진료데이터 전처리

In [3]:

```
# 필요 패키지 로딩

import pandas as pd
import re
import numpy as np
```

In [4]:

```
# 2016년 진료데이터 확인

medical_2016 = pd.read_csv('../../lawdata/medical/NHIS_OPEN_T20_2016.csv', encoding = 'cp949')

print(medical_2016.info())
print(medical_2016.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12540347 entries, 0 to 12540346
Data columns (total 19 columns):
 #   Column        Dtype
---  ------        -----
 0   기준년도          int64
 1   가입자일련번호       int64
 2   진료내역일련번호      int64
 3   성별코드          int64
 4   연령대코드(5세단위)   int64
 5   시도코드          int64
 6   요양개시일자        int64
 7   서식코드          int64
 8   진료과목코드        int64
 9   주상병코드         object
 10  부상병코드         object
 11  요양일수          int64
 12  입내원일수         int64
 13  심결가산율         float64
 14  심결요양급여비용총액    int64
 15  심결본인부담금       int64
 16  심결보험자부담금      int64
 17  총처방일수         int64
 18  데이터공개일자       int64
dtypes: float64(1), int64(16), object(2)
memory usage: 1.8+ GB
None
   기준년도  가입자일련번호  진료내역일련번호  성별코드  연령대코드(5세단위)  시도코드    요양개시일자  서식코드  진료과목코드  \
0  2016        1  34153116     2            2    45  20160817     3       14
1  2016        2  46374827     2            8    41  20160202     3        1
2  2016        2   3115110     2            8    41  20160407     3        1
3  2016        2  50348269     2            8    41  20160104     3        1
4  2016        2  15822952     2            8    41  20160130     3       13

  주상병코드  부상병코드  요양일수  입내원일수  심결가산율  심결요양급여비용총액  심결본인부담금  심결보험자부담금  총처방일수  \
0  L209  L0888     1      1   0.15       14410     4300     10110      1
1  J209   J304     1      1   0.15       10300     3000      7300      3
2  J209   J304     1      1   0.15       10300     3000      7300      2
3  A099   K291     1      1   0.15       14410     4300     10110      2
4   J40   J304     1      1   0.15       17980     5300     12680      5

    데이터공개일자
0  20171218
1  20171218
2  20171218
3  20171218
4  20171218
```

In [5]:

```
# 2017년 진료데이터 확인

medical_2017 = pd.read_csv('../../lawdata/medical/NHIS_OPEN_T20_2017.csv', encoding = 'cp949')

print(medical_2017.info())
print(medical_2017.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12568532 entries, 0 to 12568531
Data columns (total 19 columns):
 #   Column       Dtype
---  ------       -----
 0   기준년도         int64
 1   가입자 일련번호     int64
 2   진료내역일련번호     int64
 3   성별코드         int64
 4   연령대코드        int64
 5   시도코드         int64
 6   요양개시일자       int64
 7   서식코드         int64
 8   진료과목코드       int64
 9   주상병코드        object
 10  부상병코드        object
 11  요양일수         int64
 12  입내원일수        int64
 13  심결가산율        float64
 14  심결요양급여비용총액   int64
 15  심결본인부담금      int64
 16  심결보험자부담금     int64
 17  총처방일수        int64
 18  데이터 기준일자     int64
dtypes: float64(1), int64(16), object(2)
memory usage: 1.8+ GB
None
```

	기준년도	가입자 일련번호	진료내역일련번호	성별코드	연령대코드	시도코드	요양개시일자	서식코드	진료과목코드	주상병코드
0	2017	1	4661608	1	5	11	20170316	3	13	J060
1	2017	1	14468123	1	5	11	20170123	3	1	R51
2	2017	1	22980223	1	5	11	20170404	3	13	J0190
3	2017	1	23050697	1	5	11	20170407	3	13	J0190
4	2017	1	29517148	1	5	11	20170516	3	14	L309

	부상병코드	요양일수	입내원일수	심결가산율	심결요양급여비용총액	심결본인부담금	심결보험자부담금	총처방일수	데이터 기준일자
0	R51	1	1	0.15	16420	4900	11520	3	20181126
1	K297	1	1	0.15	16420	4900	11520	3	20181126
2	J060	1	1	0.15	14650	4300	10350	3	20181126
3	J060	1	1	0.15	10620	3100	7520	3	20181126
4	L301	1	1	0.15	21760	6500	15260	7	20181126

In [6]:

```
# 2018년도 데이터 확인 및 병합

medi_2018_1 = pd.read_csv('../../lawdata/medical/NHIS_OPEN_T20_2018_PART1.csv', encoding = 'cp949')
medi_2018_2 = pd.read_csv('../../lawdata/medical/NHIS_OPEN_T20_2018_PART2.csv', encoding = 'cp949')
medi_2018_3 = pd.read_csv('../../lawdata/medical/NHIS_OPEN_T20_2018_PART3.csv', encoding = 'cp949')

medi_2018 = pd.concat([medi_2018_1,medi_2018_2,medi_2018_3], axis=0, ignore_index = True)

print(medi_2018.info())
print(medi_2018.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12974000 entries, 0 to 12973999
Data columns (total 19 columns):
 #   Column       Dtype
---  ------       -----
 0   기준년도         int64
 1   가입자일련번호      int64
 2   진료내역일련번호     int64
 3   성별코드         int64
 4   연령대코드        int64
```

```
 5   시도코드          int64
 6   요양개시일자       int64
 7   서식코드          int64
 8   진료과목코드       int64
 9   주상병코드         object
 10  부상병코드         object
 11  요양일수          int64
 12  입내원일수         int64
 13  심결가산율         float64
 14  심결요양급여비용총액   int64
 15  심결본인부담금       int64
 16  심결보험자부담금     int64
 17  총처방일수         int64
 18  데이터기준일자       int64
dtypes: float64(1), int64(16), object(2)
memory usage: 1.8+ GB
None
```

