Antoine Chen (r0863011) Charlotte Vercammen (r0611329) Sander Van Wayenbergh (r0620963)

Assignment
Statistical Analysis of Reliability and Survival Data

Introduction

The US Veterans Administration conducted a study whereby male patients with advanced lung cancer were given two different treatments (trt): the standard therapy (1) or test chemotherapy (2). Another categorical covariate was the cell type (celltype). There are four different cell types: squamous cell carcinoma (squamous), small cell carcinoma (smallcell), adenocarcinoma (adeno) and large cell carcinoma (large). Key variables are the survival time (time) and censoring status (status). The data is right-censored which is the most common type of censoring in survival datasets. The value is 1 when the person died due to lung cancer and 0 when censored. Censoring occurs when for example when the person is alive at the end of the study or drops out of the study. The Karnofsky performance score (karno) describes the overall status of patients at the beginning of the study. It ranges from 0 to 100 with 0 representing dead and 100 meaning normal / no complaints. Inbetween scores are determined by the percentage of time a day the patient is restricted to his bed. It thus stands for the level of physical impairment and lower scores usually go along with a higher mortality risk. In addition, the time from diagnosis to randomisation (diagtime) is given in months and the age of the patients (age) is given in years. The last three mentioned variables are continuous. Lastly, it was indicated if the patient had received another therapy before the current one (prior): 0=no, 10=yes. This variable is categorical. The veteran data set can be found in the survival package in R.

Methodology

Part A

The first step was to make a descriptive analysis of the variables in the data. When interpreting the data, one should keep in mind that some observations are censored.

Part B

In the second part, the focus is on the response variable (time), the censoring indicator (status) and a single categorical variable as covariate. Two categorical variables were investigated: treatment (trt) and cell type (celltype). First, the survival distribution was computed for each of the levels of the categorical variable. In doing so, the survfit() function was used to compute the Kaplan-Meier estimator. The result was plotted with the corresponding confidence intervals, risk table and legend. Next, the Kaplan-Meier estimator and confidence interval for the three quartiles of the survival distribution were obtained for each level. Because of the skewed distribution in the data, the median was taken instead of the mean. Subsequently, a single test of differences between the survival curves was conducted. The following hypotheses were proposed: "H_0: There is no difference between the survival functions" and "H_a: There is a difference between the survival functions". The survdiff() function was used for the purpose of hypothesis testing. This function can be used to perform the log-rank test or the Gehan-Wilcoxon test. The Gehan-Wilcoxon test gives more weight to deaths at early time points while the log-rank test gives equal weight to all time points. The log-rank test was chosen since it is the most commonly used among the two regarding survival analysis. There was also no compelling reason to opt for unequal weights.

Part C:

The Cox Proportional Hazard (CoxPH) model is needed to consider multiple covariates. The veteran data set contains continuous covariates worth investigating, such as the age of the patient and the Karnofsky performance score for functional impairment. Non-parametric survival functions lack the ability to deal with continuous covariates when they are not categorized and do not function properly when a large number of covariates is included in the model. The use of a semiparametric Cox proportional hazards model solves both issues. The general cumulative hazard function of this model is given below:

$h_i(t) = h_0(t) \exp(\beta^t x_i)$

The hazard function of subject i is expressed in terms of the baseline hazard function in which all covariates are set to zero multiplied with the exponent of a linear combination of coefficients and covariates. Only the last part depends on specified parameters. The hazard function thereby calculates hazard ratios with regard to a reference category. A hazard ratio (HR) larger than 1 signifies an increased risk of dying while a HR lower than 1 suggests the opposite. This can be done in Rstudio with the coxph() function of the Survival package. First of all, a backward variable selection procedure is applied to only include variables that have sufficient predictive value. Variables lacking significance in subsequent z-tests were removed to avoid the use of a model that is too complex.

This model relies on an important assumption that needs to be fulfilled or else the interpretation of the outcome may not be valid. As the name suggests, the model only functions properly when the ratio of the hazards remains constant over time. Major evidence in favor of the violation of this assumption is present when the survival functions of two separate groups cross. This would be due to time-dependent covariates for which the value of the covariate has a different effect on the hazard depending on the time point where we are situated in the survival curve. Regardless of whether this is the case, the Cox proportional hazard model will be used and discussed while ignoring the validation of the assumption since it is stated like that in the project description. Afterwards, a short paragraph will be devoted to the validation of the assumption and the consequences of possible violation.

