08-10-21-Survival-Analysis

October 8, 2021

1 Analysis of NCIT:C3052 Digestive System Neoplasm - Merve Güngör, Angela Topic

1.1 Setup

[372]:
```
import numpy  as np
import pandas as pd
import sys
#!{sys.executable} -m pip install lifelines
from lifelines import KaplanMeierFitter
import csv
import matplotlib.pyplot as plt
#!{sys.executable} -m pip install -U notebook-as-pdf
```

1.2 Import Dataset

[373]:
```
df = pd.read_csv('Documents/GitHub/UZH-BIO392/course-material/2021/survival/ds.
 ↪csv', sep = ',')
#print(df.head)
```

1.3 Housekeeping

[374]:
```
selection = df.loc[:,['_id','histologicalDiagnosis.id','provenance.geoLocation.
 ↪properties.country','info.tnm',
                      'info.death','info.followupMonths','sex','info.
 ↪cnvstatistics.cnvfraction']]
print(selection.shape)
print(selection.describe())
#print(selection.head)
print(list(selection))
selection.rename(columns = {'_id':'ID','histologicalDiagnosis.id':
 ↪'NCIT','provenance.geoLocation.properties.country':'Country','info.tnm':
 ↪'TNM',
                      'info.death':'Death','info.followupMonths':
 ↪'Followup','sex':'Sex','info.cnvstatistics.cnvfraction':
 ↪'CNVfraction'},inplace = True)
```

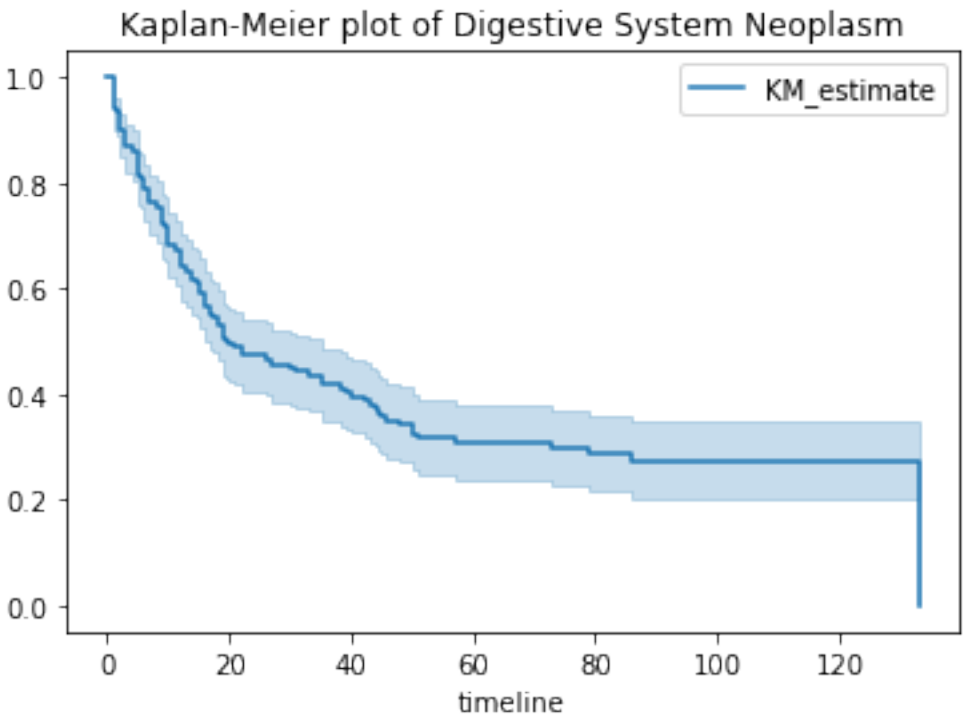
```
print(list(selection))
#print(selection.head)
```

```
(359, 8)
       info.death  info.followupMonths  info.cnvstatistics.cnvfraction
count  359.000000           338.000000                      359.000000
mean     0.607242            28.854172                        0.189128
std      0.489045            30.954017                        0.158355
min      0.000000             1.000000                        0.000000
25%      0.000000             7.000000                        0.059000
50%      1.000000            16.000000                        0.162000
75%      1.000000            40.750000                        0.293500
max      1.000000           134.000000                        1.000000
['_id', 'histologicalDiagnosis.id', 'provenance.geoLocation.properties.country',
'info.tnm', 'info.death', 'info.followupMonths', 'sex',
'info.cnvstatistics.cnvfraction']
['ID', 'NCIT', 'Country', 'TNM', 'Death', 'Followup', 'Sex', 'CNVfraction']
```

