Scientific Research Publishing

Advances in Lung Cancer, 2020, 9, 1-15
https://www.scirp.org/journal/alc
ISSN Online: 2169-2726
ISSN Print: 2169-2718

Survival Analysis of Lung Cancer Patients from TCGA Cohort

Ruibin Lyu

Shanghai Shangde Experimental School, Shanghai, China
Email: LyuRuiBin02@gmail.com

How to cite this paper: Lyu, R.B. (2020) Survival Analysis of Lung Cancer Patients from TCGA Cohort. *Advances in Lung Cancer*, **9**, 1-15.
https://doi.org/10.4236/alc.2020.91001

Received: October 9, 2019
Accepted: March 7, 2020
Published: March 10, 2020

Abstract

Lung cancer is one of the leading causes of death worldwide, accounting for an estimated 2.1 million cases in 2018. To analyze the risk factors behind the lung cancer survival, this paper employs two main models: Kaplan-Meier estimator and Cox proportional hazard model [1]. Also, log-rank test and wald test are utilized to test whether a correlation exists or not, which is discussed in detail in later parts of the paper. The aim is to find out the most influential factors for the survival probability of lung cancer patients. To summarize the results, stage of cancer is always a significant factor for lung cancer survival, and time has to be taken into account when analyzing the survival rate of patients in our data sample, which is from TCGA. Future study on lung cancer is also required to make improvement for the treatment of lung cancer, as our data sample might not represent the overall condition of patients diagnosed with lung cancer; also, more appropriate and advanced models should be employed in order to reflect factors that can affect survival rate of patients with lung cancer in detail.

Keywords

Lung Cancer, Survival Analysis, Kaplan-Meier Estimator, Cox Proportional Hazard Model

1. Introduction

Lung cancer, also called lung carcinoma, is a type of cancer that causes uncontrolled rate of cell growth in lung tissues, and it is the leading death-causing cancer among all types of cancer [2]. The two major types of lung cancer are small cell lung cancer and non-small cell lung cancer; both contain different stages regarding the seriousness of the disease. About 85% of the case of lung cancer can be attributed to the non-small cell lung cancer [3]. There are various

risk factors of lung cancer, such as air pollution, personal characteristics, genetics. The dominant and the well-known cause is cigarette smoking, which accounts for about 85% of lung cancer [4], because cigarette contains hazardous chemical components such as nicotine, which speeds up the cell growth and eventually results in tumor and potential malignant lung cancer [5].

To explore and better understand lung cancer, the study of genes is extremely important. Cancer genomics is to provide better treatment via structural genomics, which "measures the activity of genes encoded in our DNA in order to understand which proteins are abnormally active or silenced in cancer cells" [6]. With the huge amount of data on genome, drugs invented can thus be more effective and specific, since they can target those abnormal genes or proteins precisely, instead of killing all cells like chemotherapy [7]. As a result, the survival probability of lung cancer patients would be largely boosted.

In this paper, we studied the impact of several risk factors on the survival of a lung cancer patient cohort, including genomic factors. The data used is from TCGA, which is an organization that gathers tons of gene data of cancer sequence, endeavoring to make contributions to cancer treatment. TCGA's data is relatively convincing, because the teams in TCGA classify cancers, or tumors, into subgroups that can be better analyzed by experts and investigators in the field of lung cancer [8].

2. Methodology

The dataset consists of data of 1145 patients, who were all diagnosed with different types of lung cancer. Important variables in the data include diagnosis age, sex, smoking history, stage of lung cancer, fraction genome altered, and mutation count. Specifically, diagnosis age, fraction genome altered, and mutation count are continuous variables; smoking history, sex, and stage are categorical variables [9]. However, in order to better capture the correlation between these variables and the survival, some continuous variables would be processed and transformed into categorical data in different analyses. For example, the smoking history could be divided into several categories based on the length of smoking. In order to get a better sense of data, tables and histograms are first employed before analysis.

Two main kinds of statistical models are involved in the analysis of dataset.

2.1. Kaplan-Meier Estimator

To understand the relationship between categorical covariables and survival, Kaplan-Meier estimator is used, which is one of the most widely-used non-parametric measures in survival analysis and in medical research. The formula used is:

$$\hat{S}(t) = \prod_{j:\tau_j \le t} \frac{r_j - d_j}{r_j} = \prod_{j:\tau_j \le t} \left(1 - \frac{d_j}{r_j}\right) \tag{1}$$

where t_j is the time; d_j is the number of deaths at t_j; and r_j is the number of indi-

viduals "at risk" right before the jth death time (everyone dead or censored at or after that time). Censorings tied at t_j are included in c_j [10].

2.2. Cox Proportional Hazard Model (Cox PH Model)

On the other hand, to study the effect of multiple factors simultaneously, Cox PH model is a better approach. The formula to measure the hazard ratio between the two groups is:

$$\lambda_i(t, Z_i) = \lambda_0(t)\exp\left(\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi}\right) \tag{2}$$

where $\lambda_0(t)$ is the hazard rate for the control group and $\lambda_i(t)$ is the hazard rate for the treatment group. Z is a vector of covariates, including continuous factors, indicators for categorical factors, and possible interactions (e.g. age by sex interaction). β is the coefficient for each covariate. As a result, with the use of Cox PH model, two groups of people with different condition can be analyzed and thus find out the hazard ratio θ [1] [11]:

$$\theta = \frac{\lambda_i(t)}{\lambda_0(t)} = \exp\left(\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi}\right) \tag{3}$$

To formally draw inference of the relationships, we use log-rank test and wald test for hypothesis testing. Hypothesis testing includes the comparison between p-value (the possibility that data matches null hypothesis) and alpha level (the possibility to reject null hypothesis given the null hypothesis is true). Typically, we used 0.05 value for alpha level in our data analysis, because 0.05 corresponds to the confidence interval of 95% (the most common one). If p-value is smaller than the alpha level, that means we have enough statistical evidence to reject the null hypothesis, and vice versa. Hence, a p-value < 0.05 is required to show a statistically significant effect of a variable on the survival [12] [13] [14].

