Project 2: Survival Analysis of VA Lung Cancer Dataset

Neil Scheidwasser-Clow

08 August, 2020

1 Introduction

Cancer is the second cause of death in the United States and in the world [1]. Although declining, lung cancer remains the deadliest form of cancer for both men and women [2]. The most frequent symptoms include dyspnea (i.e. shortness of breath), constant coughing and weight loss [3]. Generally speaking, lung cancers are first divided by tumour cell size. Tumour cells of small-cell lung carcinomas (SCLC) appear smaller than those of non-small-cell lung carcinomas [4] [5] [6]. In the latter category, three subtypes are typically distinguished by cell appearance. While adenocarnicoma tumours have a glandular structure, squamous-cell carcinoma have flatter cells. Finally, large-cell carcinoma cells are simply larger than normal cells.

To get a deeper understanding of such chronic diseases, it is of significant importance to investigate the influence of risk factors. While most analyses tend to focus on disease epidemiology, similar studies can be conducted to evaluate current treatments or clinical trials. In this project, we examined the trial from the Veterans Administration of two treatment regimens for lung cancer [6] [7]. Although the main interest was to investigate differences between both treatment protocols, the importance of the following auxiliary variables was assessed: type of cancer (among the four types described above), age, functional ability [8], number of months from diagnosis to entry (or diagnostic-entry interval, DEI), and the existence of prior treatment.

For this survival analysis, exploratory data analysis was first performed to examine the main characteristics of the dataset. Given that some individuals were censored, standard regression procedures could not apply to analyse the dataset of interest. Thus, Kaplan-Meier estimates were first used to analyse differences in survival for the two therapies at stake. To investigate further the trial data, the model was enriched with the other risk factors, and Cox regression was performed to fit the hazard function of the resulting model. For all statistical tests, a significance level of 0.05 was used.

2 Data

In the experiment led by the Veterans Administration Lung Cancer Study Group [6], 137 male cancer patients were randomized to a standard or a test chemotherapy protocol. The response variable is the survival time (or censoring time for censored individuals), expressed in days. The corresponding event is whether an individual was dead or censored. Subsidiary quantitative risk factors include age (in years), duration between diagnosis and entry (DEI, in months), and the Karnofsky Performance Status (KPS). The latter, expressed in %, evaluates the functional abilities of cancer patients. 100% indicates normal function, with no visible symptoms, whereas 0% corresponds to death. Finally, both the lung cancer type (adenocarcinoma, small-cell, large-cell or squamous-cell) and the existence of prior treatment (yes or no) were grouping factors of the dataset.

3 Exploratory data analysis

Boxplots were realized for quantitative variables (age, KPS, failure (or censoring) time and time from diagnostic to entry). As can be seen in Figure 1, median age of participants was 63 years. The skewness of age distribution indicated that older patients were more represented. In contrast, Karnofsky index data was fairly equally distributed. Finally, both failure time and DEI contained numerous outliers. One can note that all these outliers originated from non-censored trials that did not include adenocarcinoma. While half of outliers from failure time data occurred in squamous-cell cancer trials (six with tested therapy, two with standard therapy), outliers from DEI data equally appeared in squamous- and small-cell lung cancers (two with standard, two with test therapy for each cell).

