Analysing Lung Cancer Survival Time

Mohammad Saad Khan

May 8, 2020

Contents

Description

Data source

The data is taken from the data challenge *Predicting lung cancer survival time by OWKIN* on Challenge data. (Link)

Data Description

The data is based on CT scans of the lungs of patients diagnosed with lung cancer.
The data consists of four elements:
1. Survival time and censorship (1=death observed, 0=escaped from study)
2. Images (one scan and one mask per patient)
3. Radiomics features (an ensemble of 53 quantitative features per patient, extracted from the scan)
4. Clinical data (contains basic meta-information for each patient)

Note: Since the radiomics features were extracted from the scans, I will not be using the images in the analysis.

```
# Loading the packages
library(survival)
library(mgcv)
library(survminer)
```

Reading the data

```
surv <- read.csv("survival_data.csv")
radiomics <- read.csv("radiomics.csv")
clinical <- read.csv("clinical_data.csv", na.strings = c("nan"))
# Converting entries to lower case in column Histology
clinical$Histology <- as.factor(tolower(clinical$Histology))
# Converting M,N and T stage to categorical variables
clinical$Mstage <- as.factor(clinical$Mstage)
clinical$Nstage <- as.factor(clinical$Nstage)
clinical$Tstage <- as.factor(clinical$Tstage)
# Merging into a single dataframe
df <- merge(surv, clinical)
df <- merge(df, radiomics)
# Displaying other features as radiomics features are too many
head(merge(surv,clinical))
```

```
##   PatientID SurvivalTime Event               Histology Mstage Nstage
## 1         2          638     0          adenocarcinoma      0      0
## 2         3          421     0 squamous cell carcinoma      0      3
## 3         4          465     1 squamous cell carcinoma      0      0
## 4         5         1295     1                     nos      0      0
## 5         7         1393     0 squamous cell carcinoma      0      0
## 6         8         1076     1                     nos      0      0
##   SourceDataset Tstage     age
## 1            11      3 59.4223
```

```
## 2            l1      1 77.0986
## 3            l1      3 84.5722
## 4            l1      2 71.8439
## 5            l1      4 60.7283
## 6            l1      1 82.4093
```

Understanding features: Brief Descriptive Satistics

```
cat("Features:",length(df),"\n")
```

```
## Features: 62
```

```
cat("Sample size:",nrow(df),"\n")
```

```
## Sample size: 300
```

```
# Survival Time
summary(df$SurvivalTime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.0   302.5   644.0   867.1  1298.5  3500.0
```

Histology

```
# Non Small Cell Lung Cancer can itself be split into four major subtypes
# based on histology observations:
# squamous cell carcinoma, large cell carcinoma, adenocarcinoma and a mixture of all
table(df$Histology, useNA = "always")
```

```
##
##                     adenocarcinoma                         large cell
##                                101                                 61
##                                nos nsclc nos (not otherwise specified)
##                                 38                                  2
##             squamous cell carcinoma                              <NA>
##                                 78                                 20
```

```
# Event (1=death observed, 0=escaped from study)
table(df$Event)
```

```
##
##   0   1
## 138 162
```

TNM staging (known prognostic biomarker for survival)

```r
# Mstage
summary(df$Mstage)
```

```
##   0   1   3
## 294   3   3
```

```r
# Nstage
summary(df$Nstage)
```

```
##   0   1   2   3
## 167  17  76  40
```

```r
# Tstage
summary(df$Tstage)
```

```
##   1   2   3   4   5
## 105 111  33  50   1
```

```r
# SourceDataset
summary(df$SourceDataset)
```

```
##  l1  l2
## 199 101
```

```r
# Age
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   42.51   62.98   69.95   68.77   76.20   87.13      16
```

Radiomics features

Details over each computed feature can be found at link

Since there are a lot of features (53) I will describe a few important ones which are used by the baseline model in the challenge

```r
# Sphericity: a measure of the roundness of the shape of the tumor region relative to a sphere
summary(df$original_shape_Sphericity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3129  0.5566  0.6227  0.6187  0.6929  0.8661
```

```r
# SurfaceVolumeRatio: measure of the compactness of the tumor, related to its size
summary(df$original_shape_SurfaceVolumeRatio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09271 0.20097 0.29629 0.33253 0.42629 0.82871
```

```r
# Maximum3DDiameter: The biggest diameter measurable from the tumor volume
summary(df$original_shape_Maximum3DDiameter)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.45   33.92   57.82   63.74   82.93  216.72
```

```r
# JointEntropy: specifying the randomness in the image pixel values
summary(df$original_glcm_JointEntropy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.616   5.958   7.448   7.335   8.775  10.362
```

```r
# Inverse Difference Normalized (IDN): a measure of the local homogeneity of the tumor
summary(df$original_glcm_Idn)
```

