Survival analysis of lung cancer patients from the Veterans' Administration Lung Cancer study

Deniz Pekin[1], Felipe Lopez Velez[1] and Gustavo Mauricio Chinchayan Bernal[1]

[1] DSTI- S21 Applied MSc in Data Science & AI
(authors are listed in alphabetical order of the first name. All authors contributed equally to this work).

E-mails: deniz.pekin@edu.dsti.institute; felipe.lopez.velez@edu.dsti.institute,
gustavo.mauricio.chinchayan@edu.dsti.institute

Submitted: 15/11/2021

Abstract

In this report we evaluate the survival outcome of patients with lung cancer who received two different types of treatment. In a randomized experiment between two treatment groups for lung cancer, 137 male patients with advanced inoperable lung cancer were given either a standard therapy (control group) or a test chemotherapy (test group) observations of 7 variables: type of treatment, cell-type, survival time, Karnofsky performance score (which quantifies the patient's general well-being and activity level), time of diagnosis, age, and whether if the patient has received any prior treatment were also noted. **Our goal was to determine if the test treatment was beneficial to the patients' survival.** We found that the type of treatment does not significantly influence patient survival (p-value=0.93) while the cell-type and the Karnofsky Performance have a significant impact on patient survival (p-values 10^{-5} and 2.10^{-13} respectively). However, a generalized linear model showed that only the Karnofsky performance score had a statistically significant influence on the survival.

Keywords: Survival analysis, Kaplan-Meier model, Cox proportional hazard model, Log-rank test, linear model.

1. Introduction

Survival analysis applies to datasets where we measure events occurring over time, and gathers a set of methods to answer some questions such as:

- How much time does it take for an event to occur?
- What is the probability for a patient to survive a certain amount of time, given a condition?
- Are there statistically significant differences in survival time between different patient groups?

Here we use the 'veteran' dataset [1]. The data comes from a study conducted by the US Veterans Administration. Male patients with advanced inoperable lung cancer were given either a standard therapy or a test chemotherapy. Time to death was recorded for 137 patients, while 9 left the study before death. Various covariates were also documented for each patient:

- **trt:** treatment, 1 = standard, 2 = test
- **celltype:** factor describing the type of cell: 1=squamous, 2=smallcell, 3=adeno, 4=large
- **time:** survival time (from start of study to death), in days
- **status:** censoring status, 0 = patient death was not observed (survival time was censored), 1 = patient death was observed.
- **karno:** Karnofsky performance score (quantifies cancer patients' general well-being and activities of daily life, 0 = Dead to 100 = Normal)
- **diagtime:** time from diagnosis to randomisation, in months
- **age:** age of the patient in years
- **prior:** 0 = no prior therapy, 10 = prior therapy

In order to facilitate the analysis, we have transformed trt (treatment type) into a factor ("standard" & "test" instead of 1 & 2). The same operation was done for the prior variable ("no" & "yes" instead of 0&10). We have then divided the patients into 2 age groups: under 65 years old ("young") and over 65 years old ("old") and we also segmented the Karnofsky performance score into four groups, labelling scores 0, 25, 50, 75, 100 as “very bad”,“bad”,“good”, “very good” respectively.

We have first investigated whether the tested treatment improved patient survival. Then we proceeded to determine which factors provided a significant difference in survival.

1.1 Kaplan-Meier model

Kaplan-Meier (KM) statistic measures the probability that a patient will survive past a specific point in time. At t = 0, the statistic is 1 (or 100%). When t increases infinitely, the statistic becomes 0 (see equation below.)

$$\hat{S}(t) = \prod_{i:t\leq t}(1 - \frac{d_i}{n_i})$$

The plot of the KM estimator is a series of decreasing horizontal steps, approaching the true survival function based on conditional probabilities; each new proportion is conditional on the previous proportions.

In our study we performed a Log-rank test to see the influence of the treatment on the survival of the patients with the hypotheses below:

H0: the survival curves for both groups (test and control/standard treatment) are identical

H1: there is significant evidence that the curves are different.