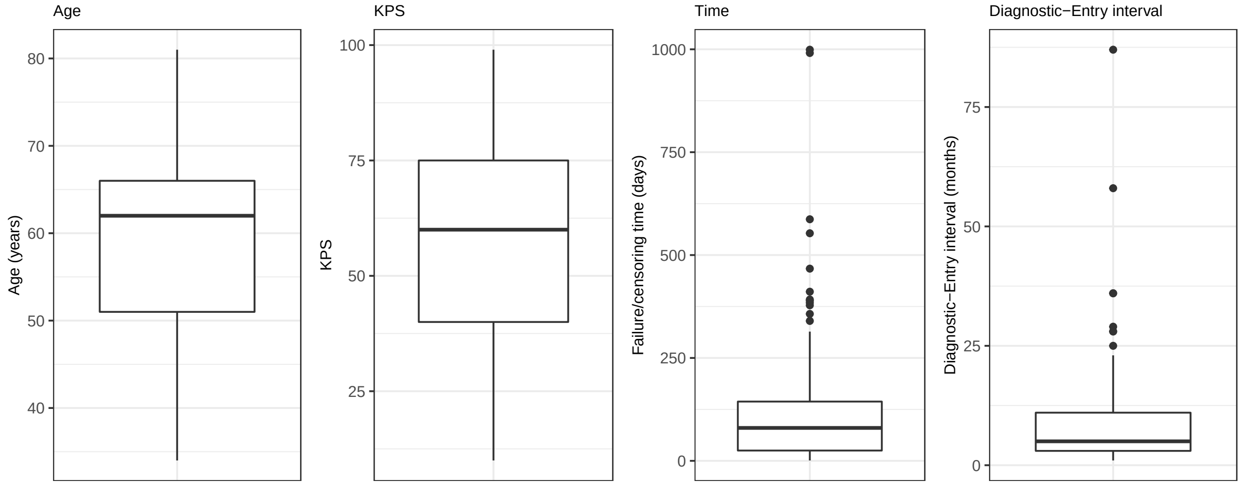

Figure 1: Boxplots of quantitative variables: age, KPS, survival time, and diagnostic-entry interval.

Categorical variables (type of cancer, type of therapy, status and existence of prior treatment) were analysed using histograms. As shown in Figure 2, the trial comprised less patients with small- and squamous-cell carcinomas. Besides, more than two-thirds of patients had no prior treatment before the trial. Given that less than 10% of patients were censored, one could choose to ignore censored data to perform linear or logistic regression. That said, discarding censored data could bias the analysis.

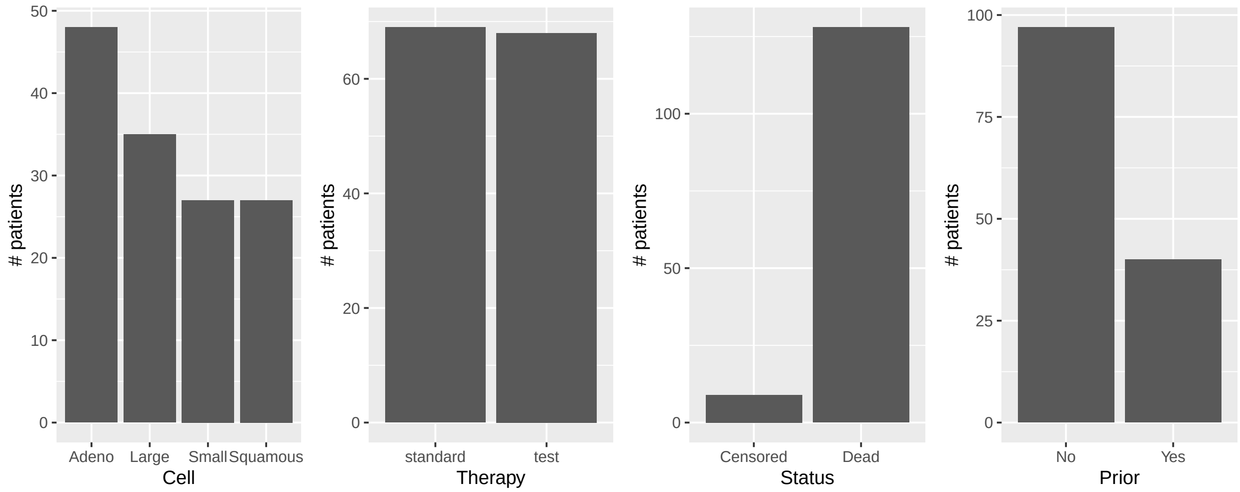

Figure 2: Histograms of categorical variables: type of cancer, type of therapy, status, and existence of prior treatment.

Following univariate data analysis, quantitative variables were also analysed in a bivariate fashion, both through their correlation matrix (Figure 3) and scatter plots (Figure 4). The correlation coefficients r_{ij} in the matrix are a measure of linear association between two variables. While $r_{ij} \approx \pm 1$ indicates a linear correlation, $r_{ij} \approx \pm 0$ denotes no linear association. In this dataset, no linear association arose between the quantitative variables of interest (maximum correlation was between KPS and time, $r = 0.38$). The scatter plots confirmed the nonlinear depenencies found in the correlation matrix. For each graph, bivariate points were generally stacked around a single region, excepting plots where KPS is the explanatory variable (as KPS is an ordinal discrete variable).

