patient_data_statistics

June 12, 2018

0.1 Here are the main statistics about the data allready available from the clinic

```
In [126]: import numpy as np
          import pandas as pd
          import matplotlib as plt
          import os
          %matplotlib inline
          ##file paths##
          DIAG_PROZ = "anom_labels_DIAG_PROZ.xlsx"
          IOL = "anom_labels_IOL.xlsx"
          VISUS = "anom_labels_VISUS.xlsx"

          ##load into pandas##
          DIAG_PROZ_pd = pd.read_excel(pd.ExcelFile(DIAG_PROZ))
          IOL_pd = pd.read_excel(pd.ExcelFile(IOL))
          VISUS_pd = pd.read_excel(pd.ExcelFile(VISUS))
```

As seen in previous notebook, we have approx. 1400 patients and 7500 studies for these patients.

0.1.1 This section diplays the main statistics concerning Diagnosis and Procedures

```
In [134]: #columns available
          pd.DataFrame({"Columns":DIAG_PROZ_pd.columns})
```

```
Out[134]:                          Columns
          0                             ID
          1          Datum_der_Diagnose_DAT
          2   Katalog_des_Diagnosecode_DKAT
          3               Diagnosecode_DKEY
          4   Lokalisation_des_Diagnose_LOK
          5                    Date_from_Pr
          6  Lokalisation_des_Prozedure_LOK
          7             Prozedure_code_ICPMK
          8           ID_des_OP_Katalog_ICPML
```

0.1.2 The different DKAT and DKEY values available

```
In [137]: pd.DataFrame({"DKAT":pd.unique(DIAG_PROZ_pd['Katalog_des_Diagnosecode_DKAT'])})
```

```
Out[137]:    DKAT
          0    10
          1     5
          2     6
          3     8
          4    17
          5    12
          6    13
          7     7
          8     9
          9    11
          10   GA
          11   16
          12  NaN
          13    3
          14   18
          15   AS
          16    4
          17   A1
          18   A2
          19   A3
          20    2
          21    1
          22   ME
```

```
In [139]: pd.DataFrame({"DKEY":pd.unique(DIAG_PROZ_pd['Diagnosecode_DKEY'])})
```

```
Out[139]:        DKEY
          0     H26.9
          1     H35.5
          2     H20.0
          3     H25.1
          4     Z96.1
          5    E13.30
          6     H36.0
          7     H01.0
          8    E14.30
          9    E11.30
          10    Z01.0
          11    H35.3
          12    T85.2
          13    H27.1
          14    H43.1
          15    H27.0
          16  GAEP-B2
```

```
17      I10.90
18      E11.90
19       Z85.9
20         NaN
21     GAEP-E1
22         H46
23       H47.2
24       H25.9
25       H52.0
26       H52.2
27       H35.0
28       H11.1
29       D31.3
..         ...
821     H35.07
822    H36.068
823     H25.05
824      D68.8
825     E14.50
826      C18.2
827     C79.88
828     H47.39
829      R07.2
830      A49.9
831      K21.9
832     M79.66
833     H40.13
834        E14
835      E03.8
836      I74.3
837        B07
838      H11.8
839      G23.1
840     H40.01
841     M61.16
842     M61.15
843     M08.20
844      Z44.2
845     J96.01
846      J45.8
847      M32.8
848      D31.0
849      G35.0
850      Q87.8

[851 rows x 1 columns]
```

