The survival analysis of patients with lung cancer

Donghyeok Lee(1523856)

ldh9509@naver.com

I, Donghyeok Lee, hereby declare that the presented report is solely my own and I am able to defend it by myself.

1. Introduction

The three most common cancers are breast, colon and prostate cancer, however, it is reported that patients with lung cancer die more than the combined number of patients who died due to the three mentioned cancer every year, according to the American Cancer Society.[i] Therefore, it is of great interest for researchers to ensure whether a new treatment is significantly more helpful for increasing a survival probability of a patient with lung cancer than a standard treatment. In this paper, the data of the Phase II trial and the Phase III trial are analyzed statistically, using the Kaplan Meier estimator, log-rank test, and cox proportional hazards model.

In the subsection 2.1, based on the data of the Phase II trial, the effect of the new treatment, compared to the standard treatment, is addressed by the Kaplan Meier estimators and log-rank Renyi type test. Then, in the section 2.2, the analysis of the Phase III trial is conducted. In the subsection 2.2.1, it is addressed how a follow-up time and a sample size affect the power of the log-rank test, along with the coverage rate of the median survival time. Finally, in the subsection 2.2.2, the effect of the each variable, given in the data(age, gender, smoking history, type of the tumor cells, CDP), is estimated based on the cox proportional hazards model. (CDP denotes the number of chronic drug prescriptions in the past 12 months.)

2. Survival analysis

2.1 Data analysis of the Phase II trial

The 137 number of the patients are recruited in the study immediately after being diagnosed with lung cancer and randomized to the two groups: the new treatment group and standard treatment group. The characteristics of each patient are the type of the tumor cell, age at diagnosis, a measure of general medical status(categorized as ‘completely hospitalized’, ‘partial confinement to a hospital’, ’able to care for self’). Plus, this data is right-censored. Each of the patients is marked as 1 if a death is observed during the study. If not observed during the study, he or she will be marked as 0.

Underlying assumptions for the analysis in this section is that the samples are independent of each other, and the right censoring is independent of a survival time. Note that this data is also left-censored since the patients are recruited when they are diagnosed with lung cancer, but not when they actually have it. There is a big chance that patients would visit a hospital to be diagnosed only when they have obvious symptoms. Thus, the survival time in this paper indicates the time between the time of being diagnosed and the death.

Before conducting the statistical analysis, it is important to check whether the two treatment groups are well balanced regarding the characteristics of the patients. The below plots show the distribution of each variable per the different treatment group.

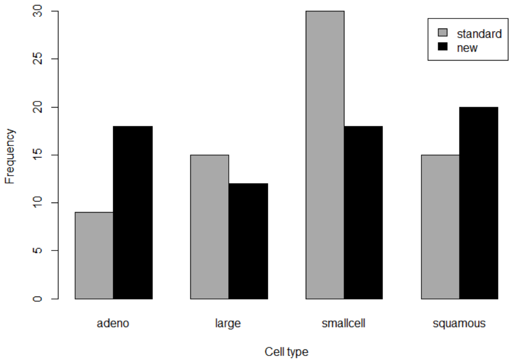
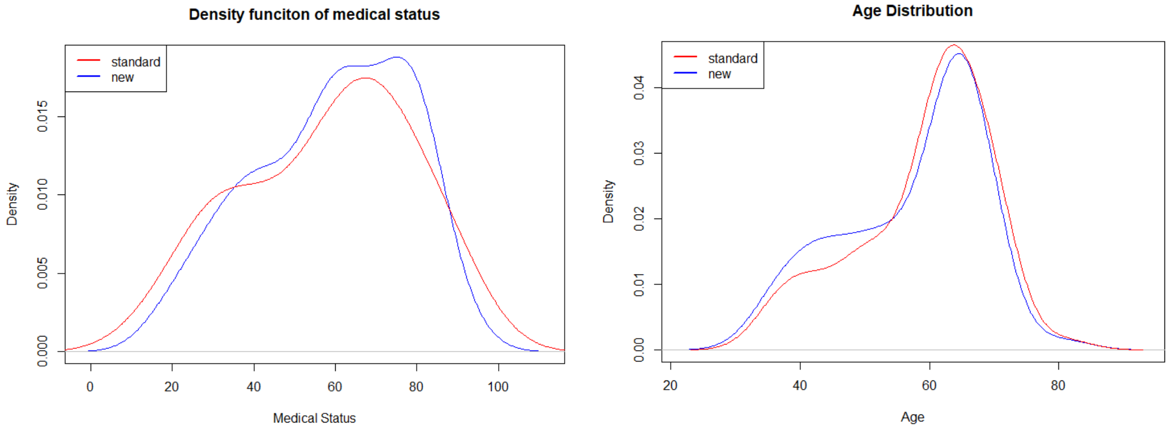


Figure 1 : The bar chart per each cell type per the treatment group



(a) : Medical status

(b) : Age

Figure 2 : The density plots of medical status and age per the treatment group

In Figure 1, it is observed that the number of each cell type per the treatment group is different, especially in Adeno and small cell type, while, from Figure 2 (a) and (b), medical status and age are quite equally distributed across the two groups. Since the patients are not well balanced with regard to the cell type, the results of the hypothesis test that the survival curves between the two treatment groups are different might be misleading. For example, assume that the small cell type increases a hazard a lot. In the new treatment group, there are much less patients with the small cell type, therefore, the survival probabilities of the new treatment group would be higher than the other group. As a result, the log-rank test might produce a result that the survival curves between the two treatment groups are different. We will leave this issue to the experts of the corresponding domain knowledge. The following analysis is given under the assumption that the randomization across the two treatment groups is well conducted. Firstly, the summary statistics with the Kaplan Meier estimator are presented, and conduct a non-parametric hypothesis to test whether the survival curves between the two groups are significantly different. Secondly, parametric models are fitted to the data, and the best fitted model is selected according to the AIC(Akaike information criterion).

- Survival curve

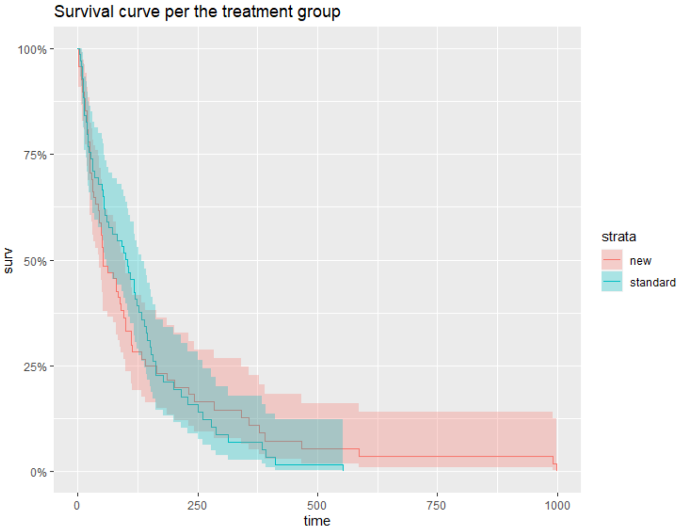
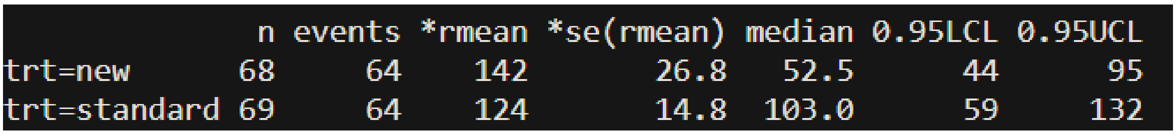


Figure 3 : Survival curve per the treatment group

In Figure 3, for each of the treatment groups, the solid line represents the survival probability estimated by the Kaplan-Meier estimator, and the shaded area shows the 95% linear confidence interval. We can find that the two shaded areas are overlapped a lot at almost every time points, which indicates that there is no significant difference in the survival curves between the treatment groups.



```
                n events *rmean *se(rmean) median 0.95LCL 0.95UCL
trt=new        68     64    142       26.8   52.5      44      95
trt=standard 69     64    124       14.8  103.0      59     132
```

Figure 4 : restricted mean and median survival time

- Median survival time

The two survival curves reach 0.5 survival probability, thus, the median can be obtained. For the new treatment group, the survival probability at the time 52(days) is exactly 0.5. For the standard treatment group, it is 0.4863 at the time 103(days). If one wishes to obtain the more accurate median survival time, the linear interpolation can be used.

- Restricted mean survival time

If it is assumed that the longest survival time is equal to the longest survival time in the data, this restricted mean time would be close to the true mean survival time since the survival probabilities in each group are being close to 0 as the time increases. However, the mean survival times are much larger than the medians, which means that the outliers with very long survival time affect the means by being unusually large. Thus, the median survival time would be more informative for patients than the mean.

- Non-parametric hypothesis to test the difference between the two treatment groups

It is possible to quantitatively check the difference in the survival curves between the treatment groups by conducting a hypothesis test. Since there is a crossing point between the two survival curves, as seen in Figure 3, the Renyi type test should be used rather than the log-rank test. Firstly, the hypotheses are constructed as follows;

H0 : The two groups have identical hazard functions.