3. Results

3.1. Kaplan-Meier

First, we used Kaplan-Meier estimator to find out the effect of cancer type, cancer stage, patient sex, and smoking history on survival rate. Summary of data and results are shown in **Figures 1-3**.

```
> # summary data together
> summary(data)
      time            death          smoking         cigarette      Diagnosis.Age       Sex         stage
 Min.   :  0.00   Min.   :0.0000   Min.   :0.0000   Min.   :  0.00   Min.   :38.00   Female:374   Length:927
 1st Qu.:  1.55   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.: 30.00   1st Qu.:60.00   Male  :553   Class :character
 Median :  7.80   Median :0.0000   Median :1.0000   Median : 42.00   Median :67.00                Mode  :character
 Mean   : 18.99   Mean   :0.2816   Mean   :0.9083   Mean   : 48.24   Mean   :66.36
 3rd Qu.: 27.30   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 60.00   3rd Qu.:73.00
 Max.   :224.10   Max.   :1.0000   Max.   :1.0000   Max.   :240.00   Max.   :90.00
                                                    NA's   :200
 Oncotree.Code Fraction.Genome.Altered Mutation.Count
 LUAD:465      Min.   :0.0000          Min.   :   1.0
 LUSC:462      1st Qu.:0.1567          1st Qu.: 120.0
               Median :0.3257          Median : 203.0
               Mean   :0.3308          Mean   : 265.0
               3rd Qu.:0.4793          3rd Qu.: 324.5
               Max.   :0.9373          Max.   :2361.0
```

Figure 1. A summary of all data.

```
> ### logrank test for cancer type: LUAD vs. LUSC
> survdiff(Surv(time, death) ~ Oncotree.Code, data=data)
Call:
survdiff(formula = Surv(time, death) ~ Oncotree.Code, data = data)

                      N Observed Expected (O-E)^2/E (O-E)^2/V
Oncotree.Code=LUAD 465      109      114     0.208     0.374
Oncotree.Code=LUSC 462      152      147     0.161     0.374

 Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

Figure 2. The result of the log-rank test.

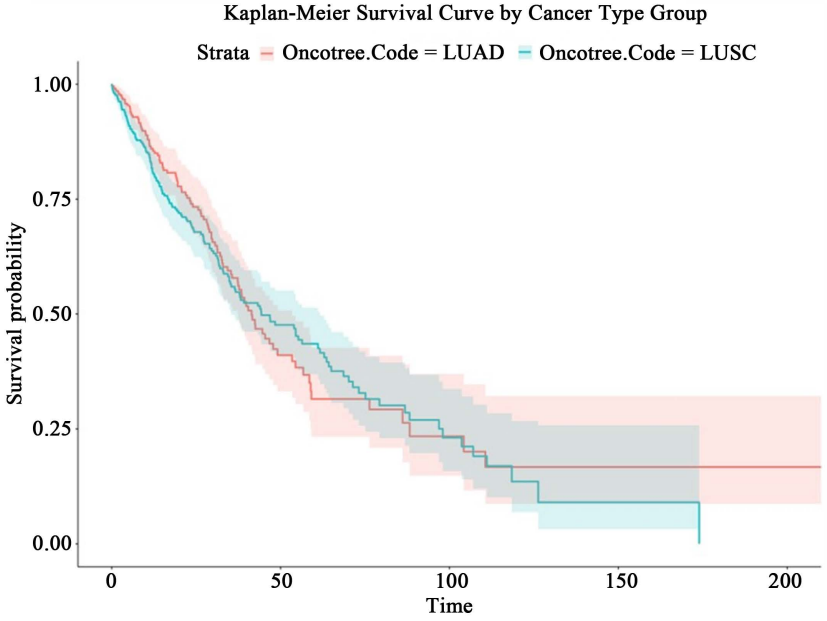

Figure 3. Kaplan-meier survival curve by cancer type group.

3.1.1. Cancer Type

There are two types of cancer involved: LUAD and LUSC. From the Kaplan-Meier survival plot of the two groups, there seems no big difference between the survival rate (survival probability in the y axis) of LUAD and that of LUSC.

After we applied log-rank test, we found the p-value to be 0.5, which is far greater than the alpha level (0.05). Hence, we do not have the statistical evidence to reject the null hypothesis and conclude that the survival rates of patients of different types of lung cancer are roughly the same.

3.1.2. Stage of Cancer

Stage of cancer is divided into four stages, which are I, II, III, and IV, where IV is the worst stage that the cancer cell spreads to different organs. As shown in Figures 4-6, the survival probability of stage IV is the lowest, then III, II, and I, meaning that patients in stage IV have shorter survival compared to the other three stages even with treatments. This is reasonable considering the categorization of the four stages.

Here, the p-value equals to $2e^{-6}$, which is far smaller than the alpha level of 0.05, indicating that the survival probability for different stages of cancer is statistically different from each other.

```
>
> # stage - 4 missing
> table(data$Stage, useNA = "ifany")

   I   IA   IB   II  IIA  IIB  III IIIA IIIB   IV <NA>
   8  214  266    4  110  154    3  129   30   32    4
> # recode stage - stage 1,2,3,4
> data$stage[data$Stage=='IA' | data$Stage=='IB' | data$Stage=='I'] = 'I'
> data$stage[data$Stage=='IIA' | data$Stage=='IIB' | data$Stage=='II'] = 'II'
> data$stage[data$Stage=='IIIA' | data$Stage=='IIIB' | data$Stage=='III'] = 'III'
> data$stage[data$Stage=='IV'] = 'IV'
> data <- data[complete.cases(data$stage),]
> table(data$stage, useNA = "ifany")

  I  II III  IV
488 268 162  32
> # Onotree Code - no missing
> table(data$Oncotree.Code, useNA = "ifany")

LUAD LUSC
 479  471
>
> # Fraction Genome Altered - No missing
> summary(data$Fraction.Genome.Altered)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1507  0.3238  0.3284  0.4776  0.9373
>
> # Mutation Count - No missing
> summary(data$Mutation.Count)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0   119.0   202.0   264.0   323.2  2361.0
> # Smoking History - 23 missing
> table(data$Smoking.History, useNA = "ifany")
```

Figure 4. The survival probability of cancer patients in four stages.