	기준년도	가입자일련번호	진료내역일련번호	성별코드	연령대코드	시도코드	요양개시일자	서식코드	진료과목코드	주상병코드
0	2018	1	29474228	1	5	47	20181120	3	1	J209
1	2018	2	40229620	2	4	28	20180927	3	13	J42
2	2018	2	26253257	2	4	28	20181204	3	13	J42
3	2018	2	42606436	2	4	28	20180807	3	14	L238
4	2018	2	31225816	2	4	28	20181002	3	13	J42

	부상병코드	요양일수	입내원일수	심결가산율	심결요양급여비용총액	심결본인부담금	심결보험자부담금	총처방일수	데이터기준일자
0	K291	1	1	0.15	21010	6300	14710	3	20191217
1	J303	1	1	0.15	15310	4500	10810	4	20191217
2	J303	1	1	0.15	15310	4500	10810	4	20191217
3	K297	1	1	0.15	15310	4500	10810	3	20191217
4	J303	1	1	0.15	10950	3200	7750	5	20191217

In [7]:

```
# 2018년도 병합 데이터 저장

medi_2018.to_csv('../../lawdata/medical/NHIS_OPEN_T20_2018.csv', encoding = 'cp949', index = False)
```

In [8]:

```
# 각 년도별 데이터 필요 변수 추출 및 저장

for i in [2016,2017,2018]:
    df = pd.read_csv(f'../../lawdata/medical/NHIS_OPEN_T20_{i}.csv', encoding = 'cp949')

    # 칼럼명 통합
    if i == 2016:
        df.rename(columns = {'연령대코드(5세단위)': '연령대코드'}, inplace = True)
        pass

    if i == 2017:
        df.rename(columns = {'가입자 일련번호': '가입자일련번호'}, inplace = True)
        pass

    # 필요변수 추출
    df = df[['가입자일련번호','성별코드','연령대코드','시도코드','주상병코드','요양개시일자']]
    pass

    # 데이터 저장
    df.to_csv(f'../../lawdata/medical/NHIS_{i}.csv', encoding = 'utf8', index = False)
    pass
```

In [9]:

```
# 시도코드 이름 부여

data ={
    '시도코드' : [42,41,43,44,30,47,48,45,46,11,28,27,31,29,26,49,36] ,
    '시도' : ['강원','경기','충북','충남','대전','경북','경남','전북','전남','서울','인천','대구','울산','광
','부산','제주','세종']
      }

sido = pd.DataFrame(data, columns = ['시도코드','시도'])
```

In [10]:

```
# 질병코드 분류 함수

def ill(x):
    str = x[:3]
    de = re.findall(r'([A-Z])([_]?)(\d{2})?', str)
    for n, d, t in de:
        if t == '':
            t = 0
        if n == 'A' or n == 'B':
            return 'AB'
        elif n == 'C' or (n == 'D' and int(t) <= 48):
            return 'CD48'
        elif n == 'D':
            return 'D50'
        elif n == 'H':
            if int(t) < 60:
                return 'H59'
            elif int(t) >= 60:
                return 'H60'
        elif n == ('S' or 'T'):
            return 'ST'
        elif n == ('V' or 'Y'):
            return 'VY'
        else:
            return n
        pass
    pass
```

In [13]:

```
# 각 년도 데이터에
# 질병코드 분류 함수 적용 및
# 시도명 칼럼 추가

for i in [2016,2017,2018]:

    df = pd.read_csv(f'../../lawdata/medical/NHIS_{i}.csv')
    pass

    df = pd.merge(df,sido, on='시도코드',how='inner')

    df['질병코드'] = df['주상병코드'].apply(ill)

    df.to_csv(f'../../lawdata/medical/NHIS_{i}_Fixed.csv', index = False)

    pass
```

In [14]:

```
# 호흡기 질환 진료환자 추출 및 병합

medi_2016 = pd.read_csv('../../lawdata/medical/NHIS_2016_Fixed.csv')
medi_2017 = pd.read_csv('../../lawdata/medical/NHIS_2017_Fixed.csv')
medi_2018 = pd.read_csv('../../lawdata/medical/NHIS_2018_Fixed.csv')

medi_2016 = medi_2016[medi_2016['질병코드']=='J']
medi_2017 = medi_2017[medi_2017['질병코드']=='J']
medi_2018 = medi_2018[medi_2018['질병코드']=='J']

medi = pd.concat([medi_2016,medi_2017,medi_2018], axis=0, ignore_index = True)

medi.info()
medi.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10514532 entries, 0 to 10514531
Data columns (total 8 columns):
 #   Column    Dtype
---  ------    -----
 0   가입자일련번호  int64
 1   성별코드      int64
```

```
 1   성별코드       int64
 2   연령대코드     int64
 3   시도코드       int64
 4   주상병코드     object
 5   요양개시일자   int64
 6   시도           object
 7   질병코드       object
dtypes: int64(5), object(3)
memory usage: 641.8+ MB
```

Out[14]:

	가입자일련번호	성별코드	연령대코드	시도코드	주상병코드	요양개시일자	시도	질병코드
0	23	1	9	45	J042	20160920	전북	J
1	35	1	13	45	J029	20161119	전북	J
2	117	1	11	45	J303	20161202	전북	J
3	117	1	11	45	J209	20161201	전북	J
4	143	2	14	45	J209	20160117	전북	J

In [15]:

```
# 2016~2018년도 호흡기질환 진료환자 데이터 저장

medi.to_csv('../../lawdata/medical/NHIS_J_Cases.csv', index = False)
```