H1 : The two groups do not have identical hazard functions for some time period.

The p-value of the Renyi type test with the weight 1 is 0.26013, therefore, we can not reject H0 that the two hazard functions are identical.

- Parametric survival time

The distribution of a survival curve for patients with lung cancer would be of great interest to patients or researchers. Since it is shown by the Renyi type test that the two survival curves are not significantly different, we can use the whole data(rather than stratified data per the treatment) in order to figure out what parametric model would be a best fit. In order to estimate the parameters of each distribution fitted to the data, the maximum likelihood estimation is used.

There are commonly used parametric distributions for a survival curve such as exponential, Weibull, generalized gamma, and log-normal distribution. After fitting each of these distributions to the data, the plot and Akaike information criterion(AIC) are obtained.

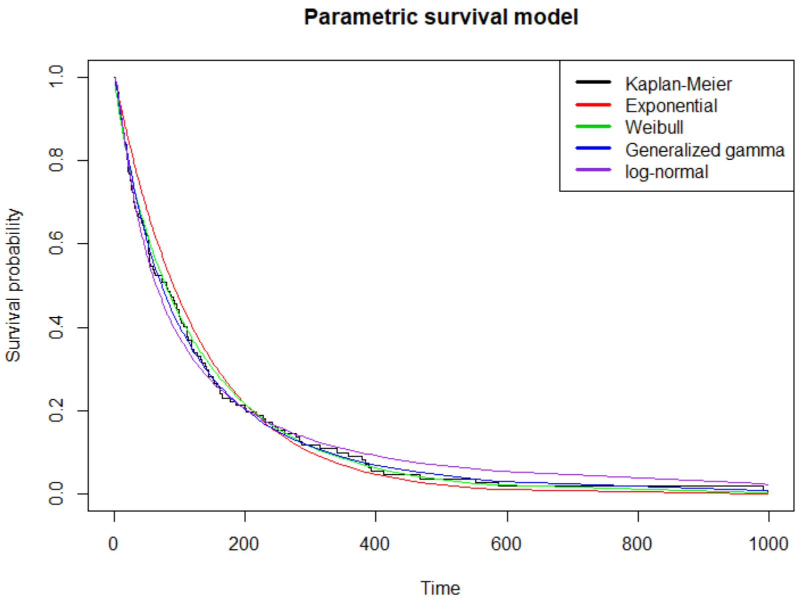


Figure 5 : Parametric survival models and Kaplan Meier estimator

The Weibull or generalized gamma distribution seems the best fit since they are more similar to the Kaplan Meier across every time points than the other distributions. In order to quantitatively investigate the goodness-of-fit, we can use the AIC, and the AIC per each parametric distribution is summarized in the table below;

Distribution	AIC
Exponential	1504.442
Weibull	1500.182
Generalized gamma	1498.942
Log-normal	1502.948

Figure 6 : AIC per the different distribution

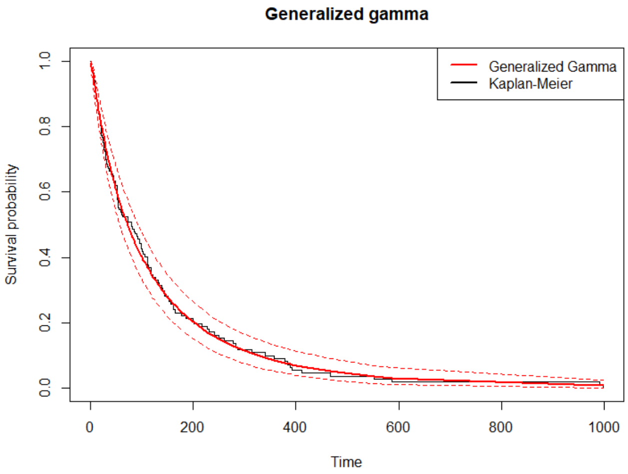
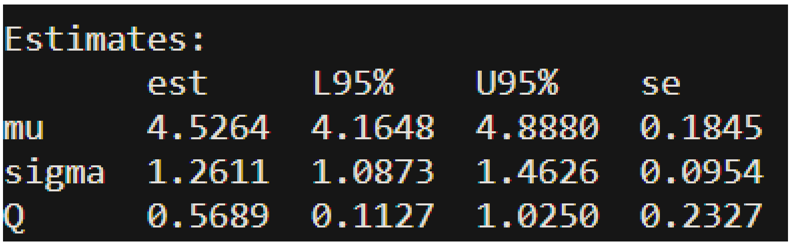


Figure 7 : Generalized gamma distribution fitted to the data

The generalized gamma distribution has the lowest AIC, therefore, it is the best fit for the survival curve of patients with lung cancer. The estimated parameters of this distribution are summarized by the R software and presented in the plot below;

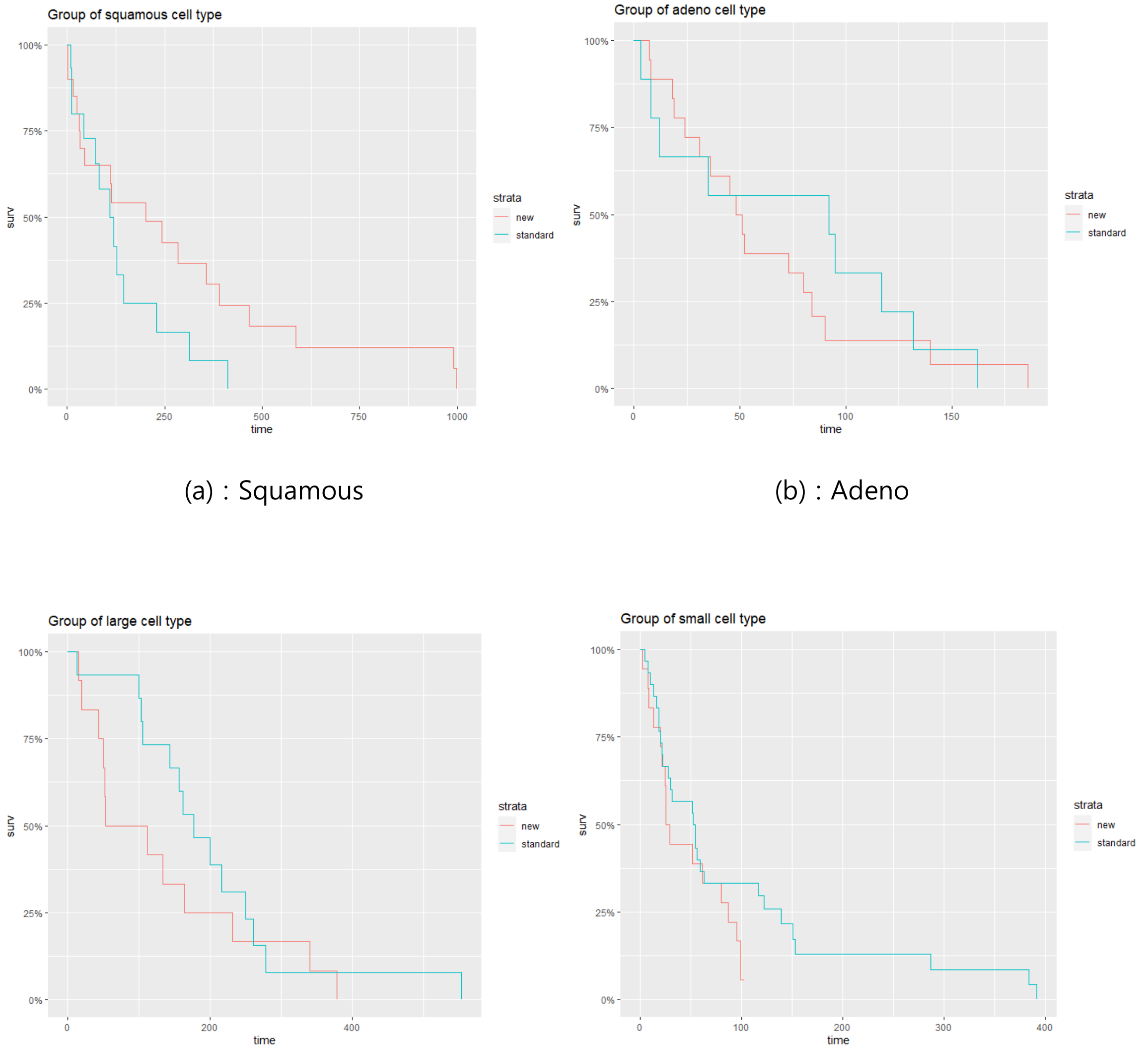


```
Estimates:
       est     L95%    U95%    se
mu     4.5264  4.1648  4.8880  0.1845
sigma  1.2611  1.0873  1.4626  0.0954
Q      0.5689  0.1127  1.0250  0.2327
```

Figure 8 : Estimated parameters of the generalized gamma distribution

Furthermore, From Figure 7, the Kaplan-Meier estimator is between the confidence interval across all time points, which indicates that the generalized gamma distribution is fairly good parametric model. If one still wishes to use the exponential or Weibull distribution by the reasons based on domain knowledge or previous research, the Weibull distribution should be chosen since the p-value of the likelihood ratio test between the exponential and Weibull model is 0.0123, which is smaller than 0.05.