1.1 Cox proportional hazard models

The Cox model is used to describe the simultaneous effect of several variables on the rate of a particular event happening at a specific point in time. In our study we use this model to analyze the simultaneous effects of treatment type, cell type, Karnofsky performance, diagnosis time, and the age of the patient on the patient's death. This model is defined with the Hazard Function, *h(t)*, which is the hazard that an event happens at a time = t, given that it has not occurred yet, prior to time t. In our case the hazard function measures the instantaneous likelihood of death at time = t without any relation to the events that happened in the past (e.g. without taking into account who died in the previous years). The hazard function returns a proportion, between 0 and 1.

We first applied a univariate Cox model to test the following hypotheses:

H0: The type of treatment doesn’t have any statistically significant coefficients

H1: The type of treatment has statistically significant coefficients.

Then we carry on with a multivariate model to determine which variables significantly affect the patients’ survival.

1.2 Machine learning models

We have performed a Random Forest model with the ranger() function in R. Using the same variables as the Cox model we generated 20 random curves and took the global average of survival for all patients. The ranger() function also allowed us to determine the most important variables according to the Random Forest model.

We have then performed predictions using the functions provided by the pROC() library.

2. Results

2.1 Summary statistics

Summary statistics are shown in Figure 1. The patients are separated evenly between the treatment and the control group. Patients’ ages range from 34 to 81 years old. We have 9 censored observations (6.5%) and we see that 93.4% patients die before the end of the study. 40 patients have received some kind of treatment prior to the study (29%). The age of patients ranges from 34 to 81 years old. We have 93 patients under 65 years of age and 44 patients over 65 years old.

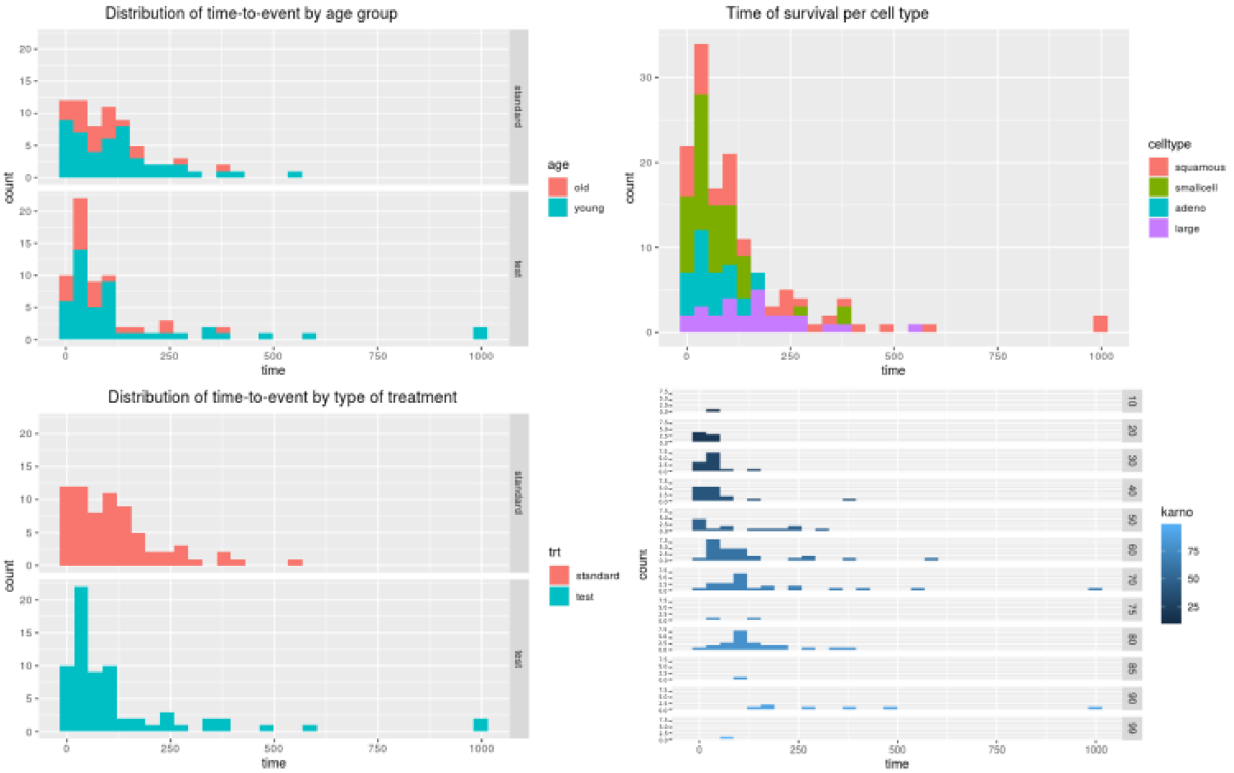
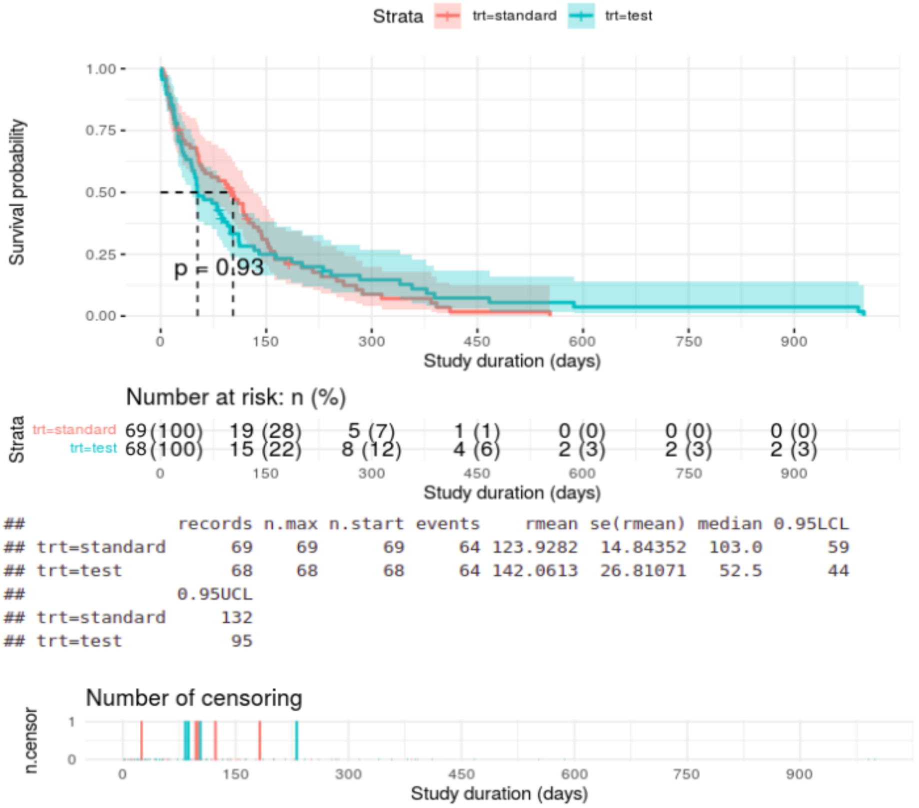