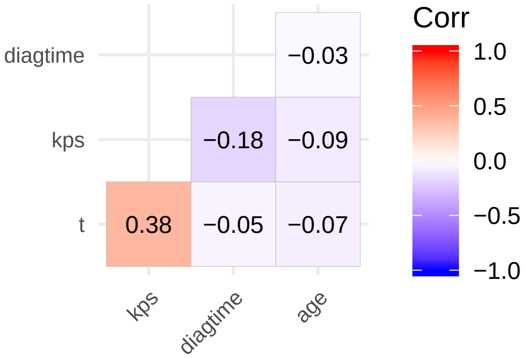

Figure 3: Correlation matrix of quantitative varirables from the dataset: age, KPS, survival time, and diagnostic-entry interval.

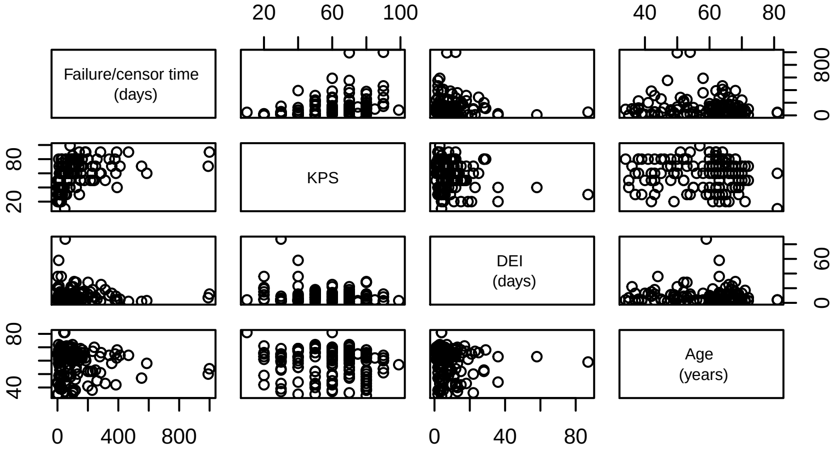

Figure 4: Scatter plots of quantitative varirables from the dataset: survival time, KPS, age, KPS, and diagnostic-entry interval (DEI).

4 Survival analysis

4.1 Kaplan-Meier estimates

To compare the effect of standard and test therapies over time, the survival function $S(t)$ is typically studied:

$$S(t) = Pr(T > t) = 1 - F(t), \tag{1}$$

where T is a continuous random variable, and F its cumulative distribution function. As our data includes censoring, individual survival as a function of therapy modality was estimated using the Kaplan-Meier estimator:

$$\hat{S}(t) = \prod_{j:t(j)\leq t} \left(1 - \frac{d_j}{r_j}\right), \tag{2}$$

where r_j is the number of individuals at risk just before $t(j)$ (including censored individuals at $t(j)$), and d_j is the number of individuals experiencing the event at time $t(j)$. From Figure 5, one could argue that patients with the test therapy seemed to survive longer. Yet, the difference between standard and test therapy curves was not statistically significant (p = 0.93) following a log-rank test, whose null hypothesis was that the survival function for both therapies were identical.

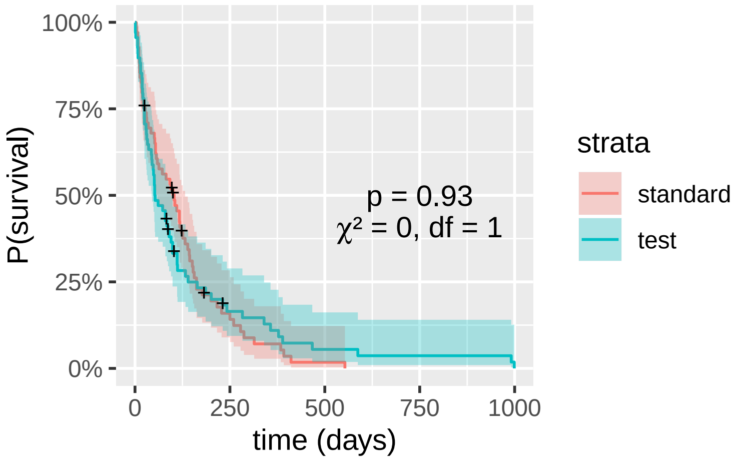

Figure 5: Kaplan-Meier estimation of the survival function for standard and test therapies. Log-rank p-value between survival curves for both therapy modes: 0.93.