1.4 Kaplan-Meier Plots

[375]:
```
kmf= KaplanMeierFitter()
#print(selection['Followup'].isnull().values.any())
selection = selection.dropna()
#print(selection['Followup'].isnull().values.any())

kmf.fit(selection['Followup'], selection['Death']).plot(title = 'Kaplan-Meier␣
↪plot of Digestive System Neoplasm')
```

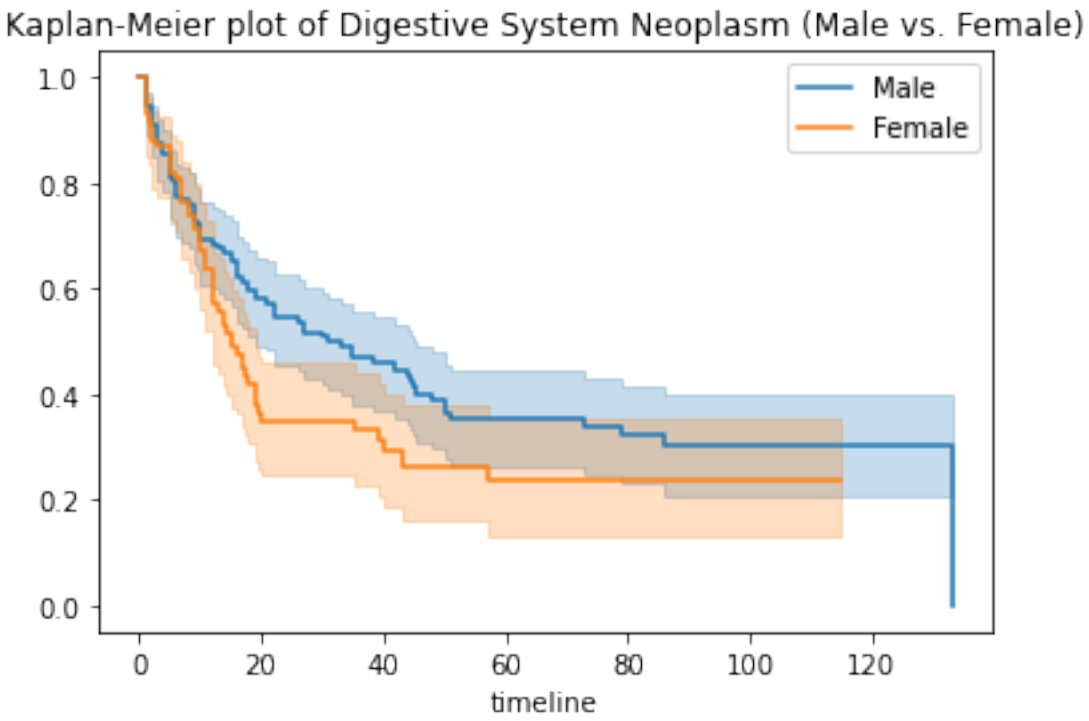
[375]:
```
<AxesSubplot:title={'center':'Kaplan-Meier plot of Digestive System Neoplasm'},
xlabel='timeline'>
```



The probability to survive is smaller in the first 20 months after the disease onset.

