Introduction

A common treatment for HIV is the use of an antiretroviral medication called Zidovudine (AZT) which is blocking the activity of the virus. It is nowadays mostly used in combination with other treatments to improve the efficacy. However, it can happen that the AZT fails as a treatment or a patient is intolerant to the medication.

This analysis is based on a study that observed patients, whose AZT therapy did not succeed. Alternatively, they were treated with Zalcitabine (ddC) or Didanosine (ddl) in a randomized clinical trial that collected data to compare the efficacy and safety of the new treatments.

A point of interest in this study was the behavior of the new treatment with respect to the overall condition of a patient immune system.

An indicator for the strength of a person's immune system is the number of CD4 cells, or T-cells, in a cubic millimeter of blood. These are white blood cells that fight infections and are important for the human immune system, which means that a higher count is implying a stronger immune system. A healthy person without HIV has a CD4 count of between 500 and 1500. The HIV virus is targeting healthy CD4 cells and becomes part of it to spread in the body. This means that an HIV-infected person has fewer working CD4 cells. With a count above 500, an infected person is usually in good health, but from a count of 200 and less, there is a high risk of developing a serious disease. [1]
In both cases, a CD4 cell count of below 200 or the development of an opportunistic disease, one is diagnosed with HIV stage III, also known as AIDS. It is the condition in which a patient's immune system is heavily damaged by the HIV infection without proper treatment or an opportunistic infection. [2]

In the following report, the different variables will be analyzed, with their influence on the survival function of the patients in the study. We will evaluate the impact the different measures have on the health condition of the patients to conclude which are the most crucial aspects to ensure the longest survival time.

Descriptive Analysis

Univariate Analysis

A total of eight variables were reported for each of the 467 patients, namely patient identifier, survival time, death indicator, CD4 cell count, treatment, gender, previous opportunistic infection, and AZT.

The **survival times** are measured in years and range from a minimum of 0,47 years to a

maximum of 21.4 years. The mean of the documented survival time is 12.63 years and the median is 13.2 years, which indicates that the data is left-skewed. Performing a Shapiro-Wilk test for normality confirms that the data is not normally distributed with a significance level of 99%.

The variable **death** indicator is a binary variable, 1 indicating death and 0 indicating that the survival time was censored. The mean of this variable is 0.4026, which means that 59.74% of the survival times are censored.

The **CD4 cell count** ranges from a minimum of 0 to a maximum of 327 with a mean of 72.97 and a median of 37. The data is strongly right-skewed as can be seen from a histogram of the variable. The shape reminds of an exponential distribution, looking at a QQ-plot with exponential theoretical quantiles, however, does not seem to confirm that which might be due to the number of outliers. The variable contains 30 outliers, which are the CD4 counts above a value of 263. Since these outliers, do not fall out of the medically possible range, no outliers are removed.

The **treatment** variable is a binary variable that indicates which medication a patient was treated with either the drug Zalcitabine (ddC) or Didanosine (ddl). The two treatments are almost equally distributed amongst the observed patients with a count of 230 for the ddl treatment and 337 for the ddC treatment.

Moreover, the patients of the study were 90.36% male and 9.64% female (**gender**). Of the observed group, 65.74% suffered from previous opportunistic infections, which is indicated with the variable **AIDS/ noAIDS**. Lastly, there is a variable that shows if patients had failed or were intolerant of zidovudine therapy.(**AZT**) 37.47% of the patients had a previous failure to the zidovudine therapy and the remaining 62.53% had an intolerance to the therapy.