A similar observation was found in Kaplan-Meier estimates stratified by cancer type or by the existence of prior treatment (Figure 6). Although the survival curves had different patterns depending on the stratification level (e.g. it could seem that squamous-cell cancer patients survived longer with the test therapy, while small-cell cancer individuals had a longer survival time with the standard therapy), the differences between therapies in terms of survival time were not statistically significant according to log-rank tests. Once again, the null hypothesis for this test is that both therapies have the same survival function, which could not be rejected for all stratification levels at a level of 0.05.

That said, such a univariate model ignores the other risk factors described in Section 1. Thus, the full dataset was fitted using Cox regression as a means to model the hazard function. This first model was denoted as the "full" model.

4.2 Cox proportional hazards model: "full" model

Whereas the survival function denotes the probability of survival beyond a given time, the *hazard function* h characterizes the probability of failure in a small time interval s over time *given survival* until time t:

$$h(t) = \lim_{s->0} \frac{P(t \leq T \leq t+s|T \geq t)}{s} \tag{3}$$

By contrast with other regression techniques, Cox regression (Equation 4) assumes linearity on the log hazard scale:

$$h(t) = h_0(t) \exp\left(\sum_{i=1}^{k} \beta_i x_i\right) \tag{4}$$

where h is the hazard function estimate, x_i the covariates of interest, β_i the coefficient for covariate x_i, and h_0 the *baseline hazard function*, which describes the common shape of survival time distribution for all individuals.

For the "full" model, Cox regression fitting was statistically significant according to Wald, likelihood ratio and log-rank tests ($p < 0.01$) (Table 1), which all evaluate the null hypothesis that all coefficients β_i in Equation