[376]:
```
groups = selection['Sex']
i1 = (groups== 'M')
i2 = (groups== 'F')


kmf.fit(selection['Followup'][i1], selection['Death'][i1], label = 'Male')
a1 = kmf.plot(title = "Kaplan-Meier plot of Digestive System Neoplasm (Male vs.␣
↪Female)")
kmf.fit(selection['Followup'][i2], selection['Death'][i2], label = 'Female')
kmf.plot(ax=a1)
```

[376]:
```
<AxesSubplot:title={'center':'Kaplan-Meier plot of Digestive System Neoplasm
(Male vs. Female)'}, xlabel='timeline'>
```

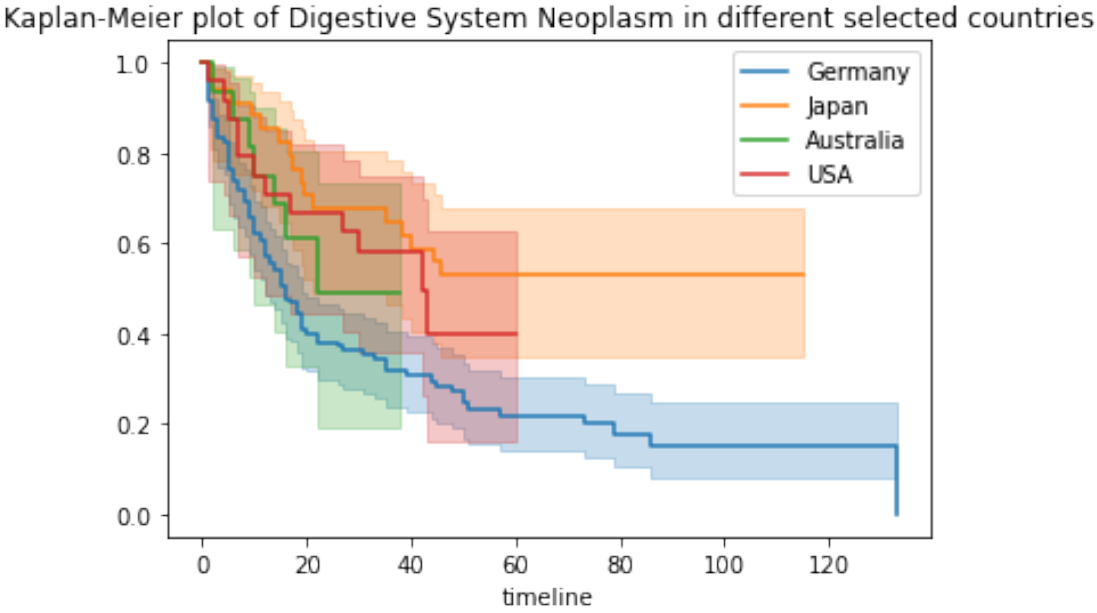


Probability of survival in females is slightly lower than in males suffering from digestive system neoplasm.

[377]:
```
groups = selection['Country']
i1 = (groups== 'Germany')
i2 = (groups== 'Japan')
i3 = (groups== 'Australia')
i4 = (groups== 'United States of America')
print(sum(i1))
print(sum(i2))
print(sum(i3))
print(sum(i4))
kmf.fit(selection['Followup'][i1], selection['Death'][i1], label = 'Germany')
a1 = kmf.plot(title = "Kaplan-Meier plot of Digestive System Neoplasm in␣
 ↪different selected countries")
kmf.fit(selection['Followup'][i2], selection['Death'][i2], label = 'Japan')
a1 = kmf.plot(ax=a1)
kmf.fit(selection['Followup'][i3], selection['Death'][i3], label = 'Australia')
a1 = kmf.plot(ax=a1)
kmf.fit(selection['Followup'][i4], selection['Death'][i4], label = 'USA')
a1 = kmf.plot(ax=a1)
```

```
152
34
```