So far, non- and semi-parametric survival functions were used. They are distribution-free to some extent, either completely in the case of non-parametric functions or partially when semi-parametric alternatives are applied. Parametric survival functions on the other hand have specified distributions with parameters that define them. Survival functions of this kind can be represented in two distinct, but equivalent, ways. First of all, the accelerated failure time model (AFT) has a survival function that can be split up in a parametric family of distributions S_0 and an acceleration factor $\exp(\theta^t x_i)$:

$S_i(t) = S_0(\exp(\theta^t x_i)t)$ (t=time)

θ is a vector of regression coefficients, which is transposed in the acceleration factor and multiplied with a vector containing values for the covariates of a particular subject i. The families of distributions that were used in the scope of our report are listed below:

Distribution	**$S_0(t)$**	**Specified parameters**
Exponential	$exp(-\lambda t)$	Rate parameter: λ>0
Weibull	$exp(-\lambda t^{\rho})$	Scale parameter: λ>0 Shape parameter: ρ>0
Log-logistic	$\frac{1}{1 + (t\lambda)^{\kappa}}$	Scale parameter: λ>0 Shape parameter: κ>0
Log-normal	$1 - F_N(\frac{log(t)-\mu}{\sqrt{\gamma}})$	Scale parameter: μ>0 Shape parameter: γ>0

Alternatively, the survival model can be represented as a linear expression with the logarithm of the survival time as response. Similarly to linear regression, the expression contains an intercept μ and a vector multiplication of the regression coefficients γ with their corresponding covariates x. Additionally, a noise parameter W is included together with a scaling parameter σ. A necessary requirement to enable the identification of the model is that the mean and the variance of W are known. The general representation of parametric survival functions in this way is given below.

$\text{Log } T = \mu + y^t x + \sigma W$

Both representations are equivalent and the following links are used to convert from one to another:

- S_0 = survival function of exp(μ + σW)
- $\theta = -\gamma$

Only the variables that remained after the backward variable selection procedure on the Cox proportional hazard model were included in the parametric survival models. Their performance was compared using the Akaike information criterion:

$\text{AIC} = -2 \log L + 2(p + 1 + k)$

In the parameter penalization term, a distinction is made between the parameters residing in the regression coefficient vector and the intercept (i.e. p+1) and the ones that depend on the distribution used (i.e. k). It simplifies the implementation of the formula because the first part is equal for all the parametric models since it arises from the number of covariates included in the model. Meanwhile, the second part differs and k is equal to the number of specified parameters in the distribution minus one.

Extra analyses:
Lastly, a pairwise analysis based on the log-rank test was performed with the pairwise_survdiff() function. The Benjamini and Hochberg (BH) method was used to correct for multiple testing. Afterwards, the cell types were clustered with the ksurvcurves() function.

Results

Part A:

Table 1: The summary

	trt	time	status	karno	diagtime	age	prior
Min.	1	1	0	10	1	34	0
1st Qu.	1	25	1	40	3	51	0
Median	1	80	1	60	5	62	0
Mean	1.496	121.6	0	58.57	8.77	58.31	2.92
3rd Qu.	2	144	1	75	11	66	10
Max.	2	999	1	99	87	81	10

Since the active control group has value 1 for treatment while the test chemotherapy group has value 2, a value close to 1.5 signifies an almost equal treatment allocation ratio. There were about the same number of patients for both groups. The survival time ranged from 1 to 999 time units with a median of 80. Since the mean of censoring status was close to 1, right censoring only occurred in a few cases. The Karnofsky performance score ranged from 10 to 99 with a median of 60. The time from diagnosis to randomisation ranged from 1 to 87 months with a median of 5 months. The age of the patients ranged from 34 to 81 years with a median of 62 years. Since the mean of prior is closer to 0 than 10, it can be derived that there were more patients without prior therapy than with (Table 1).