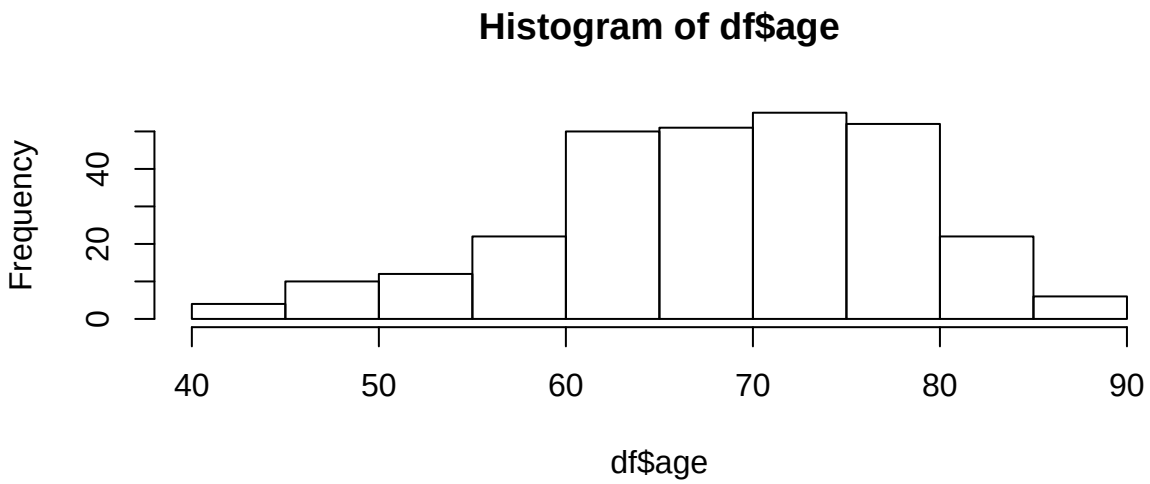
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8530  0.9185  0.9511  0.9451  0.9745  0.9943
```

```r
# Inverse Difference Moment Normalized (IDMN): another measurement of the
# local homogeneity of the tumor
summary(df$original_glcm_Idmn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9465  0.9827  0.9926  0.9888  0.9974  0.9998
```

Modelling impact of Age

```r
# Distribution
hist(df$age)
```



```r
# Applying Logistic Regression (family=binomial)
age_fit1 <- glm(Event ~ I(age/10), data = df, family = "binomial")
summary(age_fit1)
```

```
##
## Call:
## glm(formula = Event ~ I(age/10), family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4240  -1.2140   0.9812   1.1058   1.3932
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5018     0.8744  -1.718   0.0859 .
## I(age/10)     0.2370     0.1261   1.880   0.0602 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 392.57  on 283  degrees of freedom
## Residual deviance: 388.98  on 282  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 392.98
##
## Number of Fisher Scoring iterations: 4
```

```r
# Odds Ratio
exp(coef(age_fit1)[2])
```

```
## I(age/10)
##  1.267465
```

Observations

* The positive coefficient (0.02370) means a positive association with risk of death.
* The p value (0.0602) is greater than 0.05 which means that the relationship between age and risk of death is not significant
* The Odds Ratio for increase in age by a decade is 1.267465 which means about 26% increase in the risk

Checking for Non Linear effects

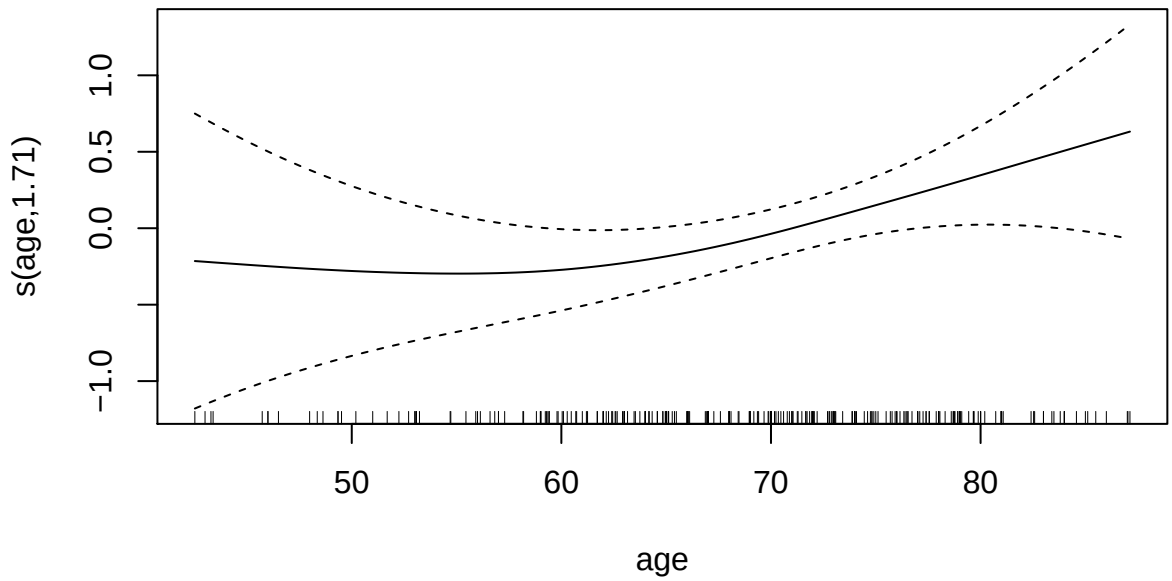
```r
# Fitting a smoothing spline for age using a Generalized additive model
age_fit2 <- gam(Event ~ s(age), data = df, family = "binomial")
summary(age_fit2)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Event ~ s(age)
```

```
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1296     0.1199   1.082    0.279
##
## Approximate significance of smooth terms:
##          edf Ref.df Chi.sq p-value
## s(age) 1.708  2.144  4.659  0.0997 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0145   Deviance explained = 1.49%
## UBRE = 0.38077  Scale est. = 1         n = 284
```

The edf(estimated degrees of freedom) is 1.7 which implies close to a quadratic relationship which is ineed confirmed when we look at the plot.