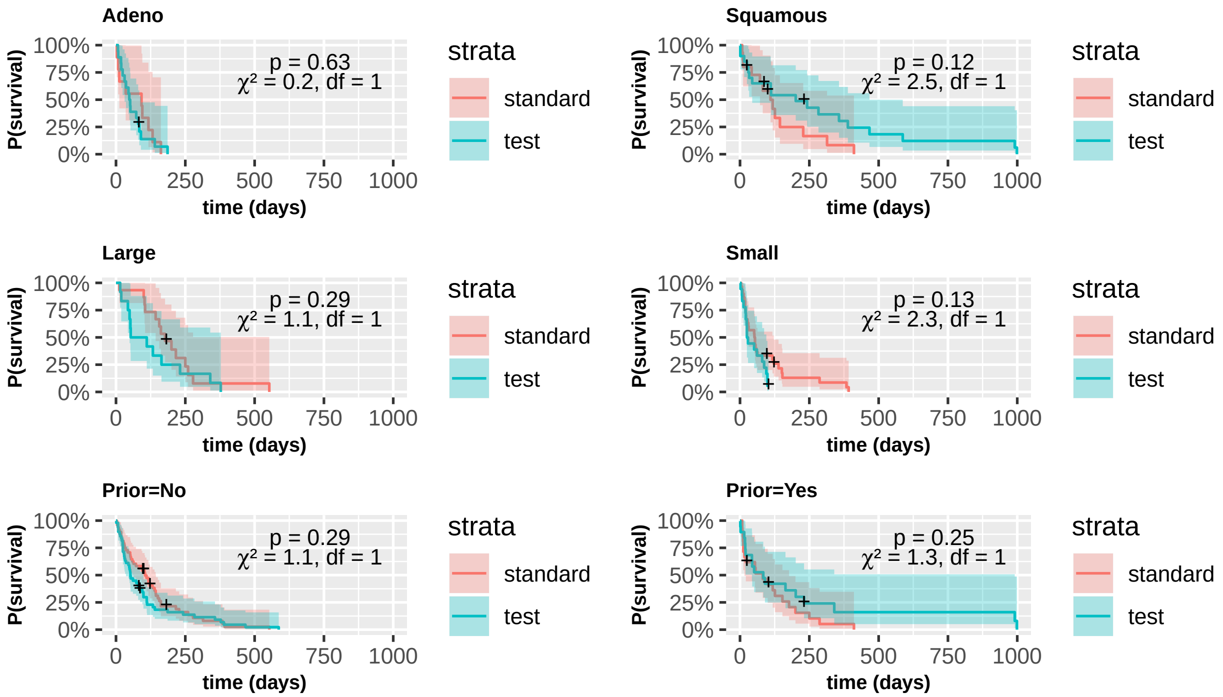

Figure 6: Kaplan-Meier estimation of the survival function for standard and test therapies, stratified either by cancer type (adenocarcinoma) or by existence of prior treatment. Log-rank p-value between survival curves for both therapy modes: 0.93.

4 are 0. In other words, the hypothesis that all coefficients β_i were null could be rejected, meaning that the model was not meaningless. Despite the inclusion of other risk factors, the difference between test and standard therapies remained statistically insignificant. Likewise, DEI, age and receiving prior treatment were also not significant to predict the hazard function. Concerning cancer types, only individuals with squamous- and large-cell lung cancers were statistically different from adenocarcinoma patients. Note that no statistical difference appeared between squamous- and large-cell subgroups when one of them was the reference group. By contrast, the KPS covariate was revealed to be an important predictor of survival time.

Nonetheless, the "full" model does not satisfy the proportional hazards assumption ("Global" variable in Table 2). Indeed, the p-value subject to the proportional hazards test was under 0.05, meaning that the null hypothesis that the modelled hazard function was time-invariant could be rejected. In particular, both the cell type and the KPS variables also failed the proportional hazards test ($p < 0.05$). To observe more in detail the time-dependencies of the latter variables, the Schoenfeld residuals were examined. More specifically, the Schoenfeld resdiual of a covariate represents the difference between the observed and the expected covariate given the risk set at that time. From the residual plots in Figure 7, one can notably observe that the Karnofsky Performance Score had a strong effect during the first 100 days of the trial. On the other hand, the "cell" variable residuals evolved in a parabolic fashion instead of being time-invariant. Thus, variables were transformed in order to satisfy the proportional hazards assumption. As the latter covariate is categorical, a stratified analysis depending on cancer type was carried out. To reduce the effect of the Karnofsky factor for failure times, all continuous variables were log-transformed. This new stratified model was thus denoted as the "log model".

Table 1: Summary of Cox regression fitting of the full model. For each covariate, estimated coefficient $\hat{\beta}$ and its 95% confidence interval are displayed. Log-rank test p-value between squamous- and large-cell cancer patients: $p = 0.40$.

	$\hat{\beta}$	$SE(\hat{\beta})$	z	$Pr(>\|z\|)$
Therapy (test vs. standard)	0.30	0.21	1.42	0.16
Cell (large-cell vs. adenocarnicoma)	−0.80	0.30	−2.62	0.01
Cell (small-cell vs. adenocarcinoma)	−0.34	0.28	−1.21	0.23
Cell (squamous-cell vs. adenocarcinoma)	−1.20	0.30	−3.98	7.05e-05 ***
Age	−0.01	0.01	−0.94	0.35
Prior treatment (yes vs. no)	0.07	0.23	0.31	0.76
DEI	0.0001	0.01	0.01	0.99
KPS	−0.03	0.07	−5.96	2.55e-09 ***
Observations	137			
Wald Test (df = 8)	62.37***			
LR Test (df = 8)	62.10***			
Score (Logrank) Test (df = 8)	66.74***			
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: Chi-square test of proportional hazards assumption of the full model. Null hypothesis: the variable (or model) is time-invariant. diagtime=DEI, chisq=Chi-square value, df=number of degrees of freedom, p=p-value.

	chisq	df	p
therapy	0.26	1	0.61
cell	15.23	3	0.00
age	1.83	1	0.18
prior	2.17	1	0.14
diagtime	0.01	1	0.91
kps	12.94	1	0.00
GLOBAL	34.55	8	0.00

4.3 Cox proportional hazards model: “log model”

Similarly to the “full” model, therapy mode, age, existence of prior treatment and diagnostic-entry time interval were not statisically significant in the “log” model (Table 3). Once again, the Karnofsky Performance Score was strongly significant. Overall, the model remained statistically significant against Wald, likelihood ratio and log-rank tests (p < 0.01), meaning that the null hypothesis that all coefficients β_i were 0 could be rejected.

