PM606 project Final Report

Mingzhi_Ye

8/9/2021

Introduction

This project is conducted by USC Public Health Data Science Department with the great help of City ofo Hope. In this project, I use the CTS data to build a model to predict whether the patient would die in the given time window.I will also analyze several variables to find out whether they are significantly related to the risk of death in the given time window

Method

Pre-processing

At first, I remove all predictors whose NA values are more than 1/4. Those predictors are not lack too much values and are probably biased. There is an exception: date_of_death_dt. All the individuals who didn't die until the end of the study shall be recorded as NA in "date_of_death_dt". So I keep it and change all individuals' "date_of_death_dt" to 4000-11-11 if and only is their "decease" is 0.

Second, I remove Some variables that is too related to other variables, which causes colinearity and lead to reduntancy. Such as smoke_statcat and cig_day_avg, or diag_icd1 and diag_icd_dsc1.I remove one of the pair.

Third, I also remove some irrelevant variables, judged by philosophy.

Forth, I also create several new variables. I create "month" recording the month of "discharge_dt". Then I also create "season" based on "month". I also create "survive time to record how long do they survive after discharge. I also create"die30","die180, "die1800" to record whether they die in 30days 180days or 1800days after discharge.

Preliminary exploration the data

Explore and process the ccs code variables

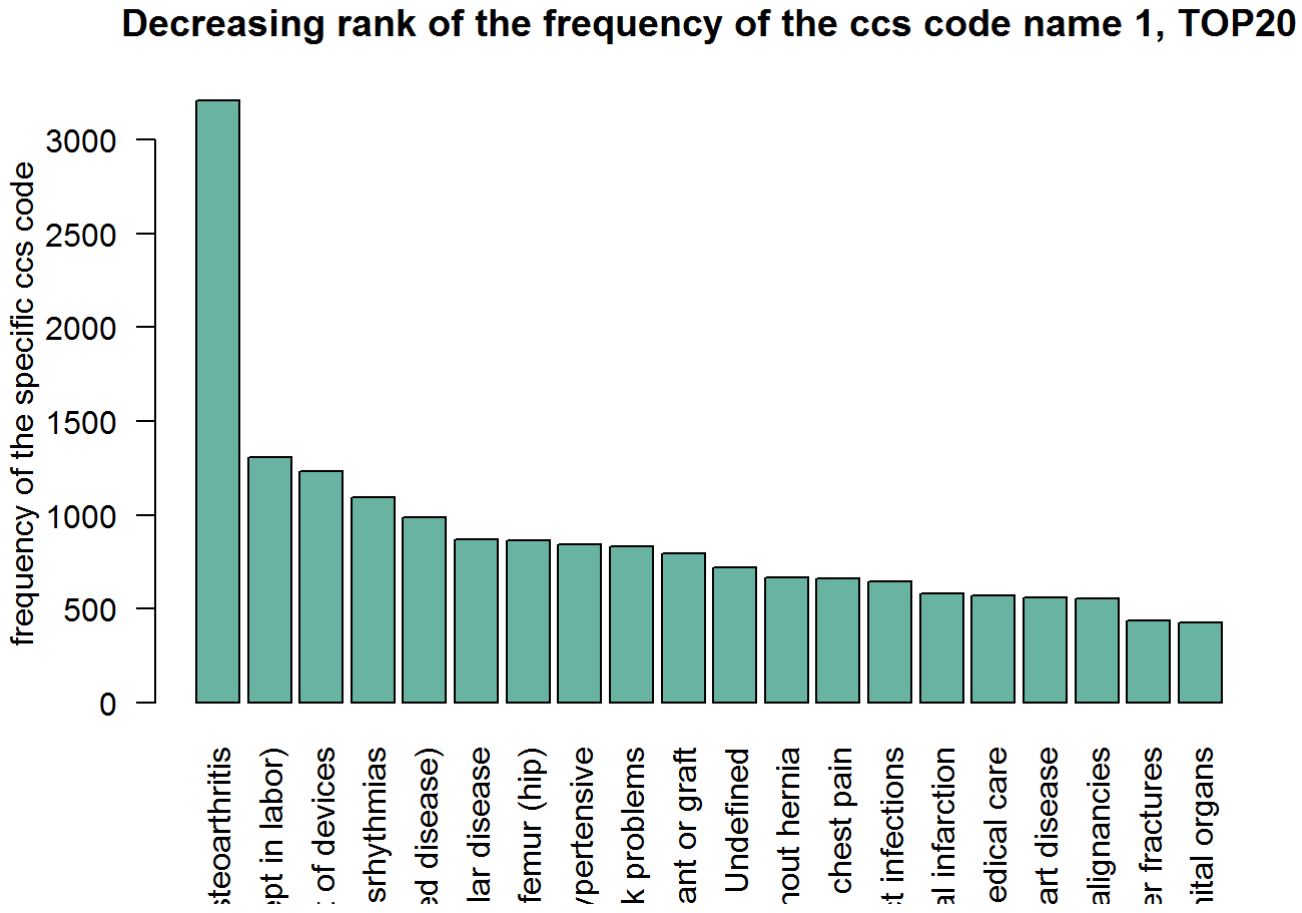
I found that each icd code have more than 2500 categories. If I use the all the ide codes as predictors, they will lead to more than 10000 dummy variables in the LASSO logistic regression and also make the calculate too complex and time consuming. Considering that the processed dataset only have less than 40000 rows, such a big amount of dummy variables will lead to overfitting in the LASSO logistic regression model. So I decide to only keep the ccs codes, which is collapsed from the icd code.

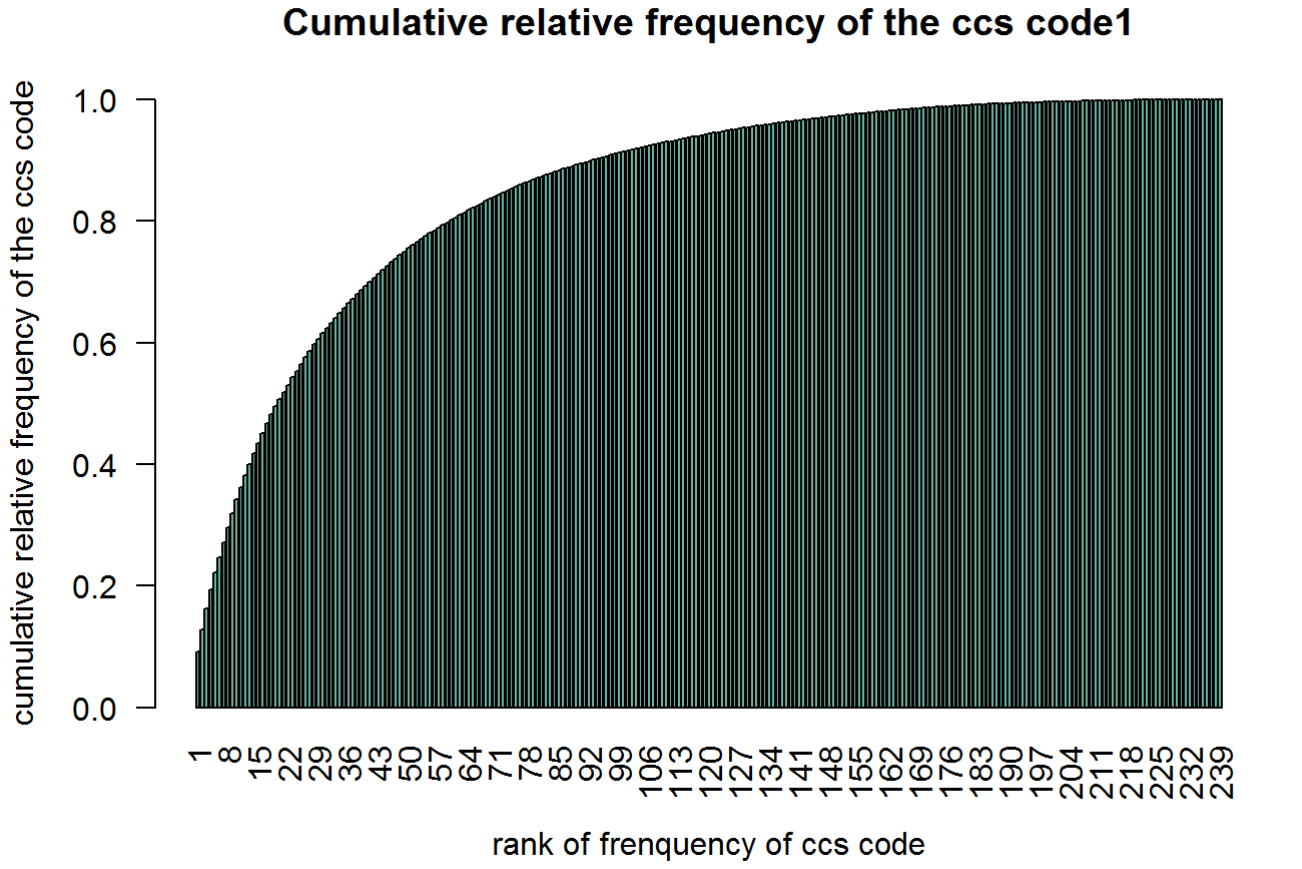
Form the plots, we can find that limited kinds of ccs codes covers most of the individuals. It is the same with ccs code1-4. We can also find that different ccs code have significantly different death time window distribution

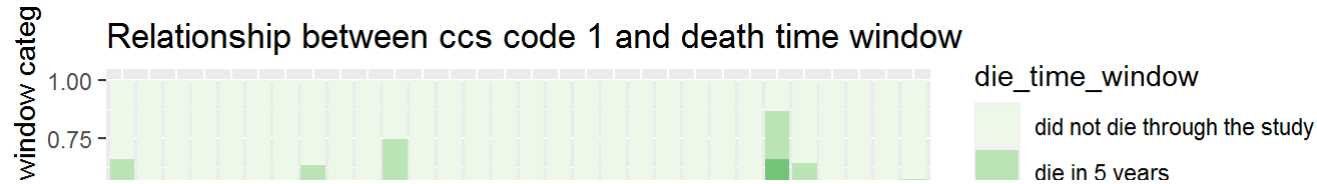
CCS code name1: diag_ccs1

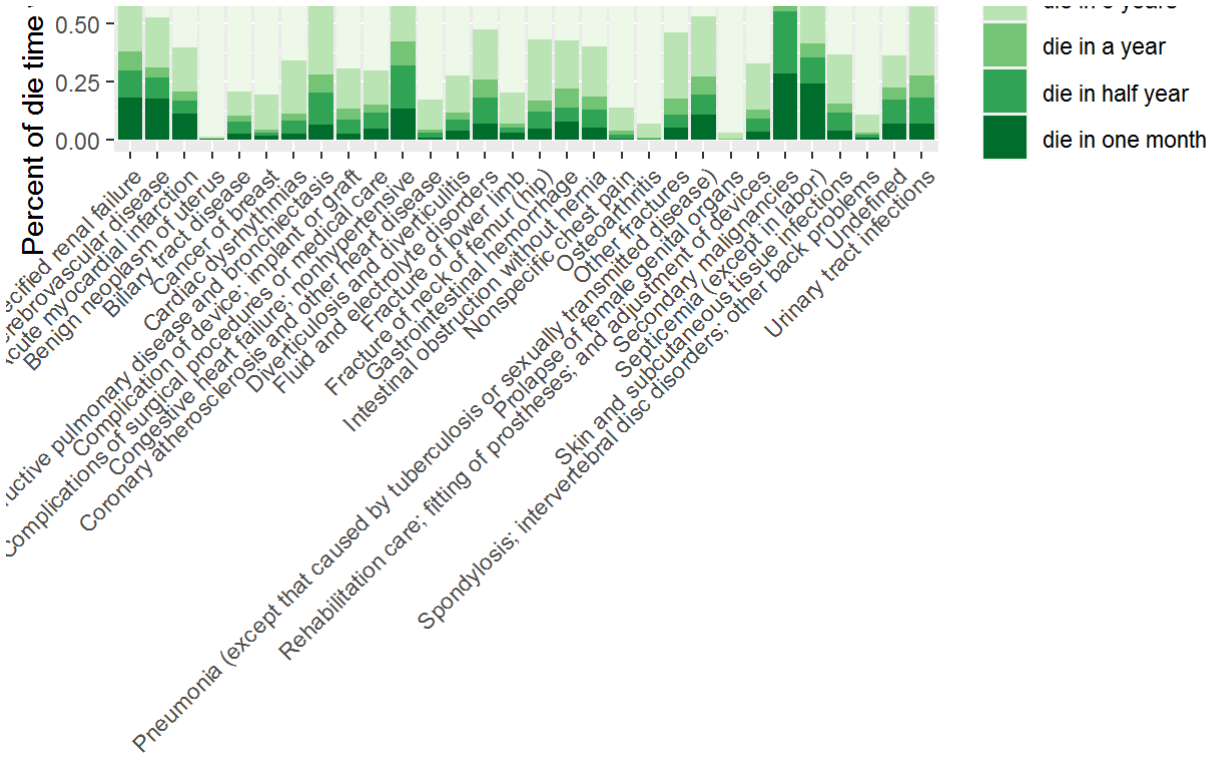
```
## [1] 239
```

```
## [1] 2597
```


Decreasing rank of the frequency of the ccs code name 1, TOP20
frequency of the specific ccs code
3000
2500
2000
1500
1000
500
0
steoarthritis
pt in labor)
of devices
srhythmias
d disease)
lar disease
femur (hip)
rtensive
k problems
ant or graft
Undefined
nout hernia
chest pain
t infections
al infarction
edical care
art disease
alignancies
r fractures
ital organs


Cumulative relative frequency of the ccs code1
cumulative relative frequency of the ccs code
1.0
0.8
0.6
0.4
0.2
0.0
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
204
211
218
225
232
239
rank of frenquency of ccs code

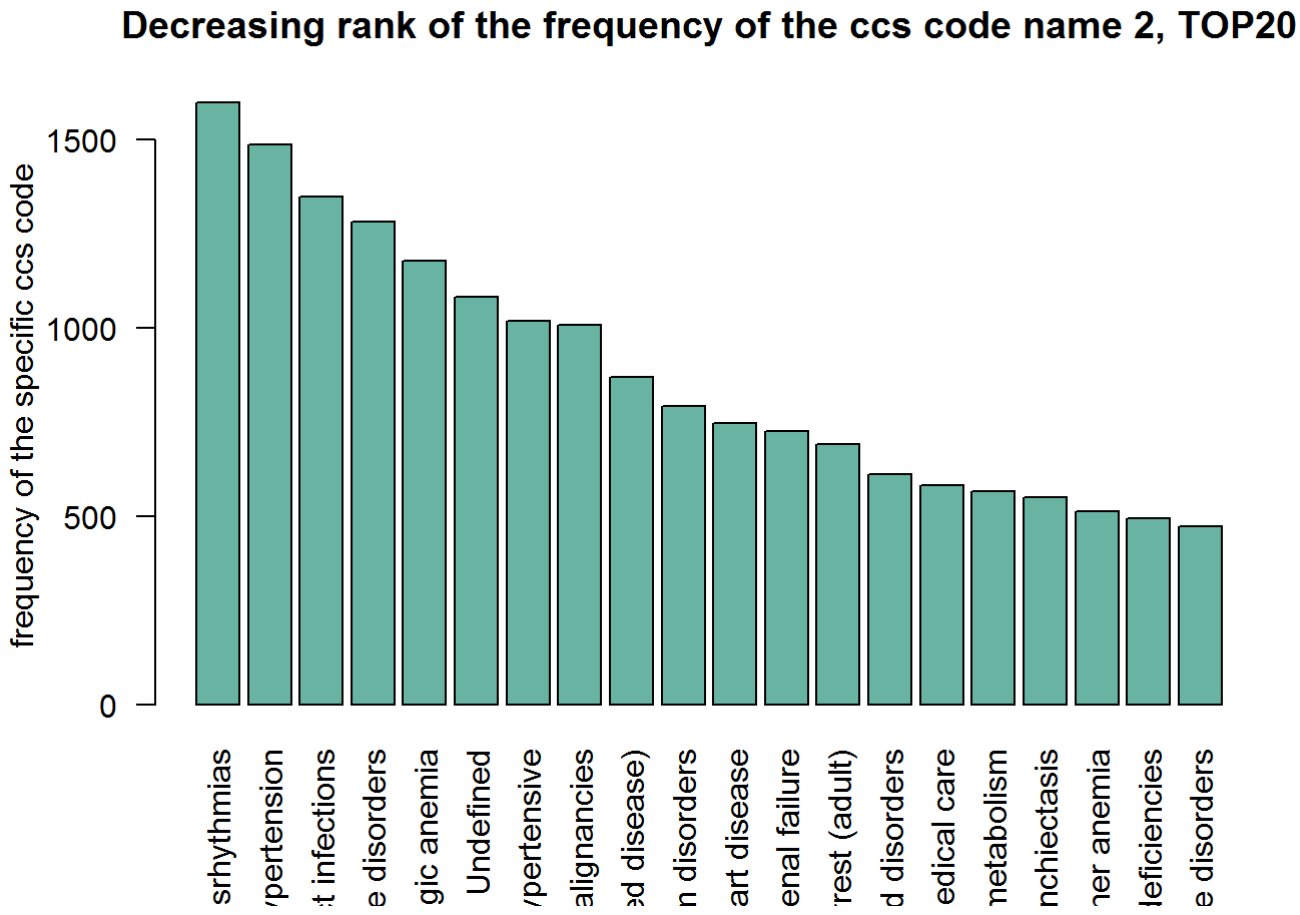

Relationship between ccs code 1 and death time window
window categ
1.00
0.75
die_time_window
did not die through the study
die in 5 years