```
plot(age_fit2)
```




```
# Checking for quadratic realtionship
age_fit3 <- glm(Event ~ age + I(age^2), data = df, family = "binomial")
summary(age_fit3)
```

```
##
## Call:
## glm(formula = Event ~ age + I(age^2), family = "binomial", data = df)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.6836  -1.1607   0.8505   1.1466   1.2521
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.854995   4.581345   1.278    0.201
## age         -0.202997   0.139374  -1.456    0.145
## I(age^2)     0.001709   0.001048   1.630    0.103
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 392.57  on 283  degrees of freedom
## Residual deviance: 386.27  on 281  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 392.27
##
## Number of Fisher Scoring iterations: 4
```

The p value for both linear and quadratic terms is similar which indicates that both are important here. This confirms our previous hypothesis

Modelling impact of Histology

```
# Distribution wrt Event
with(df, table(Event, Histology))
```

```
##      Histology
## Event adenocarcinoma large cell nos nsclc nos (not otherwise specified)
##     0             60         25  11                                    0
##     1             41         36  27                                    2
##      Histology
## Event squamous cell carcinoma
##     0                      34
##     1                      44
```

```
# Logistic Regression
hist_fit <- glm(Event ~ Histology , data = df, family = "binomial")
summary(hist_fit)
```

```
##
## Call:
## glm(formula = Event ~ Histology, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5746  -1.0876   0.8267   1.0701   1.3428
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                               -0.3808     0.2026  -1.879
```

```
## Histologylarge cell                              0.7454     0.3299   2.260
## Histologynos                                     1.2787     0.4111   3.110
## Histologynsclc nos (not otherwise specified)    14.9468   624.1939   0.024
## Histologysquamous cell carcinoma                 0.6386     0.3053   2.092
##                                                Pr(>|z|)
## (Intercept)                                     0.06022 .
## Histologylarge cell                             0.02385 *
## Histologynos                                    0.00187 **
## Histologynsclc nos (not otherwise specified)    0.98090
## Histologysquamous cell carcinoma                0.03645 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 386.73  on 279  degrees of freedom
## Residual deviance: 371.56  on 275  degrees of freedom
##   (20 observations deleted due to missingness)
## AIC: 381.56
##
## Number of Fisher Scoring iterations: 13
```

This indicates that the category "nos" (mixture of all subtypes) has a very significant relation with death. The categories "large cell" and "squamous cell carcinoma" have signifcant impact when compared to the category "adenocarcinoma"

Odds Ratio and Confidence Interval for type "nos"

```
# Odds Ratio
exp(1.2787)
```

```
## [1] 3.591967
```

```
# Confidence Interval
exp(confint(hist_fit)[3,])
```

```
##    2.5 %   97.5 %
## 1.640745 8.313460
```

So a patient having a Histology of "nos" (mixture of all subtypes) has more than 3 times the risk as compared to other sub types

Modelling impact of TNM Staging

```
# Creating a new column combining the three stages
df$TNMstage <-paste(df$Tstage,df$Nstage,df$Mstage)
df$TNMstage <- as.factor(df$TNMstage)
# Distribution wrt Event
with(df, table(Event, TNMstage))
```

```
##      TNMstage
## Event 1 0 0 1 1 0 1 2 0 1 3 0 2 0 0 2 0 1 2 1 0 2 2 0 2 3 0 3 0 0 3 1 0 3 2 0
##     0    52     1     6     2    20     0     3    10    10    11     2     2
##     1    29     3     7     5    28     1     6    22    11     7     1     8
##      TNMstage
## Event 3 3 0 3 3 3 4 0 0 4 1 0 4 2 0 4 2 1 4 2 3 4 3 0 4 3 3 5 2 0
##     0     0     0     9     1     5     0     0     3     0     1
##     1     1     1    10     0    12     2     1     6     1     0
```

```
# Logistic Regression
tnm_fit <- glm(Event ~ TNMstage , data = df, family = "binomial")
summary(tnm_fit)
```

```
##
## Call:
## glm(formula = Event ~ TNMstage, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7941  -0.9925   0.6681   1.0383   1.4823
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.5839     0.2318  -2.520   0.0117 *
## TNMstage1 1 0    1.6826     1.1777   1.429   0.1531
## TNMstage1 2 0    0.7381     0.6027   1.225   0.2207
## TNMstage1 3 0    1.5002     0.8682   1.728   0.0840 .
## TNMstage2 0 0    0.9204     0.3734   2.465   0.0137 *
## TNMstage2 0 1   17.1500  2399.5447   0.007   0.9943
## TNMstage2 1 0    1.2771     0.7441   1.716   0.0861 .
## TNMstage2 2 0    1.3724     0.4463   3.075   0.0021 **
## TNMstage2 3 0    0.6793     0.4946   1.373   0.1696
## TNMstage3 0 0    0.1320     0.5362   0.246   0.8056
## TNMstage3 1 0   -0.1092     1.2465  -0.088   0.9302
## TNMstage3 2 0    1.9702     0.8238   2.392   0.0168 *
## TNMstage3 3 0   17.1500  2399.5447   0.007   0.9943
## TNMstage3 3 3   17.1500  2399.5447   0.007   0.9943
## TNMstage4 0 0    0.6893     0.5146   1.339   0.1804
## TNMstage4 1 0  -15.9821  2399.5447  -0.007   0.9947
## TNMstage4 2 0    1.4594     0.5806   2.514   0.0119 *
## TNMstage4 2 1   17.1500  1696.7344   0.010   0.9919
## TNMstage4 2 3   17.1500  2399.5447   0.007   0.9943
## TNMstage4 3 0    1.2771     0.7441   1.716   0.0861 .
## TNMstage4 3 3   17.1500  2399.5447   0.007   0.9943
## TNMstage5 2 0  -15.9821  2399.5447  -0.007   0.9947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 413.97  on 299  degrees of freedom
## Residual deviance: 378.19  on 278  degrees of freedom
## AIC: 422.19
##
```