Although it is graphically and quantitatively shown that there is no difference in the survival curves between the treatment groups, the effect of the new treatment might be significant for a certain cell type. The four plots below show the survival curves of the two treatment groups per the cell type. In addition, the table below indicates the hypothesis test to check the significant difference of the survival curves between the groups and its corresponding p-value in each case.



(a) : Squamous

(b) : Adeno

(c) : Large

(d) : Small

Figure 9 : Survival curve per each cell type per each group

Cell type	Test	p-value
Squamous	FH(w : p=0, q=1)	0.03
Adeno	Renyi(w : p=0, q=1)	0.24
Large	Renyi(w=1)	0.27
Small	FH(w : p=0, q=1)	0.0582

Figure 10 : The type of the test and its corresponding p-value per each cell type

FH, Renyi and w denotes the 'Fleming-Harrington test', 'Renyi type test' and weight, respectively. FH test is used for the case where the difference between survival curves is remarkable in a specific time period(either early or late period), while Renyi test is often used for the case of crossing survival curves. Note that the cell type, Adeno, have both situations; therefore, Renyi type test with Fleming-Harrington weights is selected. Plus, the set of p and q in FH test - that is, p=0 and q=1 – can give more weight on the difference of survival curves at a late time.

According to the p-values, the new treatment is significantly helpful for patients with squamous cell type regarding the survival probability at the significant level 0.05, compared to the standard treatment. For the other cell types, there is no significant difference between the treatment groups. However, it is uncertain that the new groups stratified by the cell type and treatment are randomized enough, since the data itself is only randomized with regard to the treatment. For example, the group of the Adeno cell type with the new treatment would have more female patients than the group of the Adeno cell type with the standard treatment. In order to be certain about the effect of the new treatment on each cell type, additional randomized control studies should be conducted.

2.2 Data analysis of the Phase III trial

The current plan for the upcoming Phase III study is to recruit 300 patients and randomly assign them either to the new treatment group or the control group with 1:1 ratio. Each patient will be followed-up for maximum of 4 months. During the study, each patient can drop out with 5% probability. Note that the drop-out and follow-up time is independent of their health status. Then it is important to know – given sample size and follow-up time - what is a Type 1 error and power of statistical tests, along with the coverage rate that the estimated confidence interval contains the true value. In detail, the widely used test and quantity, namely the log-rank test and median survival time, are investigated with regard to a Type 1 error, power and coverage rate.

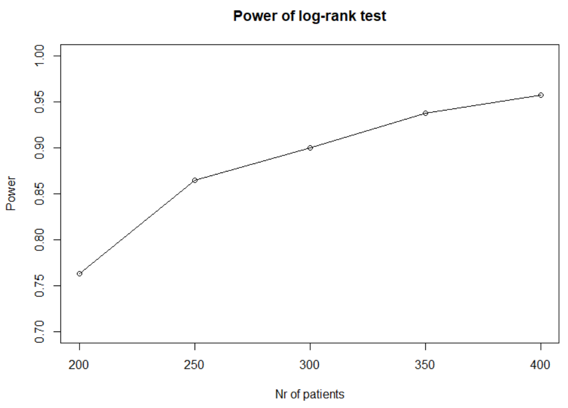
2.2.1 Power and coverage rate

The Monte-Carlo simulation study is conducted to investigate each of them as follows;

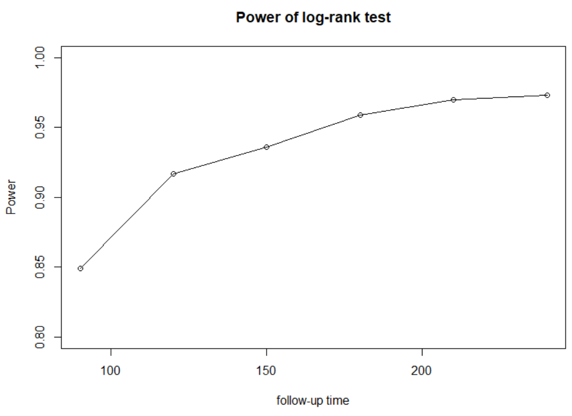
i) Power of the log-rank test

Generate a dataset of patients - who are randomly assigned to the two treatment groups with 1:1 ratio - with survival times from the proportional hazards model with Weibull baseline hazard under the alternative hypothesis of the log-rank test(that is, the effect of the new treatment is significant in reducing the hazard compared to the standard treatment, compared to the standard treatment). Additionally, survival times are right-censored by a maximum follow-up time, and random-censored by the 5% drop-out rate. Then conduct the log-rank test and repeat this procedure 1000 times to obtain the power of the test. Note that the power is the rate that the log-rank test rejects the null hypothesis at the significance level 0.05. (The detailed algorithm is attached in Appendix.)

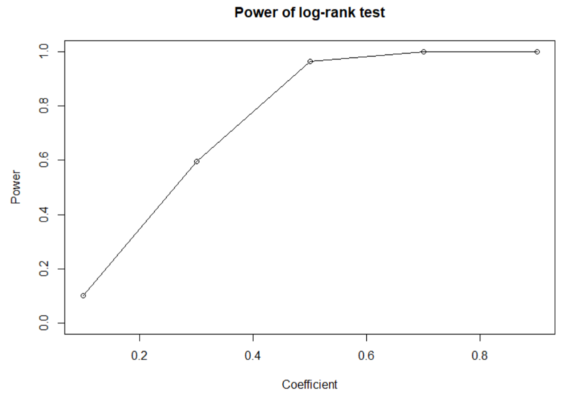
Result :



(a) : Power per the number of patients



(b) : Power per the follow-up time(days)



(c) : Power per the coefficient of the new treatment

Figure 11 : Power of the log-rank test per the number of patients(a) , the follow-up time(b) and the coefficient of the new treatment(c)

The power of the two-sample log-rank test is affected by the various factors: the significance level, the sample size, ratio of a sample to the other sample, follow-up time and hazard ratio between the two samples. As seen in Figure 11, increasing the number of patients or follow-up time contributes to the bigger power of the test. Although the bigger hazard ratio also contributes to the bigger power, researchers can not control the hazard ratio in reality. In addition, when the ratio of a sample to the other sample is 1:1, the power of the test is maximized; thus, we do not need to change the plan that the ratio is 1:1. In conclusion, there are the two ways to increase the power of the log rank test: increasing the sample size or follow-up time. For example, if we want to have the power of 0.95 for the log-rank test(according to Figure 11 (a) and (b)), we should either increase the sample size from 300 to 350 or increase the follow-up time from 120(days) to 180(days). The researchers may select the one that is more cost-effective. In general, it would be more cost-effective to increase the follow-up time than increasing the sample size.

ii) Type 1 error of the log-rank test

The procedure is same as 'the power of the log-rank test procedure'. The only difference is that a dataset is generated under the null hypothesis of the log-rank test(that is, the coefficient of the new treatment is 0.) and the Type 1 error is the rate that the log-rank test rejects the null hypothesis at the significance level 0.05.

Result :

When the sample size is 300 and the follow-up time is 120 days, the Type 1 error is 0.048, which is close to the significance level 0.05.

iii) Coverage rate of the median survival time

Generate a dataset as same as a dataset in 'Type 1 error of the log-rank test procedure', but with the sample size 100000. Then conduct the Kaplan-Meier estimator on that dataset to obtain the true median survival time. After obtaining the true median survival time, generate the 1000 datasets as same as a dataset in 'Type 1 error of the log-rank test procedure' again, but with the sample size 300. Apply the Kaplan-Meier estimator on the datasets to obtain the 95% linear confidence intervals of the median survival time and check whether each of 1000 number of the confidence intervals contains the true median.

Result :

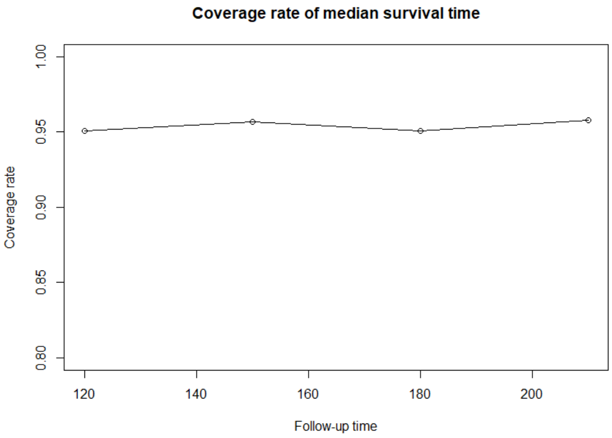


Figure 12 : Coverage rate of the median survival time per the follow-up time(days)

When the sample size is 300 and the follow-up time is 120 days, the coverage rate is around 0.95. It stays around 0.95 regardless of the follow-up time, as seen in Figure 12.