Correlation of variables

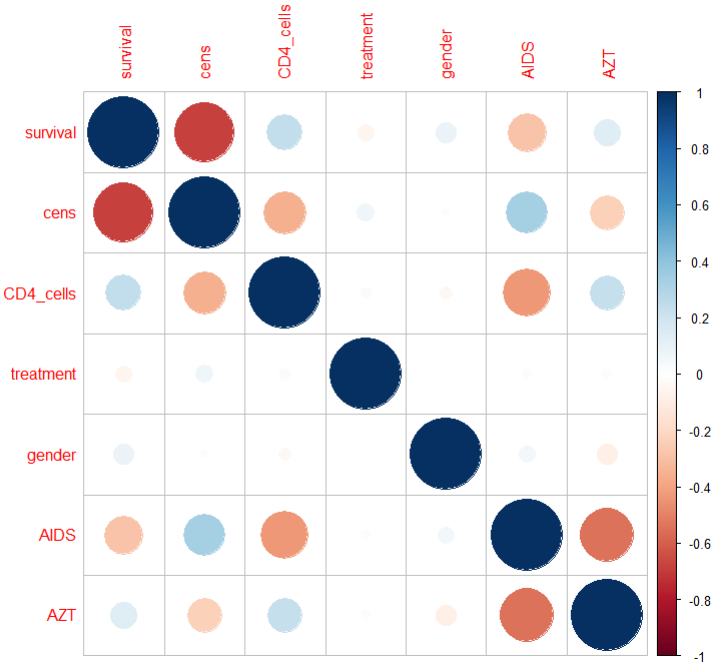

Figure 1: Correlations of the variables

As Figure 1 shows, the strongest correlation of -0.68 can be found between the variables' survival time and the censoring indicator. This is intuitive since a positive indicator value, the case of death shortens the survival time of a patient. Similarly, the AIDS variable is negatively correlated with the survival time with a correlation of -0.29.
Moreover, the survival time is positively correlated with CD4 cell count and the AZT variable with a correlation value of 0.25, and 0.14 respectively. The correlations with the other two variables are not significant.
Both variables positively correlated with survival time, are negatively correlated with the censoring variable with a correlation of -0.36 for the CD4 cell count, and -0.24 for AZT. As one would expect, the correlation between AIDS and censoring is positive with a value of 0.33. Variable AZT and AIDS are strongly negatively correlated at -0.55.
All the correlations mentioned have a significance level associated with a p-value of < 0.01. The variables gender and treatment do not show any significant correlations with other variables.

Non-parametric Analysis

The overall survival function of the study can be found in Figure 2 (left). As can be seen, the survival function seems to decrease at a constant rate from 0 to about 20 years. The median survival probability is reached at 19.07 years. At the end of the study, after 21.4 years the survival probability has reached a value of 0.49.

Further, the survival functions of different groups are looked at to detect discrepancies of the survival function with respect to the observed variables.

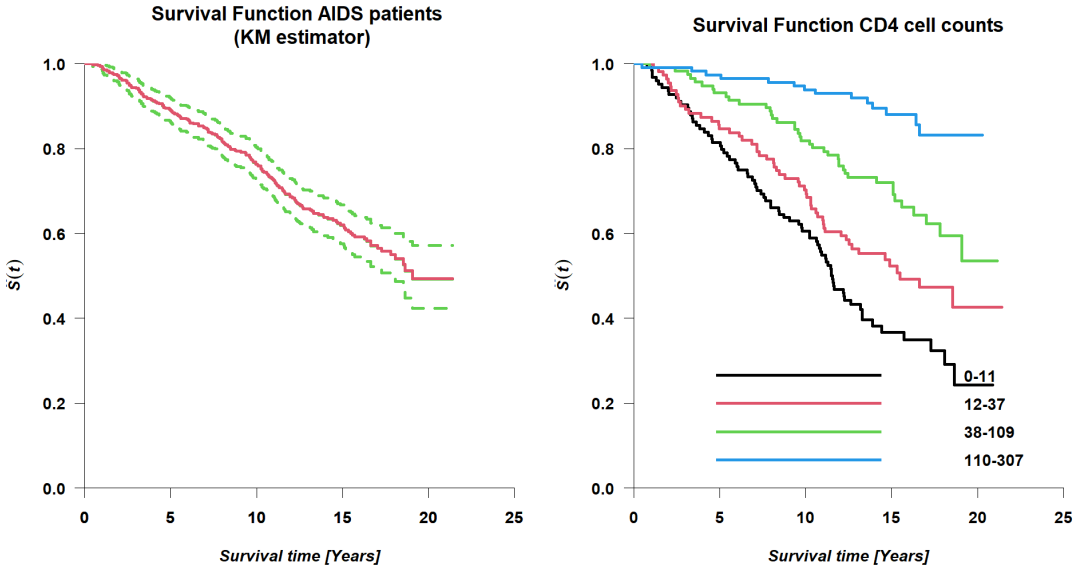

Figure 2: Overall survival function (left) and Survival function of CD4 cell count (right)

Gender

When comparing the survival function of male and female patients, it can be seen that the survival function for females seems to decrease more rapidly in the first 5 years after the start of the study. The survival probability at the end of the study reaches 0.40 women, it reaches 0.50 for men. Towards the end of the observation time, the function shows huge jumps and is overall less smooth. This can be explained by the fact that the number of female patients was much less, so there is not enough data for females to make a fair comparison between the groups. Using the log-rank test to check the hypothesis that both functions are the same, with Chi-sq= 0.5 on 1 degree of freedom, a p-value of 0.5 is obtained. Therefore, one can conclude that there is no evidence that the survival times for male and female patients are different.