```
## Number of Fisher Scoring iterations: 15
```

A few TNM stage groupings are at a higher risk: 2 0 0, 2 2 0 (highest), 3 2 0, 4 2 0

Odds ratio and Confidence Interval for TNM grouping 2 2 0

```
# Odds ratio
exp(1.3724)
```

```
## [1] 3.944807
```

```
# Confidence Interval
exp(confint(tnm_fit)[8,])
```

```
##    2.5 %   97.5 %
## 1.682069 9.796972
```

A patient with TNM grouping 2 2 0 is at 4 times more risk than other groupings.

Impact of Radiomics Features

```
# Using the 6 important features
fit <- glm(Event ~ original_shape_Sphericity + original_shape_SurfaceVolumeRatio + original_shape_Maximu
summary(fit)
```

```
##
## Call:
## glm(formula = Event ~ original_shape_Sphericity + original_shape_SurfaceVolumeRatio +
##     original_shape_Maximum3DDiameter + original_glcm_JointEntropy +
##     original_glcm_Idn + original_glcm_Idmn, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7450  -1.1029   0.7293   1.0115   1.5976
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     22.504855  31.068920   0.724   0.4688
## original_shape_Sphericity        3.162908   1.961607   1.612   0.1069
## original_shape_SurfaceVolumeRatio   2.743628   1.568977   1.749   0.0803 .
## original_shape_Maximum3DDiameter    0.009064   0.006931   1.308   0.1910
## original_glcm_JointEntropy       0.152960   0.198239   0.772   0.4404
## original_glcm_Idn               56.857238  27.938014   2.035   0.0418 *
## original_glcm_Idmn             -81.542018  54.662415  -1.492   0.1358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 413.97  on 299  degrees of freedom
## Residual deviance: 385.96  on 293  degrees of freedom
## AIC: 399.96
##
## Number of Fisher Scoring iterations: 4
```

Observations
* The only p value that indicates any impact on risk is of "original_glcm_Idn", rest are quite high meaning no significant impact
* These are tumor features which intuitively should have high impact but we see the contrary here

Impact of SourceDataset

```
fit <- glm(Event ~ SourceDataset ,data = df, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = Event ~ SourceDataset, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4357  -0.9060   0.9394   0.9394   1.4756
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.5894     0.1480   3.983 6.81e-05 ***
## SourceDataset12  -1.2677     0.2574  -4.926 8.41e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 413.97  on 299  degrees of freedom
## Residual deviance: 388.35  on 298  degrees of freedom
## AIC: 392.35
##
## Number of Fisher Scoring iterations: 4
```

Observations
* The small p values indicate a very highly significant impact on risk of death

Modelling multiple covariates

```
fit <- glm(Event ~ I(age/10) + original_glcm_Idn  + SourceDataset + Histology
           + TNMstage , data = df, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = Event ~ I(age/10) + original_glcm_Idn + SourceDataset +
##     Histology + TNMstage, family = "binomial", data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05569  -0.98051   0.00047   0.93836   1.89204
##
## Coefficients:
##                                                 Estimate Std. Error z value
## (Intercept)                                     -8.93265    6.32441  -1.412
## I(age/10)                                        0.39058    0.15673   2.492
## original_glcm_Idn                                6.74019    6.65045   1.013
## SourceDataset12                                 -1.16354    0.49433  -2.354
## Histologylarge cell                             -0.64502    0.49622  -1.300
## Histologynos                                    -0.12201    0.56112  -0.217
## Histologynsclc nos (not otherwise specified)   17.02341 1429.09748   0.012
## Histologysquamous cell carcinoma                -0.10112    0.42041  -0.241
## TNMstage1 1 0                                    1.47047    1.24231   1.184
## TNMstage1 2 0                                    0.89982    0.68493   1.314
## TNMstage1 3 0                                    0.96825    0.97252   0.996
## TNMstage2 0 0                                    0.71823    0.46898   1.531
## TNMstage2 0 1                                   18.32083 2399.54476   0.008
## TNMstage2 1 0                                    0.34293    0.94966   0.361
## TNMstage2 2 0                                    1.03148    0.56129   1.838
## TNMstage2 3 0                                    0.24287    0.64434   0.377
## TNMstage3 0 0                                   -0.09032    0.65418  -0.138
## TNMstage3 1 0                                   -0.20186    1.35259  -0.149
## TNMstage3 2 0                                    1.68653    0.98509   1.712
## TNMstage3 3 0                                   17.69026 2399.54478   0.007
## TNMstage3 3 3                                   16.74690 2399.54477   0.007
## TNMstage4 0 0                                    0.27018    0.64787   0.417
## TNMstage4 1 0                                  -16.08666 2399.54476  -0.007
## TNMstage4 2 0                                    1.09645    0.70579   1.554
## TNMstage4 2 1                                   18.35689 2399.54477   0.008
## TNMstage4 2 3                                   17.41998 2399.54480   0.007
## TNMstage4 3 0                                    1.15143    0.86825   1.326
## TNMstage4 3 3                                   16.24798 2399.54478   0.007
##                                                Pr(>|z|)
## (Intercept)                                      0.1578
## I(age/10)                                        0.0127 *
## original_glcm_Idn                                0.3108
## SourceDataset12                                  0.0186 *
## Histologylarge cell                              0.1936
## Histologynos                                     0.8279
## Histologynsclc nos (not otherwise specified)     0.9905
## Histologysquamous cell carcinoma                 0.8099
## TNMstage1 1 0                                    0.2366
## TNMstage1 2 0                                    0.1889
## TNMstage1 3 0                                    0.3194
## TNMstage2 0 0                                    0.1257
## TNMstage2 0 1                                    0.9939
## TNMstage2 1 0                                    0.7180
```