2.2.2 Cox proportional hazards model

The patients in the Phase III trial are also recruited immediately after being diagnosed with lung cancer, and randomly assigned with the ratio 1:1 to either the new treatment group or the standard treatment group. Each patient can drop out the study with the 5% probability and the maximum follow-up time is 120 days. The characteristics of the patients are as follows;

age : numerical value

gender : (1, female; 2, male),

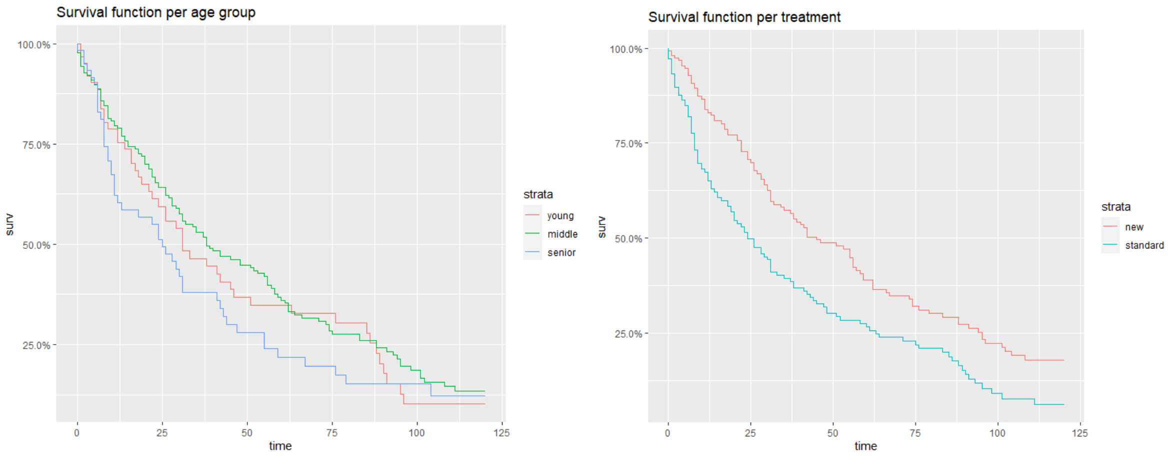
death : (1, death; 0, censored),

smoking history : (0, non-smoker; 1, smoker),

cell type : (1, squamous; 2, small cell; 3, adeno; 4, large cell),

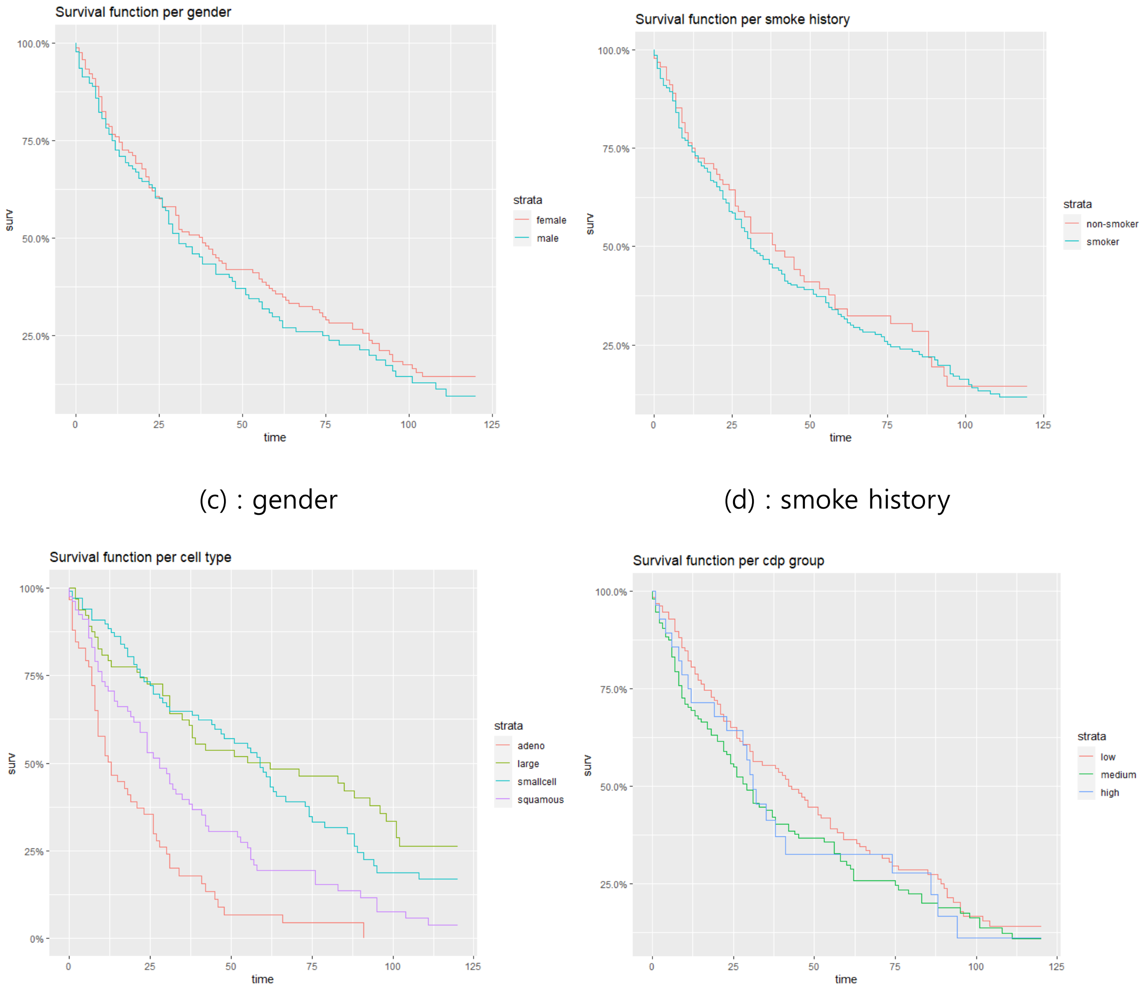
CDP : the number of chronic drug prescriptions in the past 12 months

Before constructing the cox model based on these covariates, the non-parametric tests are conducted in order to find the significant risk factors. The p-value of the coefficients from the cox model is dependent on the form of covariates(e.g., linear or polynomial form) and the cox model is only valid when several assumptions are being met. Thus, a non-parametric test such as the log-rank test is more robust to find the risk factors in the first place. The plots below show the survival curves per the covariate.



(a) : age

(b) : treatment



(c) : gender

(d) : smoke history

(e) : cell type

(f) : CDP

Figure 13 : Survival curves per each covariate

Covariate	Test	p-value
Age	Trend test(w=1)	0.3154
Treatment	Log-rank test	0.0002
Gender	Log-rank test	0.3107
Smoke history	Renyi type test(w=1)	0.7238
Cell type	Test for the four samples	0
CDP	Trend test(w=1)	0.0259

Figure 14 : The type of the test and its corresponding p-value per each covariate

The numerical variables(age and CDP) are categorized in three ordered groups: (young, middle and senior) and (low, medium, high), respectively. If age or CDP proportionally increases the hazard of patients with lung cancer, there will be a trend in hazard functions between three ordered groups. The test for the trend is conducted with the null and alternative hypothesis as follows.

H0 : $h_{group1}(t) = h_{group2}(t) = h_{group3}(t)$ for all t

H1 : $h_{group1}(t) \le h_{group2}(t) \le h_{group3}(t)$ for some t

Plus, for the cell type variable, the test for the four samples is conducted with these hypotheses;

H0 : $h_{Adeno}(t) = h_{large}(t) = h_{squamous}(t) = \hbar_{small}(t)$ for all t

H1 : At least one of the hazard functions is different in some t.

From Figure 14, the covariates with the p-value less than 0.05 are treatment, cell type and CDP. Thus, they are the significant risk factors.

- Cox proportional hazards model

According to the tests conducted before, the risk factors that significantly affect the survival probability are the treatment, cell type and CDP. However, in order to use the non-parametric test for the continuous covariates such as age and CDP, each of them has to be split into the n groups. But the p-value of the non-parametric test is affected by the choice of the n. Therefore, to ensure more whether the continuous variable is significant, we can construct the cox proportional hazards model with these variables and check the p-value of the coefficients of them.

Plus, the cox model does not make the assumption of the distribution of the survival time. Thus, when the distribution of the survival time is not known, the cox proportional hazards model is more robust choice than the accelerated failure time(AFT) model. The cox proportional hazards model is fitted to the data using r software:

```
cox_model = coxph(Surv(time, death) ~ age + treatment + cell_type + cdp,
                  data=data2, ties=c("efron"))
```

Note that the Efron method is selected as handling tied data as this method is more accurate than Breslow approximation method and has less computational cost than Exact method.

Before making an inference about the coefficients of each variable, we should check whether the underlying assumptions of the cox proportional hazards are met. The underlying assumptions of the cox model are as follows;

1) : The natural log of the hazard function is a linear function of covariates.

This is investigated by the plot of the martingale residual against fitted values.