```
> ### logrank test for cancer stage
> survdiff(Surv(time, death) ~ stage, data=data)
Call:
survdiff(formula = Surv(time, death) ~ stage, data = data)

             N Observed Expected (O-E)^2/E (O-E)^2/V
stage=I    481      109   146.10     9.421    21.578
stage=II   259       71    64.88     0.577     0.771
stage=III  155       67    42.13    14.678    17.566
stage=IV    32       14     7.89     4.736     4.905

 Chisq= 29.6  on 3 degrees of freedom, p= 2e-06
```

Figure 5. The calculation of p-value about cancer.

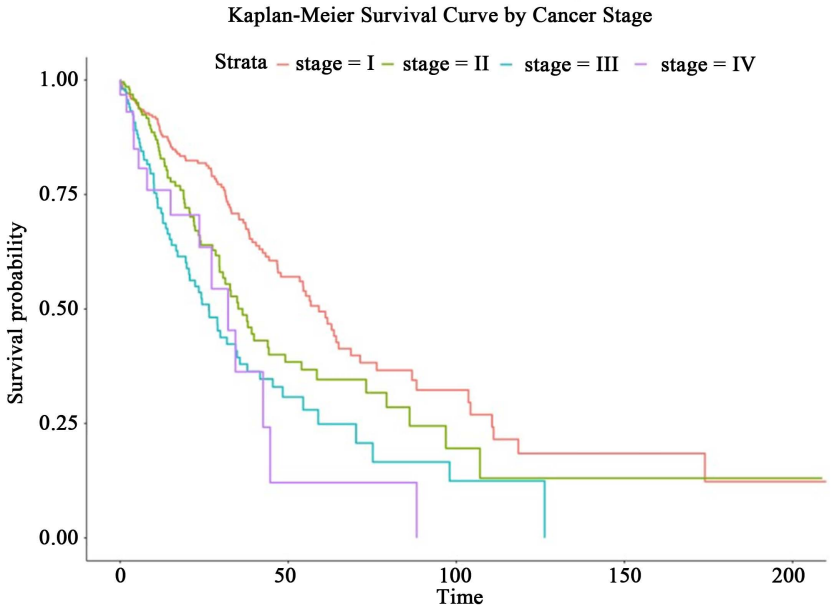

Figure 6. Kaplan-meier survival curve by cancer stage.

3.1.3. Sex

For the gender of patients, there seems no big difference between the survival probability between females and males, as their survival point estimates and confidence intervals largely overlap.

Moreover, the p-value of the log-rank test is 0.7, which is far greater than the alpha level, indicating that there is indeed no difference between the survival of females and males in this cohort.

3.1.4. Smoking

For smoking, we categorized patients in two main groups: smoking = 1 represents patients who have ever smoked during lifetime, and smoking = 2 represents patients who have never smoked during lifetime. However, contrary to our common understanding towards the harm of smoking on health, patients who smoked in our data have a better survival probability than those who never smoked based on Figure 7 and Figure 8. However, since the p-value (0.3) is bigger than the alpha-level (0.05), the differences in survival between smokers and non-smokers are not significant. The wide confidence interval of "smoking = 0" which covers that of "smoking = 1" also indicates the same conclusion. We may lack the power to test the underlying difference due to insufficient sample size in non-smokers.

```
> ### logrank test for sex: female vs. male
> survdiff(Surv(time, death) ~ Sex, data=data)
Call:
survdiff(formula = Surv(time, death) ~ Sex, data = data)

             N Observed Expected (O-E)^2/E (O-E)^2/V
Sex=Female 374      100      103    0.1106     0.185
Sex=Male   553      161      158    0.0726     0.185

 Chisq= 0.2  on 1 degrees of freedom, p= 0.7
```

Figure 7. The calculation of p-value about sex.

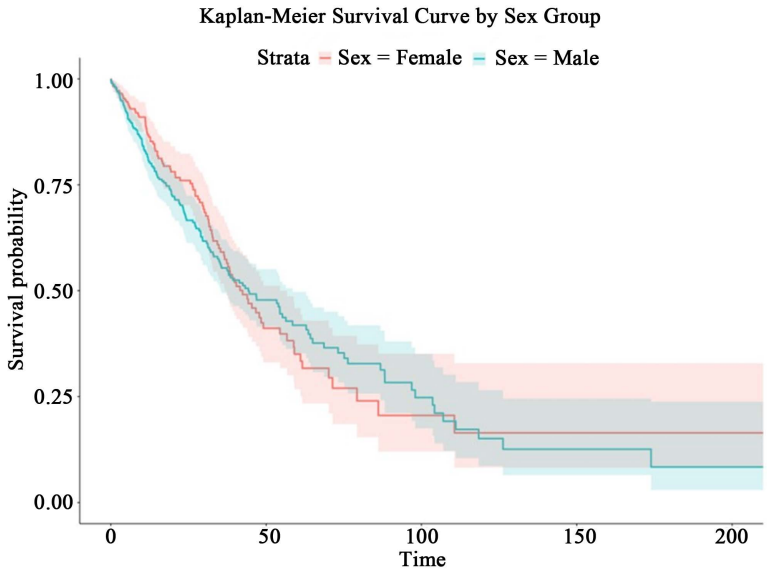

Figure 8. Kaplan-Meier survival curve by sex group.

3.1.5. Fraction of Genome Mutated

For fraction of genome altered (fga), we created two categorical variables based on it. "fga-binary = 1" corresponds to patients whose genome mutated fraction is greater than the average level of this cohort, while "fga-binary = 0" corresponds to those whose genome mutated fraction is smaller than the average level. Based on Figures 9-11, patients with more genome mutated fraction seem to have close survival rate with those with lower genome mutated fraction. As shown in Figure 12 and Figure 13, similar to previous variables, the differences between the two groups are not significant, since the p-value (0.9) is greater than the alpha-level.

3.2. Cox Proportional Hazard Model

To analyze the impact of multiple variables on the survival rate of patients, cox proportional hazard model is utilized. Figure 14 presents the regression output of a baseline cox model including the six variables we are interested in, which are three continuous variables (sex, smoking, and stage), and three continuous variables (diagnosis age, fraction genome altered, and mutation count).