```
## TNMstage2 2 0                                    0.0661 .
## TNMstage2 3 0                                    0.7062
## TNMstage3 0 0                                    0.8902
## TNMstage3 1 0                                    0.8814
## TNMstage3 2 0                                    0.0869 .
## TNMstage3 3 0                                    0.9941
## TNMstage3 3 3                                    0.9944
## TNMstage4 0 0                                    0.6767
## TNMstage4 1 0                                    0.9947
## TNMstage4 2 0                                    0.1203
## TNMstage4 2 1                                    0.9939
## TNMstage4 2 3                                    0.9942
## TNMstage4 3 0                                    0.1848
## TNMstage4 3 3                                    0.9946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 373.57  on 269  degrees of freedom
## Residual deviance: 312.43  on 242  degrees of freedom
##   (30 observations deleted due to missingness)
## AIC: 368.43
##
## Number of Fisher Scoring iterations: 15
```

Observations
1. The risk associated with age and some TNM groupings becomes stronger
2. The risk associated with various sub groups of Histology is not at all significant now. This seems to be an interesting oberservation as it suggests that type of Histology does not impact the risk of death.

```
# Odds ratio for age
exp(0.4097)
```

```
## [1] 1.506366
```

The Odds Ratio for decade increase in age now becomes 1.5 (previously 1.26)

```
# Odds ratio for TNM grouping 2 2 0
exp(1.6823)
```

```
## [1] 5.377911
```

The Odds Ratio for TNM is 5.37 now (previously 3.94)

Survival Analysis

Kaplan-Meyer Estimator

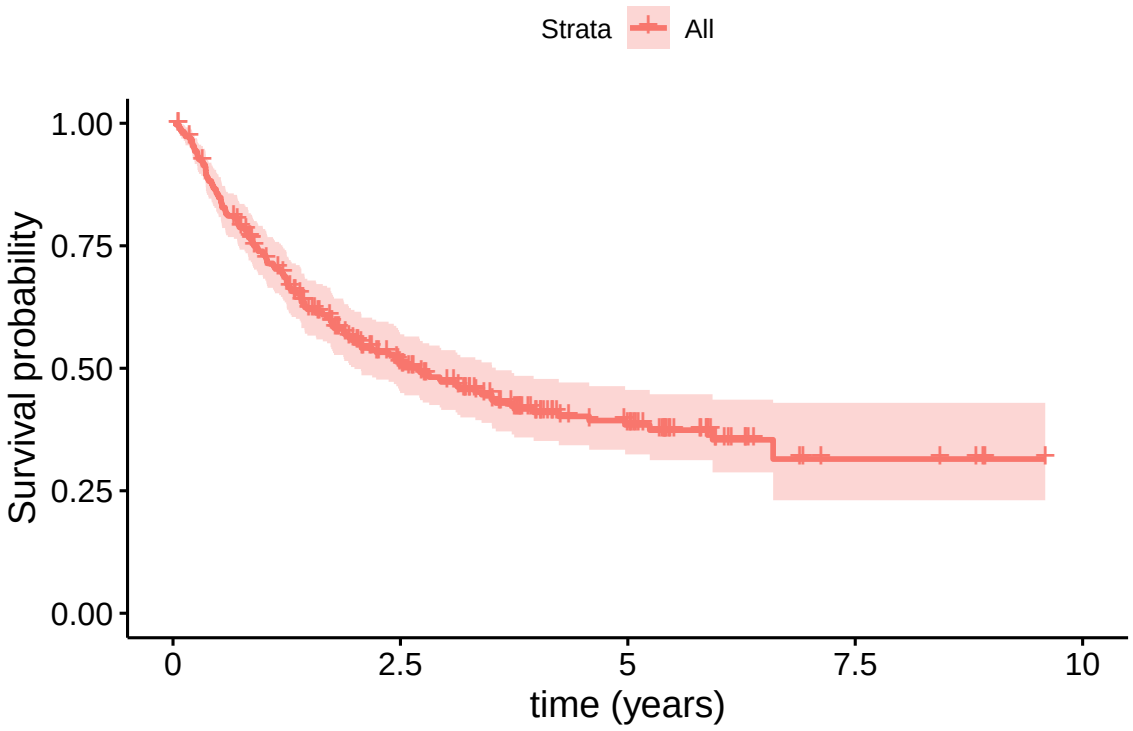
```r
# Taking SurvivalTime in years
KM <- survfit(Surv(SurvivalTime/365, Event)~1, data=df)
KM
```