Figure 1 Distribution of patients according to age group, cell-type, treatment type and Karnofsky performance score

2.2 Effect of treatment on patient survival

The Kaplan Meier curves for standard treatment and test treatment groups are shown in Figure 2 with the summary statistics. We can observe that the last patient receiving the standard treatment dies after day 450 and 6 patients receiving the test treatment continue to survive until the end of the study. However, the p-value obtained for the Log-rank test is 0.93, meaning that we fail to reject the null hypothesis. There is no significant difference between the two curves, in other words the treatment doesn't impact the survival duration.



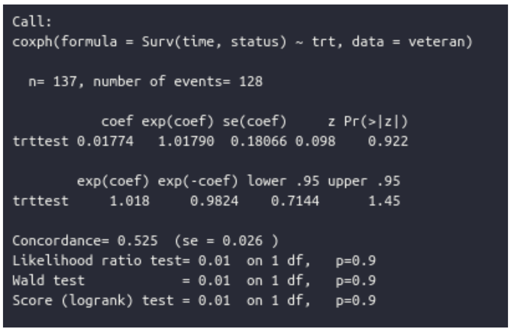
```
##              records n.max n.start events    rmean se(rmean) median 0.95LCL
## trt=standard      69    69      69     64 123.9282  14.84352  103.0      59
## trt=test          68    68      68     64 142.0613  26.81071   52.5      44
##              0.95UCL
## trt=standard     132
## trt=test          95
```