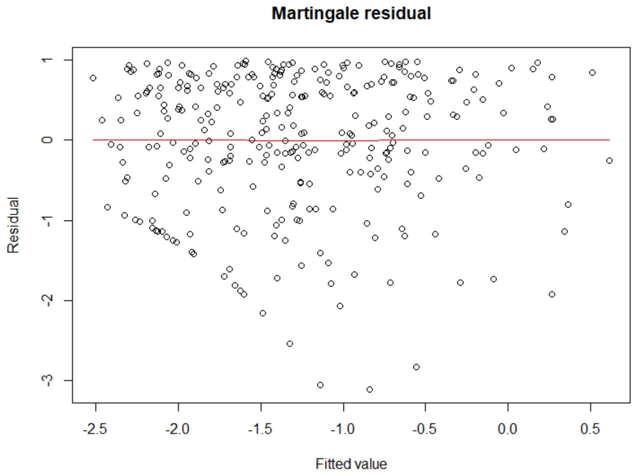


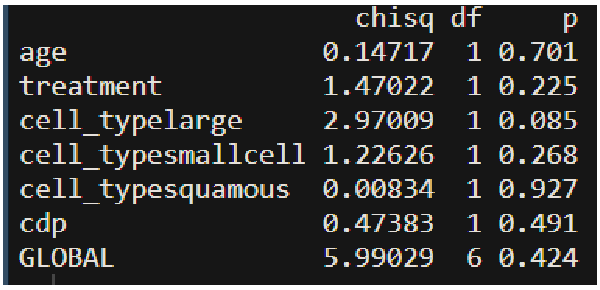
Figure 15 : Martingale residual against fitted values

The red line represents the smoothed line from the residuals, and it is almost horizontal line at 0.

Thus, this assumption is being met.

2) : Proportional hazards assumption

This assumption can be tested by the Schoenfeld test.



```
                      chisq df     p
age                 0.14717  1 0.701
treatment           1.47022  1 0.225
cell_typelarge      2.97009  1 0.085
cell_typesmallcell  1.22626  1 0.268
cell_typesquamous   0.00834  1 0.927
cdp                 0.47383  1 0.491
GLOBAL              5.99029  6 0.424
```

Figure 16 : The result of the Schoenfeld test

None of these variables have the p-value less than 0.05. Thus, this assumption is also met.

3) : The coefficient of covariates is not time-dependent.

The plots below show the effect of each covariate over time. The black solid line is the smoothed line from the points, and the black dashed line shows the corresponding 95% confidence interval. The red line is the time-fixed effect estimated by the cox model. If the black solid line is fairly horizontal and straight or the red line is between the two dashed lines, this assumption is met.

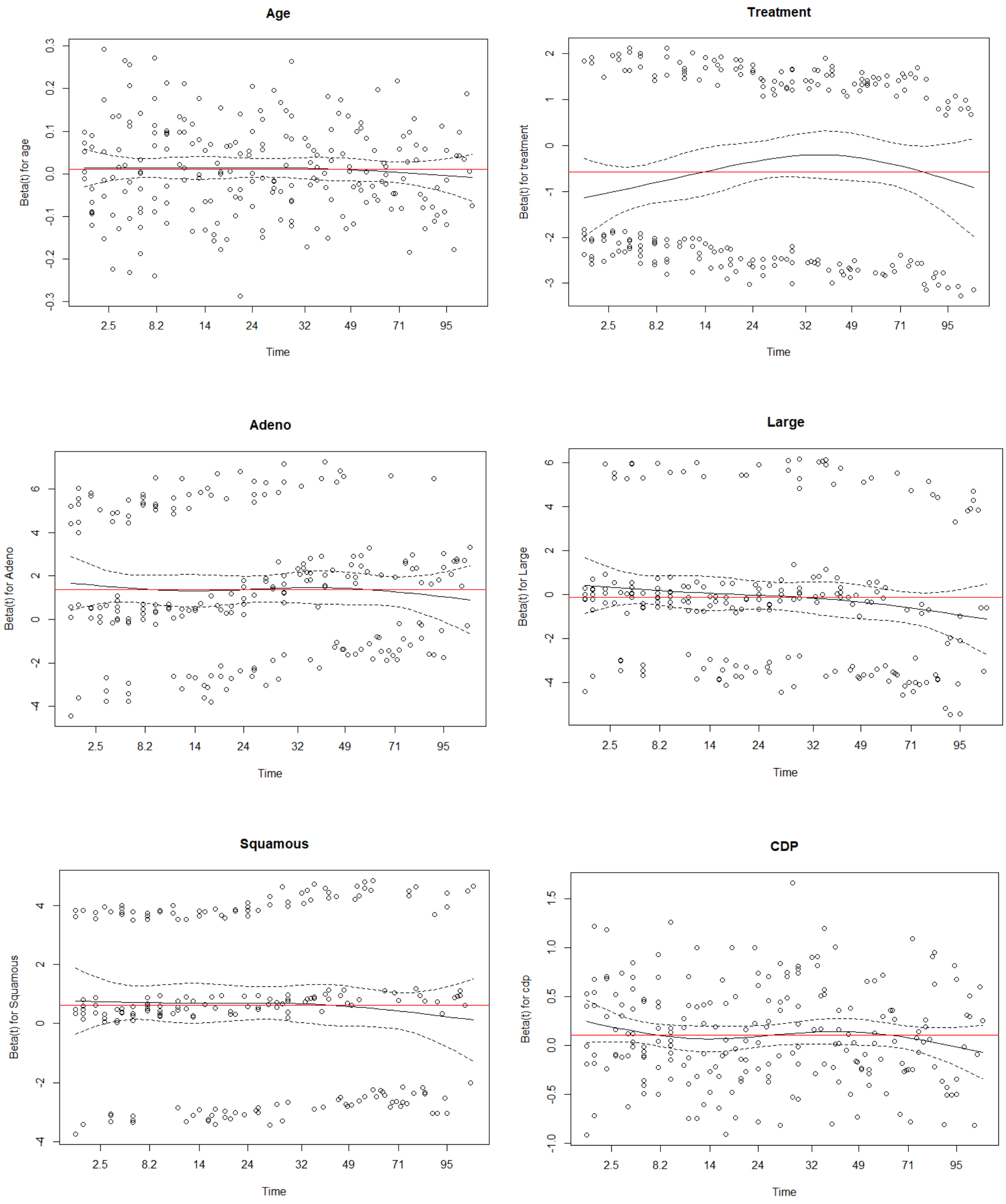


Figure 17 : Effect of each covariate over time

The red line is between the dashed lines for all of these covariates, therefore, this assumption is met for every covariates.

4) : The covariate is not time-dependent

The categorical variables such as treatment and cell type does not change over time, and CDP is already determined in the past 12 months, which means that it does not change since the Phase III trial started. Although age increases linearly with time, it does not change the form of the partial likelihood. Therefore, age can be treated as a time-independent variable. The detailed reasoning for that is formulated as below;

Denote L(B), X, and t as the partial likelihood, age, and time.

Then,

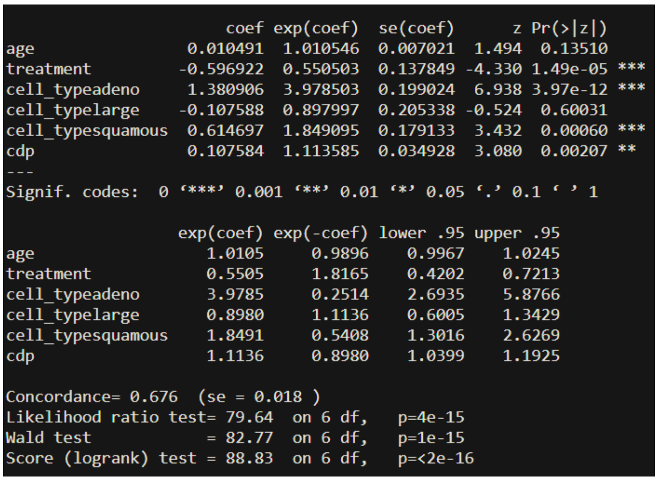
$$X_i(t) = X_i(0) + t$$

$$\text{L(B)} = \prod_{i=1}^{D} \frac{\exp(BX_i(t_i))}{\sum_{j \in R} exp\left(BX_j\,(t_i)\right)} = \prod_{i=1}^{D} \frac{\exp(BX_i(0)) * \exp(t)}{\sum_{j \in R} exp\left(BX_j\,(0)\right) * \exp(t)} = \prod_{i=1}^{D} \frac{\exp(BX_i(0))}{\sum_{j \in R} exp\left(BX_j\,(0)\right)}$$

(the D unique death times are ordered increasingly $t_1 < \cdots < t_d$, and R is the set of the indices of patients at the risk.)

As all assumptions are met, we can make an inference about the coefficients of the each variable.