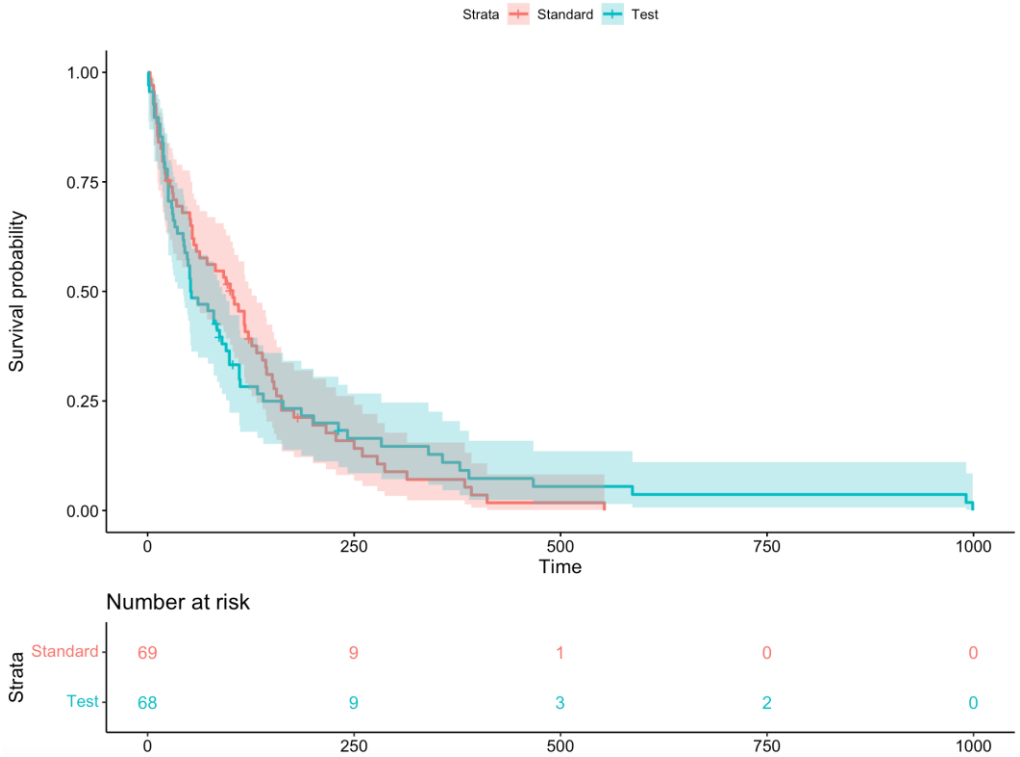
Table 2: The numbers of censoring status, treatment and cell type

Censoring status		Treatment		Cell type			
0	1	1	2	Squamous	Small cell	Adeno	Large
9	128	69	68	35	48	27	27

Only 9 out of 137 patients were censored. As previously mentioned, the subjects were equally divided among treatment groups with 68 patients assigned to the test group and an active control group consisting of 69 patients. Small cell lung cancer was the most frequent type of lung cancer among the veterans with 48 cases. The second most prevalent type was squamous cell carcinoma with 35 cases, which was followed by a tie between the two remaining cancer cell types. Both large cell and adenocarcinoma had 27 cases (Table 2).