16
24

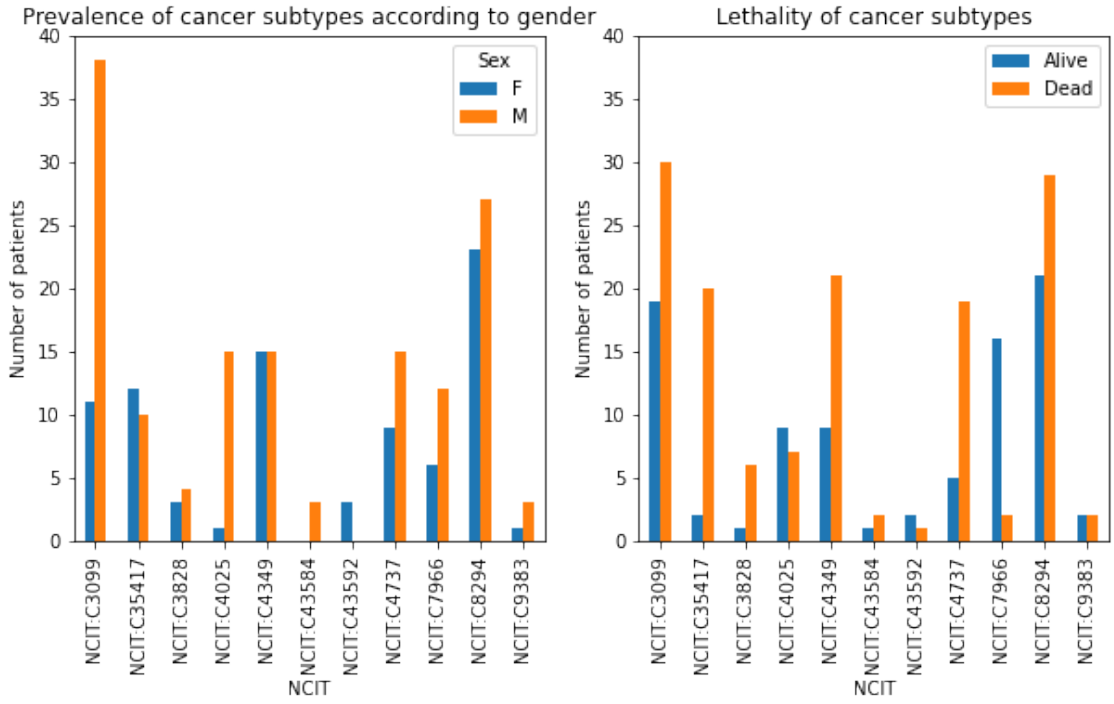


Looking at the plot it seems that the patients in Germany have the highest lethality but it was also the biggest patient group with the longest follow up.

1.5 Grouping according to NCIT subtypes

[378]:
```
fig, axes = plt.subplots(1,2, figsize=(10,5))
grouped_S = selection.groupby(['NCIT','Sex']).count().unstack('Sex')['ID'].plot.
 ↪bar(title = 'Prevalence of cancer subtypes according to gender',ylabel =␣
 ↪'Number of patients',

                                                                          ␣
 ↪    ax = axes[0],ylim=(0, 40))
grouped_D = selection.groupby(['NCIT','Death']).count().unstack('Death')['ID'].
 ↪plot.bar(title = 'Lethality of cancer subtypes',ylabel = 'Number of␣
 ↪patients',ax = axes[1],ylim=(0, 40))
grouped_D.legend(['Alive','Dead'])
```

[378]:
```
<matplotlib.legend.Legend at 0x7f9f2e2671c0>
```



Hepatocellular Carcinoma (NCIT:C3099) and Pancreatic adenocarcinoma (NCIT:C8294) are more common than the other cancer subtypes. In plot 1 can be seen that overall more male individuals are affected by digestive system neoplasm espsecially Hepatocellular Carcinoma and Esophageal adenocarcinoma seem (NCIt:C4025) to be much more frequent in male patients. Might be due to the tendency of men to have an unhealthier lifestyle than women.

Intrahepatic cholangiocarcinoma (NCIT:C35417), combined hepatocellular carcinoma and cholangiocarcinoma (NCIIT:C3828), Colon adenocarcinoma (NCIT:C4349) and Enteropathy type T.cell lymphoma (NCIT:4737) seem to have a higher lethality than others. In comparison Colon Mucinous Adenocarcinoma (NCIT:C7966) seems to have a very good prognosis.