0.1.3 How often are the different diagnosis DKAT and DKEY present?

```
In [129]: diagnoses = DIAG_PROZ_pd[['Katalog_des_Diagnosecode_DKAT']].T.drop_duplicates().T
          counts = diagnoses['Katalog_des_Diagnosecode_DKAT'].value_counts()
          print("These are the counts {}".format(counts))
```

```
These are the counts 17    51404
16    45863
12    35870
13    35143
11    21070
10    14842
8     14753
18    14367
9     14100
7     11118
6      8769
5      8179
AS     6959
4      5198
GA     4598
A1     2751
3      1970
A3     1441
A2     1199
1      1050
2       763
ME       30
Name: Katalog_des_Diagnosecode_DKAT, dtype: int64
```

```
In [130]: diagnoses = DIAG_PROZ_pd[['Diagnosecode_DKEY']].T.drop_duplicates().T
          counts_DKEY = diagnoses['Diagnosecode_DKEY'].value_counts()
          print("These are the counts {}".format(counts_DKEY))
```

```
These are the counts H35.3      54080
H35.8      24163
Z96.1      16046
H33.0      10410
Z98.8       9859
H34.8       9234
H40.1       8874
E11.30      8681
I10.90      6508
H25.8       5818
H36.0       5279
H26.9       4809
H25.1       4427
H43.1       4064
```

```
E14.30      3991
H44.2       3720
Z01.0       3684
H20.0       3603
H40.5       3598
ASA2        3524
Z46.0       3410
H52.1       2879
H47.2       2786
H35.0       2754
H44.1       2750
H20.9       2442
GAEP-E1     2356
ASA3        2097
H25.9       2080
H40.0       2075
           ...
O14.0          1
H53.19         1
H35.313        1
H32.8          1
R74.0          1
H35.5P9        1
H53.9          1
M31.3          1
H44.51         1
I67.2          1
Z11            1
L94.0          1
A49.9          1
M08.20         1
M14.8          1
M61.15         1
M61.16         1
Z94.79         1
H16.14         1
M79.66         1
C18.2          1
Z90.3          1
H31.88         1
L40.9          1
O09.3          1
T15.0          1
T15.1          1
T15.01         1
N60.1          1
A51.0          1
Name: Diagnosecode_DKEY, Length: 850, dtype: int64
```

0.1.4 One patient can have many different DKAT and DKEY, here the average amount of different values they hold

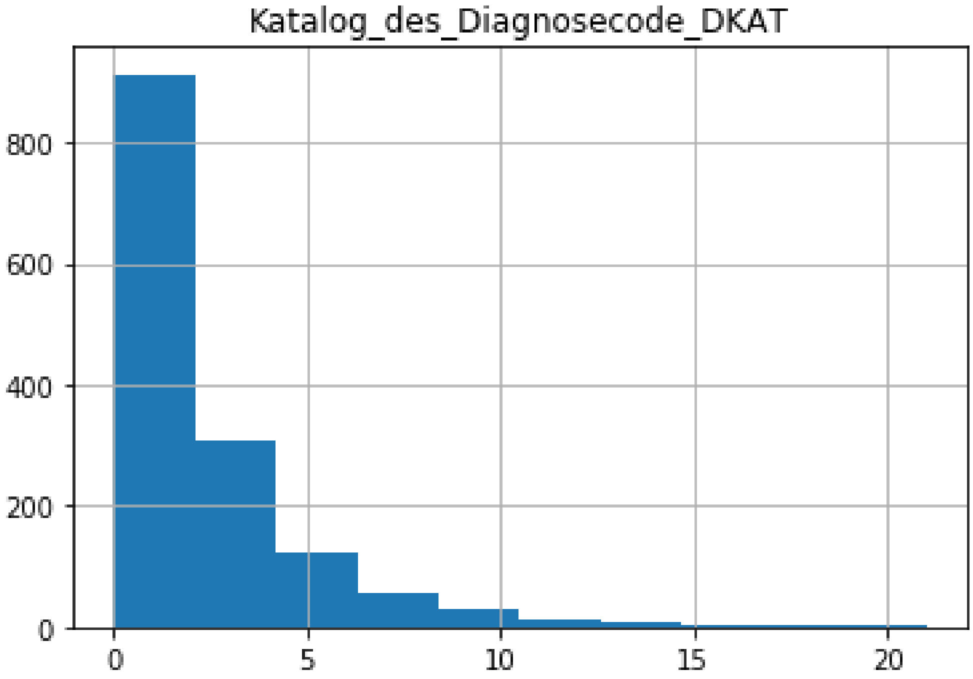
```
In [131]: num_unqie_DKAT_per_ID = pd.DataFrame(DIAG_PROZ_pd.groupby(['ID'])\
                         ['Katalog_des_Diagnosecode_DKAT'].nunique())
          num_unqie_DKEY_per_ID =pd.DataFrame(DIAG_PROZ_pd.groupby(['ID'])\
                         ['Diagnosecode_DKEY'].nunique())
          #how many different DKAT and DKEY each patient has
          avg_num_DKAT_per_patient = num_unqie_DKAT_per_ID['Katalog_des_Diagnosecode_DKAT'].mean
          avg_num_DKEY_per_patient = num_unqie_DKEY_per_ID['Diagnosecode_DKEY'].mean()

          print("The averge number of DKAT per patient is {}".format(avg_num_DKAT_per_patient))
          print("The averge number of DKEY per patient is {}".format(avg_num_DKEY_per_patient))
```

```
The averge number of DKAT per patient is 2.79234972678
The averge number of DKEY per patient is 4.5443989071
```