Figure2*- Kaplan Meier analytics for the treatment type: (top-panel) The Kaplan-Meier curve for patients who received the standard chemotherapy (pink) and the patients who received the test therapy cyan). The x-axis represents the survival time in days and the y-axis shows the probability of survival time, related to the number of days on the x-axis. The confidence intervals for both curves are represented as the area with the corresponding color. The p-value of the Log-rank test is represented on the graph p=0.93. A summary representation of the number of patients per study duration is shown in the middle panel. The bottom-panel shows the patients that have been censored (also represented as cross marks on the Kaplan-Meier curve). 9 patients left the study before death.*

2.3 Fitting Cox models

2.3.1. The univariate model on treatment type:

We have used the coxph() function provided by the survival package in R. The output is shown in Figure 3. The Wald statistic value obtained is z=0.098 and the p-value=0.922 showing that we fail to reject the null hypothesis, and the type of treatment doesn't have any statistically significant effect.

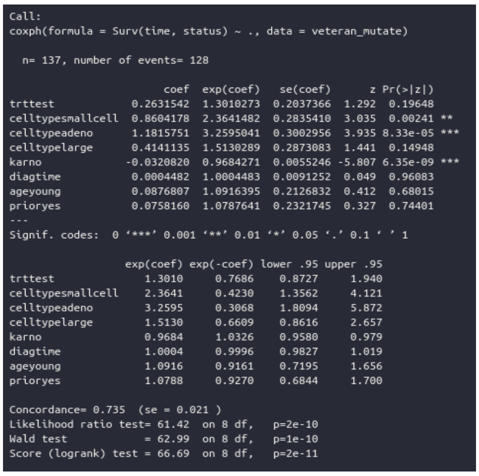


```
Call:
coxph(formula = Surv(time, status) ~ trt, data = veteran)

  n= 137, number of events= 128

           coef exp(coef) se(coef)     z Pr(>|z|)
trttest 0.01774   1.01790  0.18066 0.098    0.922

        exp(coef) exp(-coef) lower .95 upper .95
trttest     1.018     0.9824    0.7144      1.45

Concordance= 0.525  (se = 0.026 )
Likelihood ratio test= 0.01  on 1 df,   p=0.9
Wald test            = 0.01  on 1 df,   p=0.9
Score (logrank) test = 0.01  on 1 df,   p=0.9
```

Figure3*- The output of the univariate Cox model for treatment type. The p-value (0.92) shows that we fail to reject the null hypothesis according to which there is a significant difference between the two groups.*

2.3.1. The multivariate model:

With the multivariate model we observe that 2 variables are of statistical significance, namely: the cell type and the Karnofsky performance (Figure 4).



```
Call:
coxph(formula = Surv(time, status) ~ ., data = veteran_mutate)

  n= 137, number of events= 128

                       coef  exp(coef)   se(coef)      z Pr(>|z|)    
trttest           0.2631542  1.3010273  0.2037366  1.292  0.19648    
celltypesmallcell 0.8604178  2.3641482  0.2835410  3.035  0.00241 ** 
celltypeadeno     1.1815751  3.2595041  0.3002956  3.935 8.33e-05 ***
celltypelarge     0.4141135  1.5130289  0.2873083  1.441  0.14948    
karno            -0.0320820  0.9684271  0.0055246 -5.807 6.35e-09 ***
diagtime          0.0004482  1.0004483  0.0091252  0.049  0.96083    
ageyoung          0.0876807  1.0916395  0.2126832  0.412  0.68015    
prioryes          0.0758160  1.0787641  0.2321745  0.327  0.74401    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
trttest              1.3010     0.7686    0.8727     1.940
celltypesmallcell    2.3641     0.4230    1.3562     4.121
celltypeadeno        3.2595     0.3068    1.8094     5.872
celltypelarge        1.5130     0.6609    0.8616     2.657
karno                0.9684     1.0326    0.9580     0.979
diagtime             1.0004     0.9996    0.9827     1.019
ageyoung             1.0916     0.9161    0.7195     1.656
prioryes             1.0788     0.9270    0.6844     1.700

Concordance= 0.735  (se = 0.021 )
Likelihood ratio test= 61.42  on 8 df,   p=2e-10
Wald test            = 62.99  on 8 df,   p=1e-10
Score (logrank) test = 66.69  on 8 df,   p=2e-11
```

Figure4*- The output of the multivariate Cox model. The p-values for cell type (small cell and adenoma) and the Karnofsky performance score show statistical significance.*

The Cox model assumes that the covariates don’t change over time. We discuss this property on the Discussion section (see section 3.2).