1.6 Gene Analysis ERBB2

[379]:
```
df = pd.read_csv('Documents/ERBB2.csv', sep = ',')
IDs = list(df['_id'])

ERBB2 = selection
ERBB2['ERBB2']= 0
for l in ERBB2['ID']:
    if l in IDs:
        ERBB2.loc[ERBB2.ID == l,'ERBB2'] = 1
```

1.7 Gene Analysis TP53

[380]:
```
df = pd.read_csv('Documents/TP53.csv', sep = ',')
IDs = list(df['_id'])

TP53 = selection
TP53['TP53']= 0
for l in TP53['ID']:
    if l in IDs:
        TP53.loc[TP53.ID == l,'TP53'] = 1
```

1.8 Gene Analysis MYC

[381]:
```
df = pd.read_csv('Documents/MYC.csv', sep = ',')
IDs = list(df['_id'])

MYC = selection
MYC['MYC']= 0
for l in MYC['ID']:
    if l in IDs:
        MYC.loc[MYC.ID == l,'MYC'] = 1
```

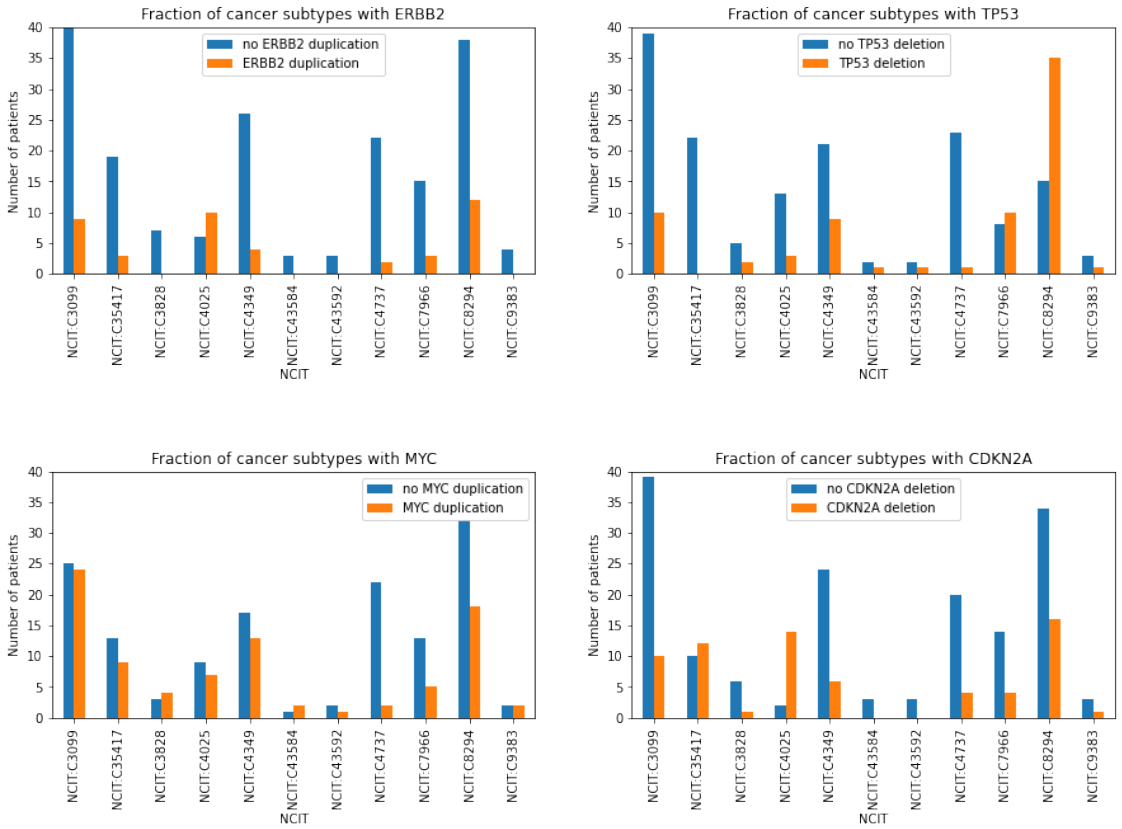
1.9 Gene Analysis CDKN2A

[382]:
```
df = pd.read_csv('Documents/CDKN2A.csv', sep = ',')
IDs = list(df['_id'])

CDKN2A = selection
CDKN2A['CDKN2A']= 0
for l in CDKN2A['ID']:
    if l in IDs:
        CDKN2A.loc[CDKN2A.ID == l,'CDKN2A'] = 1
```