For this stratified model, the proportional hazards assumption was valid. Indeed, the proportional hazards hypothesis could not be rejected for all covariates (Table 4). The p-value for the global test was also above the significance threshold (p = 0.11), meaning that the estimated hazard function could be deemed as constant over time (as all of its regressors also have a constant effect over time).

Given the numerous covariates that were statistically insignificant, the final step of this survival analysis aimed at optimizing the current model. To that end, backward elimination was performed as a means to model selection. Model selection was performed using the Akaike Information Criterion (AIC) [9]:

$$AIC = -2 \log L + 2p \tag{5}$$

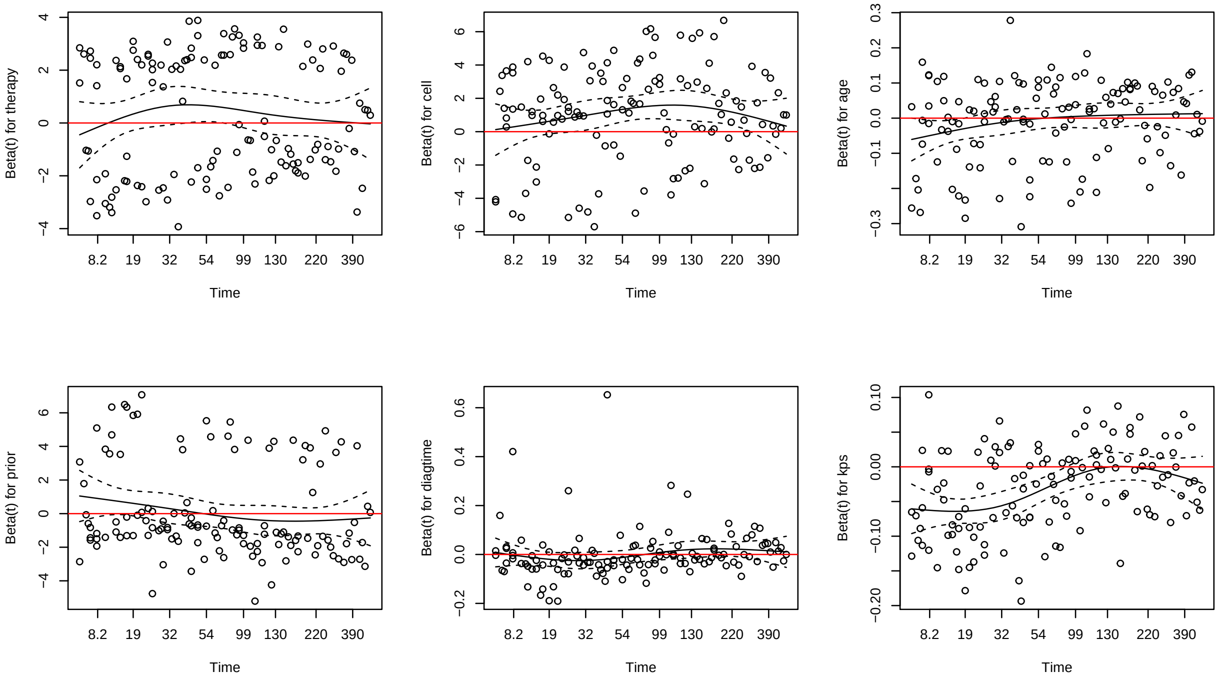

Figure 7: Scaled Schoenfeld residuals of covariates from the full model over time. diagtime=DEI. For reference, the y=0 line is shown in red.

Table 3: Summary of Cox regression fitting of the log model. For each covariate, estimated coefficient $\hat{\beta}$ and its 95% confidence interval are displayed.