```
> table(data$Smoking.History, useNA = "ifany")

    Current Reformed Smoker For < Or = 15 Years          Current Reformed Smoker For > 15 Years
                                            396                                              201
Current Reformed Smoker, Duration Not Specified                                   Current Smoker
                                              9                                              236
                            Lifelong Non-Smoker                                             <NA>
                                             85                                               23
> # create new categorical variables for smoking history
> data$smoking <- as.numeric(data$Smoking.History)
>         # We have Five categories of smoking history:
>         # 1 - Current Reformed Smoker <= 15 Years
>         # 2 - Current Reformed Smoker > 15 Years
>         # 3 - Current Reformed Smoker, Duration Not Specified
>         # 4 - Current Smoker
>         # 5 - Lifelong Non-Smoker
> # drop missingness in smoking
> data <- data[complete.cases(data$smoking),]
> ### Group people by person who ever smoked and Lifelong Non-Smoker
> data$smoking <- as.numeric(data$Smoking.History)
> # People who ever smoked - code to 1
> data$smoking[data$smoking<5] = 1
> # People who never smoked - code to 0
> data$smoking[data$smoking==5] = 0
> table(data$smoking, useNA = "ifany")

  0   1
 85 842
>
> # Cigarette Smoking: pack per year
> data$cigarette <- data$Person.Cigarette.Smoking.History.Pack.Year.Value
> summary(data$cigarette)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   30.00   42.00   48.24   60.00  240.00     200
>
>
> # keep only interesed variables in the dataset
> data <- data[c('time', 'death', 'smoking', 'cigarette', 'Diagnosis.Age', 'Sex', 'stage',
+               'Oncotree.Code', 'Fraction.Genome.Altered', 'Mutation.Count')]
```

Figure 9. The survival probability of smokers and non-smokers.

```
> ### logrank test for sex: female vs. male
> survdiff(Surv(time, death) ~ smoking, data=data)
Call:
survdiff(formula = Surv(time, death) ~ smoking, data = data)

             N Observed Expected (O-E)^2/E (O-E)^2/V
smoking=0   85       16     12.7    0.8865     0.941
smoking=1 842      245    248.3    0.0452     0.941

 Chisq= 0.9  on 1 degrees of freedom, p= 0.3
```

Figure 10. The calculation of p-value about smoking.

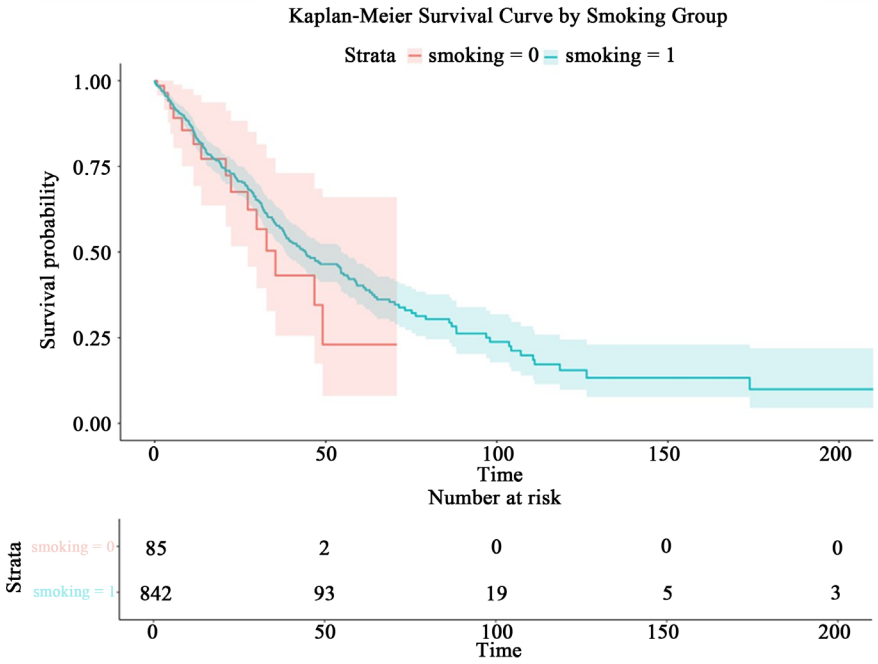

Figure 11. Kaplan-Meier survival by smoking group.

```
> ### logrank test for Fraction.Genome.Altered
> survdiff(Surv(time, death) ~ fga_binary, data=data)
Call:
survdiff(formula = Surv(time, death) ~ fga_binary, data = data)

                N Observed Expected (O-E)^2/E (O-E)^2/V
fga_binary=0 475      125      126   0.00259   0.00501
fga_binary=1 452      136      135   0.00240   0.00501

 Chisq= 0  on 1 degrees of freedom, p= 0.9
```

Figure 12. The calculation of p-value about genome mutated fraction.

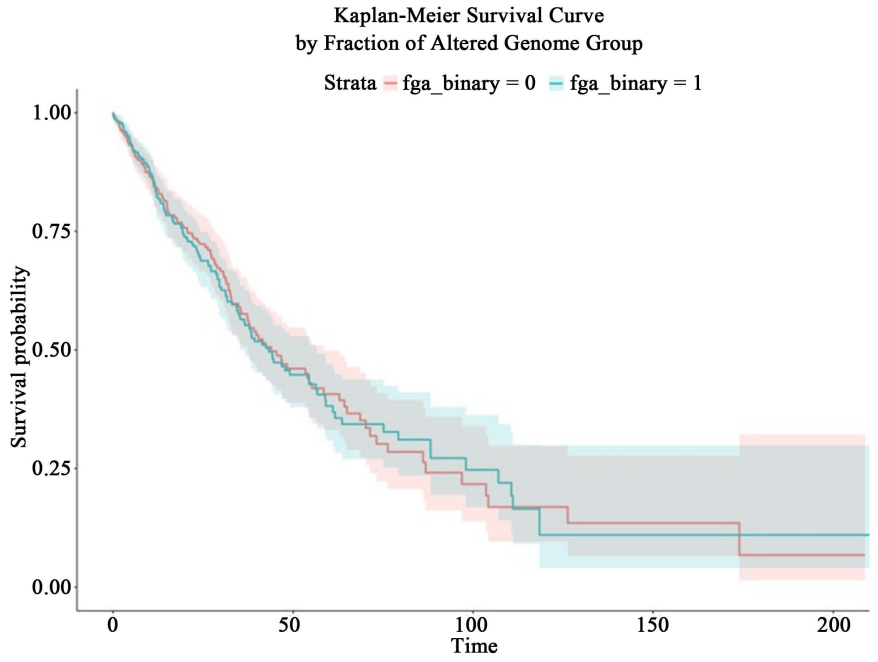

Figure 13. Kaplan-meier survival by fraction of altered genome group.