1.10 Gene vs. Tumor barplots

[383]:
```
fig, axes = plt.subplots(2,2, figsize=(15,10))
grouped_EE= ERBB2.groupby(['NCIT','ERBB2']).count().unstack('ERBB2')['ID'].plot.
 ↪bar(title = 'Fraction of cancer subtypes with ERBB2',ax = axes[0,0],ylabel =␣
 ↪'Number of patients',ylim=(0, 40))
grouped_EE.legend(['no ERBB2 duplication','ERBB2 duplication'])
grouped_TE= TP53.groupby(['NCIT','TP53']).count().unstack('TP53')['ID'].plot.
 ↪bar(title = 'Fraction of cancer subtypes with TP53',ax = axes[0,1],ylabel =␣
 ↪'Number of patients',ylim=(0, 40))
grouped_TE.legend(['no TP53 deletion','TP53 deletion'])
```

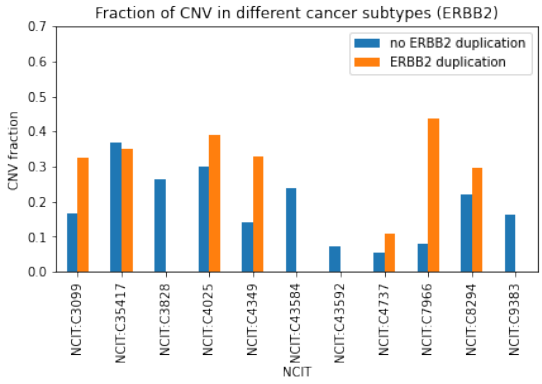
```
grouped_ME=MYC.groupby(['NCIT','MYC']).count().unstack('MYC')['ID'].plot.
 ↪bar(title = 'Fraction of cancer subtypes with MYC',ax = axes[1,0],ylabel =␣
 ↪'Number of patients',ylim=(0, 40))
grouped_ME.legend(['no MYC duplication','MYC duplication'])
grouped_CE=CDKN2A.groupby(['NCIT','CDKN2A']).count().unstack('CDKN2A')['ID'].
 ↪plot.bar(title = 'Fraction of cancer subtypes with CDKN2A',ax =␣
 ↪axes[1,1],ylabel = 'Number of patients',ylim=(0, 40))
grouped_CE.legend(['no CDKN2A deletion','CDKN2A deletion'])
plt.subplots_adjust(hspace = .8)
```

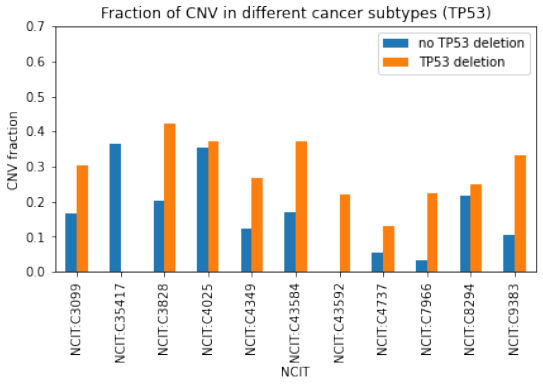


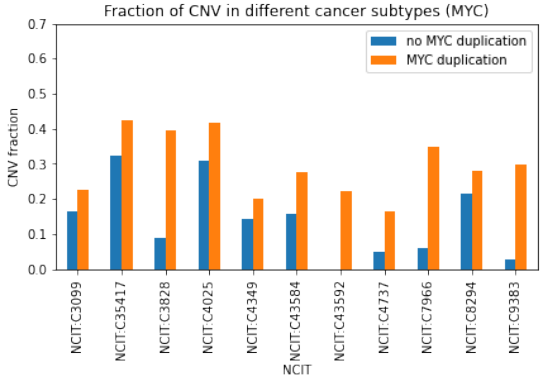
- Hepatocellular carcinoma (NCIT:C3099): Only higher occurance of MYC duplication detected (~50% of samples); other dup/del of the genes of interest are less likely to be involved in this disease.
- Enteropathy type T-cell lymphoma (NCIT:C4737): Only rare detection of the duplications or deletions in GoI
- Pancreatic adenocarcinoma (NCIT:C8294): High duplication of TP53 detected.