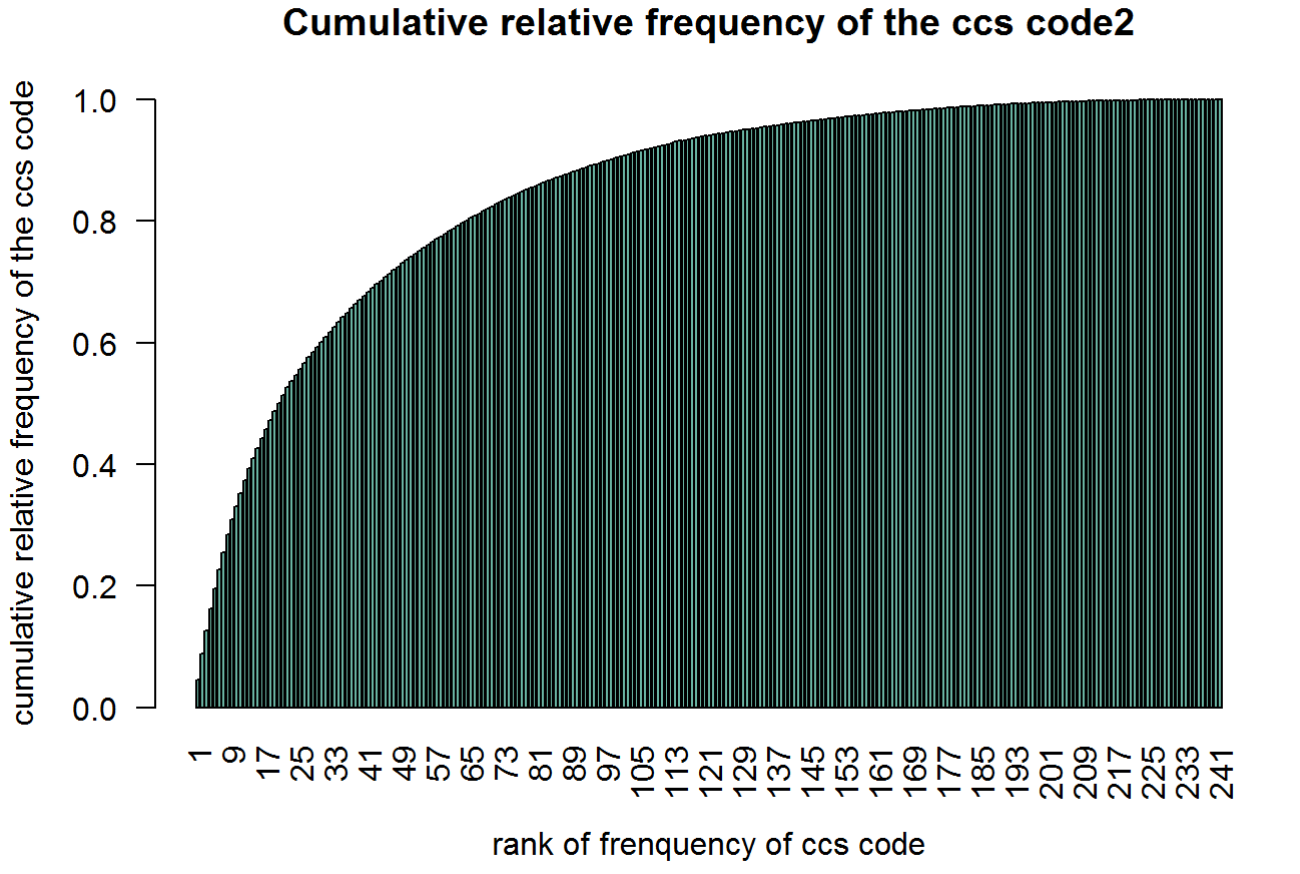


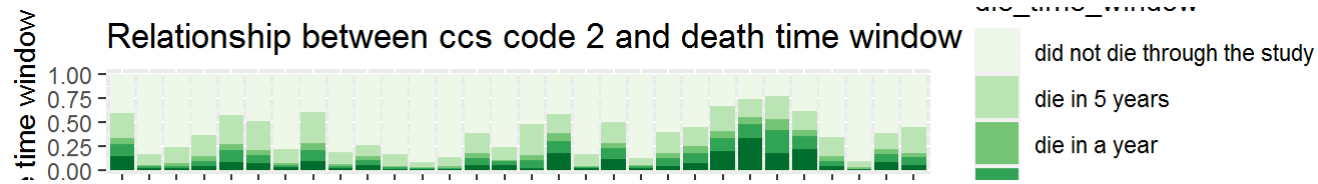
CCS code name2: diag_ccs2

```
## [1] 241
```

```
## [1] 2709
```


Decreasing rank of the frequency of the ccs code name 2, TOP20
frequency of the specific ccs code
1500
1000
500
0
srhythmias
rpertension
t infections
e disorders
gic anemia
Undefined
rpertensive
alignancies
d disease)
n disorders
art disease
enal failure
rest (adult)
d disorders
edical care
netabolism
nchiectasis
ner anemia
leficiencies
e disorders

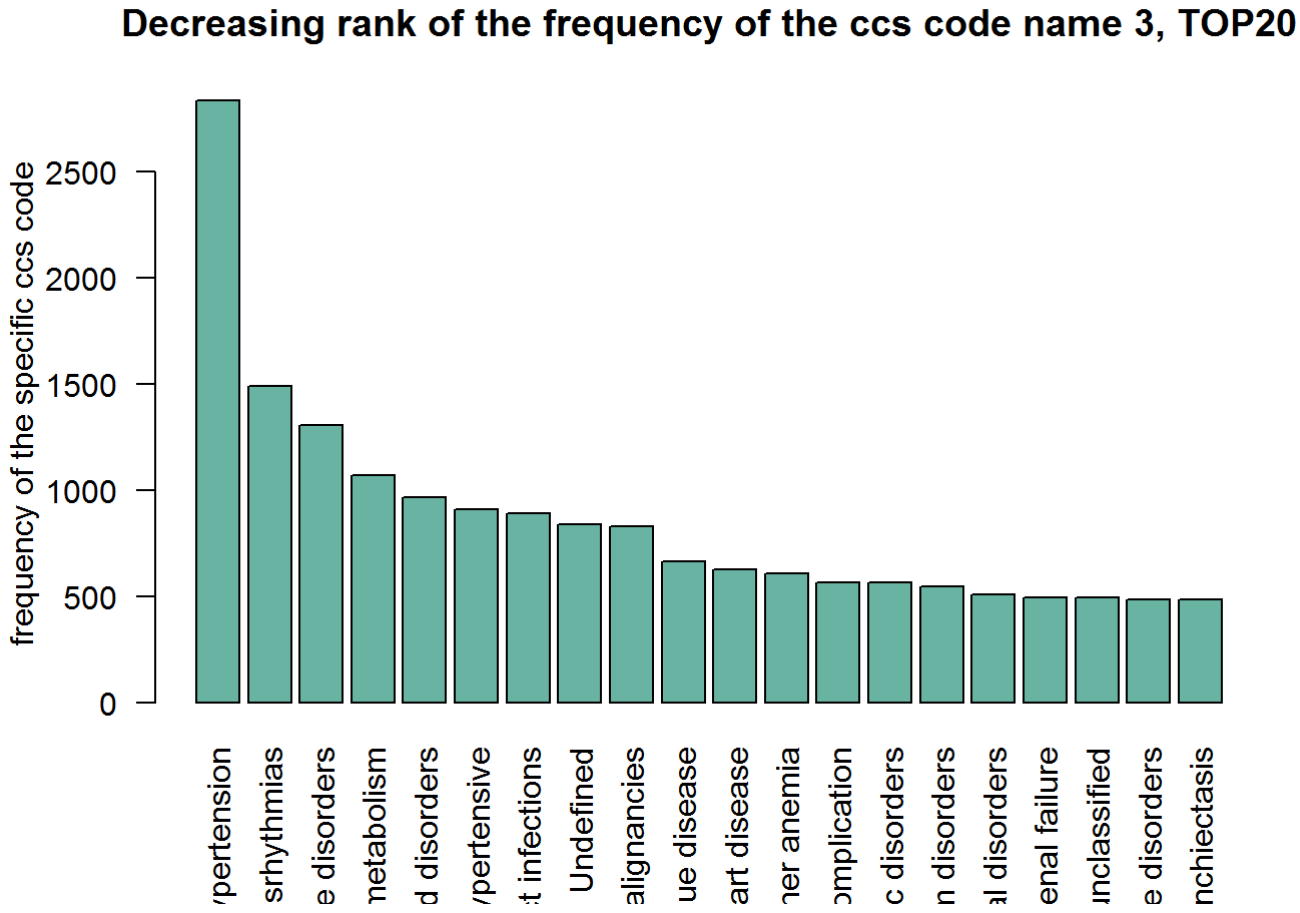

Cumulative relative frequency of the ccs code2
cumulative relative frequency of the ccs code
1.0
0.8
0.6
0.4
0.2
0.0
1
9
17
25
33
41
49
57
65
73
81
89
97
105
113
121
129
137
145
153
161
169
177
185
193
201
209
217
225
233
241
rank of frenquency of ccs code

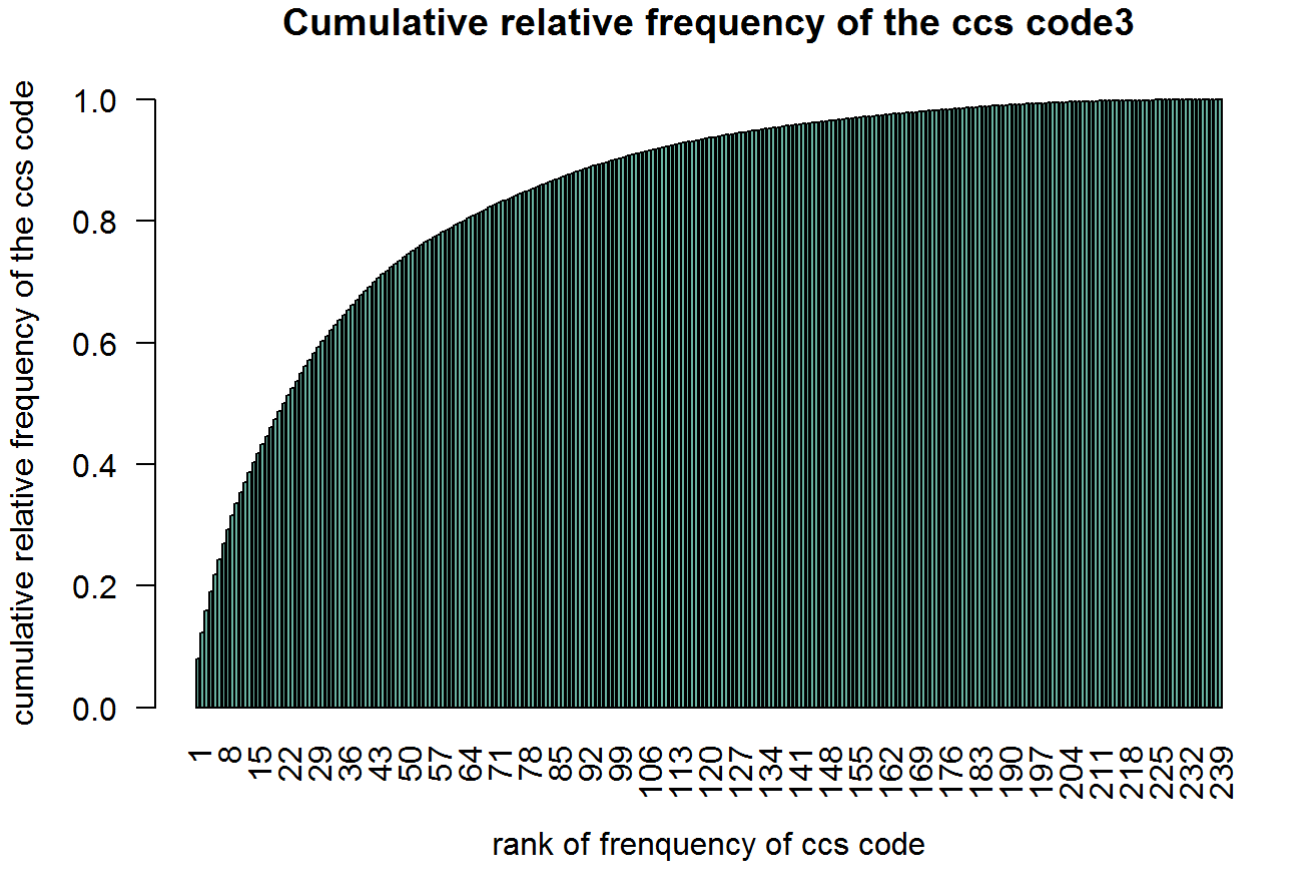

Relationship between ccs code 2 and death time window
time window
1.00
0.75
0.50
0.25
0.00
did not die through the study
die in 5 years
die in a year