```
## Call: survfit(formula = Surv(SurvivalTime/365, Event) ~ 1, data = df)
##
##       n  events  median 0.95LCL 0.95UCL
##  300.00  162.00    2.71    1.99    3.55
```

The median survival time is 988 days or 2.7 years

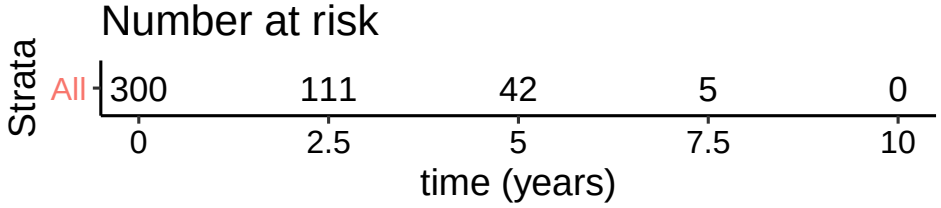
The following plot shows the survival rate of patients

```r
p1 <- ggsurvplot(KM, risk.table = TRUE, main = "Kaplan-Meier estimator",
     ylab = "Survival probability",
     xlab = "time (years)")
p1$plot
```



The following table indicates the patients still alive.

```r
p1$table
```



What's the estimated probability of surviving for (at least) 1 year?

```
summary(KM, time = 1)
```

```
## Call: survfit(formula = Surv(SurvivalTime/365, Event) ~ 1, data = df)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    208      79    0.732  0.0259        0.683        0.784
```

It is 73.2% which means that there are high chances of surviving.

Survival differences between various TNMstage groupings

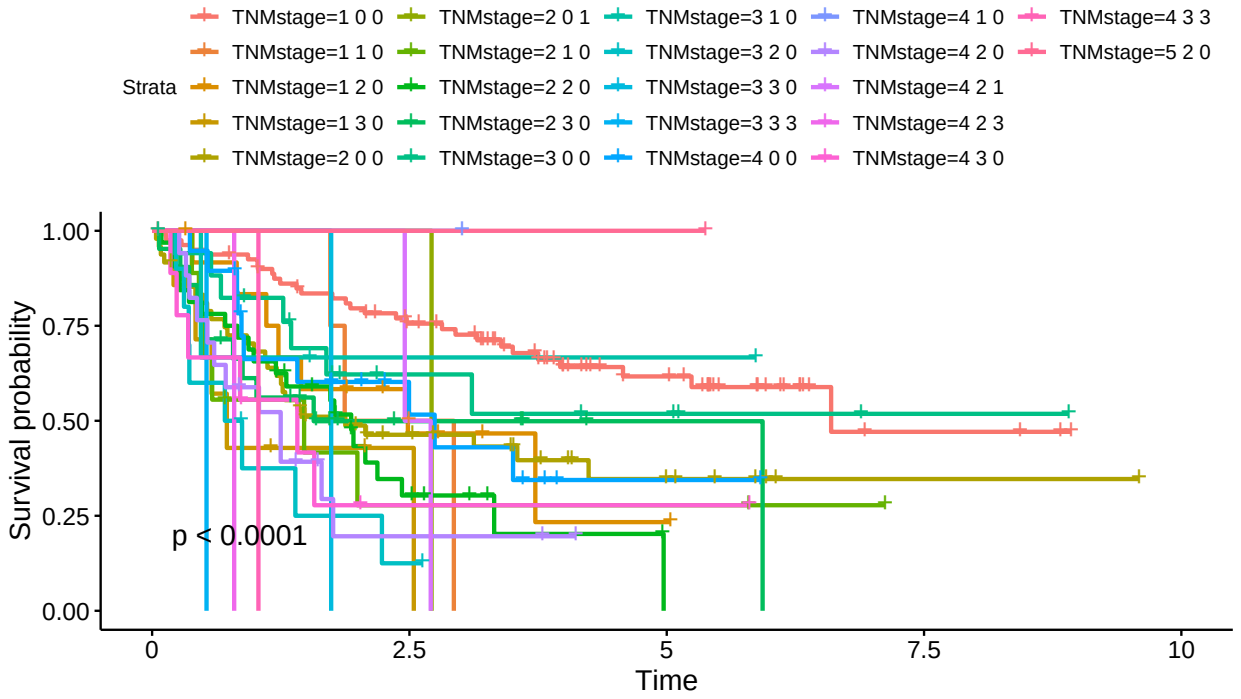
A few TNMstage groupings like "2 2 0", "4 2 0" were at higher risk. Can we confirm the same here?

```
survdiff(Surv(I(SurvivalTime/365), Event)~TNMstage, data = df)
```

```
## Call:
## survdiff(formula = Surv(I(SurvivalTime/365), Event) ~ TNMstage,
##     data = df)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## TNMstage=1 0 0 81       29   60.101  16.09419  26.52162
## TNMstage=1 1 0  4        3    2.469   0.11423   0.11642
## TNMstage=1 2 0 13        7    6.810   0.00527   0.00552
## TNMstage=1 3 0  7        5    2.238   3.41011   3.47408
## TNMstage=2 0 0 48       28   24.362   0.54326   0.64014
## TNMstage=2 0 1  1        1    0.708   0.12068   0.12146
## TNMstage=2 1 0  9        6    4.045   0.94493   0.97362
## TNMstage=2 2 0 32       22   14.397   4.01467   4.44619
## TNMstage=2 3 0 21       11    8.824   0.53635   0.56975
## TNMstage=3 0 0 18        7   10.004   0.90200   0.96559
## TNMstage=3 1 0  3        1    1.608   0.23007   0.23276
## TNMstage=3 2 0 10        8    2.853   9.28504   9.51593
## TNMstage=3 3 0  1        1    0.520   0.44264   0.44515
## TNMstage=3 3 3  1        1    0.176   3.85441   3.87240
## TNMstage=4 0 0 19       10    9.885   0.00133   0.00142
## TNMstage=4 1 0  1        0    0.748   0.74816   0.75314
## TNMstage=4 2 0 17       12    6.111   5.67497   5.94709
## TNMstage=4 2 1  2        2    1.351   0.31149   0.31504
## TNMstage=4 2 3  1        1    0.247   2.29151   2.30190
## TNMstage=4 3 0  9        6    3.230   2.37594   2.43174
## TNMstage=4 3 3  1        1    0.331   1.35168   1.35802
## TNMstage=5 2 0  1        0    0.980   0.97982   0.98880
##
##  Chisq= 57  on 21 degrees of freedom, p= 4e-05
```

The extremely low p value indicates that there is a lot of difference between the TNMstage groupings

