Relationship between Overall Survival, Clinical and Genomic data from TCGA’S Study on Pancreatic Cancer Patients via Machine Learning

Jiann-Ping Hsu College of Public Health

Georgia Southern University

Manyun Liu and Roshni Modi

Contents

- Introduction
- Data and Methods
- Results
- Discussion
- References

Introduction

- Pancreatic ductal adenocarcinoma (PDAC), the most common form of pancreatic cancer, is the fourth leading cause of cancer death in the world.
- Genes and common factors, such as age, race, smoking, alcohol consumption, obesity, and diabetes are risk factors for pancreatic cancer.
- The objective of this study:
- **Primary endpoint**: To investigate the association between risk factors and overall survival time of PDAC.
- **Secondary endpoint**: To evaluate if risk factors are associated with grade of the PDAC.

Data and Methods

- **Data**: The dataset we used is from National Cancer Institute, which is part of the TCGA's Study of PDAC containing 154 patients.
- **Methods**:

✓ Clinical data analysis

Cox proportional hazards model

Tree-based model (rpart and randomForest)

✓ Gene data analysis

T-test

Lasso regression

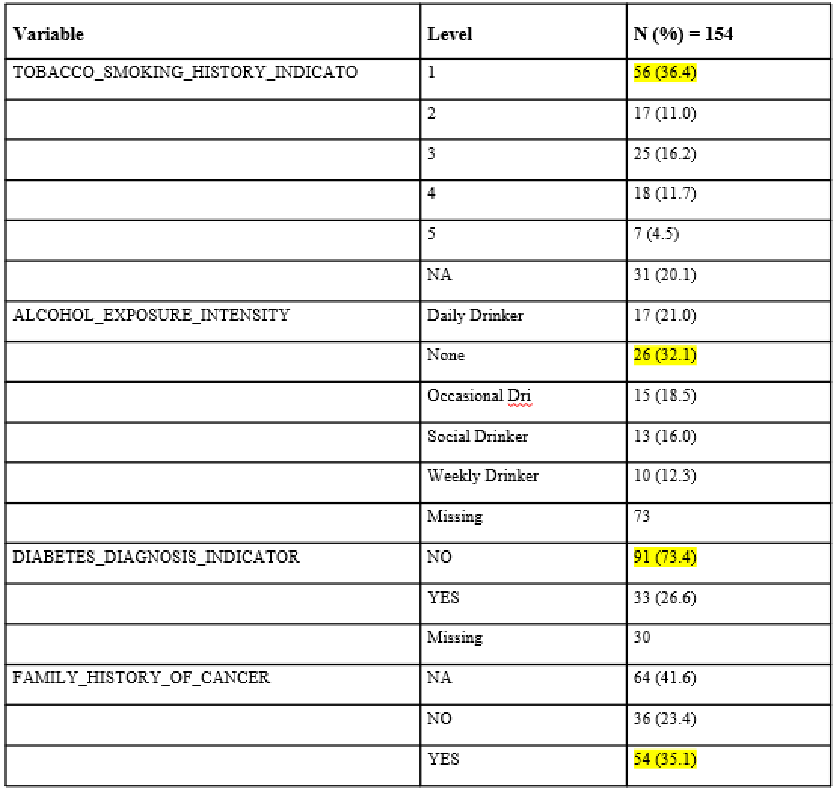
GEORGIA SOUTHERN UNIVERSITY

Results

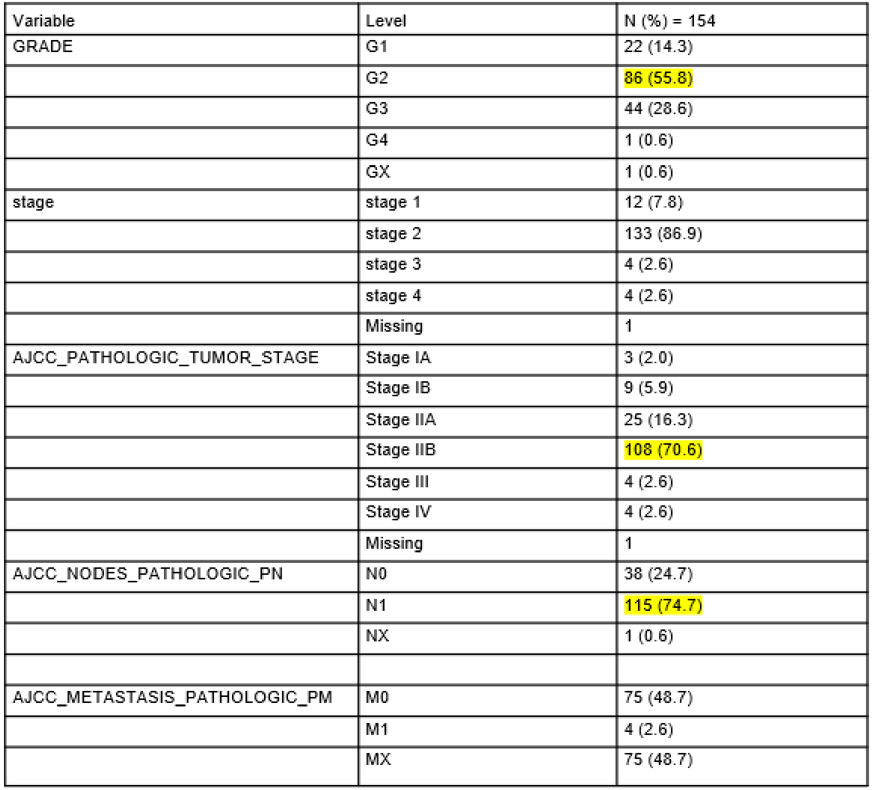
Distribution of Demographics

Variable	Level	N(%)=154
Age	Mean	65.05
	Median	65.50
	Minimum	35.00
	Maximum	85.00
	Std Dev	11.01
Sex	Female	71(46.1)
	Male	83(53.9)
Race	Asian	9(5.8)
	Black	7(4.5)
	White	133(86.4)
	NA	5(3.2)
Ethnicity	Hispanic	4(2.6)
	Not Hispanic	115(74.7)
	NA	35(22.7)

Patient and Tumor Characteristics



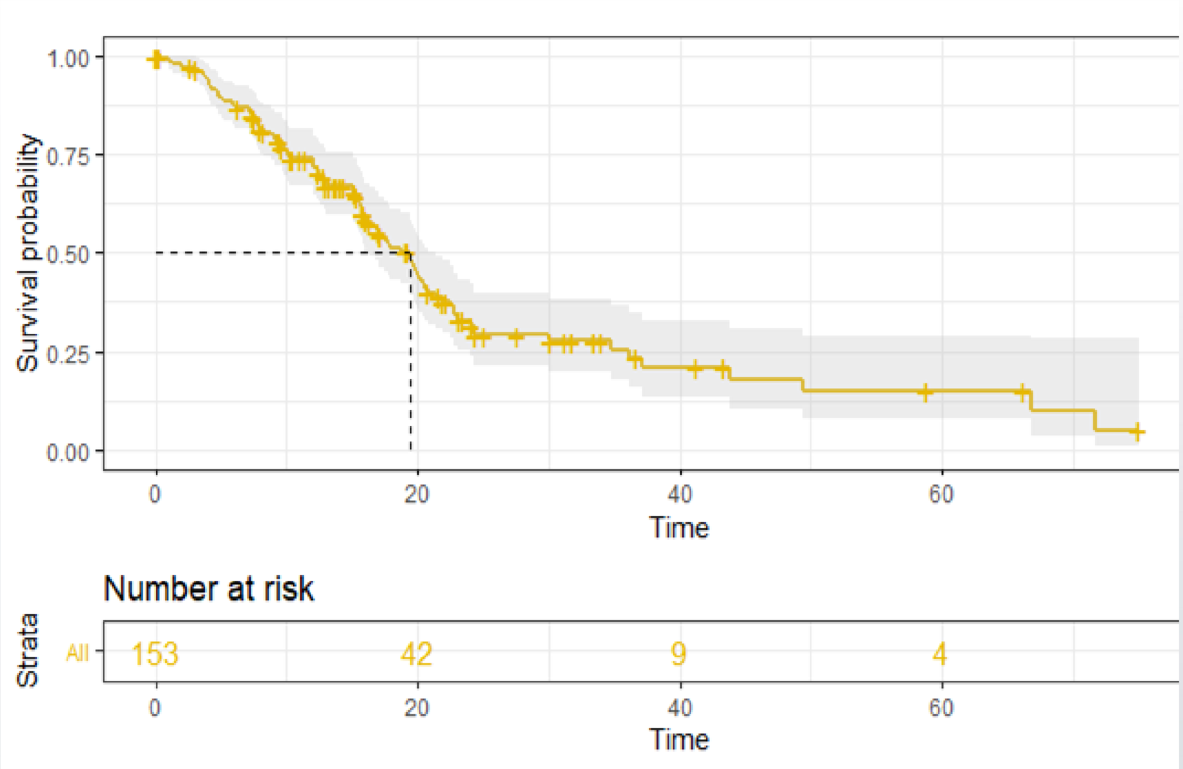
Variable	Level	N (%) = 154
TOBACCO_SMOKING_HISTORY_INDICATO	1	56 (36.4)
	2	17 (11.0)
	3	25 (16.2)
	4	18 (11.7)
	5	7 (4.5)
	NA	31 (20.1)
ALCOHOL_EXPOSURE_INTENSITY	Daily Drinker	17 (21.0)
	None	26 (32.1)
	Occasional Dri	15 (18.5)
	Social Drinker	13 (16.0)
	Weekly Drinker	10 (12.3)
	Missing	73
DIABETES_DIAGNOSIS_INDICATOR	NO	91 (73.4)
	YES	33 (26.6)
	Missing	30
FAMILY_HISTORY_OF_CANCER	NA	64 (41.6)
	NO	36 (23.4)
	YES	54 (35.1)



Variable	Level	N (%) = 154
GRADE	G1	22 (14.3)
	G2	86 (55.8)
	G3	44 (28.6)
	G4	1 (0.6)
	GX	1 (0.6)
stage	stage 1	12 (7.8)
	stage 2	133 (86.9)
	stage 3	4 (2.6)
	stage 4	4 (2.6)
	Missing	1
AJCC_PATHOLOGIC_TUMOR_STAGE	Stage IA	3 (2.0)
	Stage IB	9 (5.9)
	Stage IIA	25 (16.3)
	Stage IIB	108 (70.6)
	Stage III	4 (2.6)
	Stage IV	4 (2.6)
	Missing	1
AJCC_NODES_PATHOLOGIC_PN	N0	38 (24.7)
	N1	115 (74.7)
	NX	1 (0.6)
AJCC_METASTASIS_PATHOLOGIC_PM	M0	75 (48.7)
	M1	4 (2.6)
	MX	75 (48.7)

Kaplan-Meier Curve for Overall Survival

- The Median Survival time for overall survival is approx. 20 months.
- Some people survived up to 70 months