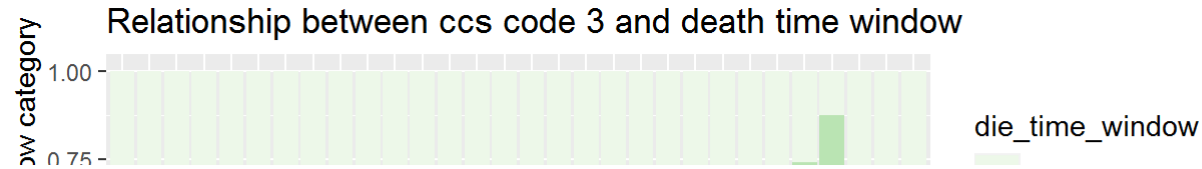
CCS code name3: diag_ccs3

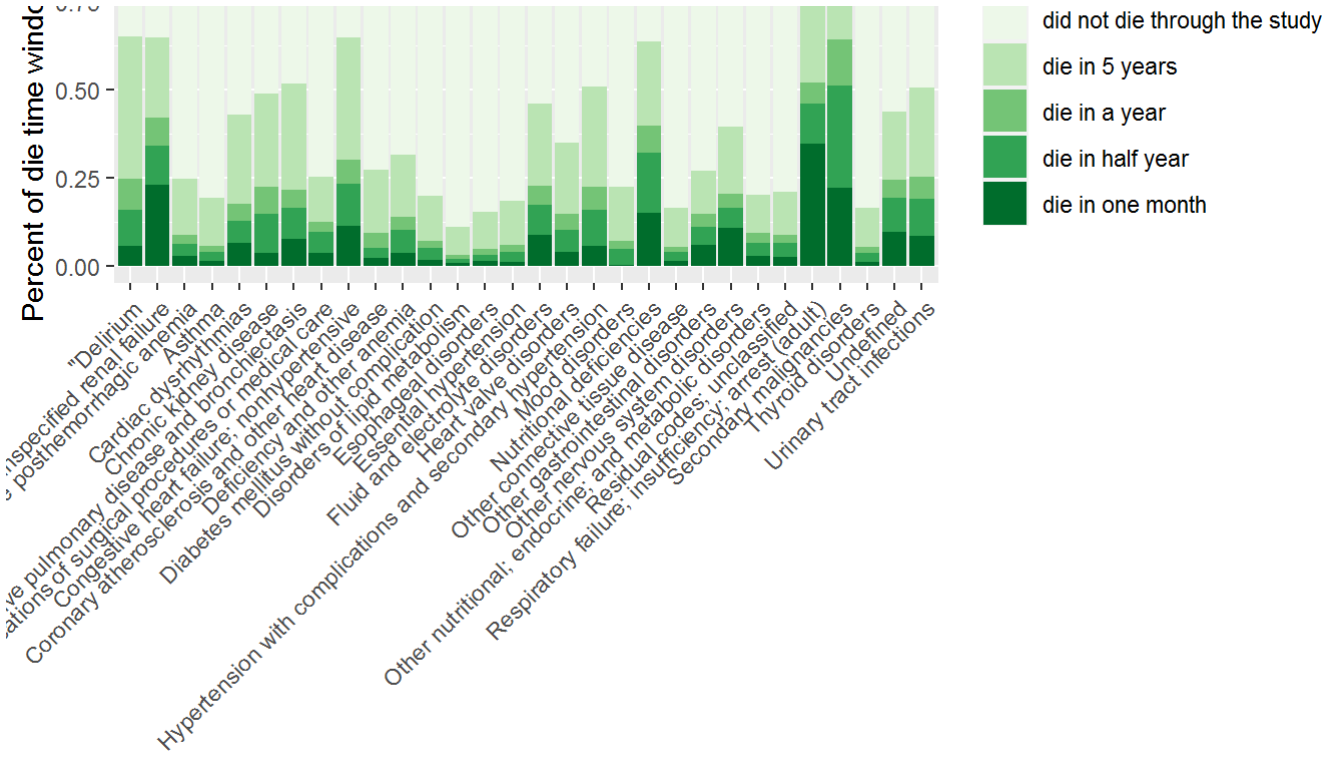
```
## [1] 239
```

```
## [1] 2725
```


Decreasing rank of the frequency of the ccs code name 3, TOP20
frequency of the specific ccs code
0
500
1000
1500
2000
2500
pertension
srhythmias
e disorders
netabolism
d disorders
pertensive
t infections
Undefined
alignancies
ue disease
art disease
ner anemia
omplication
c disorders
n disorders
al disorders
enal failure
nclassified
e disorders
nchiectasis


Cumulative relative frequency of the ccs code3
cumulative relative frequency of the ccs code
0.0
0.2
0.4
0.6
0.8
1.0
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
204
211
218
225
232
239
rank of frenquency of ccs code

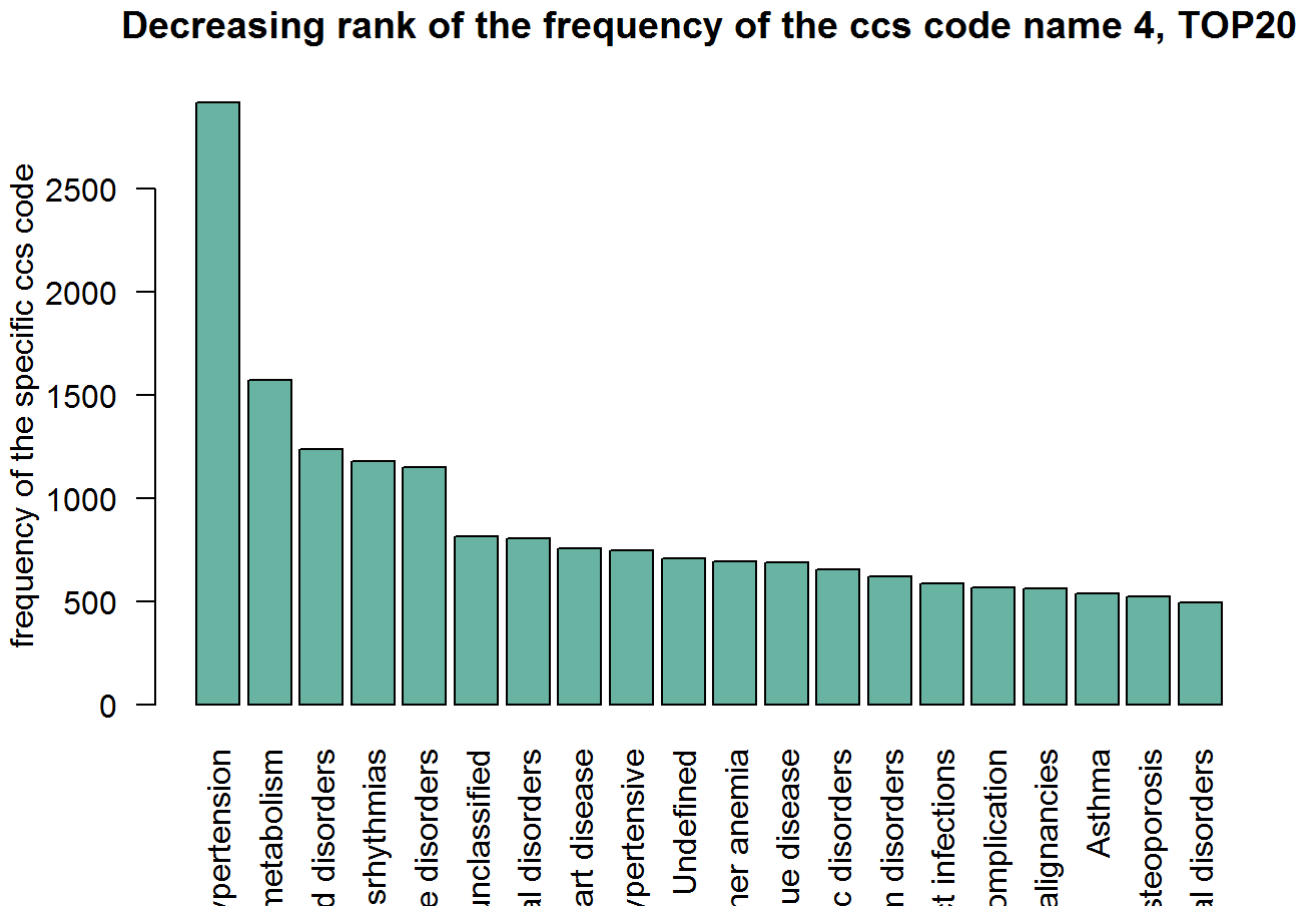

Relationship between ccs code 3 and death time window
ow category
1.00
0.75
die_time_window

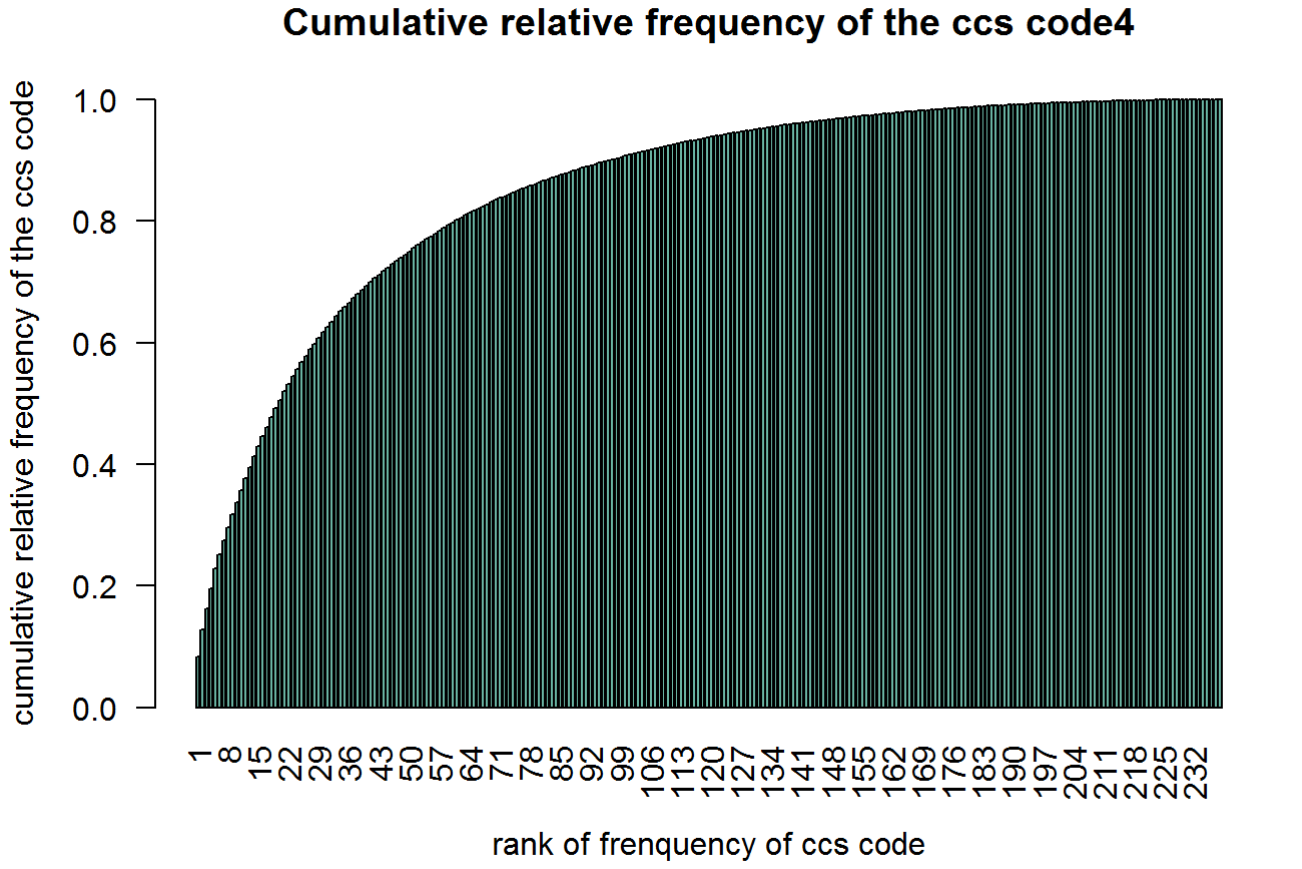


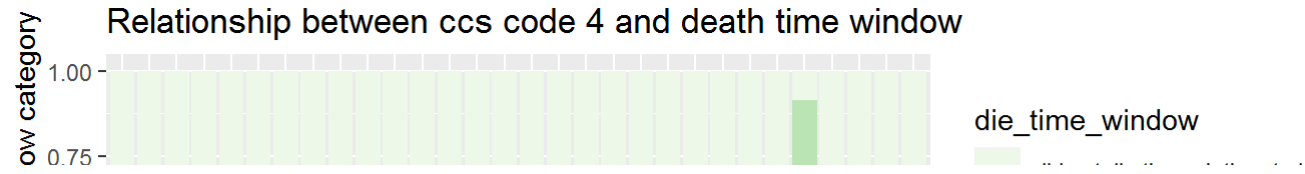
CCS code name4: diag_ccs4

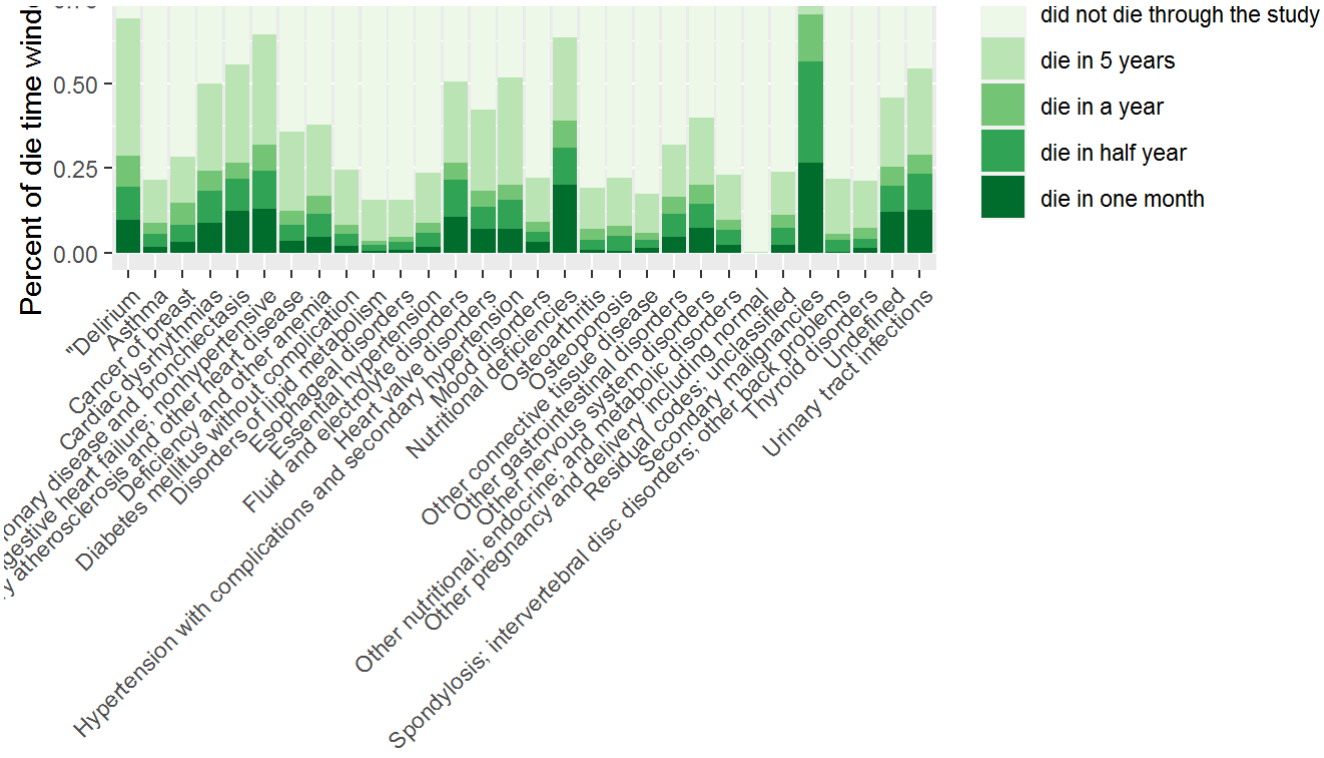
```
## [1] 238
```

```
## [1] 2653
```


Decreasing rank of the frequency of the ccs code name 4, TOP20
frequency of the specific ccs code
0
500
1000
1500
2000
2500
pertension
metabolism
d disorders
srhythmias
e disorders
nclassified
al disorders
art disease
pertensive
Undefined
ner anemia
ue disease
c disorders
n disorders
t infections
omplication
alignancies
Asthma
steoporosis
al disorders