```
p2 <- ggsurvplot(survfit(Surv(SurvivalTime/365, Event) ~ TNMstage, data=df), pval = TRUE)
p2$plot
```



Looking at the graph it seems that there are more than 2-3 TNM groupings which stand out and have higher risk but "2 2 0", "4 2 0" are not the highest risk ones

Survival differences between various Histology sub groups

From our previous analysis "nos" group is supposed to be the riskiest. Can we confirm the same here?

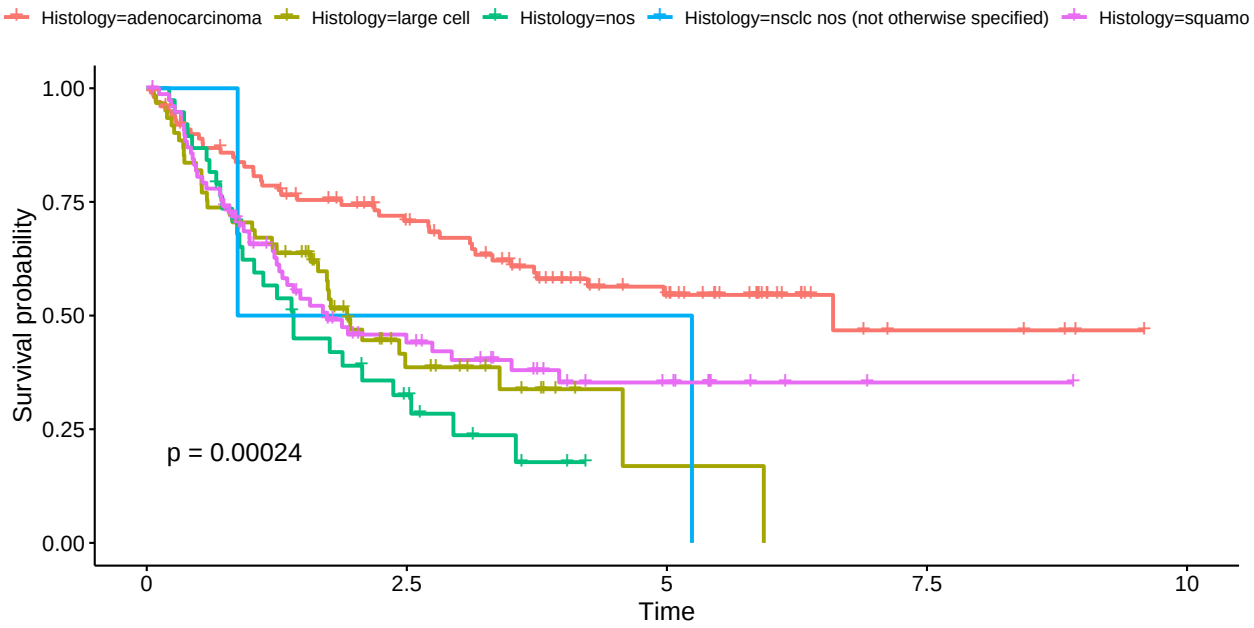
```
survdiff(Surv(SurvivalTime/365, Event)~Histology, data = df)
```

```
## Call:
## survdiff(formula = Surv(SurvivalTime/365, Event) ~ Histology,
##     data = df)
##
## n=280, 20 observations deleted due to missingness.
##
##                                                  N Observed Expected (O-E)^2/E
## Histology=adenocarcinoma                       101       41    66.86     9.999
## Histology=large cell                            61       36    27.81     2.413
## Histology=nos                                   38       27    16.48     6.714
## Histology=nsclc nos (not otherwise specified)    2        2     1.25     0.443
## Histology=squamous cell carcinoma               78       44    37.60     1.089
##                                                 (O-E)^2/V
## Histology=adenocarcinoma                           18.804
```

```
## Histology=large cell                                3.014
## Histology=nos                                       7.652
## Histology=nsclc nos (not otherwise specified)       0.448
## Histology=squamous cell carcinoma                   1.459
##
##  Chisq= 21.6  on 4 degrees of freedom, p= 2e-04
```

The low p value indicates that there is significant difference between the groups which is expected from our analysis.