Univariate Cox regression

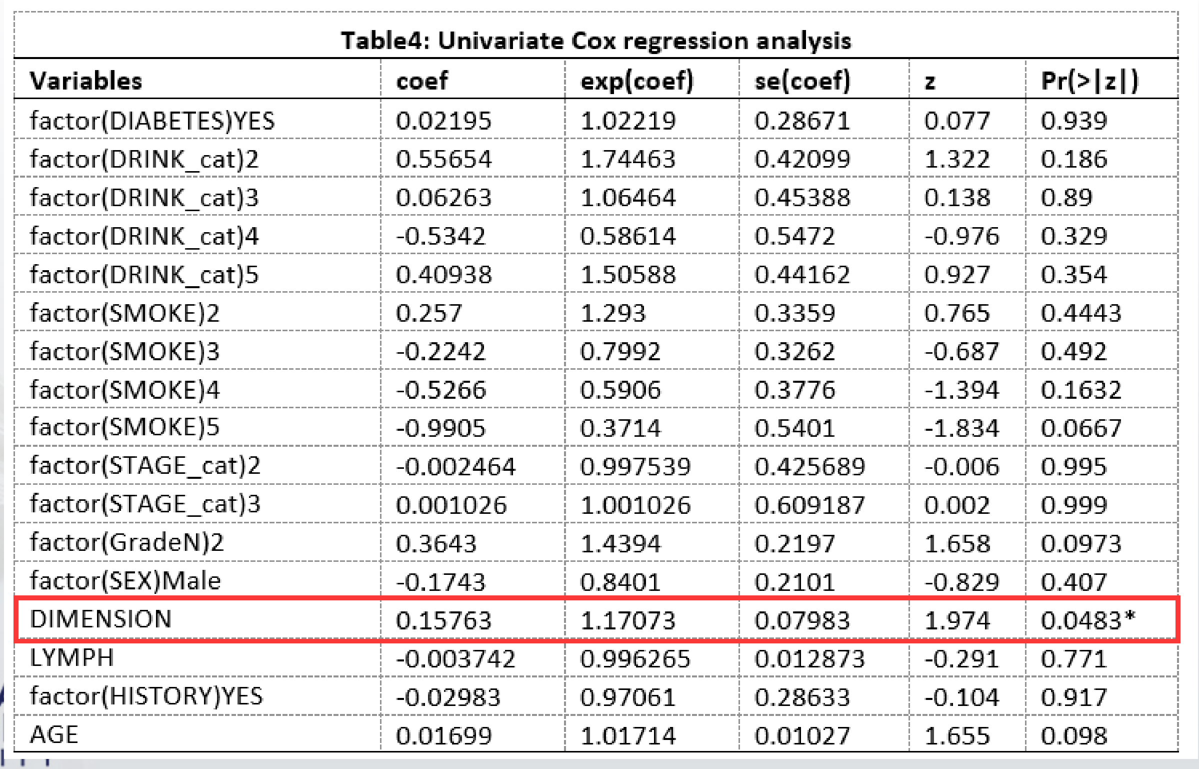


Table4: Univariate Cox regression analysis

Variables	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(DIABETES)YES	0.02195	1.02219	0.28671	0.077	0.939
factor(DRINK_cat)2	0.55654	1.74463	0.42099	1.322	0.186
factor(DRINK_cat)3	0.06263	1.06464	0.45388	0.138	0.89
factor(DRINK_cat)4	-0.5342	0.58614	0.5472	-0.976	0.329
factor(DRINK_cat)5	0.40938	1.50588	0.44162	0.927	0.354
factor(SMOKE)2	0.257	1.293	0.3359	0.765	0.4443
factor(SMOKE)3	-0.2242	0.7992	0.3262	-0.687	0.492
factor(SMOKE)4	-0.5266	0.5906	0.3776	-1.394	0.1632
factor(SMOKE)5	-0.9905	0.3714	0.5401	-1.834	0.0667
factor(STAGE_cat)2	-0.002464	0.997539	0.425689	-0.006	0.995
factor(STAGE_cat)3	0.001026	1.001026	0.609187	0.002	0.999
factor(GradeN)2	0.3643	1.4394	0.2197	1.658	0.0973
factor(SEX)Male	-0.1743	0.8401	0.2101	-0.829	0.407
DIMENSION	0.15763	1.17073	0.07983	1.974	0.0483*
LYMPH	-0.003742	0.996265	0.012873	-0.291	0.771
factor(HISTORY)YES	-0.02983	0.97061	0.28633	-0.104	0.917
AGE	0.01699	1.01714	0.01027	1.655	0.098

Multivariate Cox regression

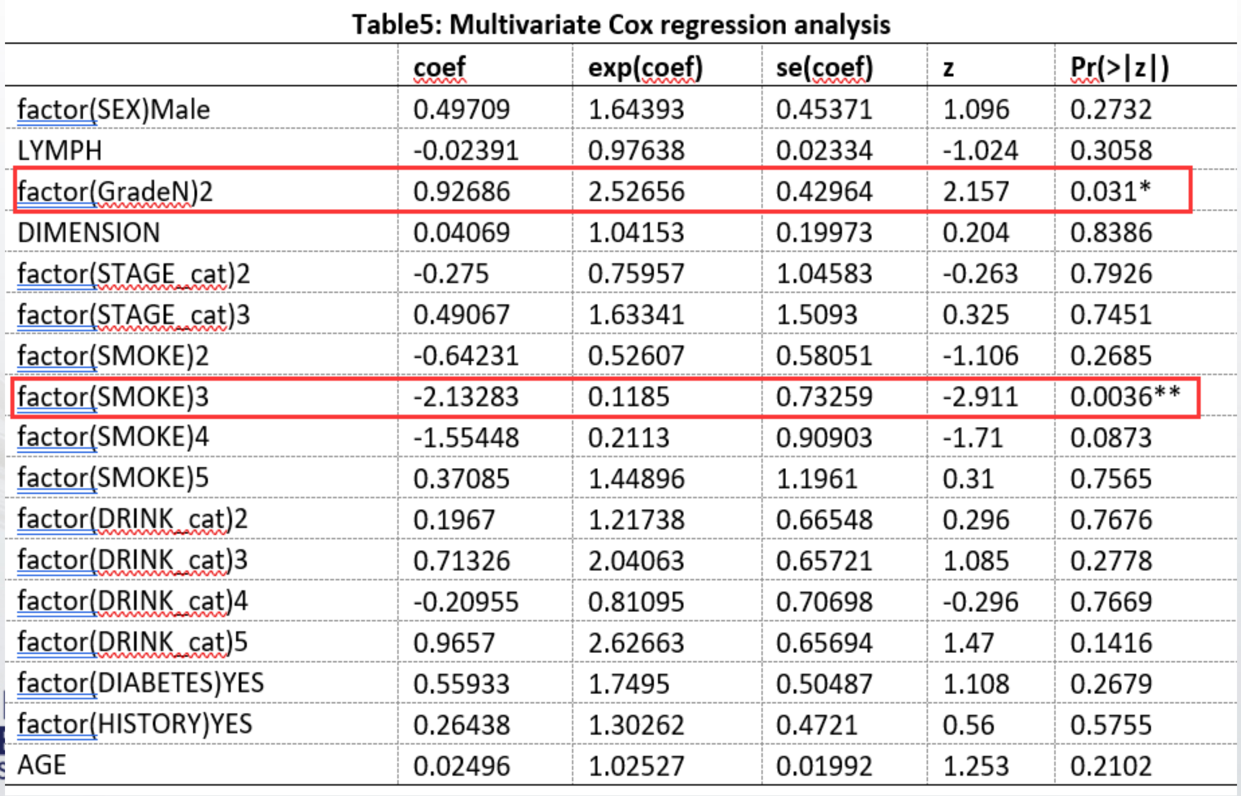
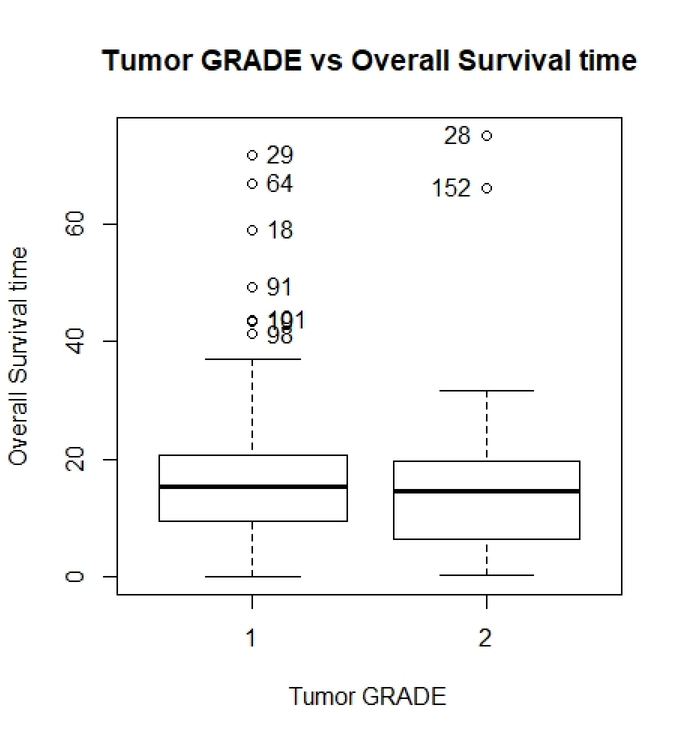


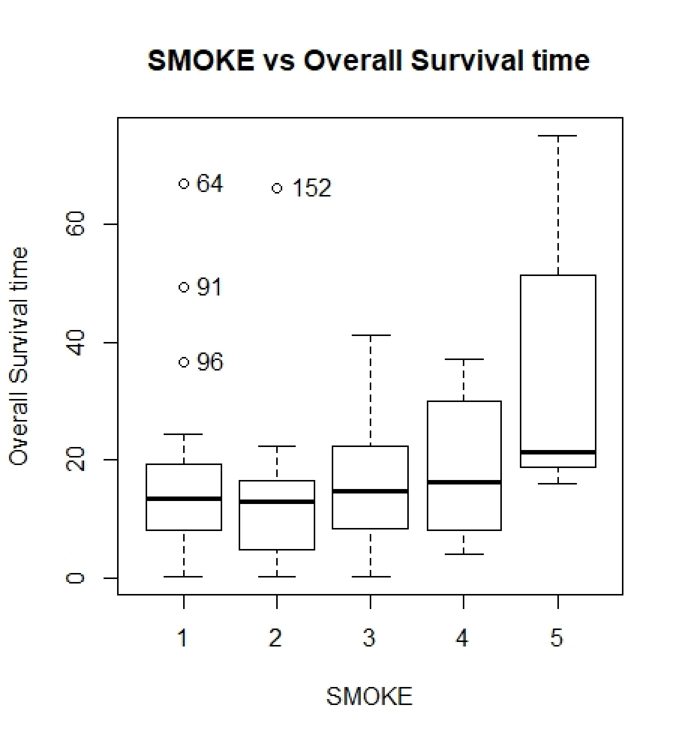
Table5: Multivariate Cox regression analysis