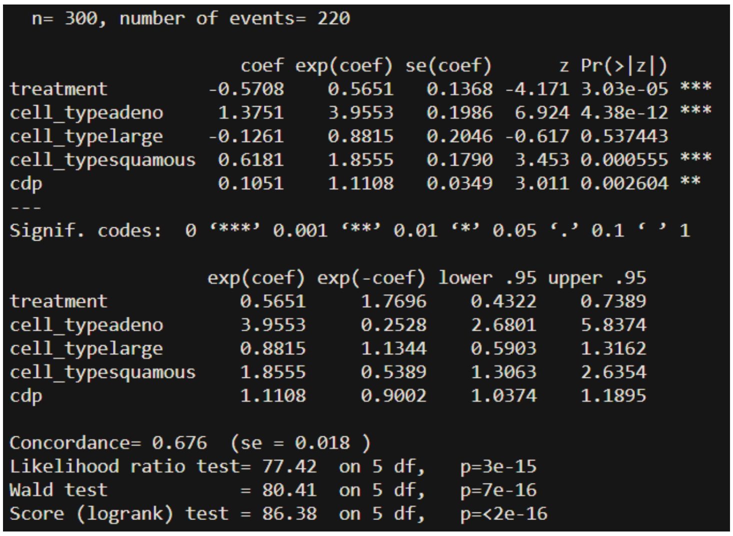
Note that the references for the categorical variables(treatment and cell type) are the standard treatment and the small cell, respectively. The below figure shows the result of the cox model.



```
                        coef exp(coef)  se(coef)      z Pr(>|z|)    
age                 0.010491  1.010546  0.007021  1.494  0.13510    
treatment          -0.596922  0.550503  0.137849 -4.330 1.49e-05 ***
cell_typeadeno      1.380906  3.978503  0.199024  6.938 3.97e-12 ***
cell_typelarge     -0.107588  0.897997  0.205338 -0.524  0.60031    
cell_typesquamous   0.614697  1.849095  0.179133  3.432  0.00060 ***
cdp                 0.107584  1.113585  0.034928  3.080  0.00207 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
age                  1.0105     0.9896    0.9967    1.0245
treatment            0.5505     1.8165    0.4202    0.7213
cell_typeadeno       3.9785     0.2514    2.6935    5.8766
cell_typelarge       0.8980     1.1136    0.6005    1.3429
cell_typesquamous    1.8491     0.5408    1.3016    2.6269
cdp                  1.1136     0.8980    1.0399    1.1925

Concordance= 0.676  (se = 0.018 )
Likelihood ratio test= 79.64  on 6 df,   p=4e-15
Wald test            = 82.77  on 6 df,   p=1e-15
Score (logrank) test = 88.83  on 6 df,   p=<2e-16
```

Figure 18 : The result of the cox model with the age variable

The age variable is not only identified as insignificant by the non-parametric test, but also by the cox proportional hazards model since the p-value of the age is larger than 0.05. Let’s check whether the age variable can be omitted in the model, using the likelihood ratio test(LRT). The null hypothesis of the likelihood ratio test is that the full model with the age variable is not significantly different from the simple model that does not have the age variable. Since the p-value of LRT is 0.1362, the age variable can be omitted, and the cox model can be fitted to the data again without the age variable. The coefficients of each variable from the model without the age variable is summarized in the table below.



```
  n= 300, number of events= 220

                      coef exp(coef) se(coef)      z Pr(>|z|)
treatment          -0.5708    0.5651   0.1368 -4.171 3.03e-05 ***
cell_typeadeno      1.3751    3.9553   0.1986  6.924 4.38e-12 ***
cell_typelarge     -0.1261    0.8815   0.2046 -0.617 0.537443
cell_typesquamous   0.6181    1.8555   0.1790  3.453 0.000555 ***
cdp                 0.1051    1.1108   0.0349  3.011 0.002604 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
treatment             0.5651     1.7696    0.4322    0.7389
cell_typeadeno        3.9553     0.2528    2.6801    5.8374
cell_typelarge        0.8815     1.1344    0.5903    1.3162
cell_typesquamous     1.8555     0.5389    1.3063    2.6354
cdp                   1.1108     0.9002    1.0374    1.1895

Concordance= 0.676  (se = 0.018 )
Likelihood ratio test= 77.42  on 5 df,   p=3e-15
Wald test            = 80.41  on 5 df,   p=7e-16
Score (logrank) test = 86.38  on 5 df,   p=<2e-16
```

Figure 19 : The result of the cox model without the age variable

Compared to the standard treatment, the new treatment reduces the hazard around 43%(because exp(treatment=0.5651)). Since exp(cell_typeadeno) and exp(cell_typesquamous) are around 3.96 and 1.86, respectively, the patients with the adeno and squamous cell type are around 296% and 86% as likely to die than compared to the patients with the small cell type. However, the large cell type variable is not significant at the significance level 0.05.

A goodness-of-fit test of this model can be conducted by the Cox-Snell residual plot.

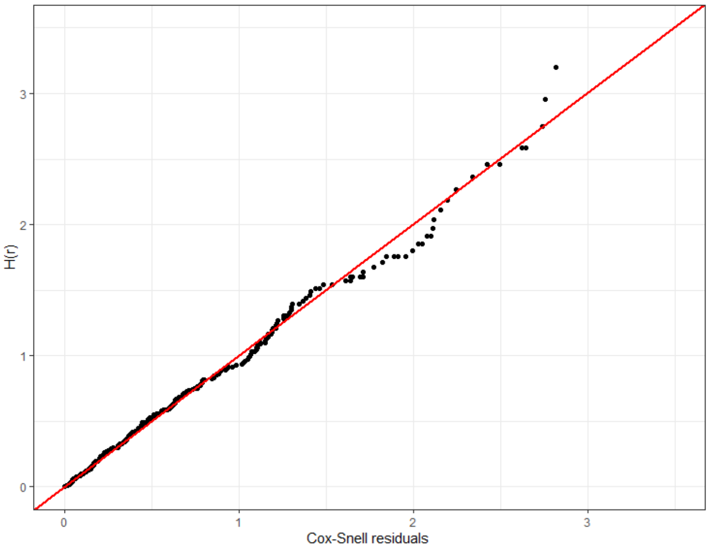


Figure 20 : The Cox-Snell residual plot.

The points are aligned with the red line, which has the slope 1. Thus, this model is appropriate in terms of the fitting.

Finally, we can make a prediction of survival probabilities based on this cox model. For example, a particular patient - with these characteristics(60 years old male, smoker, who had a cell type of squamous, and 4 drug prescriptions in the past 12 months) – would have the survival curves as shown in Figure 21.

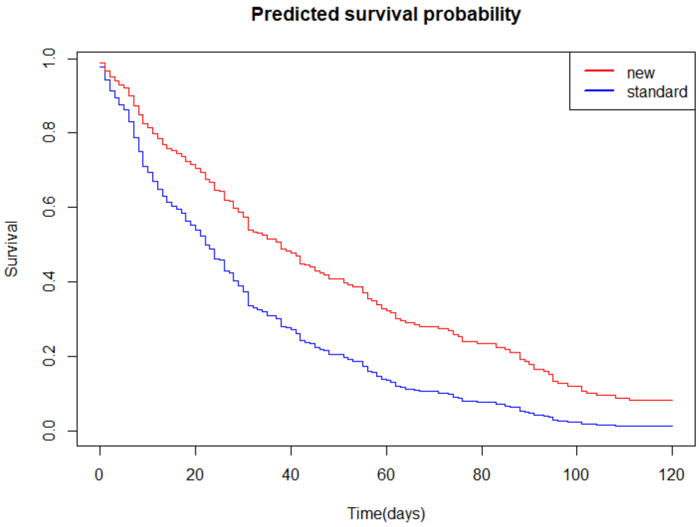


Figure 21 : The predicted survival curve per the treatment

If that particular patient had the new treatment, the three-month and six-month survival probabilities are 0.1793 and 0. For the case of the standard treatment, the three-month and six-month survival probabilities are 0.0478 and 0. If obtaining the accurate prediction is the only purpose(thus, the coefficients or inferences are not needed), it would be better to include every covariates in the cox model and make a prediction. Alternatively, one may use a machine learning model to increase the prediction accuracy by capturing the non-linearity effects of covariates.

3. Conclusion

In the subsection 2.1, from the data of the phase II trial, it is shown that there is no difference of the survival curves between the new and standard treatment group, using the Kaplan-Meier estimator and the Renyi type test(with weight 1). However, there should be extra caution about this result, as the distribution of the cell type is different across the groups. Plus, the survival curve is estimated with the parametric models, and the generalized gamma is selected as the best fit regarding the AIC.

In the subsection 2.2.1, by the Monte-Carlo experiments, the coverage rate of the median survival time and the power and Type 1 error of the log-rank test, in a situation where the sample size is 300 and the follow-up time is 120 days, are obtained. Furthermore, the relationship between the power of the log rank test and the sample size or the follow-up time is given by Figure 11.

Finally, in subsection 2.2.2, the risk factors are found by the non-parametric tests for the sake of the robustness. Then, the cox model with the significant risk factors(treatment, cell type, and CDP) is fitted to the data and all the underlying assumptions of the model are being tested by the various methods(e.g., Martingale residual, Cox-Snell residual and Schoenfeld test). Since all the assumption of the model are being met, the effect of each risk factor is measured by the coefficients of the covariates.