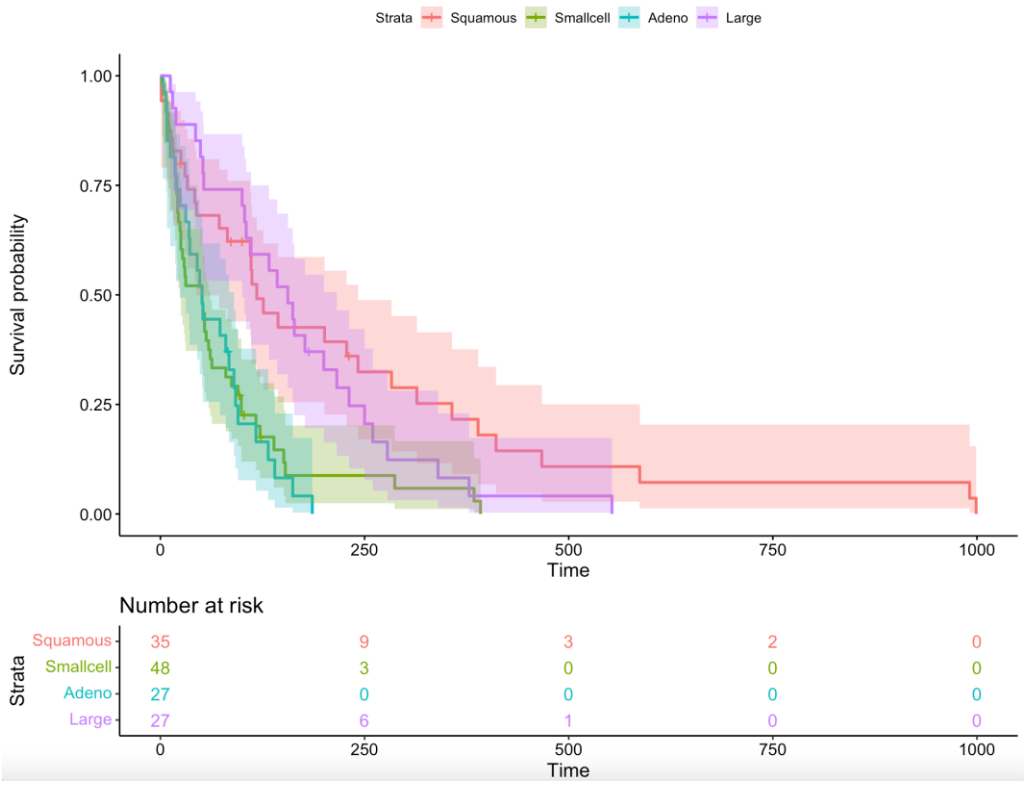
Part B

Figure 1: The survival distribution of the two treatments



p=0.93

Figure 2: The survival distribution of the four cell types



p<0.0001

When comparing the treatments, the survival curves for the standard and test group look similar. The curves also cross about halfway through. There is no significant difference according to the log-rank test (p-value = 0.93)(Figure 1). In contrast, differences between the cell types were apparent which was confirmed by the significance shown by the corresponding log-rank test (p-value < 0.001). At first sight, two clusters can be observed. Squamous and large cell carcinoma share a similar course of events on the one hand while small cell and adenocarcinoma can be clustered on the other hand. Within such an arbitrary cluster, an intersection between the two relevant survival curves is present in both cases, which has important implications for later on (Figure 2).

Table 3: Kaplan-Meier estimator for the three quartiles of survival time with confidence intervals for the survival functions of different treatments

Trt	**n**	**events**	**1st quartile (75%)**	**2nd quartile (median)**	**3th quartile (25%)**
Standard	69	64	27; 95%CI[12,54]	103; 95%CI[54,126]	162; 95%CI[132,250]
Test	68	64	25; 95%CI[15,33]	53; 95%CI[43,90]	140; 95%CI[99,283]

The Kaplan-Meier estimator revealed results regarding the median survival time that are quite contradictory at first. The median survival time of patients receiving the experimental chemotherapy was only 53 (95% CI [43,90]) time units. Thereby it seemed that they had a lower life expectancy than the standard group, which had a median survival time of 103 (95% CI [54,126]) time units (Table 3). But the survival functions cross halfway through as mentioned earlier. Moreover, the functions stay close to each other prior to the intersection given that the confidence intervals continuously overlap (Figure 1). This could indicate that the treatment was only successful for a small fraction of patients, two patients to be precise, who survived longer. Both of them had squamous cell carcinoma and this could be a lead to establishing the specificity of the test treatment, but does not prove anything since it could also be a coincidence.

Table 4: Kaplan-Meier estimator for the three quartiles of survival time with confidence intervals for the survival functions of different cell types

Celltype	**n**	**events**	**1st quartile (75%)**	**2nd quartile (median)**	**3th quartile (25%)**
Squamous	35	31	33; 95%CI[10,110]	118; 95%CI[44,242]	347; 95%CI[201,587]
Smallcell	48	45	20; 95%CI[10,25]	51; 95%CI[24,61]	99; 95%CI[59,151]
Adeno	27	26	19; 95%CI[7,36]	51; 95%CI[25,90]	92; 95%CI[73,140]
Large	27	26	53; 95%CI[15,111]	156; 95%CI[100,216]	231; 95%CI[164,340]

After applying the Kaplan-Meier estimator to obtain an estimate of the median survival time for different cell types, the same split into two clusters was suggested. Apparently, small cell and adenocarcinoma reduced the median survival time of the veterans to a greater extent than large and squamous cell carcinoma. Small cell and adenocarcinoma both had a median survival time of 51 time units, with slightly different boundaries for their respective confidence intervals. Within the alleged less harmful cluster, patients suffering from large cell carcinoma had the longest median survival time 156 (95% CI [100,216]) and veterans with squamous cell carcinoma had the second longest life expectancy with a median of 118 (95% CI [44,242]). Note however that the confidence interval for the latter disease type was rather wide and it covers the median of the two deadliest types of lung cancer. Last but not least, the right censoring seems to be equally divided among the categories when comparing their sample size with the number of events (Table 4).