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(SEX)Male	0.49709	1.64393	0.45371	1.096	0.2732
LYMPH	-0.02391	0.97638	0.02334	-1.024	0.3058
factor(GradeN)2	0.92686	2.52656	0.42964	2.157	0.031*
DIMENSION	0.04069	1.04153	0.19973	0.204	0.8386
factor(STAGE_cat)2	-0.275	0.75957	1.04583	-0.263	0.7926
factor(STAGE_cat)3	0.49067	1.63341	1.5093	0.325	0.7451
factor(SMOKE)2	-0.64231	0.52607	0.58051	-1.106	0.2685
factor(SMOKE)3	-2.13283	0.1185	0.73259	-2.911	0.0036**
factor(SMOKE)4	-1.55448	0.2113	0.90903	-1.71	0.0873
factor(SMOKE)5	0.37085	1.44896	1.1961	0.31	0.7565
factor(DRINK_cat)2	0.1967	1.21738	0.66548	0.296	0.7676
factor(DRINK_cat)3	0.71326	2.04063	0.65721	1.085	0.2778
factor(DRINK_cat)4	-0.20955	0.81095	0.70698	-0.296	0.7669
factor(DRINK_cat)5	0.9657	2.62663	0.65694	1.47	0.1416
factor(DIABETES)YES	0.55933	1.7495	0.50487	1.108	0.2679
factor(HISTORY)YES	0.26438	1.30262	0.4721	0.56	0.5755
AGE	0.02496	1.02527	0.01992	1.253	0.2102

Boxplot(Factors vs Survival time)



Tumor GRADE vs Overall Survival time

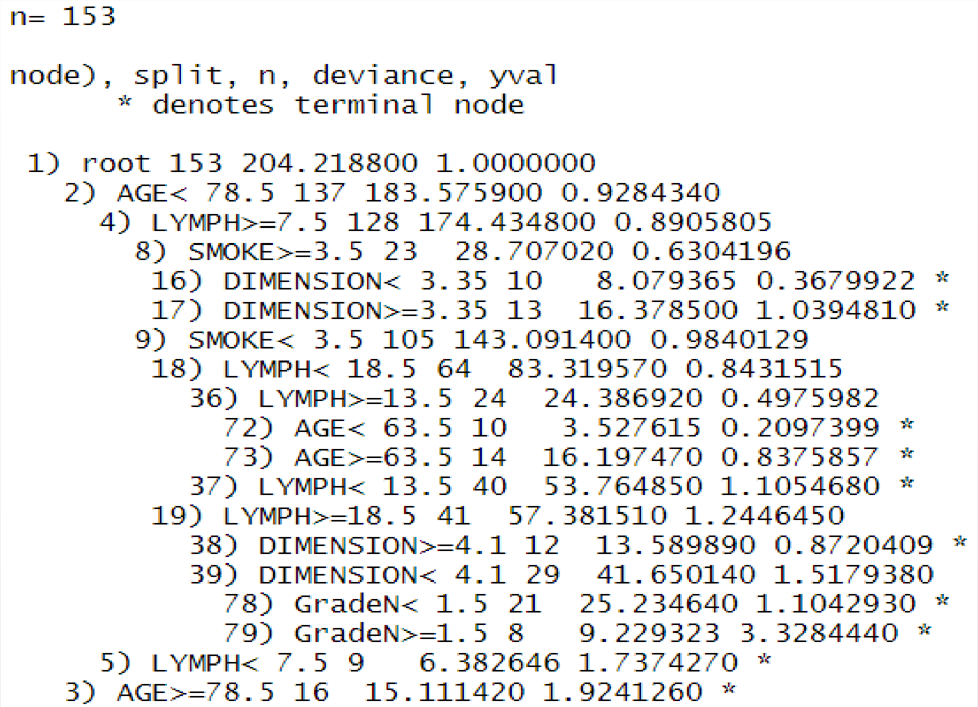


SMOKE vs Overall Survival time

Recursive partitioning (Rpart)

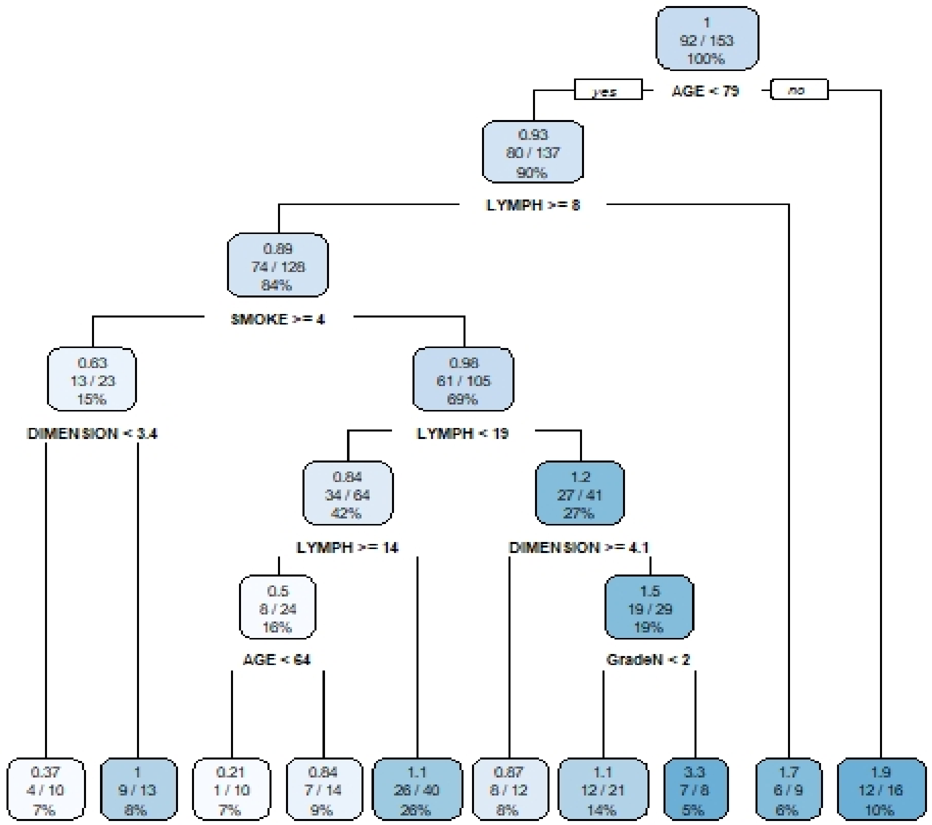
1.AGE is the most significant factor to define the subgroup.

2. The subgroup can be further divided into smaller subgroup by LYMPH,SMOKE, DIMENSION, and GRADE.



```
n= 153

node), split, n, deviance, yval
      * denotes terminal node

 1) root 153 204.218800 1.0000000
   2) AGE< 78.5 137 183.575900 0.9284340
     4) LYMPH>=7.5 128 174.434800 0.8905805
       8) SMOKE>=3.5 23  28.707020 0.6304196
        16) DIMENSION< 3.35 10   8.079365 0.3679922 *
        17) DIMENSION>=3.35 13  16.378500 1.0394810 *
       9) SMOKE< 3.5 105 143.091400 0.9840129
        18) LYMPH< 18.5 64  83.319570 0.8431515
          36) LYMPH>=13.5 24  24.386920 0.4975982
            72) AGE< 63.5 10   3.527615 0.2097399 *
            73) AGE>=63.5 14  16.197470 0.8375857 *
          37) LYMPH< 13.5 40  53.764850 1.1054680 *
        19) LYMPH>=18.5 41  57.381510 1.2446450
          38) DIMENSION>=4.1 12  13.589890 0.8720409 *
          39) DIMENSION< 4.1 29  41.650140 1.5179380
            78) GradeN< 1.5 21  25.234640 1.1042930 *
            79) GradeN>=1.5 8   9.229323 3.3284440 *
     5) LYMPH< 7.5 9   6.382646 1.7374270 *
   3) AGE>=78.5 16  15.111420 1.9241260 *
```

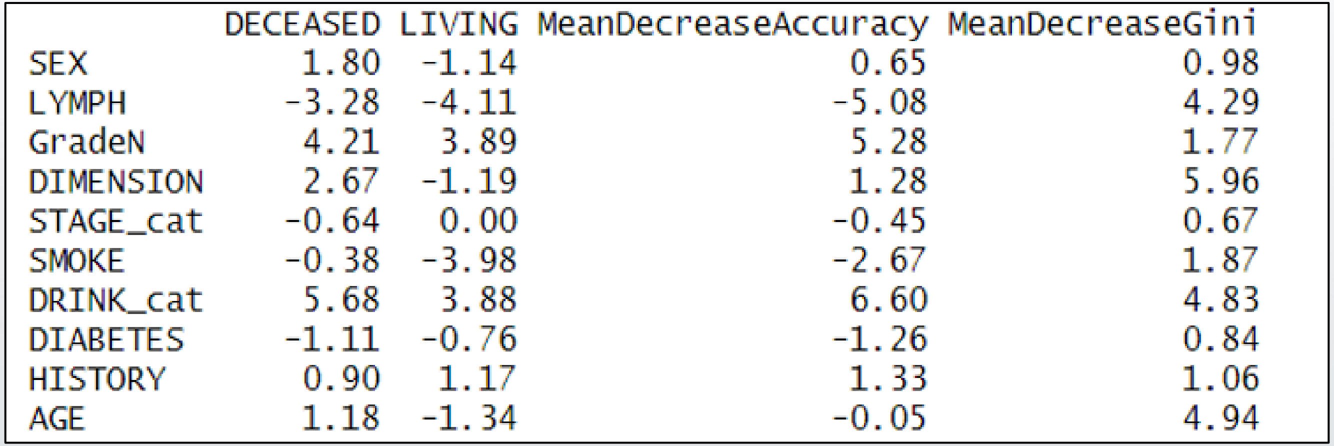
Rpart



1
92 / 153
100%
yes
AGE < 79
no
0.93
80 / 137
90%
LYMPH >= 8
0.89
74 / 128
84%
SMOKE >= 4
0.63
13 / 23
15%
0.98
61 / 105
69%
DIMENSION < 3.4
LYMPH < 19
0.84
34 / 64
42%
1.2
27 / 41
27%
LYMPH >= 14
DIMENSION >= 4.1
0.5
8 / 24
16%
1.5
19 / 29
19%
AGE < 64
GradeN < 2
0.37
4 / 10
7%
1
9 / 13
8%
0.21
1 / 10
7%
0.84
7 / 14
9%
1.1
26 / 40
26%
0.87
8 / 12
8%
1.1
12 / 21
14%
3.3
7 / 8
5%
1.7
6 / 9
6%
1.9
12 / 16
10%

Random Forest

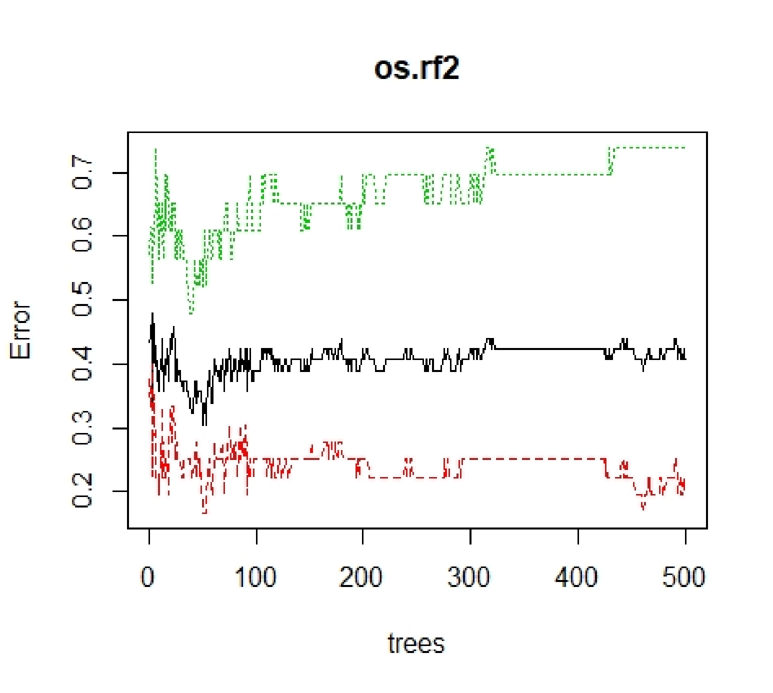
DRINK is the most important factor, follow by Tumor GRADE, then HISTORY, Tumor DIMENSION.

