TCGA Lung Cancer Data Analysis

Elishua Shumpert

12/9/2020

Introduction

Lung cancer is the most common cause of cancer related mortality worldwide causing more than 1 million deaths per year. This cancer begins in the lungs and may spread to the lymph nodes and other organs in the body such as the brain. Lung adenocarcinoma (LUAD) preys mostly on individuals who have a history of smoking although it is possible for it to attack individuals who have never smoked. In 2019, the CDC reports that nearly 14 of every 100 U.S. adults aged 18 or older currently smokes cigarettes which means an estimated 34.1 million adults in the United States currently smoke cigarettes. It is to no surprise why lung cancer is one of the leading causes of cancer-related deaths. Therefore, it is important to understand its devastating effects and what factors can contribute to it so that clinicians and researchers can better diagnose it and increase chances of survival through different treatments.

The data that was analyzed is courtesy of the cancer genome atlas (TCGA) database. The data set included genomic and clinical data for 515 lung cancer patients and incorporated variables including mRNA expressions, number of cigarette packs smoked per year, survival time of the patient (in days), the stage of cancer the patient was in, gender, race, and whether the patient received radiation therapy or not. The essential goals of this data analysis was to analyze which mRNA gene expressions were associated with the survival time of a lung cancer patient, identify the effects of potential factors on survival time of lung cancer patients, and to investigate if radiation therapy is an effective means of treatment for lung cancer.

Methods

First, the data set was pre-processed. There were missing values for the following variables: age, the number of cigarette packs smoked, the pathologic stage of cancer, whether a patient went through radiation therapy or not, race and survival time. Therefore, the missing values for age and the number of cigarette packs smoked were imputed with their mean values and the factor variables (the pathologic stage of cancer, radiation therapy, and race) were imputed with the mode value. One limitation of this data set was that the missing values for survival time were correlated perfectly with the patients whose vital status was alive. In other words, some of the records on survival time for patients who were still alive were censored contributing to selection bias in the samples. Since survival time was a target of interest for this analysis, the missing values for this variable had to be omitted updating the sample size to 182 observations. Finally, survival time was transformed to be on the log scale so that the data for our target variable is normal for linear regression.

To see which mRNA gene expressions were associated with the survival time of a lung cancer patient, the number of gene expressions had to be reduced to a smaller subset since there were 20,500 mRNA gene expressions in the data set. Therefore, a two-stage variable selection process was implemented within 5-fold cross validation. For the initial stage, the mRNA gene expressions were first screened by regressing each gene against the survival time, and the gene expressions that had significant p-values with significance level $\alpha = 0.05$ were kept in the data set while the genes expressions that were not significant were discarded from the data set. The next step was to fit a lasso model on all the genes expressions left from the first stage to get a more sparse set of genes that could be potentially associated with survival time. We then evaluated our approach post-selection by fitting a linear model with the set of gene expressions selected from our procedure. Secondly, smoking has been shown to be the leading cause of lung cancer. So, the effects of smoking on survival time were compared to other factors including race, gender, and age to see if smoking had a greater effect on survival time or not when compared to these other variables. This was accomplished by running a

multiple linear regression model with all of these factors and reporting their p-values for significance and then separately fitting univariate linear regression models for each factor and evaluating their model performance on a test sample. Lastly, the Scheffe test was used to evaluate the effectiveness of radiation therapy by comparing the group means for survival time between patients who went through radiation therapy versus those who did not thus testing if there was a difference in survival time. The Scheffe test was chosen because it can work with unequal group sizes and it gives narrower confidence intervals.

Results

Analysis Question 1: What mRNA gene expressions are significant or associated with the survival time of a lung cancer patient?

After running our two-stage procedure for selecting gene expressions out of a candidate set of 20,500 gene expressions, we reduced the set of genes to 134. A linear model was then fitted post-selection using these 134 selected mRNA gene expressions to further see how these selected gene expressions fit the data for survival time of a lung cancer patient. Out of all the 116 mRNA gene expressions that were selected, 17 of the mRNA gene expressions remained significant in association with the survival time of a lung cancer patient namely `C7orf71`, `CCDC149`, `CSPG5`, `ETAA1`, `HTRA4`, `IL1F6`,`IMPG1`, `LOC168474`, `LOC401387`, `LST.3TM12`, `P2RY6`, `PSPN`, `PTPRZ1`, `R3HDML`, `SDC1`, `SKINTL`, and `ZNF781`. Below is a table showing their estimates and p-values.