	$\hat{\beta}$	$SE(\hat{\beta})$	z	$Pr(>\|z\|)$
Therapy (test vs. standard)	0.22	0.21	1.04	0.30
log(age)	−0.73	0.52	−1.34	0.16
Prior treatment (yes vs. no)	0.16	0.25	0.63	0.53
log(DEI)	−0.02	0.13	−0.14	0.89
log(KPS)	−1.53	0.24	−6.45	1.16e-10 ***
Observations	137			
Wald Test (df = 5)	43.96 ***			
LR Test (df = 5)	38.64 ***			
Score (Logrank) Test (df = 5)	50.12 ***			

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Chi-square test of proportional hazards assumption of the log model. Null hypothesis: the variable (or model) is time-invariant. diagtime=DEI, chisq=Chi-square value, df=number of degrees of freedom, p=p-value.

	chisq	df	p
therapy	0.09	1	0.77
log(age)	2.88	1	0.09
prior	1.91	1	0.17
log(diagtime)	0.04	1	0.83
log(kps)	0.91	1	0.34
GLOBAL	9.00	5	0.11

This metric assures a good balance between model fit (characterized by the $-2\log L$ term, with L the maximum likelihood of the data using the model) and model simplicity (penalized by the $2p$ term, with p the number of parameters of the model). In addition, likelihood ratio tests (LRT) tests were computed to ensure that the nested submodels were not statistically different from the initial "log" model. For this test, the null hypothesis is that the reduced model is true, whereas the alternative hypothesis is that the larger model is true. The statistic is given by

$$LRT = -2\ln\frac{L_R}{L_F} \sim \chi^2_{df}, \tag{6}$$

where L_R is the log-likelihood of the reduced model, L_F the log-likelihood of the full model, and df is the number of degrees of freedom.

For this model selection scheme, the best cell-stratified model only comprised the KPS covariate. To compare the normal "log" and simplified "log" models, an analysis-of-deviance table was implemented as a means to examine differences in fit statistics, e.g. the log-likelihood of both models. From the analysis-of-deviance summary in Table 5, the initial "log" model had a higher AIC score than the submodel obtained via backward elimination. Thus, the latter model was preferred, as a lower relative AIC value indicates a better fit. Finally, the chi-square difference test revealed that the simplified model was not different from the general "log" model at a significance level of 0.05. This result implied that the additional parameters in the complex model could justifiably be rejected, which also suggested that the simplified model should be selected.

Table 5: Analysis-of-deviance table comparing the general log model (first model) to the optimized model from backward elimination. loglik = Log-likelihood, Chisq=Chi-square value, Df = degrees of freedom. AIC index was used a model quality criterion to select the optimized model.

	loglik	Chisq	Df	P(>|Chi|)	AIC
Log model	-319.42				648.83
Simplified model: Surv(t, dead) ˜ strata(cell) + log(kps)	-320.98	3.13	4.00	0.54	643.96

In the same vein as the "log" model, the simplified submodel was statistically significant, and the Karnofsky factor remained a strong predictor albeit its log-transformation (Table 6). What's more, the differences in survival between cancer types in the "full" model (Table 1) were also observable in the survival curves of the simplified log model, shown in Figure 8. Combining these results, one could argue that squamous- and large-cell cancer patients generally showed better survival than veterans with adenocarcinoma and small-cell lung cancer. At last, the proportional hazards assumption was verified, with a higher global p-value than previous models (p = 0.22) (Table 7).

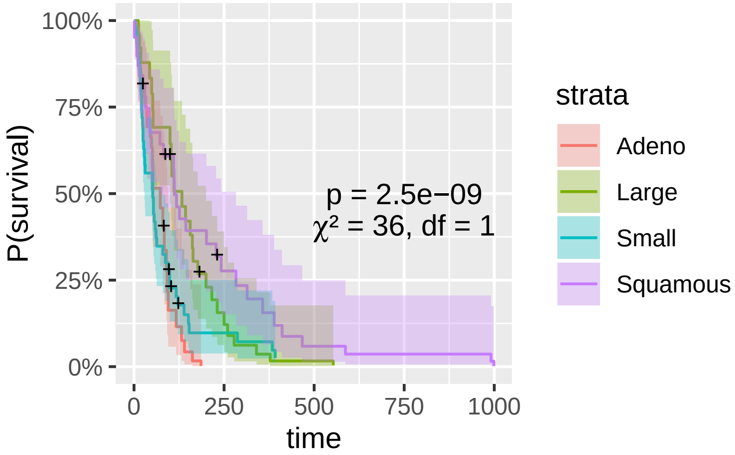

Figure 8: Estimated survival curves of the simplified log model for each strata (i.e. for each cancer type). Adeno=Adenocarcinoma.