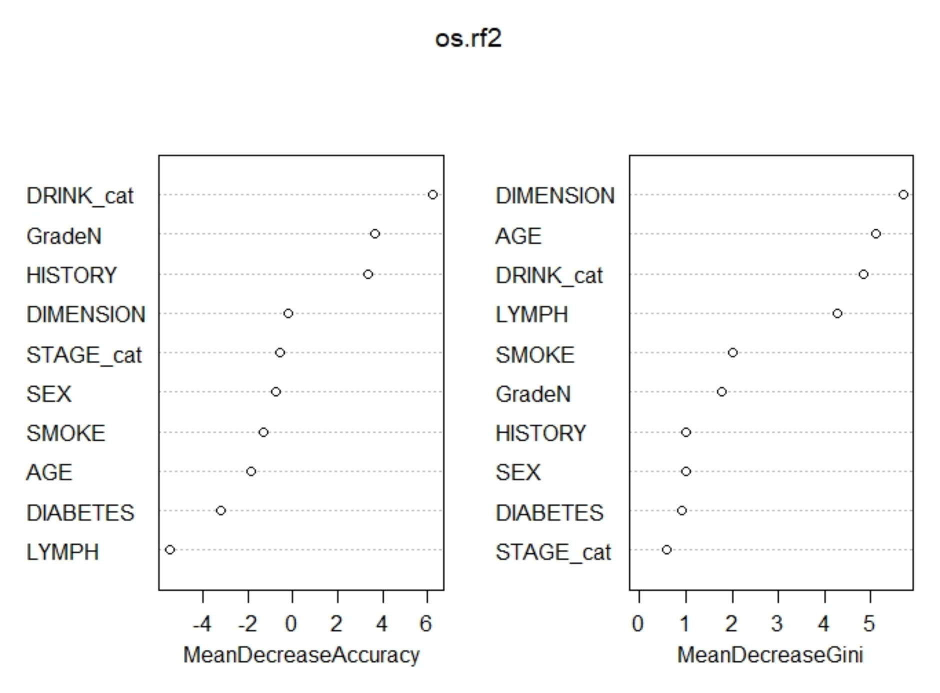


```
          DECEASED LIVING MeanDecreaseAccuracy MeanDecreaseGini
SEX           1.80  -1.14                 0.65             0.98
LYMPH        -3.28  -4.11                -5.08             4.29
GradeN        4.21   3.89                 5.28             1.77
DIMENSION     2.67  -1.19                 1.28             5.96
STAGE_cat    -0.64   0.00                -0.45             0.67
SMOKE        -0.38  -3.98                -2.67             1.87
DRINK_cat     5.68   3.88                 6.60             4.83
DIABETES     -1.11  -0.76                -1.26             0.84
HISTORY       0.90   1.17                 1.33             1.06
AGE           1.18  -1.34                -0.05             4.94
```

GEORGIA SOUTHERN UNIVERSITY

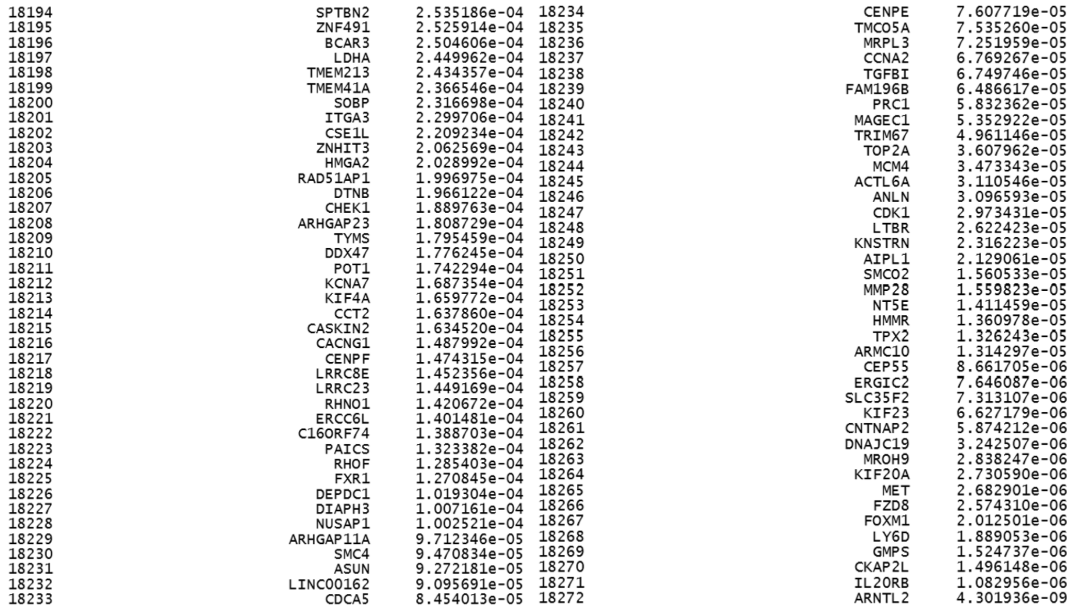
Random Forest





Gene data analysis

Cox regression(one by one): 18272 totally, 79 genes are significant.



```
18194       SPTBN2 2.535186e-04 18234        CENPE 7.607719e-05
18195       ZNF491 2.525914e-04 18235      TMCO5A 7.535260e-05
18196        BCAR3 2.504606e-04 18236       MRPL3 7.251959e-05
18197         LDHA 2.449962e-04 18237       CCNA2 6.769267e-05
18198      TMEM213 2.434357e-04 18238       TGFBI 6.749746e-05
18199      TMEM41A 2.366546e-04 18239     FAM196B 6.486617e-05
18200         SOBP 2.316698e-04 18240        PRC1 5.832362e-05
18201        ITGA3 2.299706e-04 18241      MAGEC1 5.352922e-05
18202        CSE1L 2.209234e-04 18242      TRIM67 4.961146e-05
18203       ZNHIT3 2.062569e-04 18243       TOP2A 3.607962e-05
18204        HMGA2 2.028992e-04 18244        MCM4 3.473343e-05
18205     RAD51AP1 1.996975e-04 18245      ACTL6A 3.110546e-05
18206         DTNB 1.966122e-04 18246        ANLN 3.096593e-05
18207        CHEK1 1.889763e-04 18247        CDK1 2.973431e-05
18208     ARHGAP23 1.808729e-04 18248        LTBR 2.622423e-05
18209         TYMS 1.795459e-04 18249      KNSTRN 2.316223e-05
18210        DDX47 1.776245e-04 18250       AIPL1 2.129061e-05
18211         POT1 1.742294e-04 18251       SMCO2 1.560533e-05
18212        KCNA7 1.687354e-04 18252       MMP28 1.559823e-05
18213        KIF4A 1.659772e-04 18253        NT5E 1.411459e-05
18214         CCT2 1.637860e-04 18254        HMMR 1.360978e-05
18215      CASKIN2 1.634520e-04 18255        TPX2 1.326243e-05
18216       CACNG1 1.487992e-04 18256      ARMC10 1.314297e-05
18217        CENPF 1.474315e-04 18257       CEP55 8.661705e-06
18218        LRRC8E 1.452356e-04 18258     ERGIC2 7.646087e-06
18219       LRRC23 1.449169e-04 18259    SLC35F2 7.313107e-06
18220        RHNO1 1.420672e-04 18260       KIF23 6.627179e-06
18221       ERCC6L 1.401481e-04 18261    CNTNAP2 5.874212e-06
18222     C16ORF74 1.388703e-04 18262    DNAJC19 3.242507e-06
18223        PAICS 1.323382e-04 18263       MROH9 2.838247e-06
18224         RHOF 1.285403e-04 18264     KIF20A 2.730590e-06
18225         FXR1 1.270845e-04 18265         MET 2.682901e-06
18226       DEPDC1 1.019304e-04 18266        FZD8 2.574310e-06
18227       DIAPH3 1.007161e-04 18267       FOXM1 2.012501e-06
18228       NUSAP1 1.002521e-04 18268        LY6D 1.889053e-06
18229    ARHGAP11A 9.712346e-05 18269        GMPS 1.524737e-06
18230         SMC4 9.470834e-05 18270      CKAP2L 1.496148e-06
18231         ASUN 9.272181e-05 18271      IL20RB 1.082956e-06
18232    LINC00162 9.095691e-05 18272      ARNTL2 4.301936e-09
18233        CDCA5 8.454013e-05
```