Table 5: Log-rank test for differences in survival distribution between treatments

Trt	n	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Standard	69.0	64.0	64.5	0.00388	0.00823
Test	68.0	64.0	63.5	0.00394	0.00823

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

When testing the null hypothesis with the log-rank test, there seems to be no difference in the survival function between the treatments. The null hypothesis could not be rejected ($\chi_1^2 = 0$; p-value = 0.93)(Table 5).

Table 7: Log-rank test for differences in survival distribution between lung cancer cell types

Celltype	n	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Squamous	35.0	31.0	47.7	5.82	10.53
Smallcell	48.0	45.0	30.1	7.37	10.20
Adeno	27.0	26.0	15.7	6.77	8.19
Large	27.0	26.0	34.5	2.12	3.02

Chisq= 25.4 on 3 degrees of freedom, p= 1e-05

When testing the null hypothesis with the log-rank test, there is a difference in survival function between the different cell types. The null hypothesis is rejected ($\chi_1^2 = 25.4$; p-value = 1e-05)(Table 7).

Part C:

Table 9: The full model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
treatment	2.95e-01	1.34e+00	2.08e-01	1.42	0.156
Celltype smallcell	**8.62e-01**	**2.37e+00**	**2.75e-01**	**3.13**	**1.75e-03**
Celltype adeno	**1.20e+00**	**3.31e+00**	**3.01e-01**	**3.985**	**7.05e-05**
Celltype large	4.01e-01	1.49e+00	2.83e-01	1.42	0.16
Karnofsky score	**-3.28e-02**	**9.68e-01**	**5.51e-03**	**-5.96**	**2.55e-09**
Diagnose time	8.13e-05	1.00e+00	9.14e-03	0.01	0.99
age	-8.71e-03	9.91e-01	9.30e-03	-0.94	0.35
Prior treatment	7.16e-03	1.01e+00	2.32e-02	0.31	0.76

The variable selection procedure suggested that the Karnofsky performance score was the only additional covariate with any predictive value (Table 9). Information on the age of the patient, the months from diagnosis to randomization and whether prior treatment was received were removed from the model. Once again, the effect of treatment on the survival of lung cancer patients was not apparent. The main focus of the survival analysis was therefore directed towards differences in the survival time of patients with distinct types of lung cancer.

Table 10: The reduced model

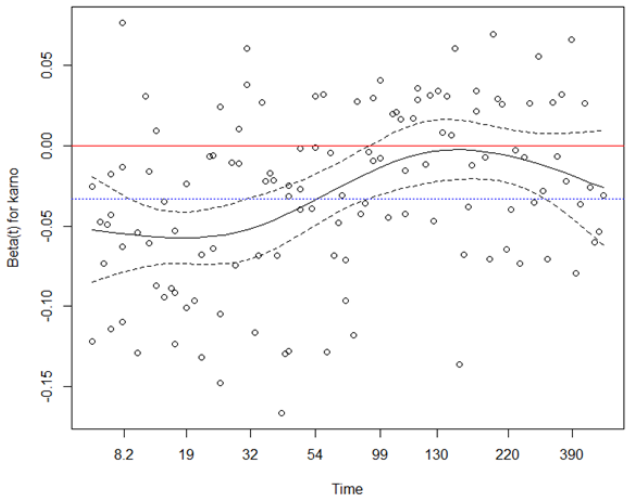
Covariate	β	s.e.(β)	HR	95% CI
Celltype smallcell	0.72	0.25	2.04	[1.24,3.36]
Celltype adeno	1.16	0.29	3.18	[1.79,5.65]
Celltype large	0.33	0.28	1.38	[0.81,2.38]
Karnofsky score	-3.10 e-02	5.18e-03	0.97	[0.96,0.98]

Results of the z-test and corresponding p-values are not shown again as they barely changed following the variable removal. The squamous lung cancer subgroup was selected as the reference category by default. This subgroup had the longest life expectancy among the cancer cell types. The hazard ratio was defined as the exponent of the parameter estimate. All three remaining lung cancer cell types had a hazard ratio larger than 1, thus indicating that they reduced the chance of surviving the disease even further than the reference category (Table 10).