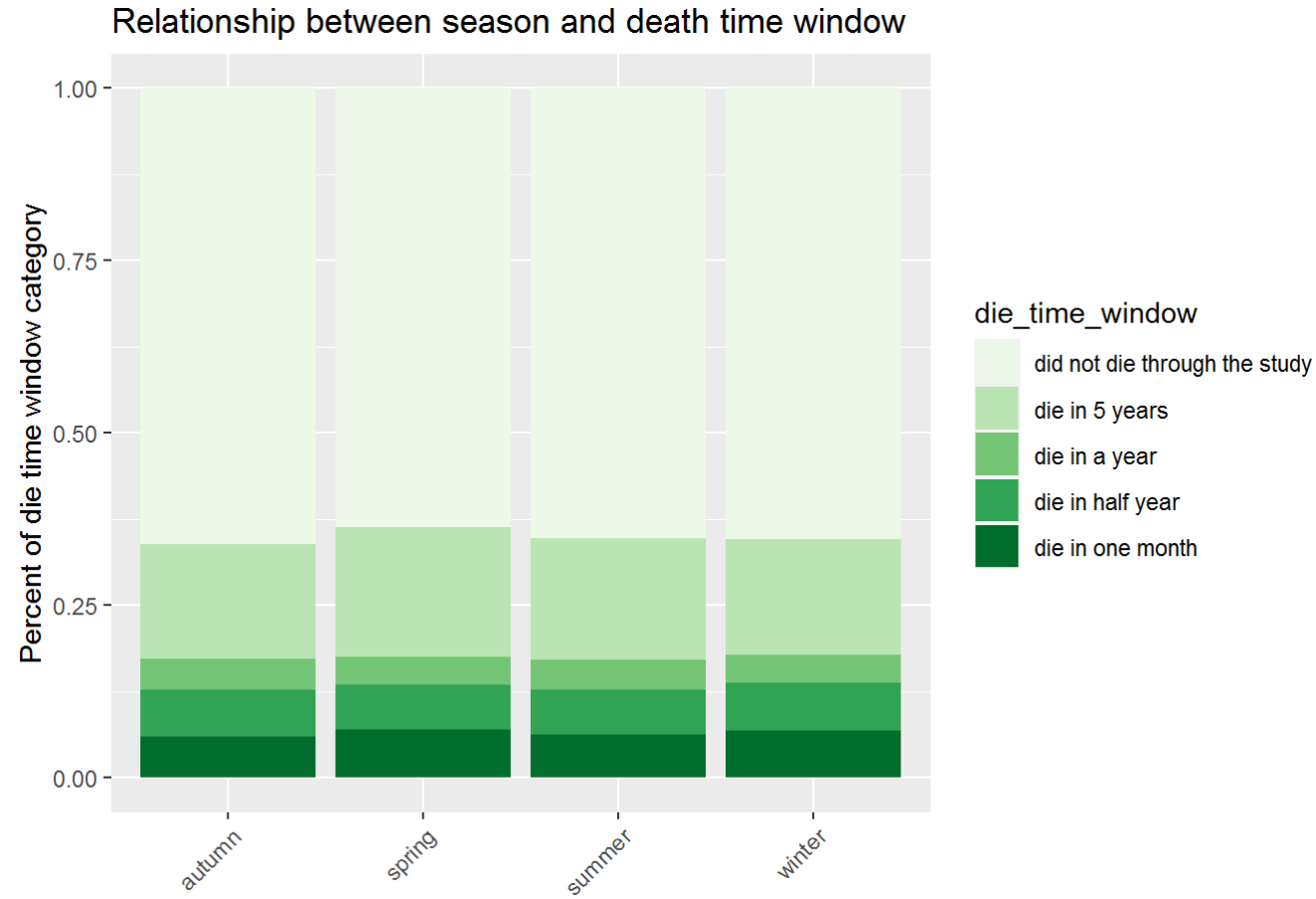

Cumulative relative frequency of the ccs code4
cumulative relative frequency of the ccs code
0.0
0.2
0.4
0.6
0.8
1.0
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
204
211
218
225
232
rank of frenquency of ccs code


Relationship between ccs code 4 and death time window
ow category
1.00
0.75
die_time_window



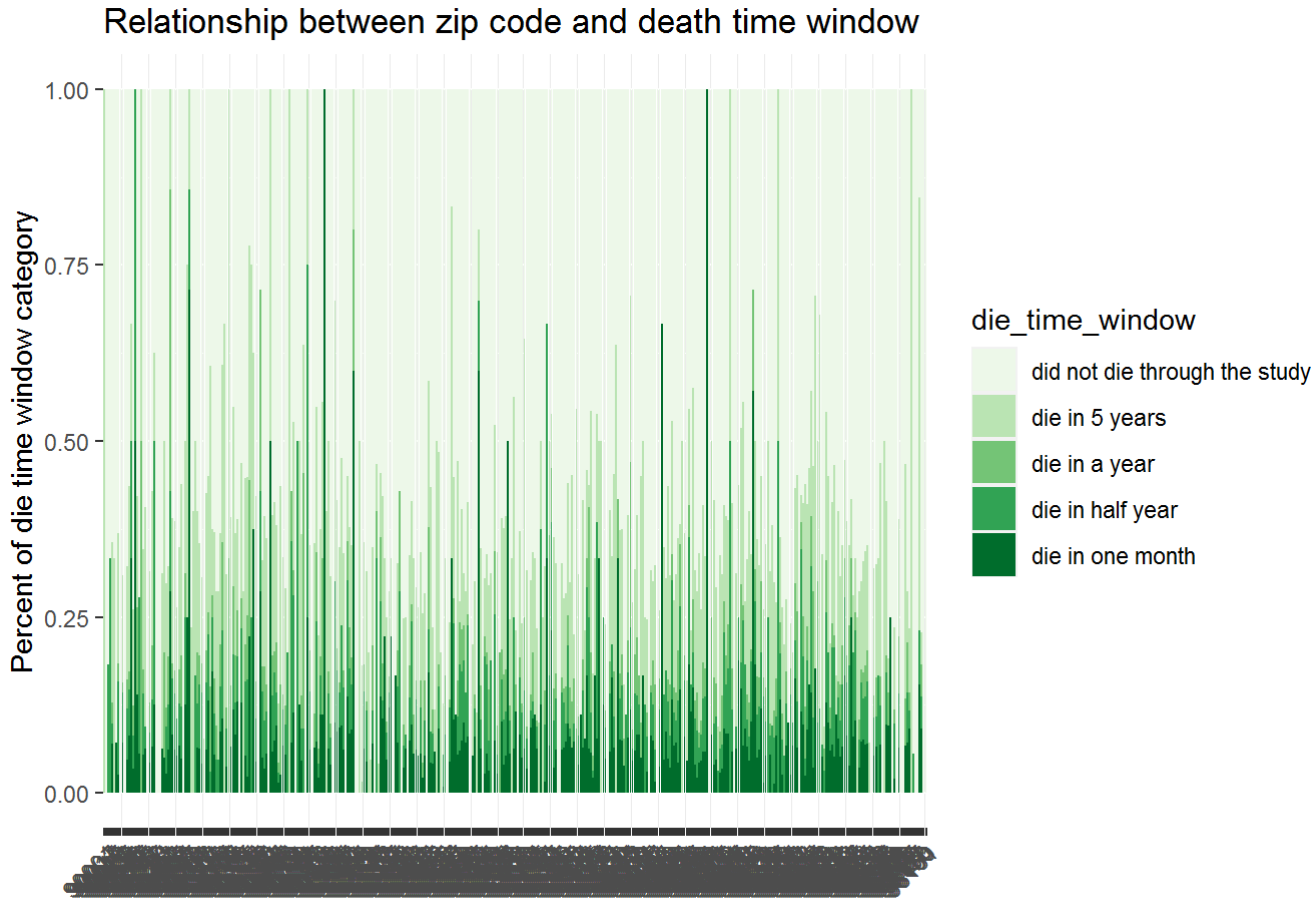
Relationship between season and death time window

It seems that season doesn't have a significant influence on death time window distribution



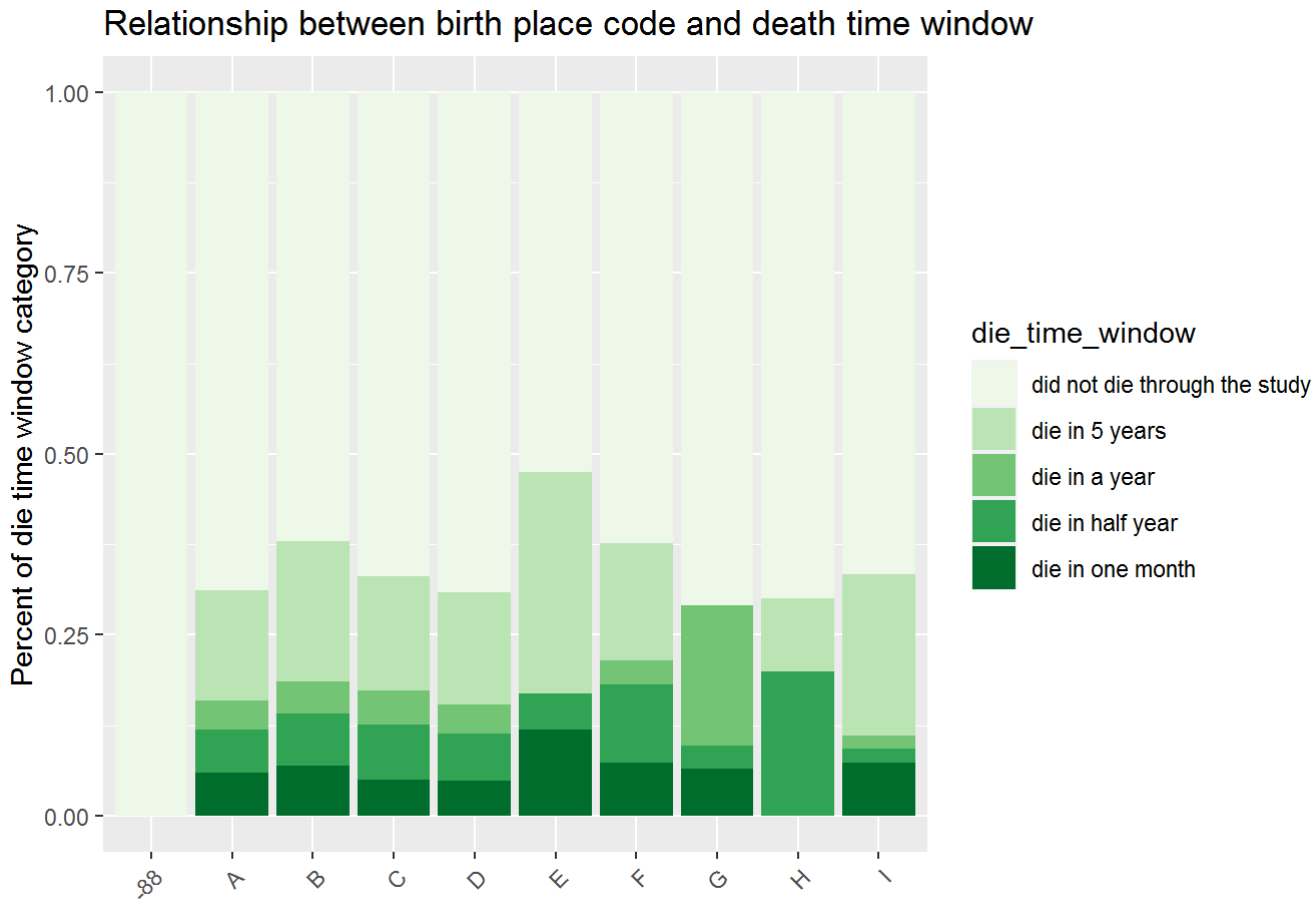
Relationship between zip code and death time window

We can find that in different zip code areas, the death time window distributions are quite different



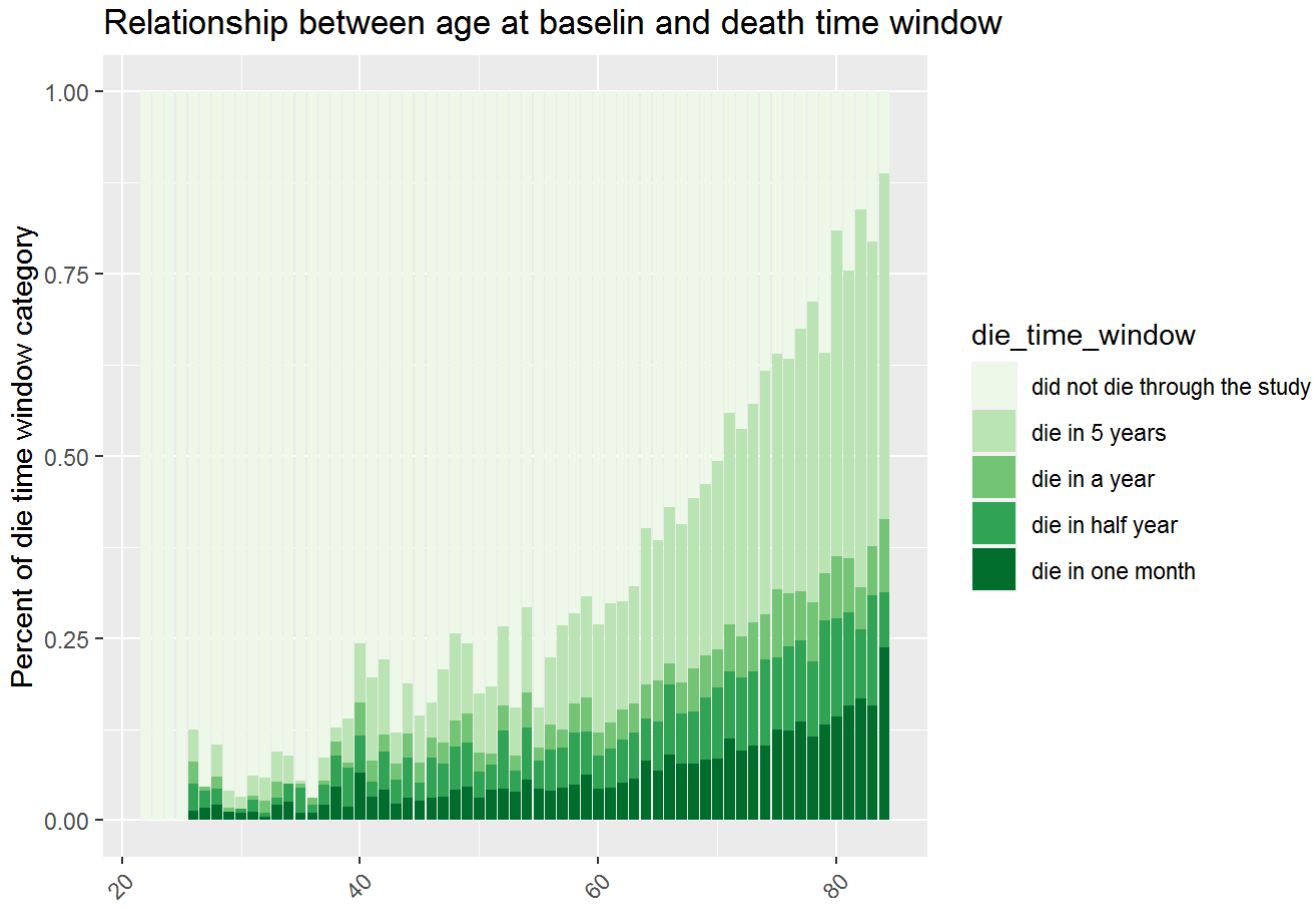
Relationship between birth place and death time window