Lasso Regression

Lasso selected 17 genes which are significantly associated with survival time

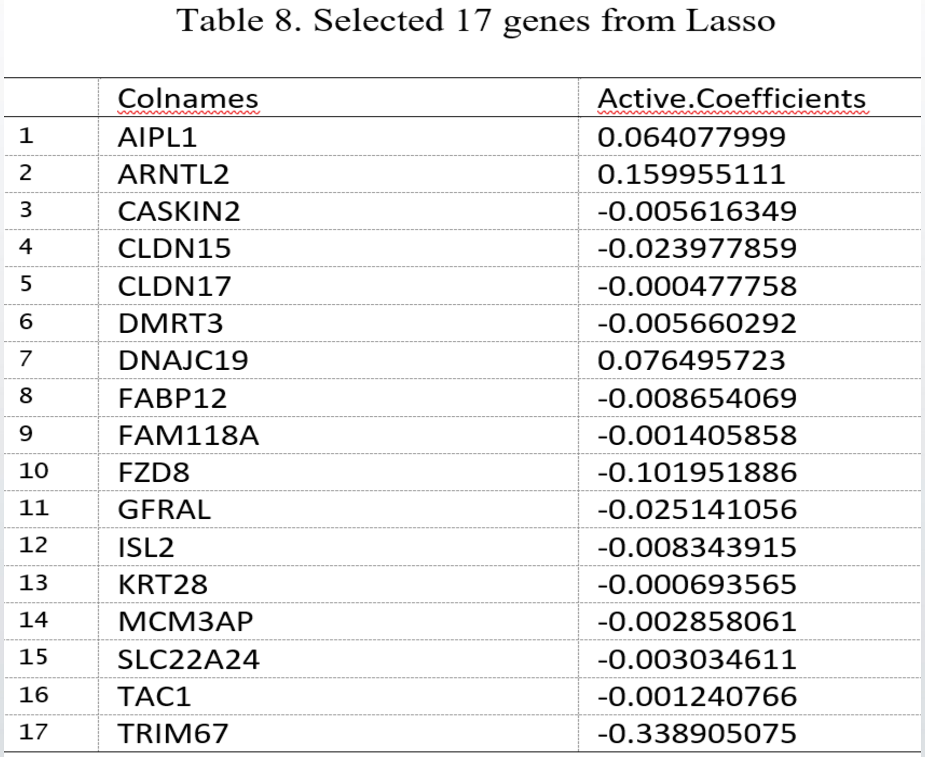
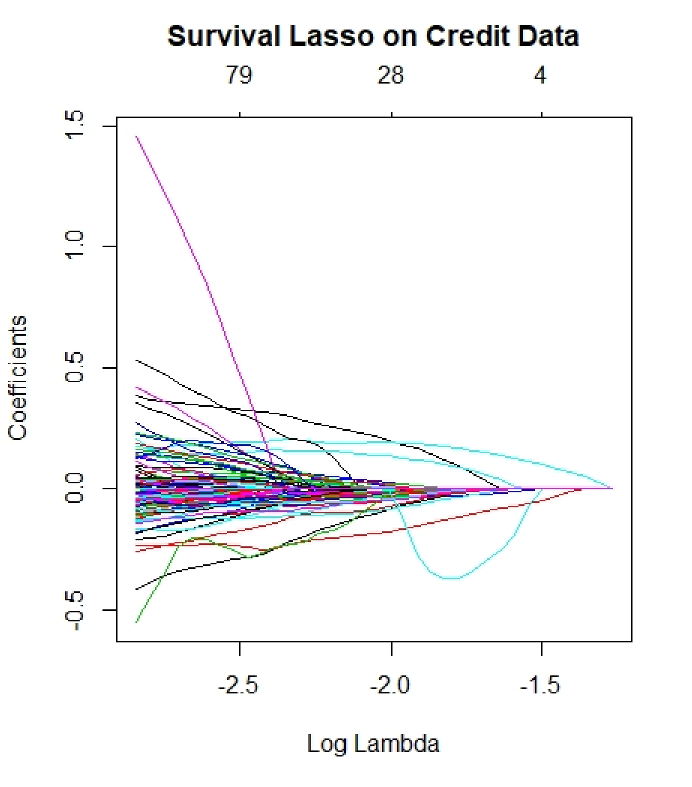
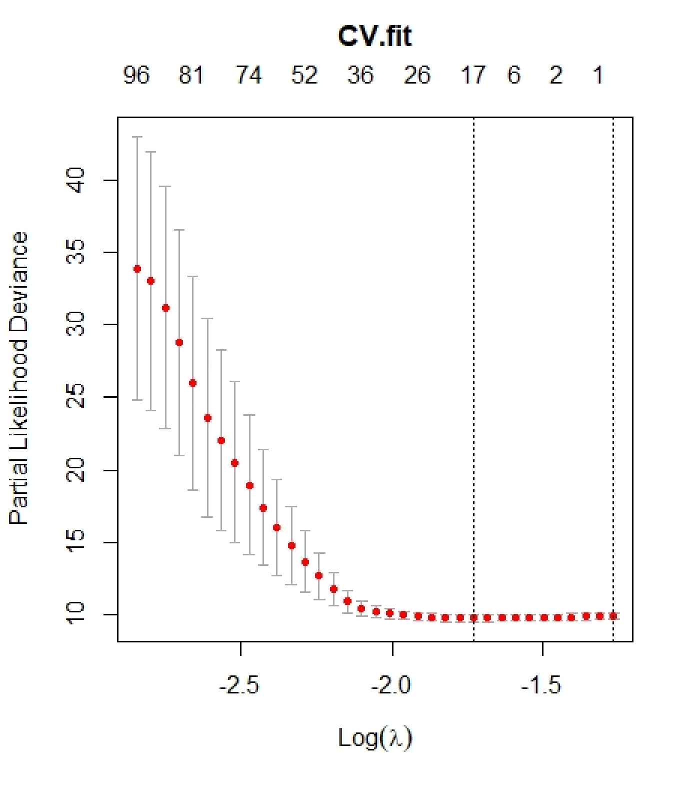


Table 8. Selected 17 genes from Lasso

	Colnames	Active.Coefficients
1	AIPL1	0.064077999
2	ARNTL2	0.159955111
3	CASKIN2	-0.005616349
4	CLDN15	-0.023977859
5	CLDN17	-0.000477758
6	DMRT3	-0.005660292
7	DNAJC19	0.076495723
8	FABP12	-0.008654069
9	FAM118A	-0.001405858
10	FZD8	-0.101951886
11	GFRAL	-0.025141056
12	ISL2	-0.008343915
13	KRT28	-0.000693565
14	MCM3AP	-0.002858061
15	SLC22A24	-0.003034611
16	TAC1	-0.001240766
17	TRIM67	-0.338905075

Lasso Regression





Final model (Cox regression)

By comparing the results, four genes are same from the two methods :

CASKIN2, TRIM67, DNAJC19, and ARNTL2.

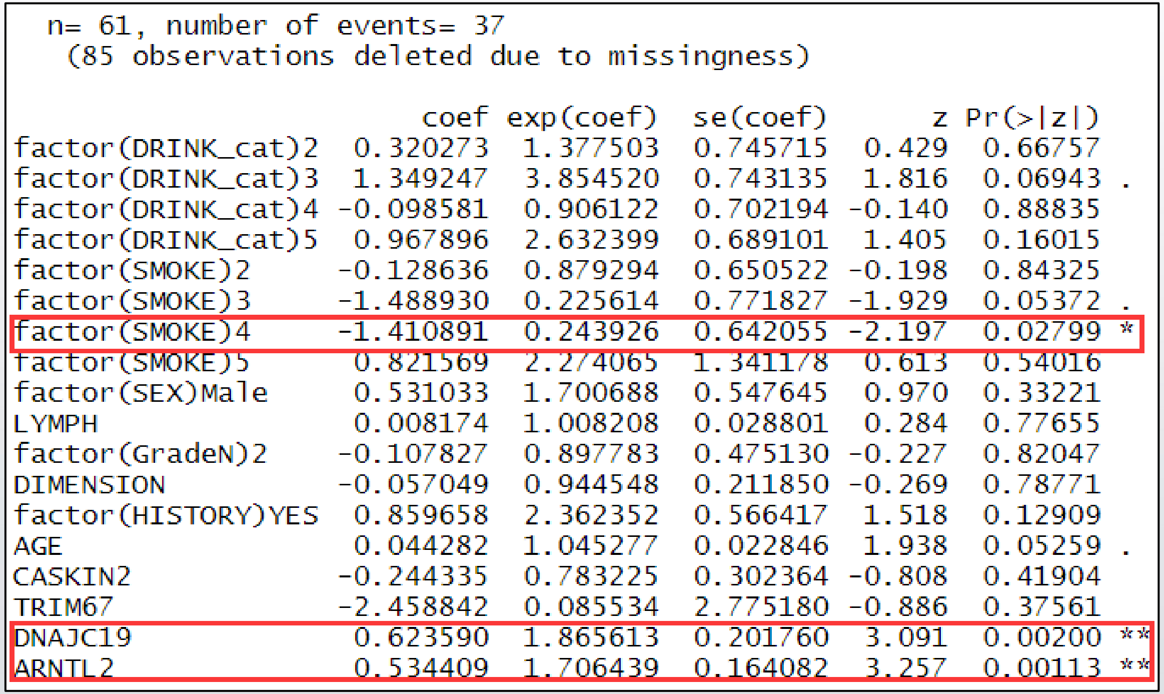
Finally, we put the four genes with the significant factors from clinical data:

GRADE, SMOKE, DRINK, add SEX, AGE, HISTORY, DIMENSION, LYMPH,

together into the survival model.

Final model

- Significant factors: **SMOKE4, DNAJC19 ARNTL2**
- Reformed SMOKE had a positive impact on Survival time.



```
  n= 61, number of events= 37
   (85 observations deleted due to missingness)

                         coef exp(coef)  se(coef)      z Pr(>|z|)
factor(DRINK_cat)2   0.320273  1.377503  0.745715  0.429  0.66757
factor(DRINK_cat)3   1.349247  3.854520  0.743135  1.816  0.06943 .
factor(DRINK_cat)4  -0.098581  0.906122  0.702194 -0.140  0.88835
factor(DRINK_cat)5   0.967896  2.632399  0.689101  1.405  0.16015
factor(SMOKE)2      -0.128636  0.879294  0.650522 -0.198  0.84325
factor(SMOKE)3      -1.488930  0.225614  0.771827 -1.929  0.05372 .
factor(SMOKE)4      -1.410891  0.243926  0.642055 -2.197  0.02799 *
factor(SMOKE)5       0.821569  2.274065  1.341178  0.613  0.54016
factor(SEX)Male      0.531033  1.700688  0.547645  0.970  0.33221
LYMPH                0.008174  1.008208  0.028801  0.284  0.77655
factor(GradeN)2     -0.107827  0.897783  0.475130 -0.227  0.82047
DIMENSION           -0.057049  0.944548  0.211850 -0.269  0.78771
factor(HISTORY)YES   0.859658  2.362352  0.566417  1.518  0.12909
AGE                  0.044282  1.045277  0.022846  1.938  0.05259 .
CASKIN2             -0.244335  0.783225  0.302364 -0.808  0.41904
TRIM67              -2.458842  0.085534  2.775180 -0.886  0.37561
DNAJC19              0.623590  1.865613  0.201760  3.091  0.00200 **
ARNTL2               0.534409  1.706439  0.164082  3.257  0.00113 **
```

Secondary endpoint analysis

- The analysis so far was using continuous response , now we are going to analyze using categorical response.