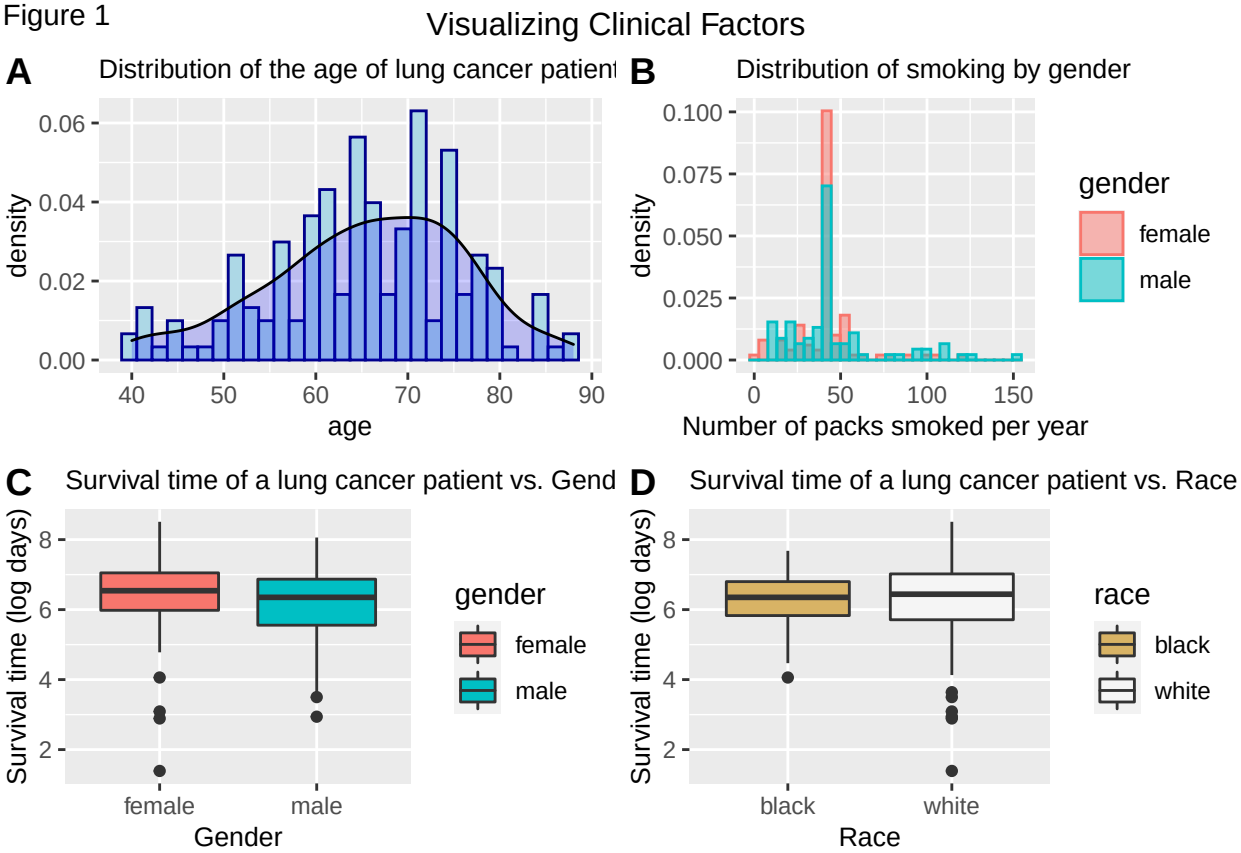
Table 1: Post-Selection mRNA Gene Expressions

mRNA Gene Expressions	estimate	p-value
C7orf71	-0.1086	0.0053
CCDC149	0.0007	0.0227
CSPG5	0.0011	0.0055
ETAA1	0.0011	0.0470
HTRA4	0.0021	0.0437
IL1F6	0.3855	0.0211
IMPG1	0.0397	0.0050
LOC168474	-0.0234	0.0337
LOC401387	0.1171	0.0115
LST.3TM12	-0.0040	0.0355
P2RY6	-0.0004	0.0010
PSPN	-0.1588	0.0046
PTPRZ1	-0.0004	0.0095
R3HDML	-0.0151	0.0061
SDC1	0.00002	0.0206
SKINTL	-0.0815	0.0449
ZNF781	0.0110	0.0044

Analysis Question 2: Is the effect of smoking on survival time of lung cancer patients greater than other factors including race, age and gender?

Let's start by first visualizing the data for some of these clinical variables.