```
> cox.model <- coxph(Surv(time, death) ~ Diagnosis.Age + Sex + smoking + factor(stage)
+                    + Fraction.Genome.Altered + Mutation.Count,
+                    data = data)
> summary(cox.model)
Call:
coxph(formula = Surv(time, death) ~ Diagnosis.Age + Sex + smoking +
    factor(stage) + Fraction.Genome.Altered + Mutation.Count,
    data = data)

  n= 927, number of events= 261

                             coef  exp(coef)   se(coef)      z Pr(>|z|)
Diagnosis.Age           0.0217242  1.0219619  0.0073272  2.965 0.003028 **
SexMale                -0.0097702  0.9902773  0.1357837 -0.072 0.942638
smoking                -0.1261007  0.8815261  0.2781694 -0.453 0.650316
factor(stage)II         0.4165980  1.5167926  0.1547457  2.692 0.007099 **
factor(stage)III        0.8050871  2.2368914  0.1577620  5.103 3.34e-07 ***
factor(stage)IV         0.9799984  2.6644519  0.2899523  3.380 0.000725 ***
Fraction.Genome.Altered -0.1971632  0.8210566  0.3025165 -0.652 0.514567
Mutation.Count          -0.0001107  0.9998893  0.0002735 -0.405 0.685525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                        exp(coef) exp(-coef) lower .95 upper .95
Diagnosis.Age              1.0220     0.9785    1.0074     1.037
SexMale                    0.9903     1.0098    0.7589     1.292
smoking                    0.8815     1.1344    0.5110     1.521
factor(stage)II            1.5168     0.6593    1.1200     2.054
factor(stage)III           2.2369     0.4470    1.6419     3.047
factor(stage)IV            2.6645     0.3753    1.5094     4.703
Fraction.Genome.Altered    0.8211     1.2179    0.4538     1.486
Mutation.Count             0.9999     1.0001    0.9994     1.000

Concordance= 0.627  (se = 0.022 )
Rsquare= 0.041   (max possible= 0.954 )
Likelihood ratio test= 39.1  on 8 df,   p=5e-06
Wald test            = 39.49  on 8 df,   p=4e-06
Score (logrank) test = 40.72  on 8 df,   p=2e-06
```

Figure 14. The regression output of a base-line cox model.

3.2.1. Diagnosis Age

Based on this output, the hazard ratio of age is 1.02, meaning a high hazard for elders. Also, the p-value (Pr > |z| in the diagram) is 0.003, which is smaller than 0.05, meaning that this association is significant. Meanwhile, stages II, III, and IV also have a significant relationship with survival, and the hazard ratios are 1.52, 2.24, and 2.66 respectively, since all the associated p-values are smaller than 0.05. The higher hazard ratio in higher stages is reasonable, because patients in higher stages tend to have a worse state of lung cancer, thereby have an increasing hazard rate. On the other hand, other variables do not have an obvious impact on hazard ratio, since their p-values are all greater than 0.05. But we should be cautious when interpreting the results. We may fail to capture the true association due to a lack of statistical power.

3.2.2. Fraction of Genome Altered

Since we failed to detect the relationship between fraction of genome altered and hazard ratio in continuous form, to further explore the association, we transform the variable from a continuous variable to a categorical variable, where there are four main groups categorized by the 25th percentile, 50th percentile, and 75th percentile: 0 - 0.157, 0.157 - 0.326, 0.326 - 0.479, and 0.479 - 0.937, as shown in **Figure 15**. However, there is still no statistical evidence suggesting an association between fraction of genome altered and hazard ratio in categorical form, as the p-values are still greater than 0.05 (0.81, 0.55, and 0.22).

3.2.3. Mutation Count

As shown in **Figure 16**, similar to the variable "fraction of genome altered", we

```
> # by quantiles of Fraction of Genome Altered
> data$fga_q <- quantcut(data$Fraction.Genome.Altered, 4)
> cox.model.fga <- coxph(Surv(time, death) ~ Diagnosis.Age + Sex + smoking + factor(stage)
+                   + factor(fga_q) + Mutation.Count,
+                   data = data)
> summary(cox.model.fga)
Call:
coxph(formula = Surv(time, death) ~ Diagnosis.Age + Sex + smoking +
    factor(stage) + factor(fga_q) + Mutation.Count, data = data)

  n= 927, number of events= 261

                                 coef  exp(coef)   se(coef)      z Pr(>|z|)
Diagnosis.Age                0.0216967  1.0219338  0.0073934  2.935  0.00334 **
SexMale                     -0.0250974  0.9752149  0.1366202 -0.184  0.85425
smoking                     -0.1110362  0.8949063  0.2789328 -0.398  0.69057
factor(stage)II              0.4205265  1.5227631  0.1546196  2.720  0.00653 **
factor(stage)III             0.7922826  2.2084317  0.1581573  5.009 5.46e-07 ***
factor(stage)IV              0.9502518  2.5863608  0.2926673  3.247  0.00117 **
factor(fga_q)(0.157,0.326] -0.0447808  0.9562070  0.1836010 -0.244  0.80731
factor(fga_q)(0.326,0.479]  0.1064746  1.1123496  0.1770239  0.601  0.54753
factor(fga_q)(0.479,0.937] -0.2185060  0.8037186  0.1785815 -1.224  0.22112
Mutation.Count              -0.0001274  0.9998726  0.0002751 -0.463  0.64323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                            exp(coef) exp(-coef) lower .95 upper .95
Diagnosis.Age                  1.0219     0.9785    1.0072     1.037
SexMale                        0.9752     1.0254    0.7461     1.275
smoking                        0.8949     1.1174    0.5180     1.546
factor(stage)II                1.5228     0.6567    1.1247     2.062
factor(stage)III               2.2084     0.4528    1.6198     3.011
factor(stage)IV                2.5864     0.3866    1.4574     4.590
factor(fga_q)(0.157,0.326]     0.9562     1.0458    0.6672     1.370
factor(fga_q)(0.326,0.479]     1.1123     0.8990    0.7862     1.574
factor(fga_q)(0.479,0.937]     0.8037     1.2442    0.5664     1.141
Mutation.Count                 0.9999     1.0001    0.9993     1.000

Concordance= 0.635  (se = 0.022 )
Rsquare= 0.045   (max possible= 0.954 )
Likelihood ratio test= 42.4  on 10 df,   p=6e-06
Wald test            = 42.82  on 10 df,   p=5e-06
Score (logrank) test = 44.16  on 10 df,   p=3e-06
```

Figure 15. The calculation of p-value about fraction of genome altered.