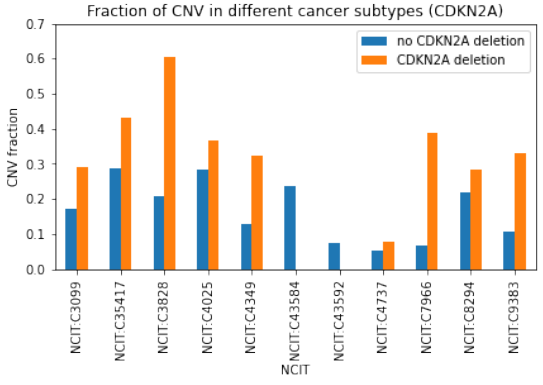
[384]:
```
fig, axes = plt.subplots(2,2, figsize=(15,10))
grouped_EC = ERBB2.groupby(['NCIT','ERBB2']).mean().
 ↪unstack('ERBB2')['CNVfraction'].plot.bar(title = 'Fraction of CNV in␣
 ↪different cancer subtypes (ERBB2)',ax = axes[0,0],
```

```
 ↪              ylabel = 'CNV fraction',ylim=(0, 0.7))
grouped_EC.legend(['no ERBB2 duplication','ERBB2 duplication'])

grouped_TC = TP53.groupby(['NCIT','TP53']).mean().
 ↪unstack('TP53')['CNVfraction'].plot.bar(title = 'Fraction of CNV in␣
 ↪different cancer subtypes (TP53)',ax = axes[0,1],ylabel = 'CNV␣
 ↪fraction',ylim=(0, 0.7))
grouped_TC.legend(['no TP53 deletion','TP53 deletion'])

grouped_MC = MYC.groupby(['NCIT','MYC']).mean().unstack('MYC')['CNVfraction'].
 ↪plot.bar(title = 'Fraction of CNV in different cancer subtypes (MYC)',ax =␣
 ↪axes[1,0],ylabel = 'CNV fraction',ylim=(0, 0.7))
grouped_MC.legend(['no MYC duplication','MYC duplication'])

grouped_CC = CDKN2A.groupby(['NCIT','CDKN2A']).mean().
 ↪unstack('CDKN2A')['CNVfraction'].plot.bar(title = 'Fraction of CNV in␣
 ↪different cancer subtypes (CDKN2A)',ax = axes[1,1],ylabel = 'CNV␣
 ↪fraction',ylim=(0, 0.7))
grouped_CC.legend(['no CDKN2A deletion','CDKN2A deletion'])

plt.subplots_adjust(hspace = .8)
```