- The grade for the subject's tumor was divided into two groups:

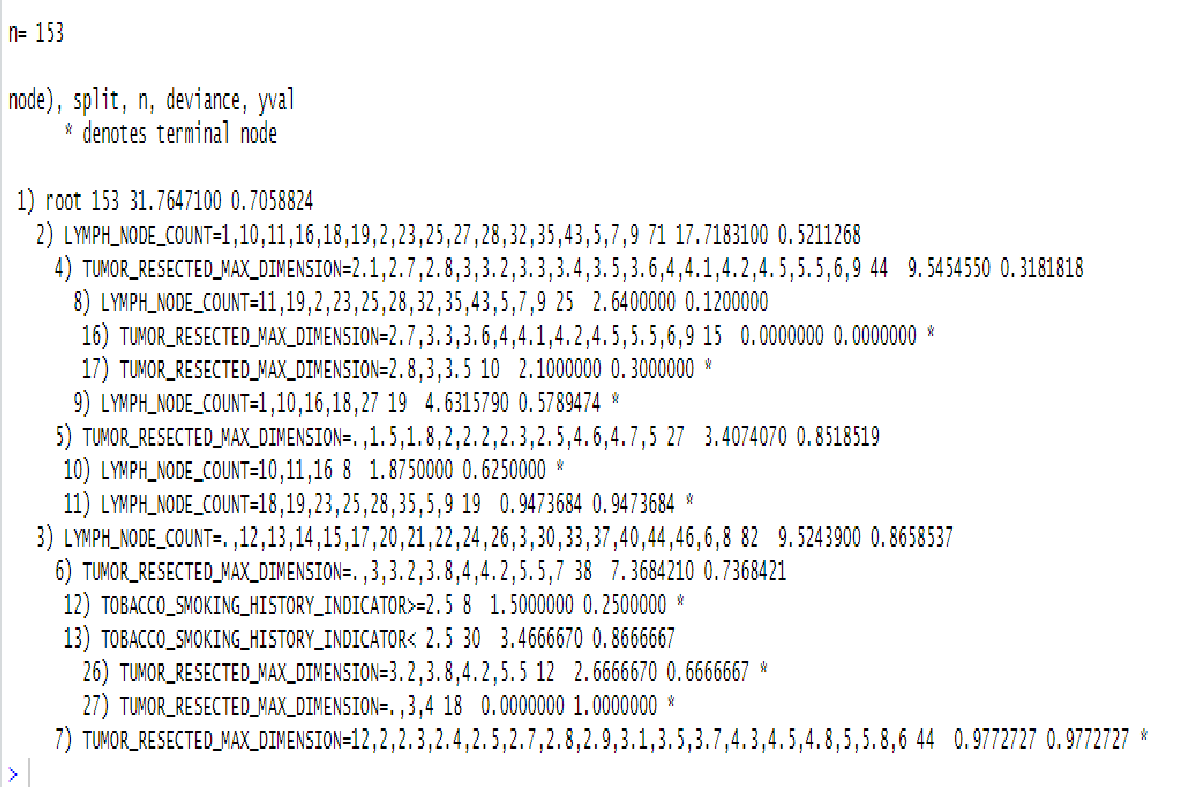
  GRADE in 1 or 2 in one group

  GRADE in 3 or 4 in another group
- Logistics regression for clinical covariates was performed to check if any clinical covariates were significantly associated with grade category.

- But results showed there were no clinical covariates show significant association with tumor grade.

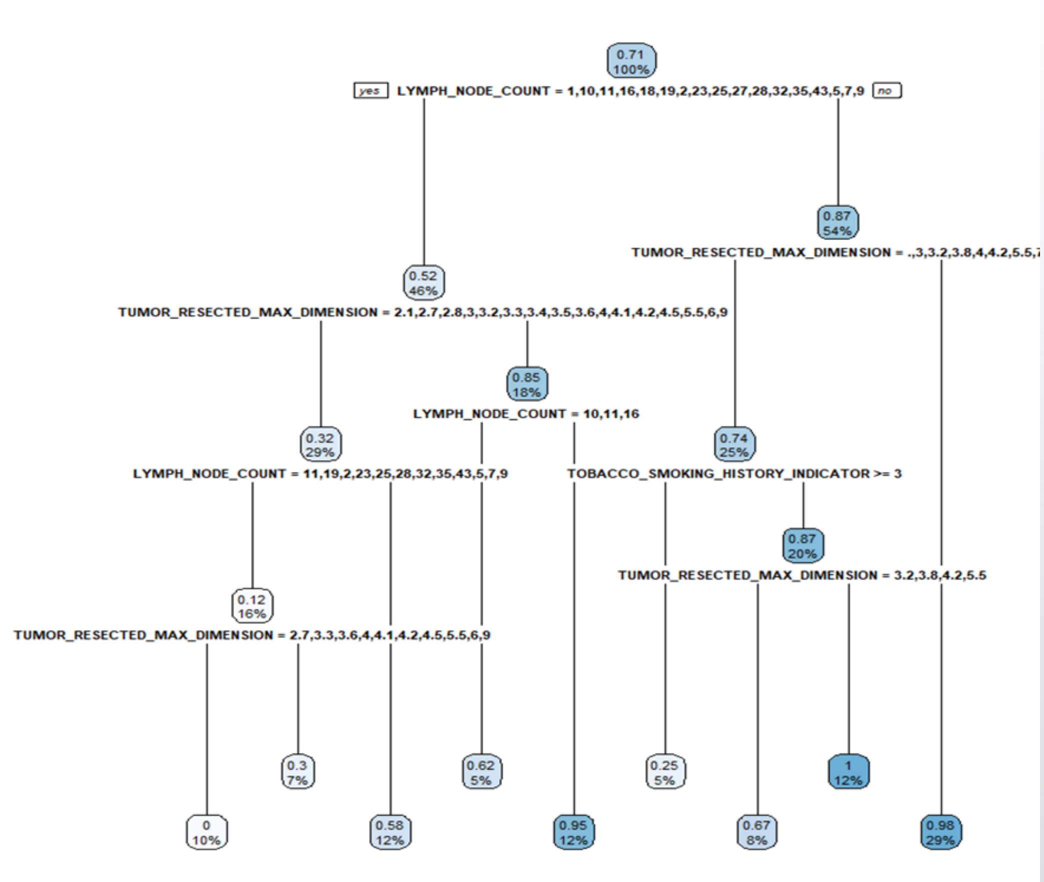
Recursive Partitioning(Rpart)

- LYMPH node count is the most significant factor to define the subgroup
- Followed by dimension of the tumor, smoke.



```
n= 153

node), split, n, deviance, yval
      * denotes terminal node

 1) root 153 31.7647100 0.7058824
   2) LYMPH_NODE_COUNT=1,10,11,16,18,19,2,23,25,27,28,32,35,43,5,7,9 71 17.7183100 0.5211268
     4) TUMOR_RESECTED_MAX_DIMENSION=2.1,2.7,2.8,3,3.2,3.3,3.4,3.5,3.6,4,4.1,4.2,4.5,5.5,6,9 44  9.5454550 0.3181818
       8) LYMPH_NODE_COUNT=11,19,2,23,25,28,32,35,43,5,7,9 25  2.6400000 0.1200000
        16) TUMOR_RESECTED_MAX_DIMENSION=2.7,3.3,3.6,4,4.1,4.2,4.5,5.5,6,9 15  0.0000000 0.0000000 *
        17) TUMOR_RESECTED_MAX_DIMENSION=2.8,3,3.5 10  2.1000000 0.3000000 *
       9) LYMPH_NODE_COUNT=1,10,16,18,27 19  4.6315790 0.5789474 *
     5) TUMOR_RESECTED_MAX_DIMENSION=.,1.5,1.8,2,2.2,2.3,2.5,4.6,4.7,5 27  3.4074070 0.8518519
      10) LYMPH_NODE_COUNT=10,11,16 8  1.8750000 0.6250000 *
      11) LYMPH_NODE_COUNT=18,19,23,25,28,35,5,9 19  0.9473684 0.9473684 *
   3) LYMPH_NODE_COUNT=.,12,13,14,15,17,20,21,22,24,26,3,30,33,37,40,44,46,6,8 82  9.5243900 0.8658537
     6) TUMOR_RESECTED_MAX_DIMENSION=.,3,3.2,3.8,4,4.2,5.5,7 38  7.3684210 0.7368421
      12) TOBACCO_SMOKING_HISTORY_INDICATOR>=2.5 8  1.5000000 0.2500000 *
      13) TOBACCO_SMOKING_HISTORY_INDICATOR< 2.5 30  3.4666670 0.8666667
        26) TUMOR_RESECTED_MAX_DIMENSION=3.2,3.8,4.2,5.5 12  2.6666670 0.6666667 *
        27) TUMOR_RESECTED_MAX_DIMENSION=.,3,4 18  0.0000000 1.0000000 *
     7) TUMOR_RESECTED_MAX_DIMENSION=12,2,2.3,2.4,2.5,2.7,2.8,2.9,3.1,3.5,3.7,4.3,4.5,4.8,5,5.8,6 44  0.9772727 0.9772727 *
>
```

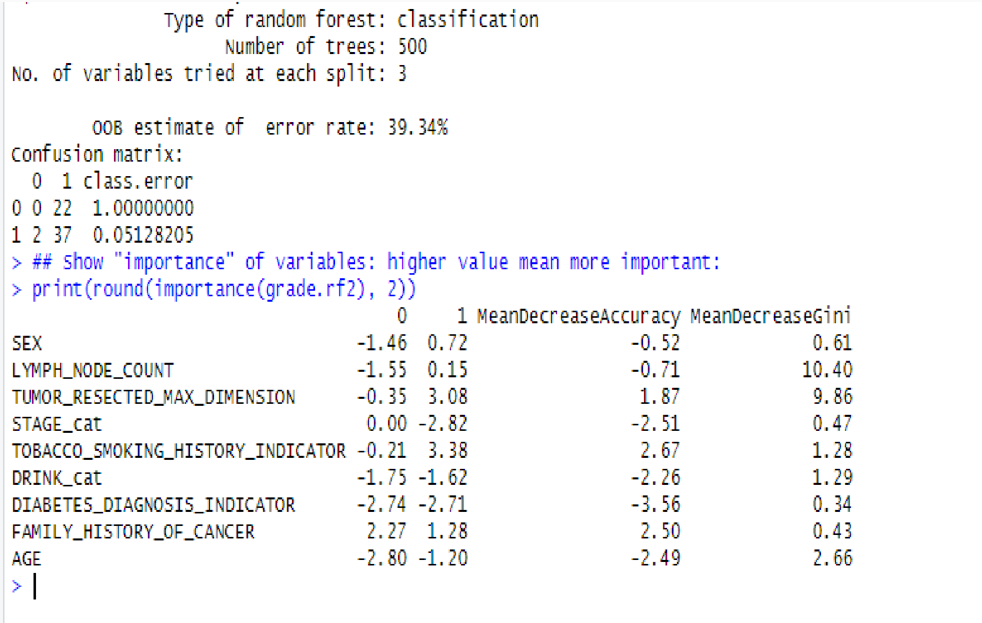
Recursive Partitioning(Rpart)


0.71
100%
yes
LYMPH_NODE_COUNT = 1,10,11,16,18,19,2,23,25,27,28,32,35,43,5,7,9
no
0.87
54%
TUMOR_RESECTED_MAX_DIMENSION = .,3,3.2,3.8,4,4.2,5.5,7
0.52
46%
TUMOR_RESECTED_MAX_DIMENSION = 2.1,2.7,2.8,3,3.2,3.3,3.4,3.5,3.6,4,4.1,4.2,4.5,5.5,6,9
0.85
18%
LYMPH_NODE_COUNT = 10,11,16
0.32
29%
0.74
25%
LYMPH_NODE_COUNT = 11,19,2,23,25,28,32,35,43,5,7,9
TOBACCO_SMOKING_HISTORY_INDICATOR >= 3
0.87
20%
TUMOR_RESECTED_MAX_DIMENSION = 3.2,3.8,4.2,5.5
0.12
16%
TUMOR_RESECTED_MAX_DIMENSION = 2.7,3.3,3.6,4,4.1,4.2,4.5,5.5,6,9
0.3
7%
0.62
5%
0.25
5%
1
12%
0
10%
0.58
12%
0.95
12%
0.67
8%
0.98
29%