```
> # by quantiles of mutation count
> data$mc_q <- quantcut(data$Mutation.Count, 4)
> cox.model.mc <- coxph(Surv(time, death) ~ Diagnosis.Age + Sex + smoking + factor(stage)
+                   + Fraction.Genome.Altered + factor(mc_q),
+                   data = data)
> summary(cox.model.mc)
Call:
coxph(formula = Surv(time, death) ~ Diagnosis.Age + Sex + smoking +
    factor(stage) + Fraction.Genome.Altered + factor(mc_q), data = data)

  n= 927, number of events= 261

                                coef exp(coef)  se(coef)      z Pr(>|z|)
Diagnosis.Age               0.021873  1.022114  0.007294  2.999 0.002710 **
SexMale                     0.002974  1.002978  0.136791  0.022 0.982656
smoking                    -0.117923  0.888764  0.289577 -0.407 0.683842
factor(stage)II             0.432369  1.540903  0.155297  2.784 0.005367 **
factor(stage)III            0.843577  2.324666  0.159265  5.297 1.18e-07 ***
factor(stage)IV             0.963760  2.621536  0.291224  3.309 0.000935 ***
Fraction.Genome.Altered    -0.215529  0.806115  0.308798 -0.698 0.485202
factor(mc_q)(120,203]      -0.158276  0.853614  0.186986 -0.846 0.397295
factor(mc_q)(203,324]       0.059212  1.061001  0.193411  0.306 0.759492
factor(mc_q)(324,2.36e+03] -0.064175  0.937841  0.194001 -0.331 0.740799
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                            exp(coef) exp(-coef) lower .95 upper .95
Diagnosis.Age                  1.0221     0.9784    1.0076     1.037
SexMale                        1.0030     0.9970    0.7671     1.311
smoking                        0.8888     1.1252    0.5038     1.568
factor(stage)II                1.5409     0.6490    1.1365     2.089
factor(stage)III               2.3247     0.4302    1.7013     3.176
factor(stage)IV                2.6215     0.3815    1.4814     4.639
Fraction.Genome.Altered        0.8061     1.2405    0.4401     1.477
factor(mc_q)(120,203]          0.8536     1.1715    0.5917     1.231
factor(mc_q)(203,324]          1.0610     0.9425    0.7262     1.550
factor(mc_q)(324,2.36e+03]     0.9378     1.0663    0.6412     1.372

Concordance= 0.626  (se = 0.022 )
Rsquare= 0.043   (max possible= 0.954 )
Likelihood ratio test= 40.58  on 10 df,   p=1e-05
Wald test            = 40.68  on 10 df,   p=1e-05
Score (logrank) test = 41.98  on 10 df,   p=8e-06
```

Figure 16. The calculation of p-value about the total mutation count.

divided the total mutation count into four categories by quantiles, where the maximum mutation count is 2360. Nevertheless, due to p-values that are greater than 0.05, the categorization of mutation count does not help to detect a significant association between mutation count and hazard ratio.

3.2.4. Stage and Smoking

Moreover, apart from analyzing effects of individual variables on hazard ratio, we add interaction terms in the model to study the effect modification between covariates.

We first focus on the interaction between stage and smoking, shown in Figure 17. We add a variable denoting the product of the two covariates in the model. However, based on the results of modeling, there is no significant interaction between the two variables, as evidenced by the p-value that is greater than 0.05.

3.2.5. Sex and Smoking

We then study the interaction between sex and smoking by adding a new variable denoting the product of the two covariates. Similarly shown in Figure 18, we also failed to capture the interaction between sex and smoking, since the p-value is 0.70, greater than 0.05.

```
> # stage and smoking
> cox.ix.ss <- coxph(Surv(time, death) ~ Diagnosis.Age + Sex + smoking + factor(stage)
+                + Fraction.Genome.Altered + Mutation.Count + factor(stage)*smoking,
+                data = data)
> summary(cox.ix.ss)
Call:
coxph(formula = Surv(time, death) ~ Diagnosis.Age + Sex + smoking +
    factor(stage) + Fraction.Genome.Altered + Mutation.Count +
    factor(stage) * smoking, data = data)

  n= 927, number of events= 261

                               coef  exp(coef)   se(coef)      z Pr(>|z|)
Diagnosis.Age             0.0217978  1.0220371  0.0073589  2.962  0.00306 **
SexMale                  -0.0093736  0.9906702  0.1356615 -0.069  0.94491
smoking                   0.1168727  1.1239763  0.5194975  0.225  0.82200
factor(stage)II           0.5592878  1.7494260  0.6460579  0.866  0.38666
factor(stage)III          1.1887492  3.2829722  0.7082393  1.678  0.09326 .
factor(stage)IV           1.9567460  7.0762633  0.8736380  2.240  0.02511 *
Fraction.Genome.Altered  -0.1944622  0.8232773  0.3026747 -0.642  0.52056
Mutation.Count           -0.0001209  0.9998791  0.0002741 -0.441  0.65928
smoking:factor(stage)II  -0.1438284  0.8660364  0.6651428 -0.216  0.82880
smoking:factor(stage)III -0.4023891  0.6687205  0.7259647 -0.554  0.57939
smoking:factor(stage)IV  -1.0709000  0.3426999  0.9242568 -1.159  0.24659
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
Diagnosis.Age               1.0220     0.9784    1.0074     1.037
SexMale                     0.9907     1.0094    0.7594     1.292
smoking                     1.1240     0.8897    0.4060     3.111
factor(stage)II             1.7494     0.5716    0.4931     6.206
factor(stage)III            3.2830     0.3046    0.8192    13.156
factor(stage)IV             7.0763     0.1413    1.2769    39.214
Fraction.Genome.Altered     0.8233     1.2147    0.4549     1.490
Mutation.Count              0.9999     1.0001    0.9993     1.000
smoking:factor(stage)II     0.8660     1.1547    0.2352     3.189
smoking:factor(stage)III    0.6687     1.4954    0.1612     2.775
smoking:factor(stage)IV     0.3427     2.9180    0.0560     2.097

Concordance= 0.63  (se = 0.022 )
Rsquare= 0.043   (max possible= 0.954 )
Likelihood ratio test= 40.45  on 11 df,   p=3e-05
Wald test            = 41.93  on 11 df,   p=2e-05
Score (logrank) test = 44.16  on 11 df,   p=7e-06
```

Figure 17. The calculation of p-value about stage and smoking.