It seems that birth place have a significant influence on death time window



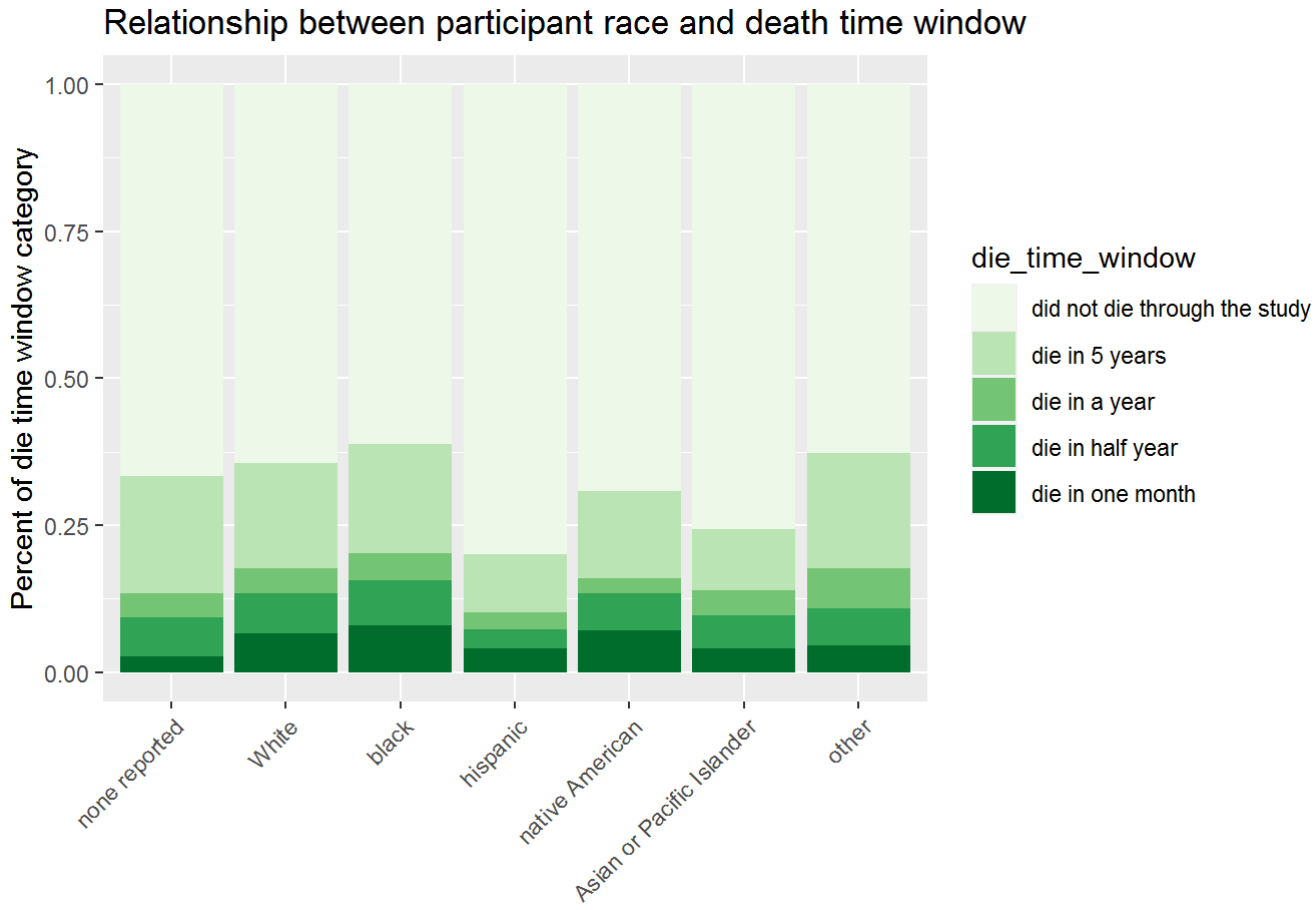
Relationship between age and death time window

It seems that age have a significant influence on death time window. The older, the more risk of death



Relationship between participant race and death time window

It seems that participant race have a significant influence on death time window distributio. The African Americans have the biggest risk of death.



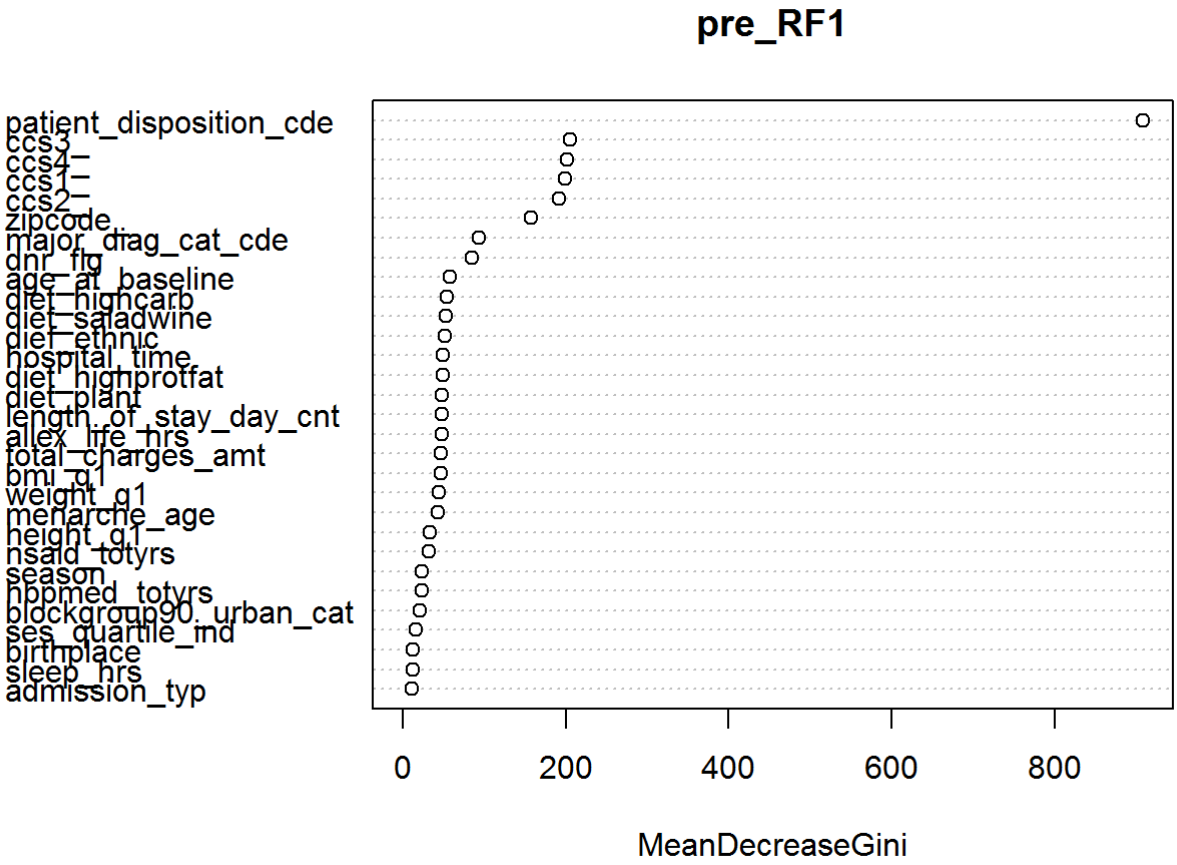
Build and select models for predicting die30

(die30 indicates whether the individual die in 30 days after discharge)

I will use four models: random forest, LASSO logistic regression, boosting and SVM. Then I will select the best model. I will also use that model to predict die1800. For some variables with more than 50 categories, I only keep the 50 most frequent categories and mark other categories as “other”.

Random forest

In this model, the most important predictors are patient_disposition_cde, CCS code 1-4 and zip code. The AUC is 0.7328, not very good. There are two reason: First, random forest is not inherently a good model. Second, the model delivered by randomForest cannot provide the possibility, only 0 or 1, which doesn’t have advantage in AUC.


pre_RF1
patient_disposition_cde
ccs3
ccs4
ccs1
ccs2
zipcode
major_diag_cat_cde
dnr_flg
age_at_baseline
diet_highcarb
diet_saladwine
diet_ethnic
hospital_time
diet_highprotfat
diet_plant
length_of_stay_day_cnt
allex_life_hrs
total_charges_amt
bmi_q1
weight_q1
menarche_age
height_q1
nsaid_totyrs
season
hppmed_totyrs
blockgroup90_urban_cat
ses_quartile_md
birthplace
sleep_hrs
admission_typ
0
200
400
600
800
MeanDecreaseGini

```
##                             MeanDecreaseGini
## ses_quartile_ind                   16.7414633
## blockgroup90_urban_cat             21.3877469
## hysterectomy_ind                    5.6186164
## bilateral_mastectomy_ind            0.5694043
## bilateral_oophorectomy_ind          5.1737893
## age_at_baseline                    58.0258426
## adopted                             0.9883099
## twin                                1.0059664
## birthplace                         12.2884274
## participant_race                    7.5773925
## menarche_age                       43.9027853
## oralcntr_ever_ql                    8.1147141
## preg_ever_ql                        5.0839590
## height_ql                          33.3848338
## weight_ql                          44.8022885
## bmi_ql                             46.4593427
## allex_life_hrs                     47.6316548
## alchl_analyscat                    11.2206874
## brca                                4.9621503
## mammo_ever_ql                       2.8490895
## hbpmed_totyrs                      23.1278744
## nsaid_totyrs                       31.7869027
## sleep_hrs                          12.0764941
## diet_plant                         48.8326855
## diet_highprotfat                   49.1342059
## diet_highcarb                      54.1840777
## dief_ethnic                        51.4565758
## diet_saladwine                     52.9618185
## smoke_statcat                       9.7935775
## asthma_q3                           3.9398022
## insulin_daily                       2.1858167
## aceinhb_daily                       4.0282047
## othhbp_daily                        5.4063048
## tamox_daily                         4.2321221
## steroid_daily                       3.5832532
## brondil_daily                       4.1292450
## cholmed_daily                       4.8079272
## antidep_daily                       4.0523751
## admission_typ                      11.7792268
## length_of_stay_day_cnt             47.6371265
## dnr_flg                            85.1105531
## major_diag_cat_cde                 93.7307453
## patient_care_typ                    4.7807270
## patient_disposition_cde           908.4115643
## total_charges_amt                  46.8404234
## diag_poa1                           2.2138213
## diag_poa2                           6.8399736
## diag_poa3                           6.3980379
## diag_poa4                           5.4213020
## hospital_time                      49.3259845
## season                             23.8409083
## ccs1_                             199.1781398
## ccs2_                             191.4893347
## ccs3_                             205.1382918
## ccs4_                             201.3648095
## zipcode_                          158.0996676
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
## Setting levels: control = 1, case = 2
```

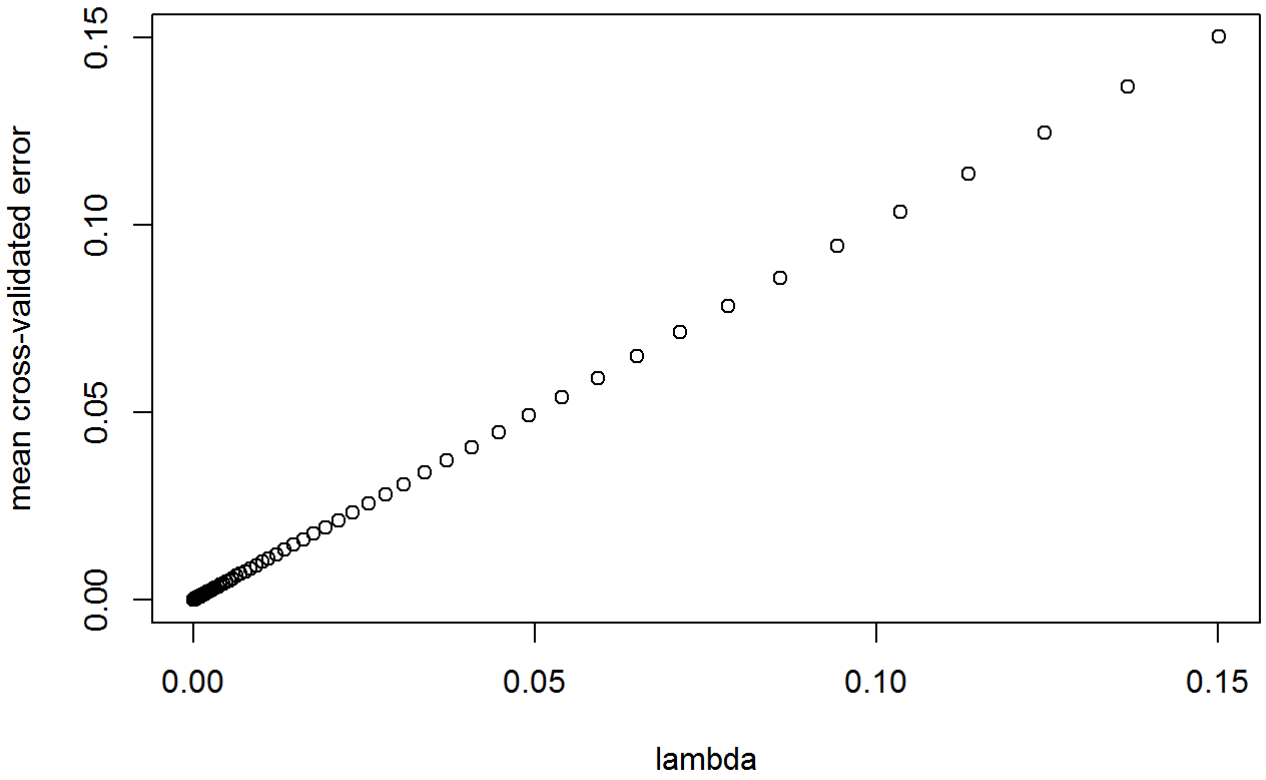
```
## Setting direction: controls < cases
```

```
## [1] "AUC for LASSO in the test set:"
```

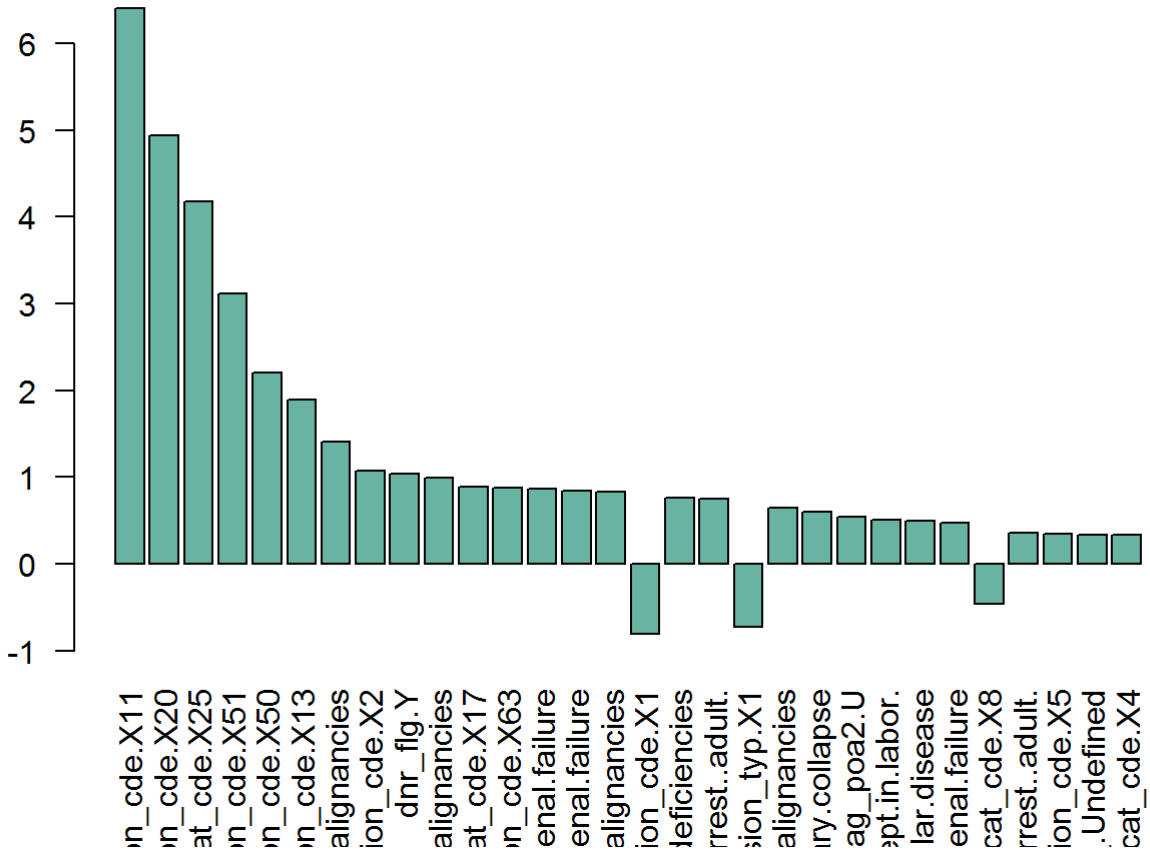
```
## Area under the curve: 0.7356
```

LASSO logistic classification

The optimal lambda is 0.0008202554, so it is quite similiar to the logistic regression without any penalty on more variables. And the most important predictors are the dummy variables from patient_disposition_cde. LASSO logistic regression have a big advantage: It can provide the Beta of every category of factor variables. However, it can not provide the p-value of the Beta. The AUC of this model is 0.9212, which is considerably good.



```
##                                                 names       Beta
## 1                            patient_disposition_cde.X11  6.4072442
## 2                            patient_disposition_cde.X20  4.9337482
## 3                                 major_diag_cat_cde.X25  4.1710452
## 4                            patient_disposition_cde.X51  3.1185925
## 5                            patient_disposition_cde.X50  2.1982525
## 6                            patient_disposition_cde.X13  1.8960046
## 7                          ccs1_.Secondary.malignancies  1.4033413
## 8                            patient_disposition_cde.X2  1.0710238
## 9                                               dnr_flg.Y  1.0356694
## 10                         ccs4_.Secondary.malignancies  0.9863471
## 11                                major_diag_cat_cde.X17  0.8886343
## 12                           patient_disposition_cde.X63  0.8790254
## 13          ccs1_.Acute.and.unspecified.renal.failure  0.8633028
## 14          ccs4_.Acute.and.unspecified.renal.failure  0.8431186
## 15                         ccs2_.Secondary.malignancies  0.8329342
## 16                           patient_disposition_cde.X1 -0.8042216
## 17                       ccs2_.Nutritional.deficiencies  0.7558373
## 18 ccs1_.Respiratory.failure..insufficiency..arrest..adult.  0.7458320
## 19                                        admission_typ.X1 -0.7306056
## 20                         ccs3_.Secondary.malignancies  0.6507283
```




```
## [1] "optimal lambda:"
```

```
## [1] 0.0009002293
```

```
## Setting levels: control = 1, case = 2
```