GEORGIA SOUTHERN UNIVERSITY

Random Forest

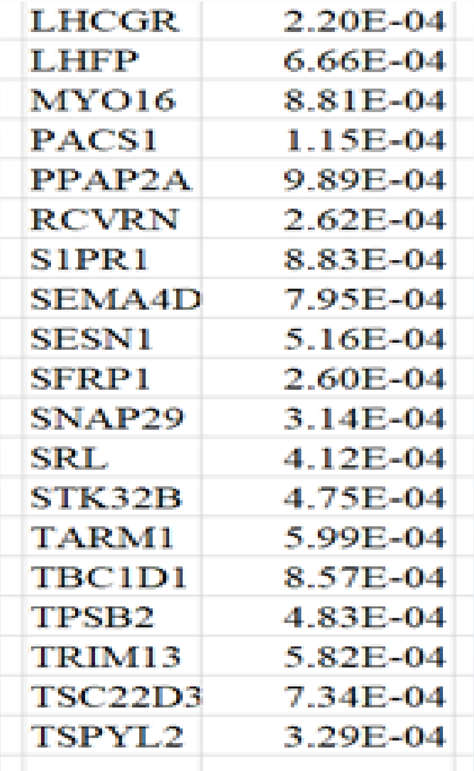
- Smoking is important factor along with family history of cancer and dimension of tumor



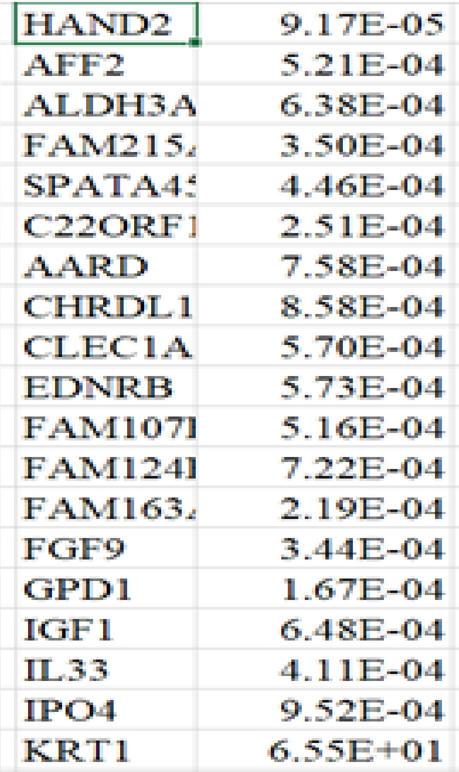
```
              Type of random forest: classification
                    Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 39.34%
Confusion matrix:
  0  1 class.error
0 0 22  1.00000000
1 2 37  0.05128205
> ## Show "importance" of variables: higher value mean more important:
> print(round(importance(grade.rf2), 2))
                                     0     1 MeanDecreaseAccuracy MeanDecreaseGini
SEX                              -1.46  0.72                -0.52             0.61
LYMPH_NODE_COUNT                 -1.55  0.15                -0.71            10.40
TUMOR_RESECTED_MAX_DIMENSION     -0.35  3.08                 1.87             9.86
STAGE_cat                         0.00 -2.82                -2.51             0.47
TOBACCO_SMOKING_HISTORY_INDICATOR -0.21  3.38                 2.67             1.28
DRINK_cat                        -1.75 -1.62                -2.26             1.29
DIABETES_DIAGNOSIS_INDICATOR     -2.74 -2.71                -3.56             0.34
FAMILY_HISTORY_OF_CANCER          2.27  1.28                 2.50             0.43
AGE                              -2.80 -1.20                -2.49             2.66
> |
```

T-Test

- Association between grade category and gene (one by one) using T-test: 18272 totally, 39 genes are significant between two tumor grade categories at P-value<0.001.



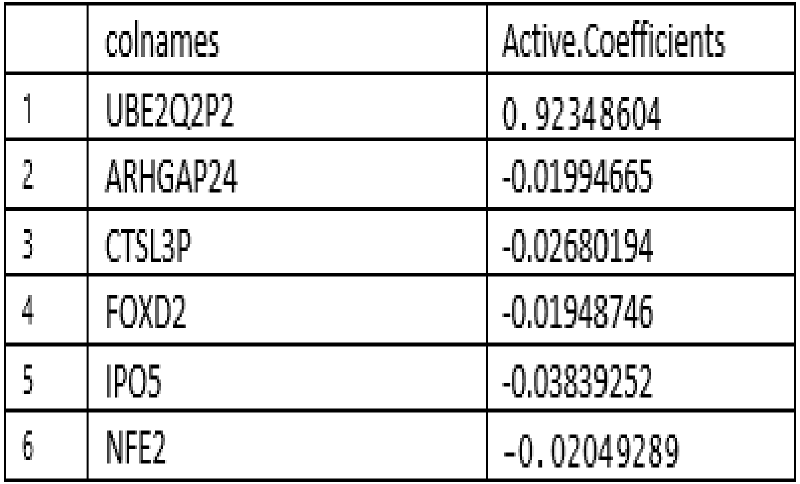
LHCGR	2.20E-04
LHFP	6.66E-04
MYO16	8.81E-04
PACS1	1.15E-04
PPAP2A	9.89E-04
RCVRN	2.62E-04
S1PR1	8.83E-04
SEMA4D	7.95E-04
SESN1	5.16E-04
SFRP1	2.60E-04
SNAP29	3.14E-04
SRL	4.12E-04
STK32B	4.75E-04
TARM1	5.99E-04
TBC1D1	8.57E-04
TPSB2	4.83E-04
TRIM13	5.82E-04
TSC22D3	7.34E-04
TSPYL2	3.29E-04



HAND2	9.17E-05
AFF2	5.21E-04
ALDH3A	6.38E-04
FAM215.	3.50E-04
SPATA4:	4.46E-04
C22ORF1	2.51E-04
AARD	7.58E-04
CHRDL1	8.58E-04
CLEC1A	5.70E-04
EDNRB	5.73E-04
FAM107I	5.16E-04
FAM124I	7.22E-04
FAM163.	2.19E-04
FGF9	3.44E-04
GPD1	1.67E-04
IGF1	6.48E-04
IL33	4.11E-04
IPO4	9.52E-04
KRT1	6.55E+01

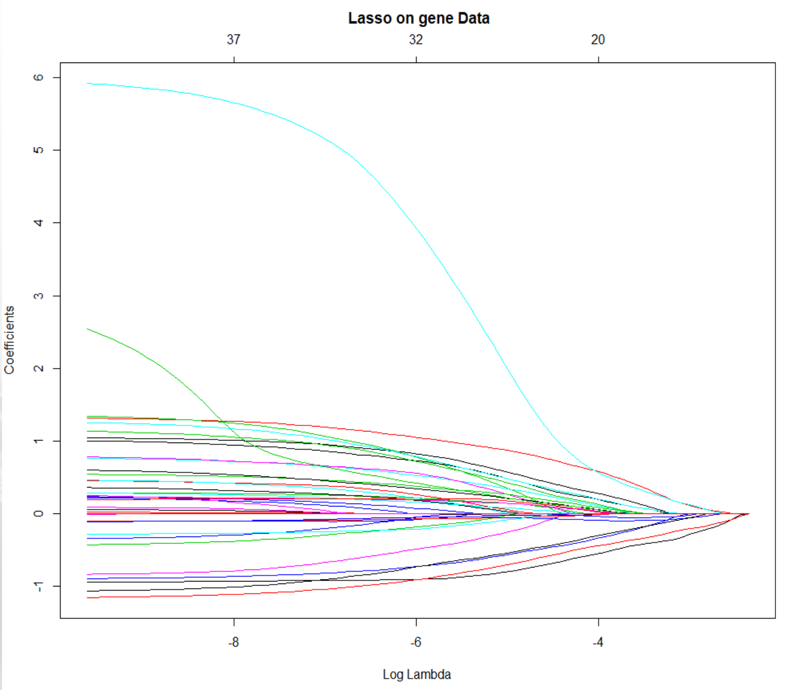
Lasso regression for grade category

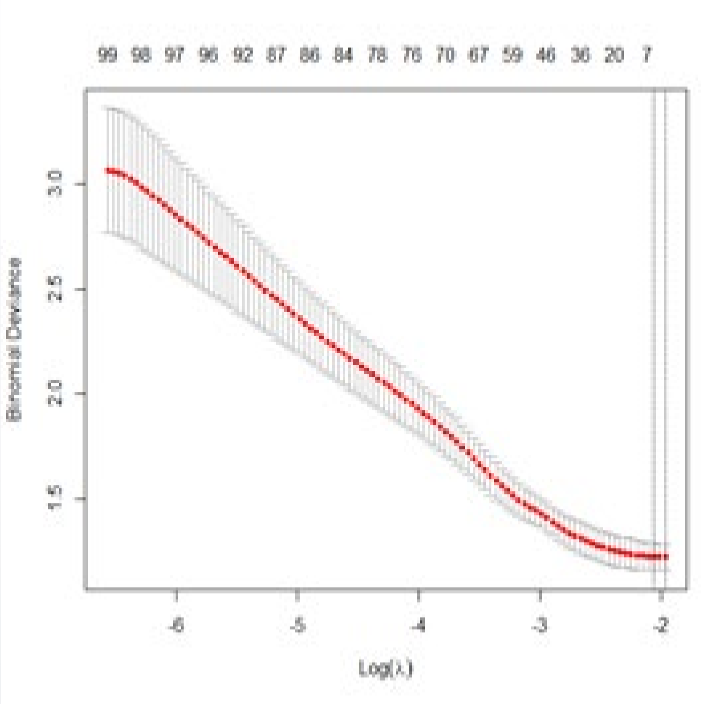
- Later, we performed lasso regression to identify which gene is significantly associated with the grade category.
- We found out there are six genes related to the grade of the tumor.
- The result obtained from T-test and Lasso regression was different. Hence, a detailed investigation is needed.



	colnames	Active.Coefficients
1	UBE2Q2P2	0. 92348604
2	ARHGAP24	-0.01994665
3	CTSL3P	-0.02680194
4	FOXD2	-0.01948746
5	IPO5	-0.03839252
6	NFE2	-0. 02049289

