 | 1.463766 |
| Long term (current) use of aspirin | 25497 | 1.411774 |
| Anxiety disorder, unspecified | 23335 | 1.292063 |
| Major depressive disorder, single episode, unspecified | 21764 | 1.205077 |
| Atherosclerotic heart disease of native coronary artery without angina pectoris | 20049 | 1.110117 |
| Fever, unspecified | 17718 | 0.981049 |
| Acute kidney failure, unspecified | 17400 | 0.963441 |

There are 16239 unique values in DIAGNOSIS\_NAME

For **DX\_CODE** and **DIAGNOSIS\_NAME**, they are not one-to-one relationships.

* + - 1. PRIMARY\_DX\_IND, frequency

|  |  |  |
| --- | --- | --- |
| value | Occurrence time | percentage (% ) |
| N | 1603246 | 87.763689 |
| Y | 223530 | 12.236311 |

* + - 1. DX\_FROM\_DATE: range

There is no missing value in DX\_FROM\_DATE

The minimum value of DX\_FROM\_DATE is ‘2011-01-01’

The maximum value of DX\_FROM\_DATE is ‘2021-01-31’

* + - 1. DX\_TO\_DATE: range

There are 140321 missing values in DX\_TO\_DATE

The minimum value of DX\_FROM\_DATE is ‘2011-01-01’

The maximum value of DX\_FROM\_DATE is ‘2021-02-16

* + - 1. Duration = DX\_TO\_DATE-DX\_FROM\_DATE

For all existed records:

The largest duration is 292 days

The smallest duration is 0 day