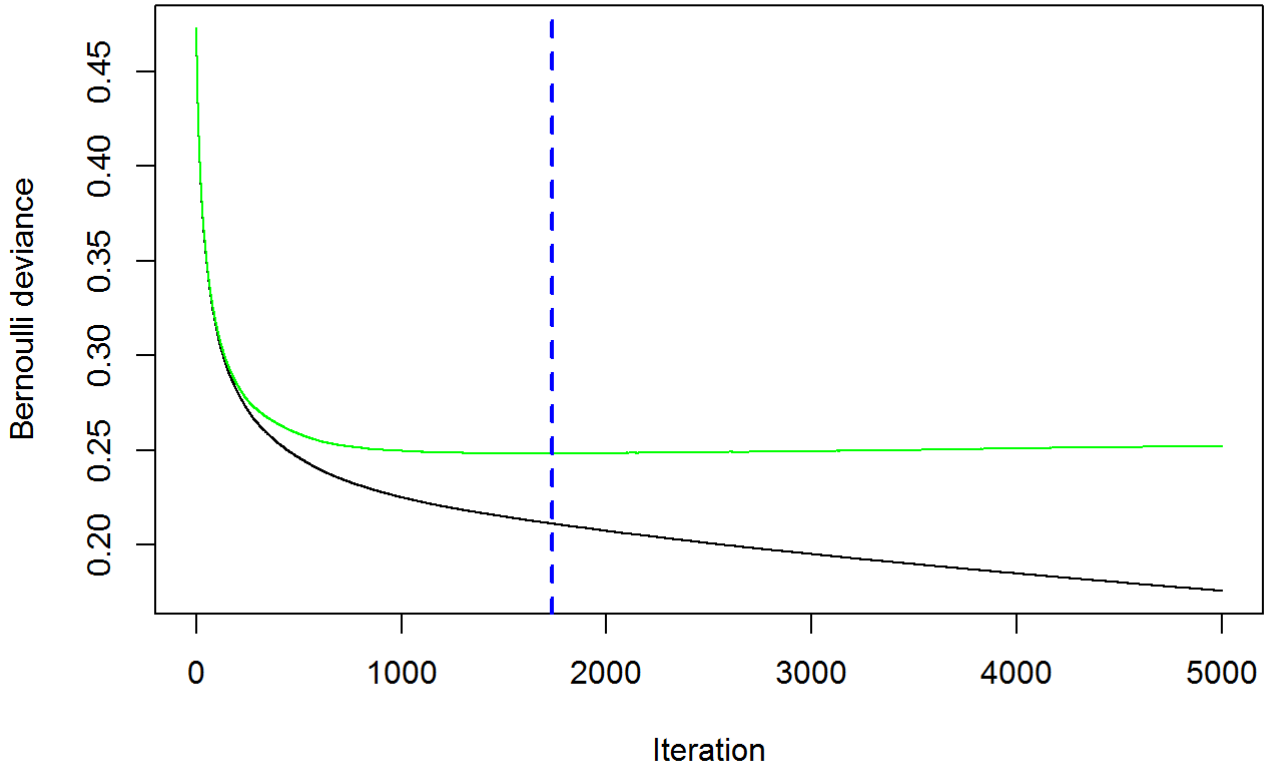
```
## Setting direction: controls < cases
```

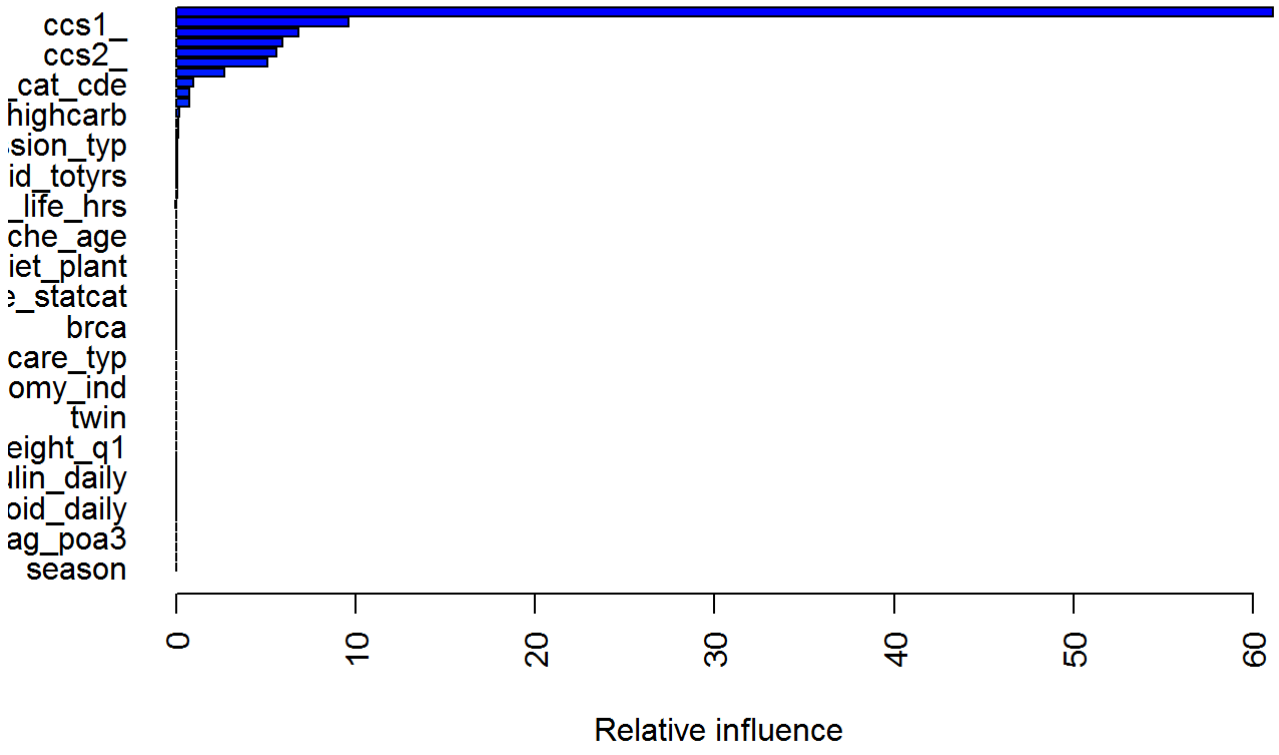
```
## [1] "AUC for LASSO in the test set:"
```

```
## Area under the curve: 0.9212
```

Boosting

The optimal tree number is 1524 and the most important predictors are patient_disposition_cde, CCS code1-4 and zip code, quite similiar to rabdom forest. But this model is much more accurate than random forest because its AUC is 0.9268.


Bernoulli deviance
0.20
0.25
0.30
0.35
0.40
0.45
0
1000
2000
3000
4000
5000
Iteration


ccs1_
ccs2_l
_cat_cde
highcarb
sion_typ
id_totyrs
_life_hrs
che_age
iet_plant
e_statcat
brca
care_typ
omy_ind
twin
eight_q1
ılin_daily
oid_daily
ag_poa3
season
0
10
20
30
40
50
60
Relative influence

```
##                                             var      rel.inf
## patient_disposition_cde       patient_disposition_cde 61.103784024
## ccs1_                                       ccs1_  9.594899812
## ccs3_                                       ccs3_  6.799722234
## ccs4_                                       ccs4_  5.929404956
## ccs2_                                       ccs2_  5.591591298
## zipcode_                                 zipcode_  5.079880705
## dnr_flg                                   dnr_flg  2.670469284
## major_diag_cat_cde               major_diag_cat_cde  0.966645063
## length_of_stay_day_cnt         length_of_stay_day_cnt  0.717910883
## age_at_baseline                 age_at_baseline  0.714453113
## diet_highcarb                     diet_highcarb  0.150305244
## diet_saladwine                   diet_saladwine  0.098389547
## sleep_hrs                           sleep_hrs  0.097658224
## admission_typ                     admission_typ  0.078648741
## bmi_q1                               bmi_q1  0.076504809
## total_charges_amt                total_charges_amt  0.050723350
## nsaid_totyrs                      nsaid_totyrs  0.047936626
## diag_poa2                            diag_poa2  0.037275287
## weight_q1                            weight_q1  0.035253287
## allex_life_hrs                   allex_life_hrs  0.033388595
## diet_highprotfat                 diet_highprotfat  0.026692564
## birthplace                         birthplace  0.023478619
## menarche_age                     menarche_age  0.020719332
## dief_ethnic                       dief_ethnic  0.019460037
## tamox_daily                       tamox_daily  0.005372312
## diet_plant                         diet_plant  0.005272608
## alchl_analyscat                 alchl_analyscat  0.004956879
## ses_quartile_ind                ses_quartile_ind  0.004685884
## smoke_statcat                     smoke_statcat  0.002912967
## hbpmed_totyrs                     hbpmed_totyrs  0.002760106
## brondil_daily                     brondil_daily  0.002388429
## brca                                       brca  0.002316795
## cholmed_daily                     cholmed_daily  0.001604349
## participant_race                participant_race  0.001487529
## patient_care_typ                patient_care_typ  0.001046506
## blockgroup90_urban_cat    blockgroup90_urban_cat  0.000000000
## hysterectomy_ind                hysterectomy_ind  0.000000000
## bilateral_mastectomy_ind  bilateral_mastectomy_ind  0.000000000
## bilateral_oophorectomy_ind bilateral_oophorectomy_ind  0.000000000
## adopted                                 adopted  0.000000000
## twin                                       twin  0.000000000
## oralcntr_ever_q1               oralcntr_ever_q1  0.000000000
## preg_ever_q1                     preg_ever_q1  0.000000000
## height_q1                           height_q1  0.000000000
## mammo_ever_q1                   mammo_ever_q1  0.000000000
## asthma_q3                           asthma_q3  0.000000000
## insulin_daily                   insulin_daily  0.000000000
## aceinhb_daily                   aceinhb_daily  0.000000000
## othhbp_daily                     othhbp_daily  0.000000000
## steroid_daily                   steroid_daily  0.000000000
## antidep_daily                   antidep_daily  0.000000000
## diag_poa1                           diag_poa1  0.000000000
## diag_poa3                           diag_poa3  0.000000000
## diag_poa4                           diag_poa4  0.000000000
## hospital_time                   hospital_time  0.000000000
## season                                 season  0.000000000
```

```
## [1] 1734
```

```
## Using 1734 trees...
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## [1] "AUC for LASSO in the test set:"
```

```
## Area under the curve: 0.9264
```

SVM

The optimal C value is 4. The AUC is only 0.7238, not very good.

```
## [Tune] Started tuning learner classif.ksvm for parameter set:
```

```
##       Type len Def  Constr Req Tunable Trafo
## C discrete   -   - 3,4,5,6   -    TRUE     -
```

```
## With control class: TuneControlGrid
```

```
## Imputation value: 1
```

```
## [Tune-x] 1: C=3
```

```
## [Tune-y] 1: mmce.test.mean=0.0385006; time: 8.1 min
```

```
## [Tune-x] 2: C=4
```

```
## [Tune-y] 2: mmce.test.mean=0.0387034; time: 8.0 min
```

```
## [Tune-x] 3: C=5
```

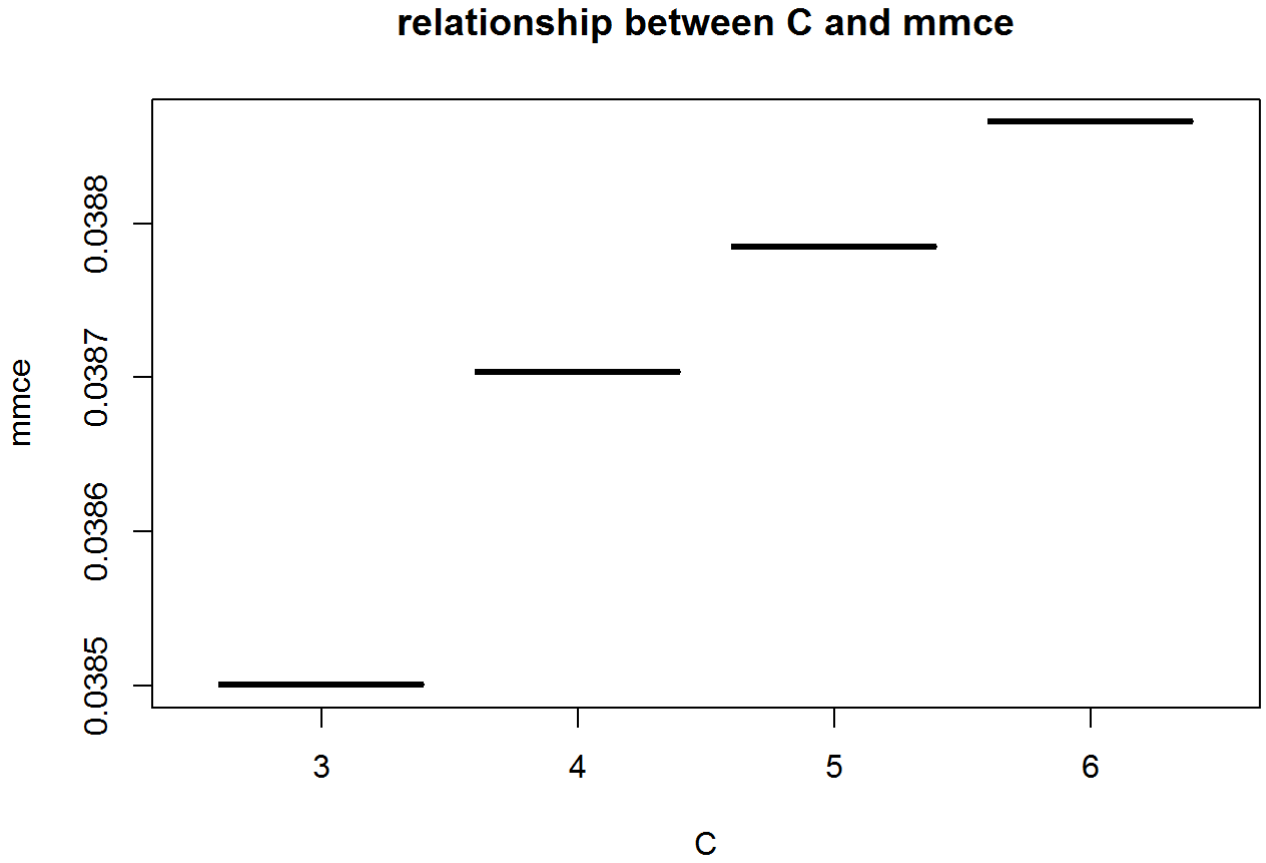
```
## [Tune-y] 3: mmce.test.mean=0.0387846; time: 8.1 min
```

```
## [Tune-x] 4: C=6
```

```
## [Tune-y] 4: mmce.test.mean=0.0388657; time: 8.5 min
```

```
## [Tune] Result: C=3 : mmce.test.mean=0.0385006
```

```
## [1] 3
```




```
##  Setting default kernel parameters
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## [1] "AUC for SVM in the test set:"
```

```
## Area under the curve: 0.7238
```