```
p3 <- ggsurvplot(survfit(Surv(SurvivalTime/365, Event) ~ Histology, data=df), pval = TRUE)
p3$plot
```



Indeed "nos" has the lowest Survival Time.

Cox Regression

On age

```
cph <- coxph(Surv(SurvivalTime, Event) ~ I(age/10) , data = df)
summary(cph)
```

```
## Call:
## coxph(formula = Surv(SurvivalTime, Event) ~ I(age/10), data = df)
##
##   n= 284, number of events= 151
##    (16 observations deleted due to missingness)
##
##              coef exp(coef) se(coef)     z Pr(>|z|)
```

```
## I(age/10) 0.15007   1.16191  0.09066 1.655   0.0979 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## I(age/10)     1.162     0.8607    0.9727     1.388
##
## Concordance= 0.546  (se = 0.025 )
## Likelihood ratio test= 2.79  on 1 df,   p=0.1
## Wald test            = 2.74  on 1 df,   p=0.1
## Score (logrank) test = 2.74  on 1 df,   p=0.1
```

Observations

- The Hazard Ratio (HR) is 1.16 with a Confidence Interval of (0.9727,1.388)
- HR is greater than 1 which indicates that age is positively associated with the event of death

- The p value is large which is in tune with our previous results indicating that age is not a highly significant factor

On Histology

```
cph <- coxph(Surv(SurvivalTime, Event) ~ Histology , data = df)
summary(cph)
```

```
## Call:
## coxph(formula = Surv(SurvivalTime, Event) ~ Histology, data = df)
##
##   n= 280, number of events= 150
##    (20 observations deleted due to missingness)
##
##                                                coef exp(coef) se(coef)     z
## Histologylarge cell                           0.7924    2.2087   0.2333 3.397
## Histologynos                                  1.0348    2.8146   0.2535 4.082
## Histologynsclc nos (not otherwise specified) 0.9575    2.6052   0.7250 1.321
## Histologysquamous cell carcinoma              0.6734    1.9610   0.2188 3.078
##                                               Pr(>|z|)
## Histologylarge cell                           0.000681 ***
## Histologynos                                  4.46e-05 ***
## Histologynsclc nos (not otherwise specified) 0.186611
## Histologysquamous cell carcinoma              0.002087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                               exp(coef) exp(-coef) lower .95
## Histologylarge cell                               2.209     0.4528    1.3983
## Histologynos                                      2.815     0.3553    1.7125
## Histologynsclc nos (not otherwise specified)      2.605     0.3839    0.6291
## Histologysquamous cell carcinoma                  1.961     0.5099    1.2771
##                                               upper .95
## Histologylarge cell                               3.489
```

```
## Histologynos                                        4.626
## Histologynsclc nos (not otherwise specified)       10.788
## Histologysquamous cell carcinoma                    3.011
##
## Concordance= 0.588  (se = 0.023 )
## Likelihood ratio test= 21.91  on 4 df,   p=2e-04
## Wald test            = 20.54  on 4 df,   p=4e-04
## Score (logrank) test = 21.64  on 4 df,   p=2e-04
```

Observations
* As previously observed, "nos" group has the highest significance * All sub groups except "adenocarcinoma" (not shown in the test) have an HR greater than 1 implying positive association with risk of death. This was expected * The p values are small which is in line with our previous findings

On TNMstage

```
cph <- coxph(Surv(SurvivalTime, Event) ~ TNMstage , data = df)
cph
```

```
## Call:
## coxph(formula = Surv(SurvivalTime, Event) ~ TNMstage, data = df)
##
##                       coef  exp(coef)   se(coef)      z        p
## TNMstage1 1 0  9.983e-01  2.714e+00  6.093e-01  1.638 0.101353
## TNMstage1 2 0  8.163e-01  2.262e+00  4.231e-01  1.929 0.053706
## TNMstage1 3 0  1.646e+00  5.189e+00  4.894e-01  3.364 0.000768
## TNMstage2 0 0  9.041e-01  2.470e+00  2.658e-01  3.401 0.000671
## TNMstage2 0 1  1.126e+00  3.082e+00  1.019e+00  1.105 0.269286
## TNMstage2 1 0  1.153e+00  3.167e+00  4.503e-01  2.560 0.010459
## TNMstage2 2 0  1.230e+00  3.422e+00  2.872e-01  4.284 1.84e-05
## TNMstage2 3 0  1.019e+00  2.771e+00  3.567e-01  2.858 0.004268
## TNMstage3 0 0  3.975e-01  1.488e+00  4.221e-01  0.942 0.346336
## TNMstage3 1 0  2.826e-01  1.327e+00  1.018e+00  0.278 0.781259
## TNMstage3 2 0  1.881e+00  6.560e+00  4.063e-01  4.630 3.66e-06
## TNMstage3 3 0  1.489e+00  4.432e+00  1.020e+00  1.459 0.144554
## TNMstage3 3 3  2.659e+00  1.428e+01  1.028e+00  2.587 0.009677
## TNMstage4 0 0  7.990e-01  2.223e+00  3.688e-01  2.167 0.030252
## TNMstage4 1 0 -1.451e+01  5.009e-07  2.394e+03 -0.006 0.995165
## TNMstage4 2 0  1.503e+00  4.496e+00  3.486e-01  4.313 1.61e-05
## TNMstage4 2 1  1.183e+00  3.263e+00  7.337e-01  1.612 0.106959
## TNMstage4 2 3  2.303e+00  1.000e+01  1.025e+00  2.247 0.024620
## TNMstage4 3 0  1.427e+00  4.165e+00  4.512e-01  3.162 0.001565
## TNMstage4 3 3  1.987e+00  7.291e+00  1.023e+00  1.942 0.052078
## TNMstage5 2 0 -1.452e+01  4.919e-07  2.037e+03 -0.007 0.994310
##
## Likelihood ratio test=52.84  on 21 df, p=0.000145
## n= 300, number of events= 162
```