0.1.5 Distribution of number of DKAT and DKEY per patient are

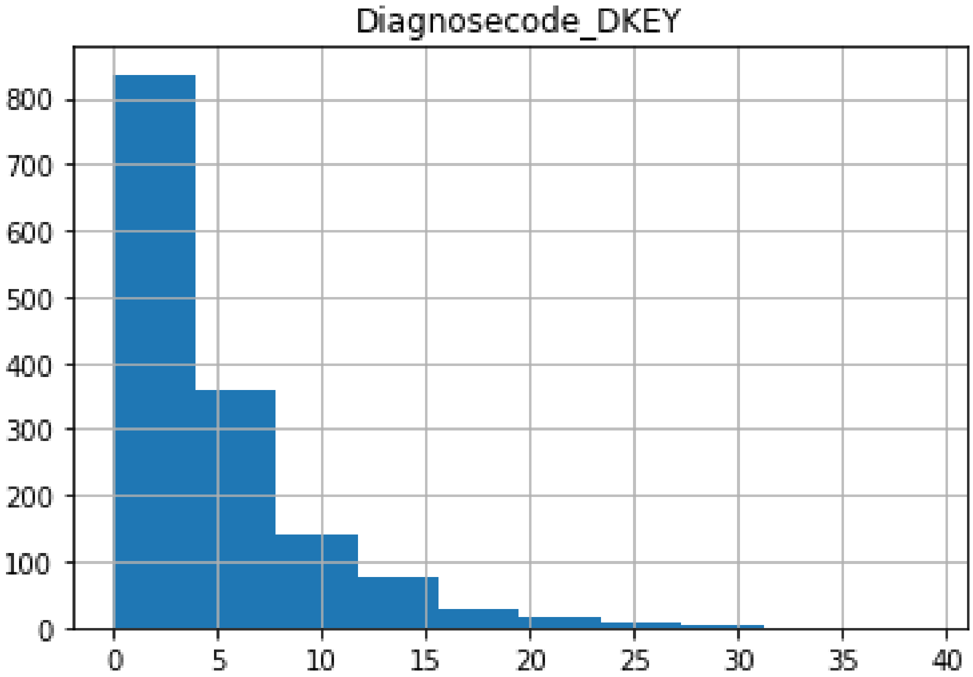
```
In [141]: num_unqie_DKAT_per_ID.hist()
```

```
Out[141]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7fdb2d0c4410>]],
                dtype=object)
```



```
In [142]: num_unqie_DKEY_per_ID.hist()
```

```
Out[142]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7fdb28b36550>]],
                dtype=object)
```



0.2 Now lookign towards the prozedures

0.2.1 The different ICPMK and ICPML are

```
In [144]: pd.DataFrame({"ICPMK":pd.unique(DIAG_PROZ_pd['Prozedure_code_ICPMK'])})
```

```
Out[144]:    ICPMK
          0     PA
          1    NaN
          2     PC
          3     PD
          4     P7
          5     P8
          6     PB
          7     PF
          8     PE
          9     PG
          10    P9
          11    P4
          12    P6
          13    P5
          14    P1
          15    P2
          16    P3
```

```
In [145]: pd.DataFrame({"ICPML":pd.unique(DIAG_PROZ_pd['ID_des_OP_Katalog_ICPML'])})
```

```
Out[145]:         ICPML
          0    5-144.3A
          1     5-156.9
          2       5-984
          3         NaN
          4     3-300.0
          5    5-144.5A
          6     5-155.3
          7    5-145.2J
          8     5-159.4
          9    5-158.00
          10    5-133.0
          11    5-985.2
          12   5-139.1X
          13   5-158.11
          14    5-154.2
          15    5-142.0
          16      3-30X
          17      3-690
          18    5-142.2
          19    5-136.1
          20   5-158.41
          21    5-137.2
          22   5-144.50
          23    5-137.7
          24    5-154.3
          25    5-155.4
          26    5-155.0
          27   5-158.22
          28    5-156.X
          29    5-133.6
          ..        ...
          340   1-207.1
          341  5-146.0A
          342   9-201.1
          343  9-401.22
          344   1-901.1
          345  8-810.W5
          346  8-800.C0
          347     1-424
          348  6-001.G0
```

```
349   5-075.0
350  5-143.10
351  5-144.XA
352   1-208.Y
353  5-131.00
354   5-134.1
355   5-119.1
356  5-138.1X
357  5-146.0G
358  5-158.14
359  5-157.21
360  5-158.33
361     1-700
362     9-607
363  9-649.13
364  9-649.60
365  9-649.31
366  9-649.52
367  9-649.10
368  9-649.33
369  9-649.80

[370 rows x 1 columns]
```