Lung adenocarcinoma was the most severe cancer cell type with a hazard ratio of 3.18 (95% CI [1.79,5.65]) with regard to the squamous cell carcinoma of the lungs as reference. Next up, small cell lung cancer was the second most deadly type of lung cancer with a hazard ratio of 2.04 (95% CI [1.24,3.36]). These findings were solid indications that both variants of lung cancer were more severe than the reference category since 1 is excluded from the confidence interval in both cases. Large cell lung carcinoma, on the other hand, did not share this conclusion. Although a hazard ratio larger than 1 was obtained, namely 1.38 (95% CI [0.81,2.38]), we could not conclude that the cancer type was more detrimental to the survival odds of the patient than squamous cell carcinoma because the confidence interval contains the ratio 1 where the two are equally harmful. The karnofsky performance score variable has a continuous nature. It was shown to reduce the risk of mortality due to lung cancer since a hazard ratio of 0.97 (95% CI [0.96,0.98]) was obtained (Table 10). The larger the score, the better the condition of the patient. This outcome is thus in line with our expectations. A one-point increase on the score scale reduces the respective risk by 3 %. However, the Karnofsky performance status scale works with steps of ten units. Plugging in the number 10 in hazard ratio calculation (i.e. exp(10 x -3.10 e-02)), we get a ratio of 0.73 for a ten-point increase. A patient residing one level higher on the scale has a reduction in the risk of mortality by 27 % when only looking at the Karnofsky performance score variable.

Up to this point, we ignored the proportional hazard assumption intrinsic to the semi-parametric survival model applied here. Validation of the assumption was done with the cox.zph() function of the 'survival' package in R. The fact that the survival functions crossed both when stratifying according to treatment or cell type served as an obvious indication that the assumption was violated. Indeed, the validation of the assumption in the reduced model, which focused on comparing the course of the survival functions of different cell types, showed that both the effect of the cancer cell type ($p = 2.82*10^{-3}$) and the Karnofsky score ($p = 1.50*10^{-4}$) were time-dependent. For the covariate of the Karnofsky score, this is shown below by means of plotting its coefficient against time (Figure 3). A high Karnofsky score, which corresponds to a low degree of physical impairment at baseline, reduces the risk of mortality since the model coefficient stays below the red line in the graph. However, the coefficient is much larger in absolute value at the start of the study and gradually approaches zero after some time. In summary, a high Karnofsky score at the start of the study reduces the risk of mortality but its negative effect on the hazard fades after a while because the disease progresses while the score is not remeasured. Baseline measurements of this covariate gradually become useless and no longer represent the state of the patient. The blue dotted line in the graph represents the coefficient estimate that was previously obtained (Figure 3). It clearly shows that this averaged value is not a robust measure. A possible solution would be to split the survival function in several time epochs and study the behavior of the cox proportional hazard model at each compartment.

Figure 3: The coefficient of the Karnofsky score plotted against time



Several distributions were assessed to find the parametric model best suited to describe the veteran data set. The survreg() function of the 'survival' package was used for this matter. The four most promising candidates are given below with their corresponding AIC values:

Table 11: Akaike Information Criterion (AIC)

Log-logistic	**Exponential**	**Log-normal**	**Weibull**
1437.19	1443.94	1444.32	1445.03

In accordance with the objective of minimizing the AIC, the log-logistic model (AIC = 1437.19) was selected as the best option to model the behavior of the survival functions (Table 11). The coefficients of the linear representation of the model were extracted from the R output. Subsequently, the coefficients of the accelerated failure time (AFT) model were derived from this by changing the sign of the coefficients.