```
> # sex and smoking
> cox.ix.ss2 <- coxph(Surv(time, death) ~ Diagnosis.Age + Sex + smoking + factor(stage)
+                   + Fraction.Genome.Altered + Mutation.Count + Sex*smoking,
+                   data = data)
> summary(cox.ix.ss2)
Call:
coxph(formula = Surv(time, death) ~ Diagnosis.Age + Sex + smoking +
    factor(stage) + Fraction.Genome.Altered + Mutation.Count +
    Sex * smoking, data = data)

  n= 927, number of events= 261

                             coef  exp(coef)   se(coef)      z Pr(>|z|)
Diagnosis.Age            0.0221351  1.0223819  0.0073953  2.993 0.002761 **
SexMale                  0.2780570  1.3205615  0.7680501  0.362 0.717330
smoking                 -0.0886358  0.9151788  0.2982190 -0.297 0.766301
factor(stage)II          0.4176624  1.5184080  0.1547948  2.698 0.006972 **
factor(stage)III         0.8073774  2.2420204  0.1578951  5.113 3.16e-07 ***
factor(stage)IV          0.9794769  2.6630627  0.2899771  3.378 0.000731 ***
Fraction.Genome.Altered -0.1892021  0.8276193  0.3031564 -0.624 0.532557
Mutation.Count          -0.0001113  0.9998887  0.0002737 -0.407 0.684226
SexMale:smoking         -0.2970790  0.7429853  0.7829313 -0.379 0.704358
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                        exp(coef) exp(-coef) lower .95 upper .95
Diagnosis.Age              1.0224     0.9781    1.0077     1.037
SexMale                    1.3206     0.7573    0.2931     5.950
smoking                    0.9152     1.0927    0.5101     1.642
factor(stage)II            1.5184     0.6586    1.1211     2.057
factor(stage)III           2.2420     0.4460    1.6453     3.055
factor(stage)IV            2.6631     0.3755    1.5085     4.701
Fraction.Genome.Altered    0.8276     1.2083    0.4569     1.499
Mutation.Count             0.9999     1.0001    0.9994     1.000
SexMale:smoking            0.7430     1.3459    0.1602     3.447

Concordance= 0.626  (se = 0.022 )
Rsquare= 0.041   (max possible= 0.954 )
Likelihood ratio test= 39.24  on 9 df,   p=1e-05
Wald test            = 39.62  on 9 df,   p=9e-06
Score (logrank) test = 40.83  on 9 df,   p=5e-06
```

Figure 18. The calculation of p-value about sex and smoking.

3.3. Time Added to Different Variables

The assumption for Cox PH model is that hazard ratio does not depend on time (t), *i.e.* the hazards of the two groups remain proportional over time, the hazard ratio between t1 and t2 is the same as that for t2 and t3 in the sample. However, this assumption seems to be violated based on the Kaplan Meier analysis and above cox models. For example, due to the cross between two lines "fga_binary = 0" and "fga_binary = 1" in the diagram below, the hazard ratio between the two groups changes, and even reverse, over time. The impact of some other covariates on survival rate also seems to change with time. As a result, time should be taken into account in modeling. Thus, we revise the Cox PH model by adding new covariates indicating the interaction between time and other covariates.

Based on the output in **Figure 19** and **Figure 20**, all variables except diagnosis age have an extra covariate with time. The diagnosis age effect is thus not significant anymore because there is a colinearity between time and age. For other variables, we obtained both significant main effect and interaction between time and sex, smoking, and stage. For example, "smoking_time" denotes the difference in hazard ratio between smokers and non-smokers as time increases. Hence, given the exp (coef), hazard ratio, is 0.763, the longer the patient was diagnosed with cancer, the smaller the difference in the hazard rate between smokers and non-smokers. The p-value is far less than 0.05, making the effect of the interaction between smoking and time on survival convincing.

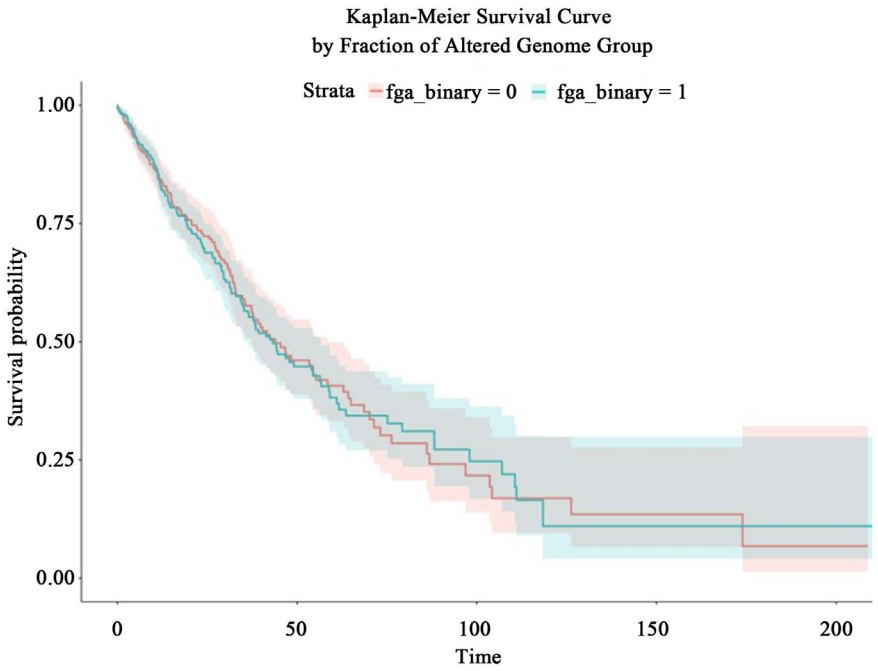

Figure 19. Kaplan-meler survival curve by fraction of altered genome group.