2.4 Data modelling and Machine Learning

2.3.1. Random forest model

We have used the ranger() function in R to fit a Random Forests Ensemble model. This function builds a model for each observation in the dataset. We have used the same variables as the Cox model (Figure 5). The importance given to each variable (on the statistical significance) is shown in Figure 6.

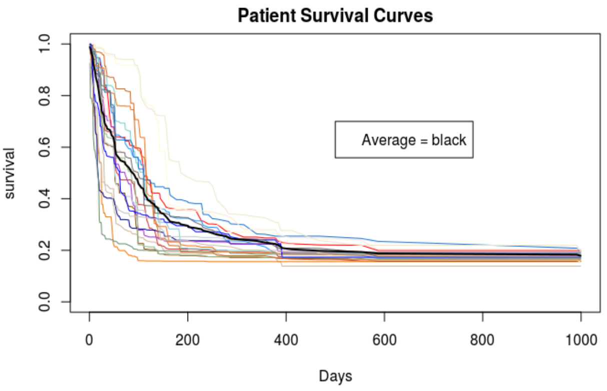
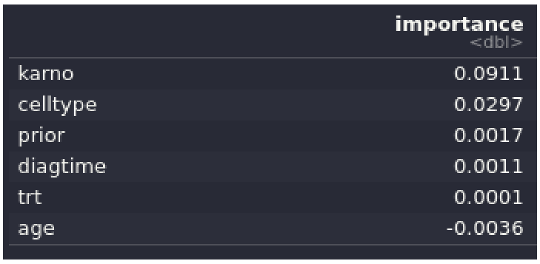


Figure 5- *The output of the Random Forests Model. Each of the variables (treatment, cell type, karno, diagtime, age and prior) were used to plot twenty random curves along with a curve that represents the global average for all patients.*



	importance <dbl>
karno	0.0911
celltype	0.0297
prior	0.0017
diagtime	0.0011
trt	0.0001
age	-0.0036

Figure 6 - *The importance given by the Random Forest Model to each variable.*

2.3.2. The predictive power (ROC & AUC)

Figure 7 shows the AUC obtained by with the linear models (see Supplementary Materials for the code).

AUC provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. The AUC was 0.731 in our case.

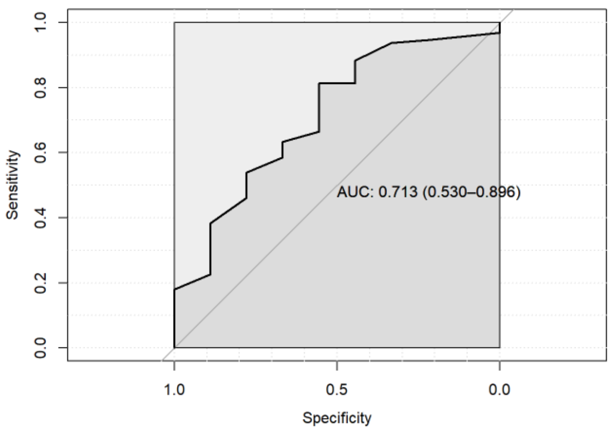


Figure 7- *The AUC obtained with the data. The AUC value obtained is 0.713*

3. Discussion

3.1 The effectiveness of the treatment

The goal of this study was to determine if the novel treatment improved patient survival. By candidly looking at the distributions, we observe that the patient which survives the longest belongs to the test treatment group (Figure 1). According to the Kaplan Meier test and the univariate Cox model however, the treatment does not have a significant effect on survival. It does not improve or worsen the patient outcome.

Besides, it is interesting to note that the median survival time for standard treatment is 103 days, against only 52.5 days for the test treatment (see Figure 2). Although this observation might raise the suspicion that the test treatment worsens the survival, the log-rank test shows a very high p-value (p=0.93), therefore even though the difference in medians seems large, there is no strong statistical evidence that the survival of both groups is different.