Table 6: Summary of Cox regression fitting of the simplified log model. For the unique covariate of this model (KPS), estimated coefficient $\hat{\beta}$ and its 95% confidence interval are respectively displayed.

	$\hat{\beta}$	$SE(\hat{\beta})$	z	$Pr(> \|z\|)$
log(KPS)	-1.43	0.22	-6.41	1.46e-10 ***
Observations	137			
Wald Test (df = 1)	41.09^{***}			
LR Test (df = 1)	35.51^{***}			
Score (Logrank) Test (df = 1)	46.93^{***}			

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: Chi-square test of proportional hazards assumption of the simplified log model. Null hypothesis: the variable (or model) is time-invariant. chisq=Chi-square value, df=number of degrees of freedom, p=p-value.

	chisq	df	p
log(kps)	1.48	1	0.22
GLOBAL	1.48	1	0.22

5 Discussion

This project aimed at analysing the survival of lung cancer patients when treated with a standard or a test chemotherapy. Estimation of the survival function with the Kaplan-Meier estimator revealed that no significant difference appeared between the standard and test therapies following a log-rank test (where the null hypothesis is that both groups follow the same hazard function). Using Cox multivariate regression to model the hazard function, there was also no statistical evidence that these treatments differed in terms of survival time. Among the registered risk factors in this dataset, only the Karnofsky Performance Score and, to some extent, the cell type, were statistically significant. To verify the proportional hazards assumption, the cell type covariate was stratified whereas continuous variables were log-transformed. Using backward elimination, the best stratified model for Cox regression only comprised the log-transformed Karnofsky index. As a matter of fact, other risk factors (including the therapy mode) increased model complexity while not significantly improving the goodness-of-fit. With that in mind, according to the estimated coefficient for KPS in Table 6, the optimal model from this project could be summarized by Equation 7:

$$h(t) = h_0(t)\exp(-1.43\log(\mathrm{KPS})) = h_0(t)\mathrm{KPS}^{-1.43}, \tag{7}$$

where h and h_0 are defined as in Equation 4.

That said, the fact that KPS had a strong impact for low survival times (Figure 7) could motivate the use of accelerated failure time models, which assume time-varying effects of covariates. Given that the population in this dataset only consisted of male patients, trials including female cancer would enable a more meaningful comparison of the two treatments. Further research could also include other socioeconomic risk factors such as smoking, which is still by far the main cause of lung cancer [10].

References

[1] H. Ritchie and M. Roser, "Causes of death," *Our World in Data*, 2020.

[2] M. Roser and H. Ritchie, "Cancer," *Our World in Data*, 2020.

[3] D. L. Longo *et al.*, *Harrison's principles of internal medicine*, vol. 2012. Mcgraw-hill New York, 2012, pp. 551–562.

[4] A. J. Alberg and J. M. Samet, "Epidemiology of lung cancer," *Chest*, vol. 123, no. 1, pp. 21S–49S, 2003.

[5] C. Zappa and S. A. Mousa, "Non-small cell lung cancer: Current treatment and future advances," *Translational lung cancer research*, vol. 5, no. 3, p. 288, 2016.

[6] J. D. Kalbfleisch and R. L. Prentice, “The statistical analysis of failure time data,” *New York: John Wiley & Sons, Inc.*, 1980.

[7] R. L. Prentice, “Exponential survivals with censoring and explanatory variables,” *Biometrika*, vol. 60, no. 2, pp. 279–288, 1973.

[8] D. Karnofsky and J. Burchenal, “Evaluation of chemotherapeutic agents,” *New York, NY, Columbia University*, p. 19, 1949.

[9] H. Akaike, “A new look at the statistical model identification,” *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[10] P. M. de Groot, C. C. Wu, B. W. Carter, and R. F. Munden, “The epidemiology of lung cancer,” *Translational lung cancer research*, vol. 7, no. 3, p. 220, 2018.