4, Further research

Overall, this paper analyzed the data of the Phase II and III trial based on the statistical methods. However, there are some parts where domain knowledge is required in order to further analyze these data in depth. First of all, in the cox model in the subsection 2.2.2, the age variable is identified as insignificant for the risk factor, which is counter-intuitive. Generally thinking, younger patients would have stronger immune system of a body and more likely to be resilient against disease. Plus, the p-value of the age variable in the cox model is around 0.14, which is not big enough so that we can not simply neglect. Secondly, the interaction terms(e.g., treatment*cell type) are not included in the cox model as domain knowledge related to health science is out of scope in this paper. It is also not wise approach to simply include interaction terms in the model and see the p-value, since there are several disadvantages of this approach as discussed in this paper[ii].

Furthermore, if you make the Kaplan Meier plot with the data of the Phase III trial, the survival probability converges to 0 in around 6 months, which is shorter than another data. Note that the median survival of patients with untreated metastatic non-small-cell lung cancer is only four to five years, according to the manager of this project. There might be a possibility that the patients, who were recruited in this project, were diagnosed with lung cancer later than other usual patients by chance or some reasons.

5. Appendix

- Monte-Carlo simulation in subsection 2.2.1

i) Data generation

Parameters of the algorithm(N, lambda, rho, beta and followup)

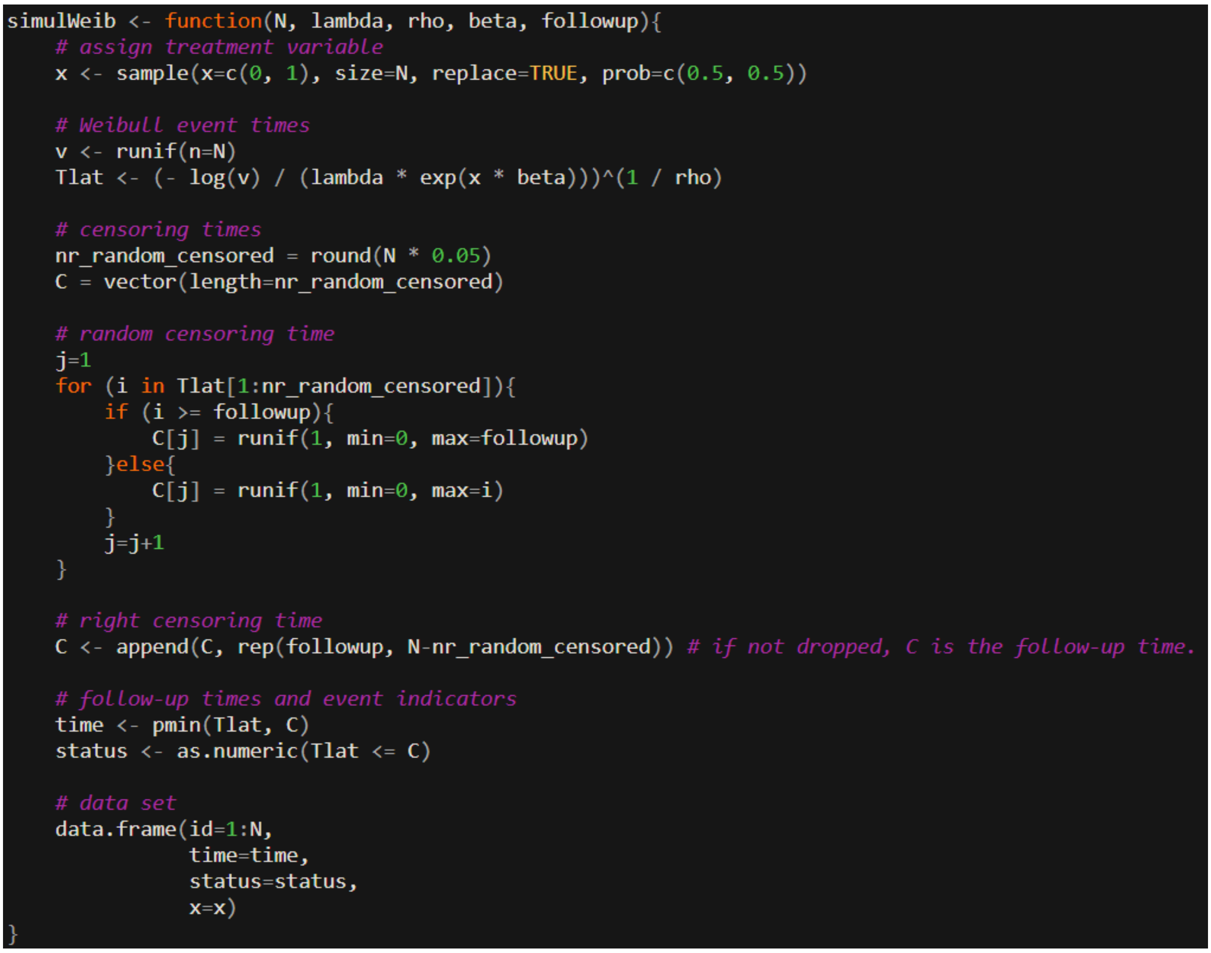
N = sample size

lambda = scale parameter in h0()

rho = shape parameter in h0()