0.2.2 How often are the different diagnosis DKAT and DKEY present?

```
In [146]: diagnoses = DIAG_PROZ_pd[['Prozedure_code_ICPMK']].T.drop_duplicates().T
          counts = diagnoses['Prozedure_code_ICPMK'].value_counts()
          print("These are the counts {}".format(counts))
```

```
These are the counts PD    50947
PC    49119
PF    39091
PE    38995
PB    31066
P9    14128
P8    13944
P6    12909
PA    12708
P7    11919
PG     8781
P5     7388
P4     4862
P3     3297
P2     2002
P1     1666
Name: Prozedure_code_ICPMK, dtype: int64
```

```
In [148]: diagnoses = DIAG_PROZ_pd[['ID_des_OP_Katalog_ICPML']].T.drop_duplicates().T
          counts_DKEY = diagnoses['ID_des_OP_Katalog_ICPML'].value_counts()
          print("These are the counts {}".format(counts_DKEY))
```

```
These are the counts 5-156.9     79782
3-300.0     61543
5-984       29190
3-30X       15559
3-690        8140
5-154.2      7303
5-144.3A     4822
5-154.3      4740
5-144.5A     4636
5-154.4      4165
5-155.4      3927
5-159.4      3907
1-220.0      3541
5-985.2      3274
5-985.6      2746
5-158.42     2743
5-155.3      2457
5-139.10     1655
5-158.40     1645
3-300        1554
8-83B.31     1550
5-142.0      1472
5-142.2      1443
3-800        1414
8-020.0      1390
5-158.43     1338
5-158.22     1298
5-158.20     1277
5-156.X      1207
5-132.2      1101
            ...
6-001.G0        7
1-424           7
8-831.5         6
8-820.09        6
3-827           6
8-903           6
1-700           6
1-491.4         6
5-794.KH        5
8-919           5
5-760.23        5
9-606.3         5
9-606.5         5
```

```
8-820.04        4
1-426.3         3
1-620.01        3
1-620.0X        3
3-05F           3
5-091.10        3
9-649.80        2
9-649.10        2
9-607           2
9-649.60        2
9-649.33        2
9-649.52        2
9-649.13        2
9-649.31        2
5-144.XA        2
1-798.0         1
1-798.X         1
Name: ID_des_OP_Katalog_ICPML, Length: 369, dtype: int64
```