Model selection

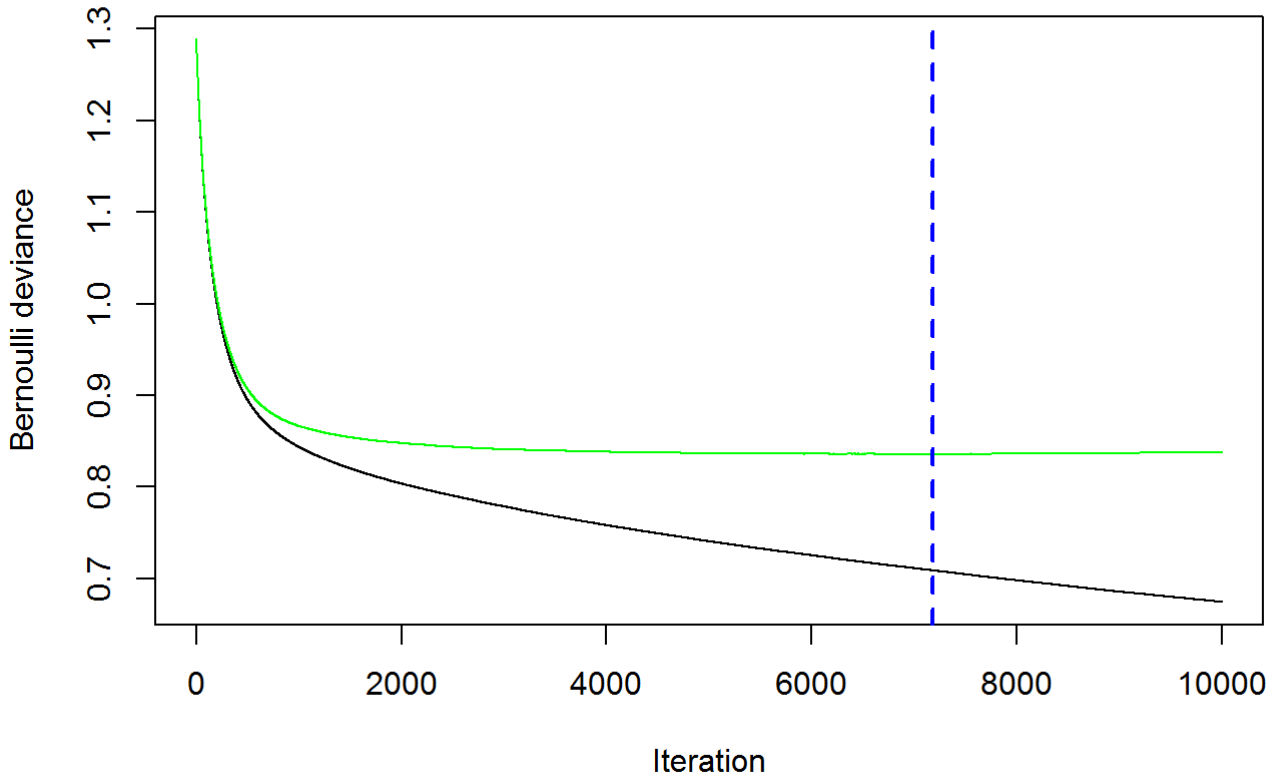
Boosting has the best AUC. And AUC is a very good measurement evaluating the models because it consider not only sensitivity but also specificity

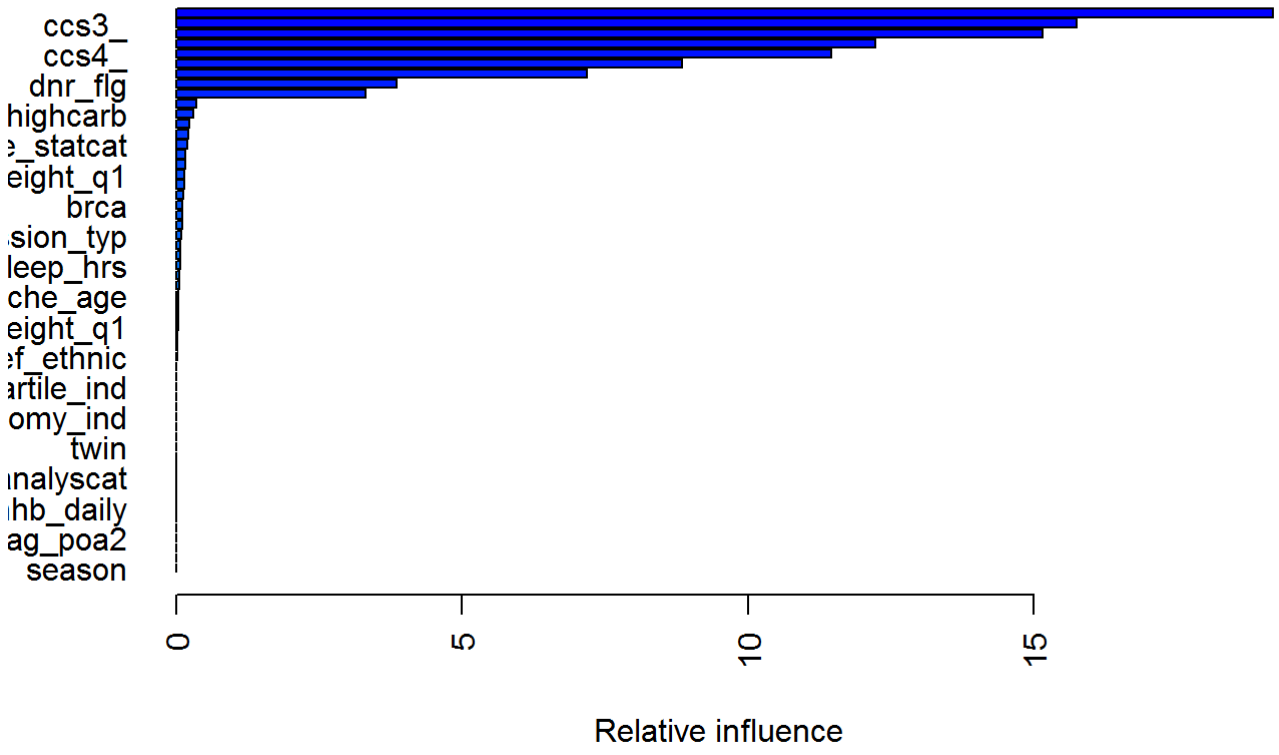
Boosting for die1800

(die1800 indicates whether the individual die in 1800 days after discharge)

I use boosting because it is the best method when predicting die30 and the model here use the same potential predictors. What's more, according to my experience, boosting is usually the best model among the four above. Maybe because of the magic of iterations. The optimal tree number is 7504, and the most important

predictors are CCS code 1-4, zip code and age_at_baseline The AUC of this model is 0.8747, which is considerably good.





```
##                                             var      rel.inf
## ccs1_                                     ccs1_ 19.185497876
## ccs3_                                     ccs3_ 15.751861861
## ccs2_                                     ccs2_ 15.163238825
## age_at_baseline                 age_at_baseline 12.236165820
## ccs4_                                     ccs4_ 11.467086200
## zipcode_                               zipcode_  8.841909603
## patient_disposition_cde patient_disposition_cde  7.192631006
## dnr_flg                                 dnr_flg  3.864884021
## major_diag_cat_cde           major_diag_cat_cde  3.320398701
## hbpmed_totyrs                   hbpmed_totyrs  0.357107085
## diet_highcarb                   diet_highcarb  0.300122182
## steroid_daily                   steroid_daily  0.230229567
## length_of_stay_day_cnt   length_of_stay_day_cnt  0.204767133
## smoke_statcat                   smoke_statcat  0.197486421
## diet_highprotfat             diet_highprotfat  0.161533873
## insulin_daily                   insulin_daily  0.157786946
## weight_q1                         weight_q1  0.150815326
## participant_race             participant_race  0.137079890
## allex_life_hrs                 allex_life_hrs  0.131342005
## brca                                     brca  0.111961191
## bmi_q1                                 bmi_q1  0.107129812
## diet_saladwine                 diet_saladwine  0.101440162
## admission_typ                   admission_typ  0.093558972
## blockgroup90_urban_cat   blockgroup90_urban_cat  0.069813331
## nsaid_totyrs                     nsaid_totyrs  0.065332789
## sleep_hrs                           sleep_hrs  0.063960929
## diet_plant                         diet_plant  0.057238518
## othhbp_daily                     othhbp_daily  0.047620069
## menarche_age                     menarche_age  0.043762707
## tamox_daily                       tamox_daily  0.040640254
## patient_care_typ             patient_care_typ  0.033492885
## height_q1                         height_q1  0.029141833
## birthplace                       birthplace  0.027383921
## total_charges_amt           total_charges_amt  0.020289031
## dief_ethnic                       dief_ethnic  0.019722639
## diag_poa1                           diag_poa1  0.005126787
## diag_poa3                           diag_poa3  0.003244455
## ses_quartile_ind             ses_quartile_ind  0.003030034
## bilateral_mastectomy_ind bilateral_mastectomy_ind  0.002869872
## antidep_daily                   antidep_daily  0.001295469
## hysterectomy_ind             hysterectomy_ind  0.000000000
## bilateral_oophorectomy_ind bilateral_oophorectomy_ind  0.000000000
## adopted                                 adopted  0.000000000
## twin                                       twin  0.000000000
## oralcntr_ever_q1             oralcntr_ever_q1  0.000000000
## preg_ever_q1                     preg_ever_q1  0.000000000
## alchl_analyscat                 alchl_analyscat  0.000000000
## mammo_ever_q1                   mammo_ever_q1  0.000000000
## asthma_q3                         asthma_q3  0.000000000
## aceinhb_daily                   aceinhb_daily  0.000000000
## brondil_daily                   brondil_daily  0.000000000
## cholmed_daily                   cholmed_daily  0.000000000
## diag_poa2                           diag_poa2  0.000000000
## diag_poa4                           diag_poa4  0.000000000
## hospital_time                   hospital_time  0.000000000
## season                                 season  0.000000000
```

```
## [1] 7183
```

```
## Using 7183 trees...
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## [1] "AUC for boosting in the test set:"
```

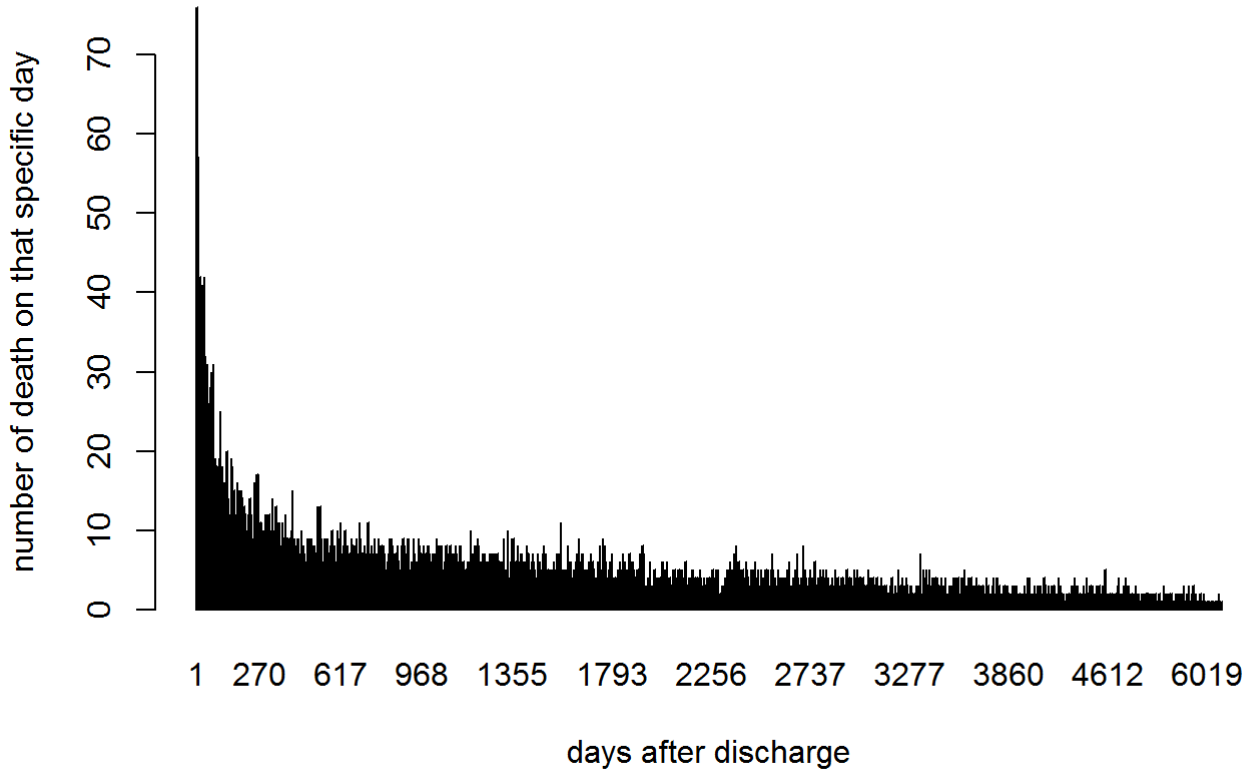
```
## Area under the curve: 0.875
```

Estimate the influence of some variables

I build classic logistic regression models for predicting die30 and die1800, so that we can find the p-values of the Beta of each variable adjusted for other variables. The classic logistic regression shall be considerably accurate because we find that the lambda in LASSO is extremely small.