Figure 1



Referring to panel A of figure 1, the mean age of lung cancer patients appears to be around 68 years old so this is a cancer that affects mostly older age individuals and the distribution of the age of lung care patients appear to be fairly normal. In panel B, we see that the distribution for the number of packs smoked per year for both genders, male and female, appear to be the same. In panel C, the median survival time for female lung cancer patients is slightly higher than that for male lung cancer patients, but there is almost no difference in survival time for both male and female patients. Lastly in panel D, there is almost no difference between the survival time of a lung cancer patient who is white versus a lung cancer patient who is African American, however, it appears that the median survival time for white lung cancer patients is slightly above that than for African American lung cancer patients. We can statistically validate if smoking has a stronger effect on survival time than the other factors by running a linear model with all the factors and comparing their p values. Then, we will separately run univariate linear regressions for each variable on a training set and evaluate their respective model performances by the mean squared error on a test set using a 60-40 split with the training set being 60% of the data and the testing set being 40% of the data.

Table 2: Multiple Linear Regression Model Results for Clinical Factors

clinical variable	estimate	p-value
number of packs smoked per year	0.0045	0.434
race	0.1659	0.662
age	-0.0172	0.126
gender	-0.2341	0.304

The report from our linear model in Table 2 suggests that none of these factors have a significant relationship with the survival time of a lung cancer patient. Let's now fit our simple linear regression models for each of the four variables seperately to see how well they can predict the survival time of a lung cancer patient.

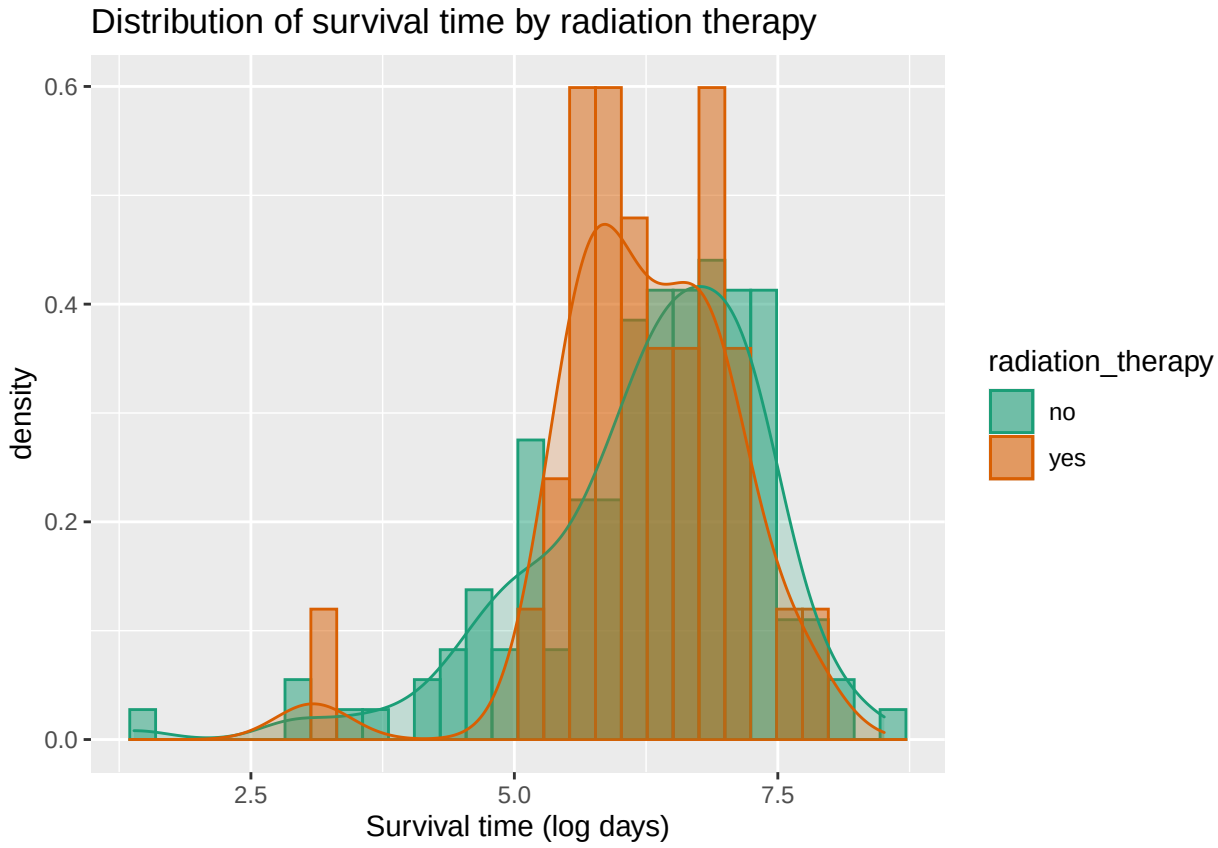
Table 3: Univariate Linear Regression Model Results for Clinical Factors