0.2.3 One patient can have many different DKAT and DKEY, here the average amount of different values they hold

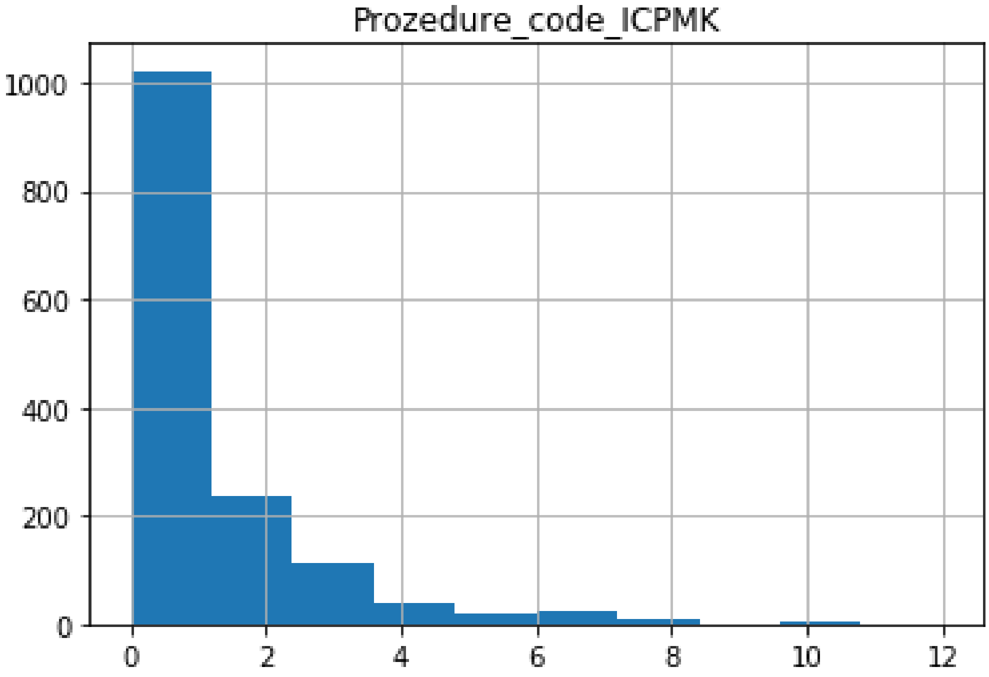
```
In [151]: num_unqie_ICPMK_per_ID = pd.DataFrame(DIAG_PROZ_pd.groupby(['ID'])\
                         ['Prozedure_code_ICPMK'].nunique())
          num_unqie_ICPML_per_ID =pd.DataFrame(DIAG_PROZ_pd.groupby(['ID'])\
                         ['ID_des_OP_Katalog_ICPML'].nunique())
          #how many different DKAT and DKEY each patient has
          avg_num_ICPMK_per_patient = num_unqie_ICPMK_per_ID['Prozedure_code_ICPMK'].mean()
          avg_num_ICPML_per_patient = num_unqie_ICPML_per_ID['ID_des_OP_Katalog_ICPML'].mean()

          print("The averge number of ICPMK per patient is {}".format(avg_num_ICPMK_per_patient)
          print("The averge number of ICPML per patient is {}".format(avg_num_ICPML_per_patient)
```

```
The averge number of ICPMK per patient is 1.3306010929
The averge number of ICPML per patient is 2.85860655738
```

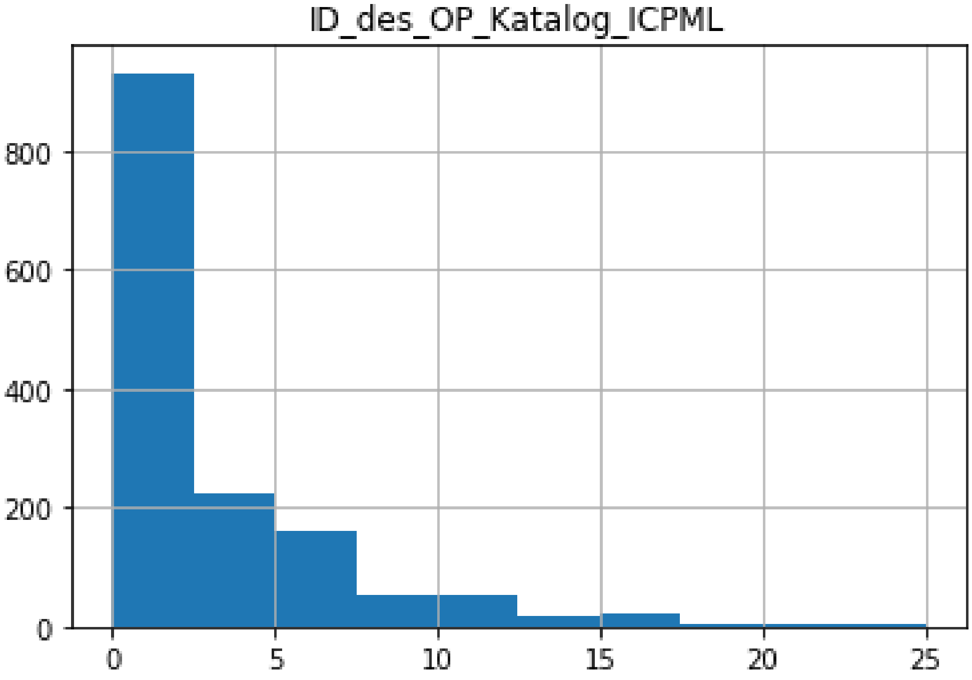
0.3 Distribution of number of ICPMK and ICPML per patient are

```
In [152]: num_unqie_ICPMK_per_ID.hist()
```

```
Out[152]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7fdb1fb2af90>]],
                dtype=object)
```



```
In [153]: num_unqie_ICPML_per_ID.hist()

Out[153]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7fdb2daa0110>]],
                dtype=object)
```


ID_des_OP_Katalog_ICPML