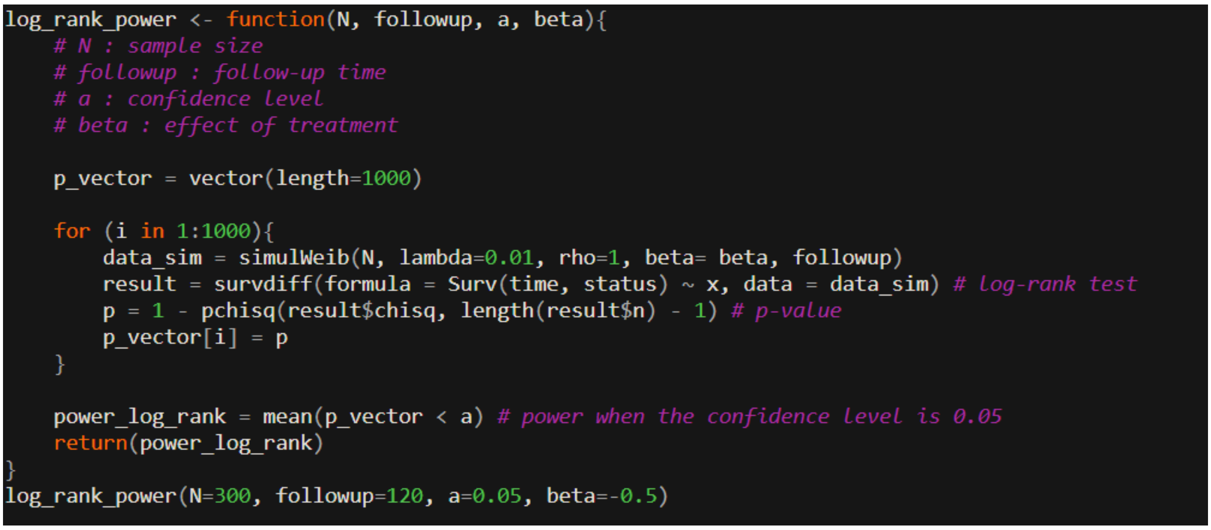
beta = fixed effect parameter

followup = follow-up time

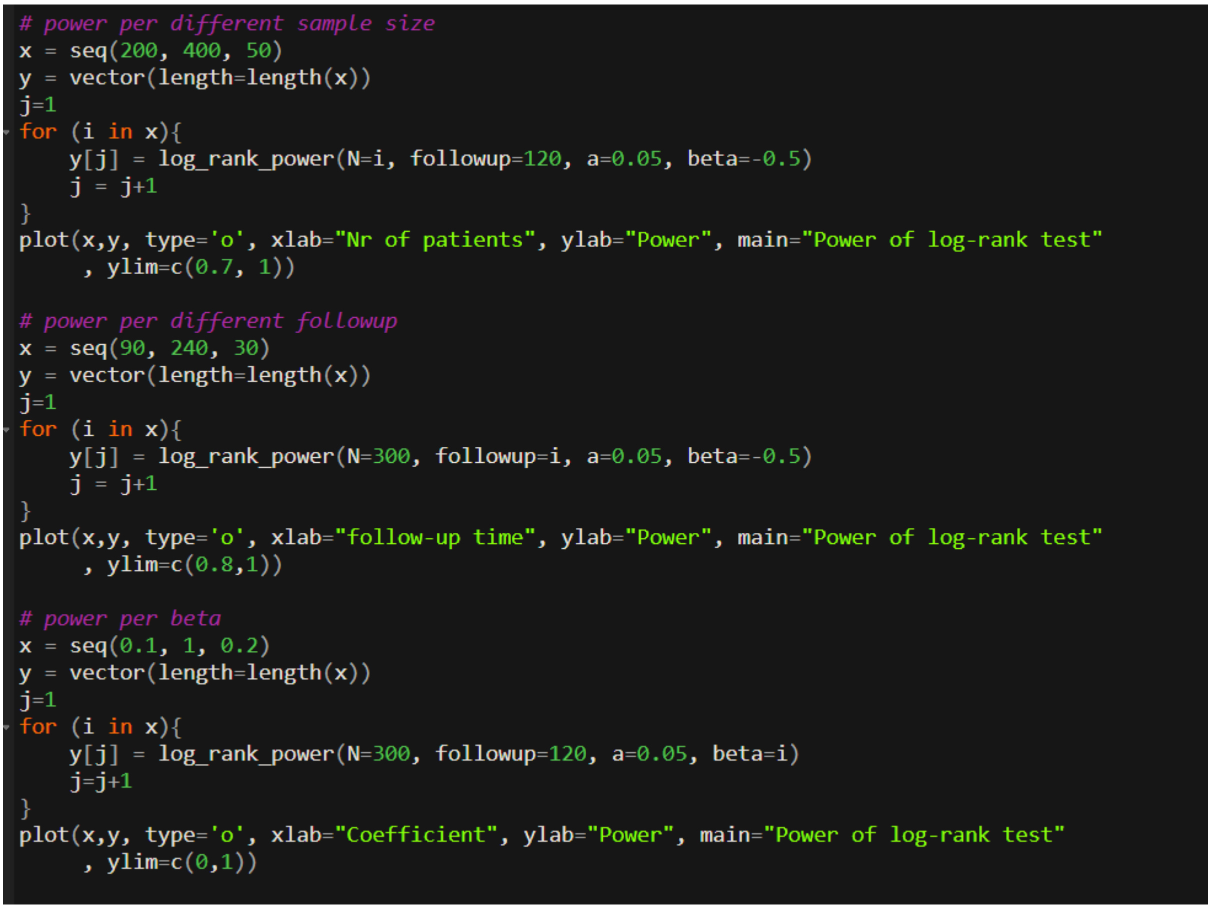


```r
simulWeib <- function(N, lambda, rho, beta, followup){
    # assign treatment variable
    x <- sample(x=c(0, 1), size=N, replace=TRUE, prob=c(0.5, 0.5))

    # Weibull event times
    v <- runif(n=N)
    Tlat <- (- log(v) / (lambda * exp(x * beta)))^(1 / rho)

    # censoring times
    nr_random_censored = round(N * 0.05)
    C = vector(length=nr_random_censored)

    # random censoring time
    j=1
    for (i in Tlat[1:nr_random_censored]){
        if (i >= followup){
            C[j] = runif(1, min=0, max=followup)
        }else{
            C[j] = runif(1, min=0, max=i)
        }
        j=j+1
    }

    # right censoring time
    C <- append(C, rep(followup, N-nr_random_censored)) # if not dropped, C is the follow-up time.

    # follow-up times and event indicators
    time <- pmin(Tlat, C)
    status <- as.numeric(Tlat <= C)

    # data set
    data.frame(id=1:N,
               time=time,
               status=status,
               x=x)
}
```

ii) Power calculation

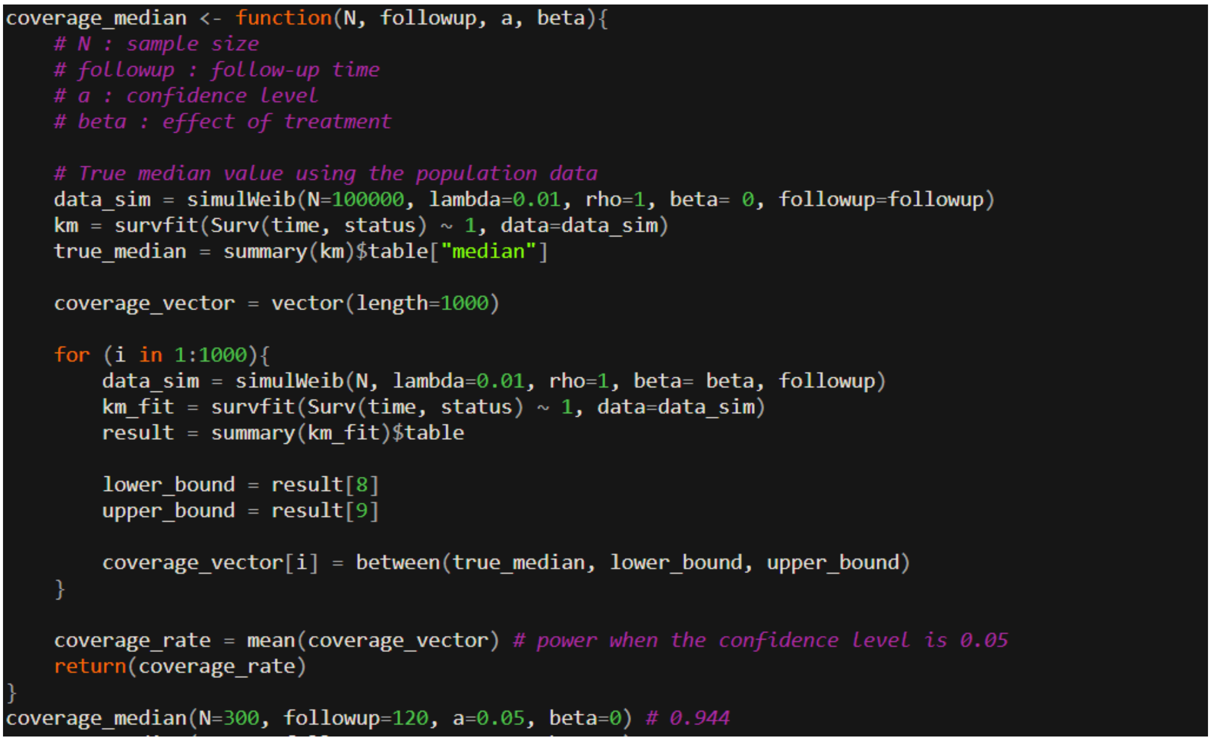


```r
log_rank_power <- function(N, followup, a, beta){
    # N : sample size
    # followup : follow-up time
    # a : confidence level
    # beta : effect of treatment

    p_vector = vector(length=1000)

    for (i in 1:1000){
        data_sim = simulWeib(N, lambda=0.01, rho=1, beta= beta, followup)
        result = survdiff(formula = Surv(time, status) ~ x, data = data_sim) # log-rank test
        p = 1 - pchisq(result$chisq, length(result$n) - 1) # p-value
        p_vector[i] = p
    }

    power_log_rank = mean(p_vector < a) # power when the confidence level is 0.05
    return(power_log_rank)
}
log_rank_power(N=300, followup=120, a=0.05, beta=-0.5)
```



```r
# power per different sample size
x = seq(200, 400, 50)
y = vector(length=length(x))
j=1
for (i in x){
    y[j] = log_rank_power(N=i, followup=120, a=0.05, beta=-0.5)
    j = j+1
}
plot(x,y, type='o', xlab="Nr of patients", ylab="Power", main="Power of log-rank test"
     , ylim=c(0.7, 1))

# power per different followup
x = seq(90, 240, 30)
y = vector(length=length(x))
j=1
for (i in x){
    y[j] = log_rank_power(N=300, followup=i, a=0.05, beta=-0.5)
    j = j+1
}
plot(x,y, type='o', xlab="follow-up time", ylab="Power", main="Power of log-rank test"
     , ylim=c(0.8,1))

# power per beta
x = seq(0.1, 1, 0.2)
y = vector(length=length(x))
j=1
for (i in x){
    y[j] = log_rank_power(N=300, followup=120, a=0.05, beta=i)
    j=j+1
}
plot(x,y, type='o', xlab="Coefficient", ylab="Power", main="Power of log-rank test"
     , ylim=c(0,1))
```

iii) coverage rate of the median survival time



```
coverage_median <- function(N, followup, a, beta){
    # N : sample size
    # followup : follow-up time
    # a : confidence level
    # beta : effect of treatment

    # True median value using the population data
    data_sim = simulWeib(N=100000, lambda=0.01, rho=1, beta= 0, followup=followup)
    km = survfit(Surv(time, status) ~ 1, data=data_sim)
    true_median = summary(km)$table["median"]

    coverage_vector = vector(length=1000)

    for (i in 1:1000){
        data_sim = simulWeib(N, lambda=0.01, rho=1, beta= beta, followup)
        km_fit = survfit(Surv(time, status) ~ 1, data=data_sim)
        result = summary(km_fit)$table

        lower_bound = result[8]
        upper_bound = result[9]

        coverage_vector[i] = between(true_median, lower_bound, upper_bound)
    }

    coverage_rate = mean(coverage_vector) # power when the confidence level is 0.05
    return(coverage_rate)
}
coverage_median(N=300, followup=120, a=0.05, beta=0) # 0.944
```

Reference

[i] American Cancer Society : https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

[ii] Małgorzata Mikucka Francesco Sarracino and Joshua Kjerulf Dubrow : Costs and Benefits of Including or Omitting Interaction Terms: A Monte Carlo Simulation