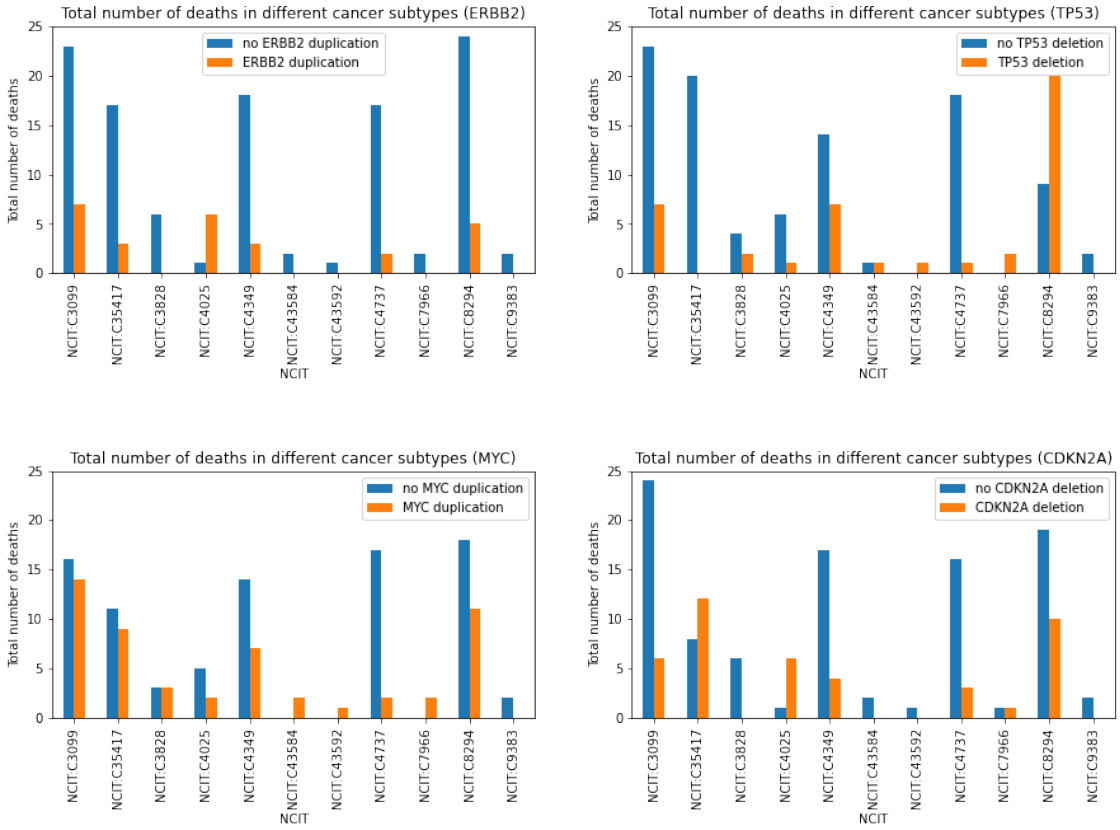






The duplication or deletion of the GoIs seem to lead to a higher CNV fraction.

[385]:
```
fig, axes = plt.subplots(2,2, figsize=(15,10))

grouped_EC = ERBB2.groupby(['NCIT','ERBB2']).sum().unstack('ERBB2')['Death'].
 ↪plot.bar(title = 'Total number of deaths in different cancer subtypes␣
 ↪(ERBB2)',ax = axes[0,0],
                                                                          ␣
 ↪              ylabel = 'Total number of deaths',ylim=(0, 25))
grouped_EC.legend(['no ERBB2 duplication','ERBB2 duplication'])

grouped_TC = TP53.groupby(['NCIT','TP53']).sum().unstack('TP53')['Death'].plot.
 ↪bar(title = 'Total number of deaths in different cancer subtypes (TP53)',ax␣
 ↪= axes[0,1],ylabel = 'Total number of deaths',ylim=(0, 25))
grouped_TC.legend(['no TP53 deletion','TP53 deletion'])

grouped_MC = MYC.groupby(['NCIT','MYC']).sum().unstack('MYC')['Death'].plot.
 ↪bar(title = 'Total number of deaths in different cancer subtypes (MYC)',ax =␣
 ↪axes[1,0],ylabel = 'Total number of deaths',ylim=(0, 25))
grouped_MC.legend(['no MYC duplication','MYC duplication'])

grouped_CC = CDKN2A.groupby(['NCIT','CDKN2A']).sum().unstack('CDKN2A')['Death'].
 ↪plot.bar(title = 'Total number of deaths in different cancer subtypes␣
 ↪(CDKN2A)',ax = axes[1,1],ylabel = 'Total number of deaths',ylim=(0, 25))
grouped_CC.legend(['no CDKN2A deletion','CDKN2A deletion'])

plt.subplots_adjust(hspace = .8)
```



- Pancreatic adenocarcinoma (NCIT:C8294): High number of deaths with this TP53 duplication