Association between time window and risk of death

It is obvious that the longer the time window is, the more people die. And risk of death just monotonously increases when the time window become longer. Generally, the farther away from the discharge, the less risk of death on that specific day



CCS code 1

CCS code 1 that is significantly associated with die30

We find that CCS code 1 is significantly associated with die30 The CCS code1 descriptions that have significant association are as follows

```
##                                                   names1     P_value
## 1                              ccs1_Biliary tract disease 0.007032301
## 2 ccs1_Other nutritional; endocrine; and metabolic disorders 0.029816690
## 3                  ccs1_Pancreatic disorders (not diabetes) 0.018741722
## 4                             ccs1_Secondary malignancies 0.025720911
```

CCS code 1 that is significantly associated with die1800

We find that CCS code 1 is significantly associated with die1800 The CCS code1 descriptions that have significant association are as follows

```
##                                                                 names2
## 1                                           ccs1_Biliary tract disease
## 2                           ccs1_Congestive heart failure; nonhypertensive
## 3                     ccs1_Coronary atherosclerosis and other heart disease
## 4                               ccs1_Intestinal obstruction without hernia
## 5                                           ccs1_Nonspecific chest pain
## 6                                                  ccs1_Osteoarthritis
## 7                                                 ccs1_Other aftercare
## 8                                ccs1_Pancreatic disorders (not diabetes)
## 9                                  ccs1_Prolapse of female genital organs
## 10 ccs1_Rehabilitation care; fitting of prostheses; and adjustment of devices
## 11                                          ccs1_Secondary malignancies
##         P_value
## 1  1.037160e-08
## 2  2.116884e-06
## 3  1.234356e-02
## 4  1.159785e-02
## 5  2.444400e-02
## 6  1.711116e-04
## 7  1.041535e-02
## 8  1.943107e-05
## 9  2.067756e-02
## 10 2.212265e-02
## 11 5.060120e-22
```

CCS code 2

CCS code 2 that is significantly associated with die30

We find that CCS code 2 is significantly associated with die30 The CCS code2 descriptions that have significant association are as follows

```
##                                                       names1      P_value
## 1 ccs2_Hypertension with complications and secondary hypertension 0.0351849126
## 2                                 ccs2_Nutritional deficiencies 0.0001661674
## 3                                   ccs2_Secondary malignancies 0.0010898445
```

CCS code 2 that is significantly associated with die1800

We find that CCS code 2 is significantly associated with die1800 The CCS code2 descriptions that have significant association are as follows

```
##
names2
## 1                                                                ccs2_Acute and unspecifi
ed renal failure
## 2                                                                    ccs2_Acute posthe
morrhagic anemia
## 3
ccs2_Asthma
## 4                                                                      ccs2_Benign ne
oplasm of uterus
## 5                                                                          ccs2_Bilia
ry tract disease
## 6                                                                           ccs2_Card
iac dysrhythmias
## 7                                          ccs2_Chronic obstructive pulmonary disease an
d bronchiectasis
## 8                                                                          ccs2_Chron
ic ulcer of skin
## 9                                                        ccs2_Coagulation and hemor
rhagic disorders
## 10                                        ccs2_Complications of surgical procedures
or medical care
## 11                                              ccs2_Congestive heart failure;
nonhypertensive
## 12                                         ccs2_Coronary atherosclerosis and oth
er heart disease
## 13                                                          ccs2_Deficiency
and other anemia
## 14                                                    ccs2_Diabetes mellitus wi
th complications
## 15                                                 ccs2_Diabetes mellitus with
out complication
## 16                                                      ccs2_Disorders of
lipid metabolism
## 17                                                                 ccs2_Esop
hageal disorders
## 18                                                                 ccs2_Essent
ial hypertension
## 19                                                ccs2_Fluid and elect
rolyte disorders
## 20                                                                 ccs2_Heart
valve disorders
## 21                                    ccs2_Hypertension with complications and second
ary hypertension
## 22                                                  ccs2_Intestinal obstructio
n without hernia
## 23                                               ccs2_Late effects of cerebro
vascular disease
## 24                                                                       ccs
2_Mood disorders
## 25                                                                 ccs2_Non-Ho
dgkin`s lymphoma
## 26
ccs2_other
## 27                                                                       ccs2
_Other aftercare
## 28                                                            ccs2_Other connectiv
```

```
e tissue disease
## 29                                                                    ccs2_Other female g
enital disorders
## 30                                                                          ccs2
_Other fractures
## 31                                                              ccs2_Other gastrointe
stinal disorders
## 32                                                                ccs2_Other nervous
system disorders
## 33                                          ccs2_Other nutritional; endocrine; and met
abolic disorders
## 34
ccs2_Paralysis
## 35 ccs2_Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or
sexually transm
## 36                                                    ccs2_Pleurisy; pneumothorax; pu
lmonary collapse
## 37                     ccs2_Pneumonia (except that caused by tuberculosis or sexually tran
smitted disease)
## 38                                                                ccs2_Residual cod
es; unclassified
## 39                                                                      ccs2_Second
ary malignancies
## 40                                                                ccs2_Septicemia
(except in labor)
## 41                                                        ccs2_Skin and subcutaneous t
issue infections
## 42                                  ccs2_Spondylosis; intervertebral disc disorders; oth
er back problems
## 43                                                                          ccs2_T
hyroid disorders
## 44
ccs2_Undefined
## 45                                                                       ccs2_Urinary
tract infections
##         P_value
## 1  1.752441e-03
## 2  1.513566e-08
## 3  1.578661e-04
## 4  1.575927e-03
## 5  4.359832e-03
## 6  8.222056e-10
## 7  4.669983e-03
## 8  2.223726e-02
## 9  5.569710e-04
## 10 1.554740e-06
## 11 7.244718e-03
## 12 4.984558e-12
## 13 2.778181e-05
## 14 4.195102e-05
## 15 9.847087e-07
## 16 2.622160e-10
## 17 8.180138e-06
## 18 6.587279e-16
## 19 1.250525e-10
## 20 1.277704e-10
## 21 4.637835e-03
## 22 3.725153e-04
```

```
## 23 1.613240e-04
## 24 8.429583e-09
## 25 6.382446e-04
## 26 2.899346e-08
## 27 4.716034e-08
## 28 3.321828e-09
## 29 7.944658e-04
## 30 6.639961e-07
## 31 4.155105e-06
## 32 4.029990e-03
## 33 4.655866e-07
## 34 3.322677e-02
## 35 3.252801e-06
## 36 1.297928e-03
## 37 7.226365e-04
## 38 2.273012e-02
## 39 8.204378e-07
## 40 1.179119e-03
## 41 6.364232e-08
## 42 1.242465e-06
## 43 1.785795e-11
## 44 4.309968e-06
## 45 3.251598e-07
```

CCS code 3

CCS code 3 that is significantly associated with die30

We find that CCS code 3 is significantly associated with die30 The CCS code3 descriptions that have significant association are as follows

```
##                                                 names1     P_value
## 1                 ccs3_Acute and unspecified renal failure 0.002350780
## 2                              ccs3_Chronic ulcer of skin 0.029750532
## 3              ccs3_Coagulation and hemorrhagic disorders 0.032344260
## 4                          ccs3_Nutritional deficiencies 0.020673844
## 5                  ccs3_Other gastrointestinal disorders 0.007578809
## 6                               ccs3_Other liver diseases 0.032103345
## 7          ccs3_Pleurisy; pneumothorax; pulmonary collapse 0.005358708
## 8  ccs3_Respiratory failure; insufficiency; arrest (adult) 0.015590451
## 9                            ccs3_Secondary malignancies 0.002902010
## 10                                        ccs3_Undefined 0.034930974
```

CCS code 3 that is significantly associated with die1800

We find that CCS code 3 is significantly associated with die1800 The CCS code3 descriptions that have significant association are as follows

```
##
names2
## 1                                                                ccs3_Acute posthe
morrhagic anemia
## 2
ccs3_Asthma
## 3                                                    ccs3_Bacterial infection;
unspecified site
## 4                                                                          ccs3_
Cancer of breast
## 5                                                                 ccs3_Card
iac dysrhythmias
## 6                                  ccs3_Chronic obstructive pulmonary disease an
d bronchiectasis
## 7                                      ccs3_Complications of surgical procedures
or medical care
## 8                                          ccs3_Coronary atherosclerosis and oth
er heart disease
## 9                                                          ccs3_Deficiency
and other anemia
## 10                                                ccs3_Diabetes mellitus with
out complication
## 11                                                   ccs3_Disorders of
lipid metabolism
## 12                                                                ccs3_Esop
hageal disorders
## 13                                                               ccs3_Essent
ial hypertension
## 14                                               ccs3_Fluid and elect
rolyte disorders
## 15                                                                ccs3_Heart
valve disorders
## 16                                 ccs3_Hypertension with complications and second
ary hypertension
## 17                                                ccs3_Intestinal obstructio
n without hernia
## 18                                                                         ccs
3_Mood disorders
## 19                                                                         ccs
3_Osteoarthritis
## 20                                                                           c
cs3_Osteoporosis
## 21
ccs3_other
## 22                                                                      ccs3
_Other aftercare
## 23                                                            ccs3_Other cir
culatory disease
## 24                                                      ccs3_Other connectiv
e tissue disease
## 25                                                    ccs3_Other gastrointe
stinal disorders
## 26                                                     ccs3_Other lower res
piratory disease
## 27                                                      ccs3_Other nervous
system disorders
## 28                                    ccs3_Other nutritional; endocrine; and met
```

```
abolic disorders
## 29 ccs3_Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or
sexually transm
## 30                                                                  ccs3_Pleurisy; pneumothorax; pu
lmonary collapse
## 31                      ccs3_Pneumonia (except that caused by tuberculosis or sexually tran
smitted disease)
## 32                                                                                ccs3_Residual cod
es; unclassified
## 33                                                                                    ccs3_Second
ary malignancies
## 34                                    ccs3_Spondylosis; intervertebral disc disorders; oth
er back problems
## 35                                                                                        ccs3_T
hyroid disorders
## 36                                                                              ccs3_Urinary
tract infections
##         P_value
## 1  3.290765e-03
## 2  1.063156e-03
## 3  1.502840e-02
## 4  6.902292e-04
## 5  2.625321e-05
## 6  2.709620e-02
## 7  3.297022e-04
## 8  3.511623e-06
## 9  1.845108e-03
## 10 3.280501e-09
## 11 1.315149e-10
## 12 1.774492e-05
## 13 1.536615e-11
## 14 1.206397e-05
## 15 1.238057e-04
## 16 3.576012e-02
## 17 6.862369e-03
## 18 1.096083e-02
## 19 8.664758e-06
## 20 2.002532e-07
## 21 1.086248e-05
## 22 1.157507e-04
## 23 4.460936e-04
## 24 2.047532e-08
## 25 1.707583e-03
## 26 2.193979e-02
## 27 1.219960e-02
## 28 1.684444e-04
## 29 1.250034e-02
## 30 1.529402e-02
## 31 6.826005e-03
## 32 4.544860e-04
## 33 4.318108e-13
## 34 5.488283e-06
## 35 1.954196e-08
## 36 3.645168e-04
```

CCS code 4

CCS code 4 that is significantly associated with die30

We find that CCS code 4 is significantly associated with die30 The CCS code4 descriptions that have significant association are as follows

```
##                                                 names1      P_value
## 1               ccs4_Acute and unspecified renal failure 0.0308402438
## 2                               ccs4_Conduction disorders 0.0459226927
## 3  ccs4_Coronary atherosclerosis and other heart disease 0.0226127697
## 4             ccs4_Diabetes mellitus without complication 0.0175385141
## 5                       ccs4_Disorders of lipid metabolism 0.0035886258
## 6                               ccs4_Esophageal disorders 0.0256542379
## 7                             ccs4_Essential hypertension 0.0005367635
## 8                                    ccs4_Osteoarthritis 0.0316035664
## 9                                      ccs4_Osteoporosis 0.0134596195
## 10                  ccs4_Other connective tissue disease 0.0028429636
## 11                            ccs4_Secondary malignancies 0.0297261224
## 12                                 ccs4_Thyroid disorders 0.0083869195
```

CCS code 4 that is significantly associated with die1800

We find that CCS code 4 is significantly associated with die1800 The CCS code4 descriptions that have significant association are as follows

```
##                                                                  names2
## 1                              ccs4_Acute and unspecified renal failure
## 2                                    ccs4_Acute posthemorrhagic anemia
## 3                                             ccs4_Allergic reactions
## 4                                              ccs4_Anxiety disorders
## 5                                                        ccs4_Asthma
## 6                           ccs4_Bacterial infection; unspecified site
## 7                                               ccs4_Cancer of breast
## 8                                            ccs4_Cardiac dysrhythmias
## 9                                          ccs4_Chronic kidney disease
## 10         ccs4_Chronic obstructive pulmonary disease and bronchiectasis
## 11                           ccs4_Coagulation and hemorrhagic disorders
## 12              ccs4_Complications of surgical procedures or medical care
## 13                                           ccs4_Conduction disorders
## 14                        ccs4_Congestive heart failure; nonhypertensive
## 15                 ccs4_Coronary atherosclerosis and other heart disease
## 16                                     ccs4_Deficiency and other anemia
## 17                               ccs4_Diabetes mellitus with complications
## 18                            ccs4_Diabetes mellitus without complication
## 19                                    ccs4_Disorders of lipid metabolism
## 20                                            ccs4_Esophageal disorders
## 21                                         ccs4_Essential hypertension
## 22                                    ccs4_Fluid and electrolyte disorders
## 23                                          ccs4_Heart valve disorders
## 24       ccs4_Hypertension with complications and secondary hypertension
## 25                                                  ccs4_Mood disorders
## 26                                                  ccs4_Osteoarthritis
## 27                                                    ccs4_Osteoporosis
## 28                                                           ccs4_other
## 29                                                 ccs4_Other aftercare
## 30               ccs4_Other bone disease and musculoskeletal deformities
## 31                                       ccs4_Other circulatory disease
## 32                                    ccs4_Other connective tissue disease
## 33                                     ccs4_Other female genital disorders
## 34                                   ccs4_Other gastrointestinal disorders
## 35                                     ccs4_Other lower respiratory disease
## 36                                      ccs4_Other nervous system disorders
## 37            ccs4_Other nutritional; endocrine; and metabolic disorders
## 38                         ccs4_Pleurisy; pneumothorax; pulmonary collapse
## 39                                       ccs4_Residual codes; unclassified
## 40               ccs4_Respiratory failure; insufficiency; arrest (adult)
## 41 ccs4_Screening and history of mental health and substance abuse codes
## 42                                          ccs4_Secondary malignancies
## 43  ccs4_Spondylosis; intervertebral disc disorders; other back problems
## 44                                                  ccs4_Thyroid disorders
## 45                                                       ccs4_Undefined
## 46                                       ccs4_Urinary tract infections
##         P_value
## 1  4.989725e-02
## 2  1.978235e-03
## 3  2.553783e-08
## 4  4.972784e-05
## 5  6.962489e-05
## 6  2.706857e-03
## 7  5.984177e-07
## 8  1.013857e-05
## 9  3.528087e-03
```

```
## 10 1.224477e-02
## 11 3.638384e-02
## 12 2.536749e-03
## 13 1.836035e-06
## 14 1.708664e-02
## 15 8.670652e-08
## 16 4.718959e-05
## 17 1.395134e-04
## 18 7.949749e-10
## 19 1.006207e-13
## 20 5.304703e-10
## 21 3.996115e-15
## 22 2.923289e-05
## 23 8.812460e-06
## 24 1.369354e-03
## 25 7.075018e-06
## 26 7.754330e-09
## 27 1.044065e-09
## 28 1.793640e-07
## 29 3.066605e-05
## 30 9.613791e-04
## 31 9.236255e-09
## 32 1.558220e-11
## 33 8.343201e-04
## 34 2.136675e-04
## 35 1.526646e-02
## 36 8.453787e-04
## 37 2.220105e-05
## 38 6.154932e-04
## 39 9.394551e-05
## 40 2.799613e-02
## 41 7.730256e-04
## 42 1.118399e-13
## 43 6.761417e-05
## 44 4.046449e-11
## 45 7.405927e-03
## 46 2.531452e-05
```

season

seasons that is significantly associated with die30

The seasons that have significant association are as follows We can find that the season winter is significantly related to the die30, which indicates whether the individual dies in the time window of 30 days after discharge

```
##         names1    P_value
## 1 seasonwinter 0.04721893
```

seaasons that is significantly associated with die1800

The seasons that have significant association are as follows We can find that season is not significantly related to the die1800, which indicates whether the individual dies in the time window of 1800 days after discharge

```
## [1] names2  P_value
## <0 rows> (or 0-length row.names)
```

ZIP code

ZIP codes that is significantly associated with die30

The ZIP codes that have significant association are as follows We can find that zip code is significantly related to the die30, which indicates whether the individual dies in the time window of 30 days after discharge

```
##           names1     P_value
## 1  zipcode_90048 0.048536423
## 2  zipcode_91105 0.022100119
## 3  zipcode_91367 0.037728114
## 4  zipcode_92037 0.012019488
## 5  zipcode_92123 0.012081555
## 6  zipcode_92653 0.018628371
## 7  zipcode_94115 0.014193484
## 8  zipcode_94596 0.001057508
## 9  zipcode_94705 0.026920659
## 10 zipcode_95119 0.037935675
```

ZIP codes that is significantly associated with die1800

The ZIP codes that have significant association are as follows We can find that zip code is not significantly related to the die1800, which indicates whether the individual dies in the time window of 1800 days after discharge

```
##          names2    P_value
## 1 zipcode_90505 0.01375748
## 2 zipcode_91360 0.01262534
## 3 zipcode_92103 0.03231716
## 4 zipcode_92373 0.04868250
## 5 zipcode_92835 0.03079069
## 6 zipcode_93720 0.00101477
```

Conclusion

In this project, I build four different models to predict die 30. At last, I found that boosting is the best and its AUC is 0.9246. I also use it to predict die 1800 and the AUC is 0.8745.

It is obvious that the longer the time window is, the more people die. And risk of death just monotonously increases when the time window become longer. Generally, the farther away from the discharge, the less risk of death on that specific day

Then I do the classic logistic regression(no penalty for more predictors) and I find that there are some CCS code 1, CCS code 2, CCS code 3, CCS code 4 and zip code that are significantly associated with die30 and die1800 adjusted for other variables. Season is only significantly associated to die 30, adjusted for other variables