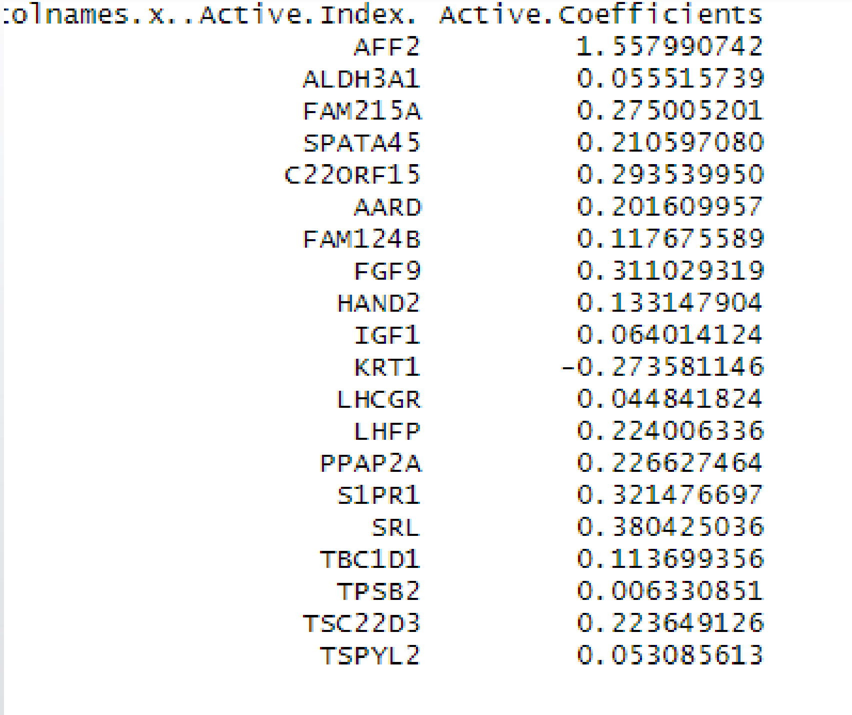
GEORGIA SOUTHERN UNIVERSITY


Lasso on gene Data
37
32
20
Coefficients
6
5
4
3
2
1
0
-1
-8
-6
-4
Log Lambda


99 98 97 96 92 87 86 84 78 76 70 67 59 46 36 20 7
Binomial Deviance
3.0
2.5
2.0
1.5
-6
-5
-4
-3
-2
Log(λ)

Lasso Regression

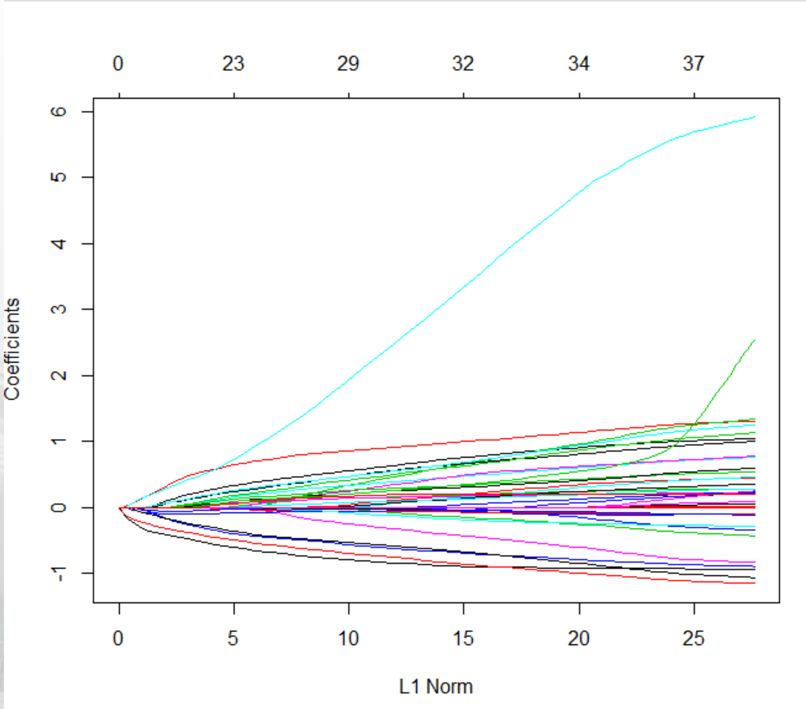
- We then used 39 significant genes obtained from T-test to see how many genes are kept in the lasso regression. We obtained 20 genes that were significant.

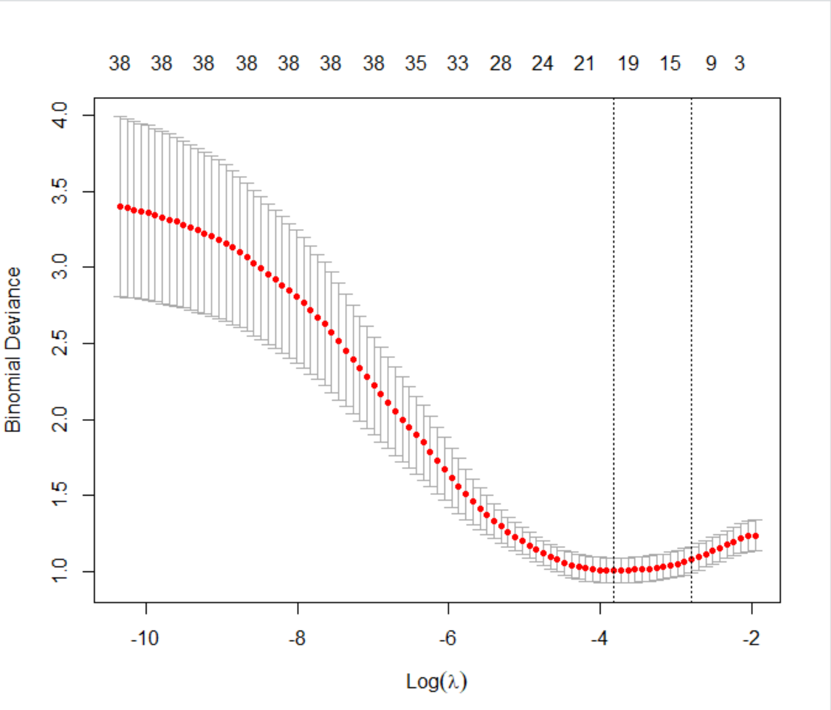


```
olnames.x..Active.Index.  Active.Coefficients
                    AFF2          1.557990742
                 ALDH3A1          0.055515739
                 FAM215A          0.275005201
                 SPATA45          0.210597080
                C22ORF15          0.293539950
                    AARD          0.201609957
                 FAM124B          0.117675589
                    FGF9          0.311029319
                   HAND2          0.133147904
                    IGF1          0.064014124
                    KRT1         -0.273581146
                   LHCGR          0.044841824
                    LHFP          0.224006336
                  PPAP2A          0.226627464
                   S1PR1          0.321476697
                     SRL          0.380425036
                  TBC1D1          0.113699356
                   TPSB2          0.006330851
                 TSC22D3          0.223649126
                  TSPYL2          0.053085613
```

GEORGIA SOUTHERN UNIVERSITY

Graphs

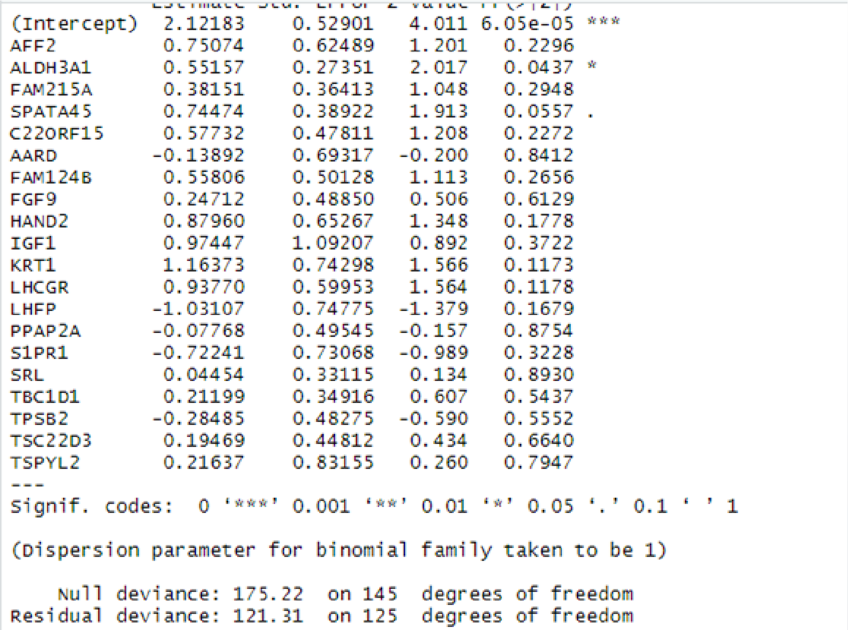




GEORGIA SOUTHERN UNIVERSITY

Final Model

- We used logistics regression in our final model and one gene was significant, which is ALDH3A1.



```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.12183    0.52901   4.011 6.05e-05 ***
AFF2          0.75074    0.62489   1.201   0.2296
ALDH3A1       0.55157    0.27351   2.017   0.0437 *
FAM215A       0.38151    0.36413   1.048   0.2948
SPATA45       0.74474    0.38922   1.913   0.0557 .
C22ORF15      0.57732    0.47811   1.208   0.2272
AARD         -0.13892    0.69317  -0.200   0.8412
FAM124B       0.55806    0.50128   1.113   0.2656
FGF9          0.24712    0.48850   0.506   0.6129
HAND2         0.87960    0.65267   1.348   0.1778
IGF1          0.97447    1.09207   0.892   0.3722
KRT1          1.16373    0.74298   1.566   0.1173
LHCGR         0.93770    0.59953   1.564   0.1178
LHFP         -1.03107    0.74775  -1.379   0.1679
PPAP2A       -0.07768    0.49545  -0.157   0.8754
S1PR1        -0.72241    0.73068  -0.989   0.3228
SRL           0.04454    0.33115   0.134   0.8930
TBC1D1        0.21199    0.34916   0.607   0.5437
TPSB2        -0.28485    0.48275  -0.590   0.5552
TSC22D3       0.19469    0.44812   0.434   0.6640
TSPYL2        0.21637    0.83155   0.260   0.7947
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.22  on 145  degrees of freedom
Residual deviance: 121.31  on 125  degrees of freedom
```

Discussion

- Pancreatic cancer is the fourth leading cause of cancer death, and it has an increasing trend of incidence and poor prognosis after diagnosis.

- Risk factors should be identified, and preventive measures should be taken accordingly.

- The genetic syndromes account for 20% of familial pancreatic cancer, there are other yet undiscovered familial pancreatic cancer genes.

- Hence more research is needed to check for association between specific genes and pancreatic cancer.

Reference

- 1) PANCREATIC CANCER PROGNOSIS & SURVIVAL. Retrieved from https://pancreatica.org/ on December 12, 2020.
- 2) Familial Pancreatic Cancer. Retrieved from https://www.cancer.net/ on December 12, 2020
- 3) Cancer Stat Facts: Pancreatic Cancer. Retrieved https://seer.cancer.gov/statfacts/html/pancreas.html on December 12, 2020
- 4) Linear Model Selection and Regularization. James et al.An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics, DOI 10.1007/978-1-4614-7
- 5) Statistical Consulting(Recursive Partitioning Modeling. Kao-Tai Tsai, Ph.D. Jiann-Ping Hsu College of Public Health. Georgia Southern University, Statesboro, GA. July 19, 2020.
- 6) Statistical Consulting(Data Analysis and Data Quality. Kao-Tai Tsai, Ph.D. Jiann-Ping Hsu College of Public Health. Georgia Southern University, Statesboro, GA. July 19, 2020.
- 7) Statistical Consulting(Examining Data Distribution. Kao-Tai Tsai, Ph.D. Jiann-Ping Hsu College of Public Health. Georgia Southern University, Statesboro, GA. July 19, 2020

Thank you !

We would like to thank Dr. Kao-Tai Tsai and Dr. Karl Peace
for guiding us!