When we fit the univariate Cox model, we have observed that the regression coefficient has a positive sign. This means that the hazard (the risk of death) is higher in patients that are in the test treatment group. The beta coefficient for the treatment group is 0.01774 which indicates that patients within this group have higher risk of death compared to the standard treatment group. But again, the p-value for the test was 0.9 which means that the difference is not statistically significant. (Figure 3).

Finally, the p-values for Likelihood-ratio test, Wald test and the log-rank test are all 0.9. These three methods test the null hypothesis that all beta values are zero and are asymptotically equivalent.

3.1 The factors that influence patient survival

After observing that the treatment type does not affect the patients’ survival, we have switched to a multivariate strategy to determine the variables that in fact influenced the survival.

When we fit the multivariate Cox model, we observed that 2 variables have a significant effect, namely the cell-type (with an emphasis on the adenomatous cells) and the Karnofsky performance score (Figure 8).

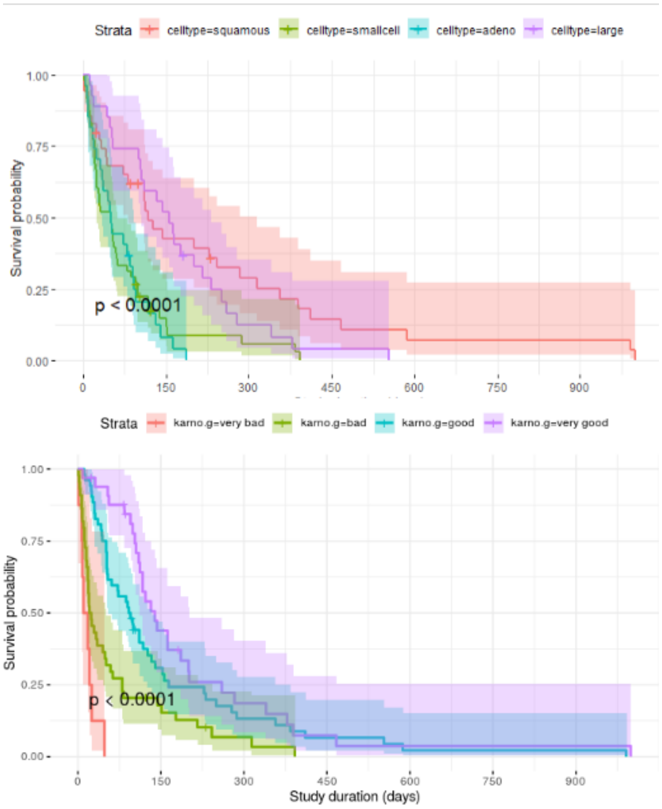


***Figure 8**- Kaplan Meier analytics for the cell type: (top-panel). The confidence intervals for both curves are represented as the area with the corresponding color code. The p-value of the Log-rank test is represented on the graph p<0.0001. A summary representation of the number of patients per study duration is shown in the middle panel.*

Furthermore, the Random Forest model flags the variables Karnofsky performance score and the cell type as most important. It is important to note however that the model doesn't address the time varying coefficients. We have found that this topic is a particular challenge in survival analysis [2].

3.2 The time dependence of variables

Despite being robust, the Cox model works under the assumption that the covariates do not vary over time. However, the Karnofsky score defines the patient's fitness and well-being and therefore is likely to be time dependent. We used the cox.zph() function of the R package to assess this time dependence (Figure 9). The plot and the p-value of the test (0.001) show that the effect of the Karnofsky performance score is not constant over time. In fact, for an illness as acute as the lung cancer, any measure of the well-being may cease to be relevant after several months.

In order to further investigate the time dependence, we have fit Aalen's additive regression model for censored data on all of the variables. Figure 10 shows how the effects of the covariates change over time.