```
> ########
> cox.model.t <- coxph(Surv(time, death) ~ Diagnosis.Age + sex + sex_time
+                    + smoking + smoking_time + factor(stage) + stage_time
+                    + Fraction.Genome.Altered + fga_time + Mutation.Count + mc_time,
+                    data = data)
> summary(cox.model.t)
Call:
coxph(formula = Surv(time, death) ~ Diagnosis.Age + sex + sex_time +
    smoking + smoking_time + factor(stage) + stage_time + Fraction.Genome.Altered +
    fga_time + Mutation.Count + mc_time, data = data)

  n= 927, number of events= 261

                              coef  exp(coef)   se(coef)       z Pr(>|z|)
Diagnosis.Age            9.770e-03  1.010e+00  7.363e-03   1.327  0.18454
sex                      4.863e-01  1.626e+00  2.258e-01   2.154  0.03125 *
sex_time                -2.762e-02  9.728e-01  8.253e-03  -3.346  0.00082 ***
smoking                  1.204e+01  1.691e+05  1.331e+00   9.041  < 2e-16 ***
smoking_time            -2.702e-01  7.633e-01  2.317e-02 -11.660  < 2e-16 ***
factor(stage)II          9.834e-01  2.673e+00  2.075e-01   4.739 2.15e-06 ***
factor(stage)III         1.707e+00  5.515e+00  2.556e-01   6.680 2.40e-11 ***
factor(stage)IV          2.273e+00  9.710e+00  3.983e-01   5.707 1.15e-08 ***
stage_time              -2.109e-02  9.791e-01  4.112e-03  -5.130 2.90e-07 ***
Fraction.Genome.Altered  6.469e-01  1.910e+00  5.133e-01   1.260  0.20752
fga_time                -1.813e-02  9.820e-01  1.491e-02  -1.216  0.22415
Mutation.Count          -4.477e-04  9.996e-01  4.754e-04  -0.942  0.34634
mc_time                  5.974e-06  1.000e+00  1.003e-05   0.596  0.55133
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                         exp(coef) exp(-coef) lower .95 upper .95
Diagnosis.Age            1.010e+00  9.903e-01 9.953e-01 1.024e+00
sex                      1.626e+00  6.149e-01 1.045e+00 2.532e+00
sex_time                 9.728e-01  1.028e+00 9.572e-01 9.886e-01
smoking                  1.691e+05  5.912e-06 1.244e+04 2.299e+06
smoking_time             7.633e-01  1.310e+00 7.294e-01 7.987e-01
factor(stage)II          2.673e+00  3.740e-01 1.780e+00 4.015e+00
factor(stage)III         5.515e+00  1.813e-01 3.342e+00 9.102e+00
factor(stage)IV          9.710e+00  1.030e-01 4.448e+00 2.120e+01
stage_time               9.791e-01  1.021e+00 9.713e-01 9.871e-01
Fraction.Genome.Altered 1.910e+00  5.236e-01 6.983e-01 5.222e+00
fga_time                 9.820e-01  1.018e+00 9.537e-01 1.011e+00
Mutation.Count           9.996e-01  1.000e+00 9.986e-01 1.000e+00
mc_time                  1.000e+00  1.000e+00 1.000e+00 1.000e+00

Concordance= 0.926  (se = 0.022 )
Rsquare= 0.595   (max possible= 0.954 )
Likelihood ratio test= 837.9  on 13 df,   p=<2e-16
Wald test            = 279.4  on 13 df,   p=<2e-16
Score (logrank) test = 284.8  on 13 df,   p=<2e-16
```

Figure 20. The calculation of p-value.

On the other hand, the fraction genome mutated (fga_time) and the mutation count (mc_time) still do not have a significant effect on hazard ratio, due to their big p-values (0.224, 0.551 respectively). However, it does not mean that the effect of these two variables on survival does not change with time. Again, the reason we failed to capture the association between survival and them may be the limited cohort and sample size of our data. More intricate statistical methods and a greater sample base might help to detect the true association in the future.

4. Conclusion

To sum up, certain variables influence the survival rate of patients with lung cancer in our data sample from TCGA. Specifically, the stage of cancer, which is divided into four stages according to the severity of lung cancer, has the most significant effect on the survival rate using both the Kaplan-Meier estimates and the Cox Proportional Hazard Model (in both time-invariant and time-variant assumptions). For other variables, such as sex, smoking, etc, only when they interact with time can they have a significant association on the survival, indicating their time-variant impact with survival. Unfortunately, we failed to detect the association between two genetic variables and survival. This research indicates that further research is required; both larger cohort and more appropriate methods are needed to study the influence of other potential factors on survival of patients with lung cancer.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cox, D.R. (2018) Analysis of Survival Data. Chapman and Hall/CRC, London.

[2] WHO (2018) Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer

[3] ACS (2019) What Is Lung Cancer? https://www.cancer.org/content/cancer/en/cancer/lung-cancer/about/what-is.html

[4] Furrukh, M. (2013) Tobacco Smoking and Lung Cancer: Perception-Changing Facts. *Sultan Qaboos University Medical Journal,* **13**, 345. https://doi.org/10.12816/0003255

[5] Popper, H.H. (2016) Progression and Metastasis of Lung Cancer. *Cancer and Metastasis Reviews,* **35**, 75-91. https://doi.org/10.1007/s10555-016-9618-0

[6] NCI (2017) Cancer Genome Research and Precision Medicine. https://www.cancer.gov/about-nci/organization/ccg/cancer-genomics-overview

[7] Stewart, B.W. and Kleihues, P. (2003) World Cancer Report.

[8] NIH (2019) The Cancer Genome Atlas Program. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[9] Campbell, J.D., *et al.* (2016) Distinct Patterns of Somatic Genome Alterations in Lung Adenocarcinomas and Squamous Cell Carcinomas. *Nature Genetics,* **48**, 607.

https://doi.org/10.1038/ng.3564

[10] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association,* **53**, 457-481. https://doi.org/10.1080/01621459.1958.10501452

[11] Collett, D. (2015) Modelling Survival Data in Medical Research. Chapman and Hall/CRC, London.

[12] Bland, J.M. and Altman, D.G. (2004) The Log Rank Test. *British Medical Journal*, **328**, 1073. https://doi.org/10.1136/bmj.328.7447.1073

[13] Fahrmeir, L., *et al.* (2013) Regression: Models, Methods and Applications. Springer Science & Business Media, New York. https://doi.org/10.1007/978-3-642-34333-9_2

[14] Goel, M.K., Khanna, P. and Kishore, J. (2010) Understanding Survival Analysis: Kaplan-Meier Estimate. *International Journal of Ayurveda Research,* **1**, 274. https://doi.org/10.4103/0974-7788.76794