Models	test MSE	R^2
Model 1: number of packs smoked per year	0.8600	0.0020
Model 2: race	0.8560	0.0002
Model 3: age	0.9003	0.0093
Model 4: gender	0.8374	0.0184

The error rates for each regression model are very close to each other, and there is not much difference between them. In terms of the test prediction error, gender seems to have the lowest prediction error when regressed upon survival time so in this sense it is the strongest predictor out of the factors affecting the survival time of a lung cancer patient. However, in general, all of these variables are poor in explaining the variation in survival time as the R^2 value for each model is extremely low.

Question 3: Is radiation therapy an effective means of lengthening the survival time of a lung cancer patient? In other words, is there a difference in survival time between those who do radiation therapy versus those who don't?

Let's start by visualizing the distribution of survival time for those who do radiation therapy for lung cancer versus those who don't do radiation therapy.



It is not easy to tell, however, it looks like the group of lung cancer patients who don't go through radiation therapy have a slightly higher mean survival time compared to the lung cancer patients who don't. Let's test to see if there is a difference by using Scheffe's test for multiple comparison of means.

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $radiation_therapy
##              diff     lwr.ci    upr.ci   pval
## yes-no -0.0304213 -0.4324577 0.3716151 0.8815
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From Scheffe's test, we can conclude that there is difference between the mean survival time of lung cancer patients who go through with radiation therapy and those who do not. The estimate from this test appears to validate that the mean survival time for lung cancer patients who do not take radiation therapy is slightly higher than those who do. We could go further by exploring if radiation therapy is effective in lengthening survival time by race or gender.

Table 4: The Effect of Radiation Therapy on Average Survival Time by Race

race	radiation therapy	average survival time (log days)
black	no	6.09
	yes	6.56
white	no	6.28
	yes	6.18

For African American lung cancer patients, radiation therapy seems to slightly lengthen the average survival time by 11.2 hours. However, we see an opposite result for white lung cancer patients. The average survival time for for white patients who do not go through radiation therapy is 2.4 hours higher than those who do through radiation therapy. So, in general, there is not too much of a significant difference in time.

Table 5: The Effect of Radiation Therapy on Average Survival Time by Race

gender	radiation therapy	average survival time (log days)
male	no	6.15
	yes	6.10
female	no	6.39
	yes	6.31

There appears to be very little difference in the mean survival time for male patients who go through radiation therapy versus those who do not. Likewise, there is hardly any difference in survival time for female patients who go through radiation therapy versus those who do not. It is difficult to say if radiation therapy is effective overall because this data set is limited to only patients who died from lung cancer, so that would be a follow-up question worth looking at.

Conclusion

This analysis found that 17 mRNA gene expressions out of 20,500 were significantly associated with the survival time of a lung cancer patient. In addition, smoking does not seem to be strongly related with the survival time of a lung cancer patient and other factors including race, gender and age seem to have the same effect on the survival time as these effects were very minimal. So, from the data we learned that there was not too much of a difference between the effects of smoking, race, gender and age on the survival time

of a patient. Lastly, we found out that radiation therapy does not seem to be that effective in lengthening survival time of a patient. Testing if there was a difference in the average survival time between patients who took radiation therapy and those who did not, Scheffe's test concluded that there was a difference. However, when looking at the estimate of the difference of means (radiation group yes - radiation group no) and the confidence interval for this estimate, the difference is minimal. We took a step further by investigating whether radiation therapy had lengthened survival time for groups of race and gender. Likewise, there were very minimal differences by race and gender. However, radiation therapy seemed to lengthen the average survival time the most for African American patients. A severe limitation of this analysis was that it was conducted using 182 samples which is extremely small. If there were many more observations included, we could have had more accurate and robust results. For example, we cannot truly know if radiation therapy was effective because individuals who survived the lung cancer are not included in the sample and if they were included we could more accurately answer this question. Therefore, a follow up study would consider the effects of lung cancer on individuals who were censored and compare these individuals to those did not survive. A further study that would be interesting to see is if the quality of life improved for individuals who successfully survived lung cancer.