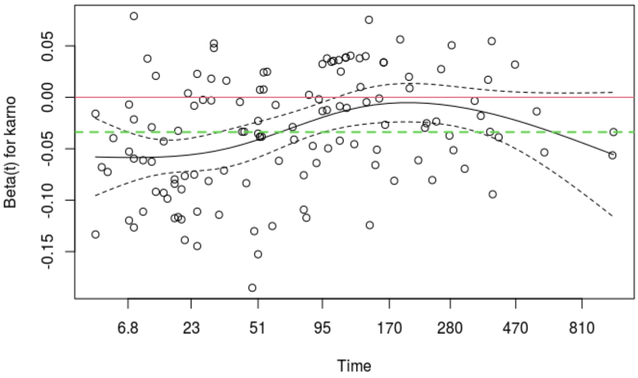


***Figure 9**- time dependence of the Karnofsky score. Early on the Karnofsky performance score has a large negative effect: the risk of someone at the first quartile is greater that of someone at the third quartile, but by 200 days this this effect approaches zero.*

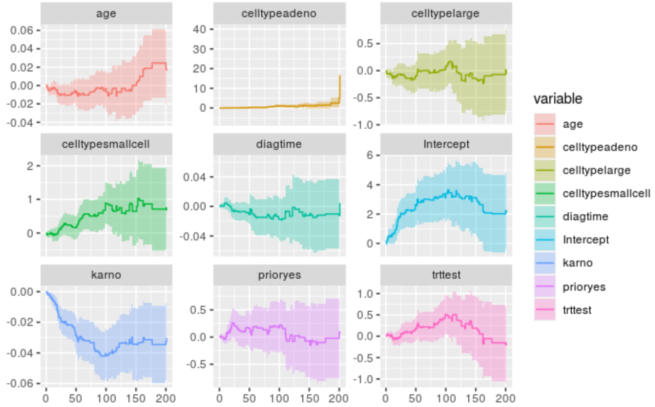


***Figure 10**- Time dependence of all variables. We can notice the steep slope and then abrupt change in slope of Karnofsky performance score.*

To tackle the time dependence of the Karnofsky score we used the survSplit() function to arbitrarily split the dataset into 3 parts: the first three months, 3-6 months and greater than 6 months. Then we performed the cox.zph() function on this time stratified dataset and obtained a p-value of 0.38 showing no longer dependency on time. A fit to the revised data showed that the effect of baseline Karnofsky score is essentially limited to the first two months (see Supplementary Materials for the code).

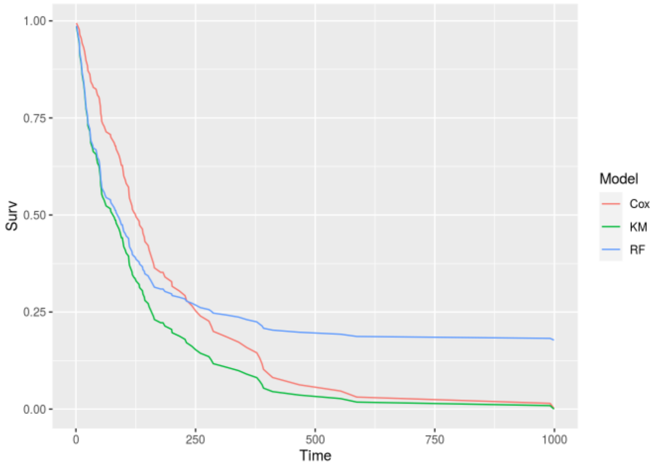


Figure 11*- Survival according to the different models tested in this study, namely the Cox model, the Kaplan-Meier (KM) and the Random Forest (RF).*

4. Conclusion

In this report we evaluate the survival outcome of patients with lung cancer who received two different types of treatment. Our initial goal was to determine if the test treatment was beneficial to the patients. Figure 11 shows the survival output of different models that we have tested (namely the Cox model, Kaplan-Meier and Random Forest).

While our results show that the patients did not receive any significant benefit from the treatment that was tested (p-value = 0.93). The Karnofsky performance score affected the patient survival on the first two months of the study (p-value <0.0001) and the patients with small cell and adenocarcinomas had a worse outcome (p-values <0.0001).

Further information on the prognostic value of the cell of origin in lung cancers can be found in this review by Sutherland [3].

References

[1] D Kalbfleisch and RL Prentice (1980), The Statistical Analysis of Failure Time Data. Wiley, New York.

[2] Bou-Hamad, I. A review of survival trees Statistics Surveys Vol.5 (2011).

[3] Sutherland KD, Berns A. Cell of origin of lung cancer. Mol Oncol. 2010

Supplementary Materials and Code

Please find our code and additional material as a single (and very short) file on the following address: https://github.com/TacticalNuclearRaccoon/DSTI-SurvivalAnalysis.git