Table 12: Coefficients for the log-logistic parametric survival model in linear and accelerated failure time (AFT) representation

coefficient	**Log-logistic**	
	Linear	**AFT**
Intercept (=squamous)	2.48; 95%CI[1.82,3.14]	-2.48; 95%CI[-1.82,-3.14]
small cell	-0.69; 95%CI[-1.17,-0.21]	0.69; 95%CI[1.17,0.21]
adeno	-0.78; 95%CI[-1.3,-0.26]	0.78; 95%CI[1.3,0.26]
large	-0.029; 95%CI[-0.55,0.49]	0.029; 95%CI[0.55,-0.49]
karnofsky	0.036; 95%CI[0.03,0.04]	-0.036; 95%CI[-0.03,-0.04]

The linear equivalent of the model sets the logarithm of the survival time as the response in function of an intercept μ, a linear combination of coefficients and variables and a noise parameter W multiplied with a scaling factor σ. In this case, W has a standard logistic distribution. The resulting model is given below:

$$\text{Log } T = \mu + y^t x + \sigma W = 2.48 - 0.69 * x_{\text{small cell}} - 0.78 * x_{\text{adeno}} - 0.03 * x_{\text{large}} + 0.04 * x_{\text{karnofsky}} + 0.58 * W$$

The variables of the cancer cell types are dummy coded and only have a value of 1 when the patient has that particular type of cancer. When the cancer cell type is squamous, all of the dummy variables related to the categorical variable are set to 0 and only the intercept determines the survival time. All three cancer cell types reduce the survival time to a greater extent than the squamous cell carcinoma reference category. Once again, lung adenocarcinoma causes the greatest reduction in the survival time considering that its coefficient has the largest negative value. The coefficient of small cell lung cancer is a close second and thus illustrates that this cancer cell type is the second most deadly disease present in the data set. Large cell lung cancer affects the survival of the veterans involved in the study in a similar fashion as the reference category given that the coefficient lies very close to zero. Note that terms regarding the Karnofsky score are transparent. They will not be discussed since the task only specifies to discuss the coefficients of the categorical variable.

Written in full the general AFT model for the log-logistic survival $S_i(t)$ from the methodology section is obtained by plugging in $exp(\theta^T x_i)t$ in the baseline survival function $S_0(t)$ with log-logistic distribution:

$$S_i(t) = S_0(exp(\theta^T x_i)t = \frac{1}{1 + (exp(\theta^T x_i)t\ \lambda)^\kappa}$$

Analogous to the linear representation, the linear combination of the coefficients and variables can be filled in the survival function. Every term of the model is known after applying the following two conversions:

- $\kappa = 1/\sigma = 1.72$
- $\lambda = \exp(-\mu/\sigma) = 0.014$

Besides having an opposite sign, the coefficients are identical and the interpretation would lead to the same conclusions. Therefore, we will not go into detail to avoid repetition.

Extra analyses:

Table 13: Pairwise analysis

	Squamous	**Smallcell**	**Adeno**
Smallcell	0.00134	-	-
Adeno	0.00134	0.75565	-
Large	0.43731	0.00331	0.00016

P value adjustment method: BH

This short extra analysis was included as formal proof for clustering, the p-values suggest that squamous and large cell carcinoma were not significantly different cell types with regard to their effect on the survival time of veterans (Table 13). They can thus be grouped into a single cluster (Table 14). The same applies to small cell -and adenocarcinoma (Table 13), which are therefore assigned to a second cluster (Table 14).

Table 14: Clustering of the cell types

Level	Cluster
Squamous	1
Smallcell	2
Adeno	2
Large	1

Conclusion

The initial primary goal in this report was to assess if the test chemotherapy treatment was beneficial or not compared to the standard treatment. The treatment type did not have a significant impact on the survival time of the veterans in a non-parametric setting. We further explored the implementation of a semi-parametric Cox proportional hazard model and a series of parametric survival models. Multiple covariates were included in the model, but only the type of lung cancer and Karnofsky score were shown to partially explain the survival time and ended up in the final reduced model after a variable selection procedure. This finding led us to shift the focus of our study towards the effects of different cell types on survival time. The log-logistic model was selected as the best-suited parametric regression model among the candidate models in our analysis according to the AIC. Both the semi-parametric and parametric method had similar results, but the conclusions obtained from the parametric model were perceived as more reliable since the proportional hazard assumption was not validated for the semi-parametric alternative. The effect of large cell lung carcinoma on survival time was quite similar to that of squamous cell carcinoma (which was the reference type). The largest reductions in survival time in veteran lung cancer patients were caused by adenocarcinoma, with small cell carcinoma as a close second with regard to its effect on mortality. Unsurprisingly, a higher Karnofsky score resulted in a lower risk of mortality, which is in line with what we expected because a higher score signifies a lower degree of physical impairment.