Treatment

For both treatment groups the survival function behaves similarly for the first 5 years, then they move apart a little, with ddc treatment having a higher survival probability until the end of the study when both functions reach approximately the same level. The log-rank test, used on both groups shows the same result, with Chi-sq= 2.1 on 1 degree of freedom, we receive p= 0.2. Therefore, there is no significant difference to be found in the behavior of both survival functions.

AIDS

The survival functions of AIDS vs. noAIDS show very different behavior. As one would expect, the survival function of people of the group noAIDS shows a vastly better survival rate, from the start of the study to the end. The group having opportunistic infections has a fast decreasing survival function which drops to a probability of 0.33 whereas the other group is at 0.80. This creates a difference of more than 6 years in the restricted mean of both groups. (AIDS. 13.99, noAIDS=19.10). The log-rank test on both groups, confirms that there is a difference in both survival functions. Using Chi-sq= 51.7 on 1 degree of freedom the p-value is < 0.001, rejecting the hypothesis that both functions behave the same.

CD4 cells

As the count of CD4 cells ranges from 0 to 370, the data is binned into 4 categories according to its quantiles. The first category includes counts from 0 to 11, the second from 12 to 37, the third one from 38 to 109, and the last one from 110 to 370. The resulting survival plot for the categories is very insightful, it is shown in Figure 2. It shows clearly that with an increased cell count, the survival function improves drastically. For the two upper categories and the 2 lower categories, the functions are similar, for approximately the first 3 years, then they move further apart. The category with the highest count in CD4 cells is also clearly the one with the best survival function, as it has a survival probability of 0.83 after the end of the study whereas the second-highest category has already a very decreased probability of survival at the end of the study with 0.54. The second-lowest and the lowest cell count category have a survival probability of 0.43 and 0.24 respectively at the end of the study. When performing a log-rank test on the data, with Chi-sq= 76.4 on 3 degrees of freedom, a p-value of < 0.001 is obtained, which confirms that the survival functions are not the same.

AZT

The two survival functions of the reactions to the AZT treatment are shown to be quite different with a restricted mean for the failure group of 13.82 years and 16.99 years for the intolerance group. After 21 years of the study, the failure group has reached a survival probability of 0.31 whereas the intolerance group has a survival probability of 0.62. Testing the similarity of both functions with a log-rank test gives a p-value of < 0.001 with Chi-sq= 24.4 on 1 degree of freedom. It can be concluded, that there is a significant difference between both survival functions.

	Chisq	degrees of freedom	p-value
Gender	0.5	1	0.5
Treatment	2.1	1	0.2
Aids	51.7	1	<0.001
CD4 cells	76.4	3	<0.001
AZT	24.4	1	<0.001

Table 1: Log-rank test of explanatory variables

Parametric survival model

Weibull Model

Starting with one of the most promising variables from the previous section, the covariate CD4 cells are fitted with a Weibull model. With this model, the impact of the variable on the survival function is checked using the following hypothesis. The null hypothesis states that the variable does not influence the survival function, the alternative hypothesis states that it does. The resulting p-value of < 0.01 confirms that there is a significant influence of the cell count on the survival function.
Similarly, the covariate AIDS and AZT are separately fitted with the Weibull model. Both of them show a p-value of < 0.01, so one can conclude that both of them impact the survival function significantly. Also, the other two variables gender and treatment are fitted to a Weibull model, but show not to influence the survival function, as the null hypothesis cannot be rejected.

Lastly, a full model, with all the available variables is used to fit a Weibull model. As the model contains the significant variables, also the full model rejects the null hypothesis. However, when the variables are looked at in more detail using the analysis of the variance, not all of them show the same impact on the model. From the five variables, only AIDS and the CD4 count have significant levels of < 0.01, so they seem to be the variables explaining the survival function best.

Logistic Model

As an alternative to the Weibull model, the data is fitted to a log-logistic model to check its fit. At first, all of the five variables are fitted separately. Similarly, as before, it becomes apparent that only AZT, AIDS, and CD4 count are significant as a single covariate for the model. When testing a full model, the same three variables are shown to be significant with a p-value of < 0.01, gender and treatment considered significant as well, but with a level of < 0.1.

Log-Normal Model

Thirdly, the data is fitted with a log-normal model to investigate the relationship between the explanatory and target variable. As with the log-logistic models, the same three variables (AZT, AIDS, and CD4 count) are significant as single covariates. In a full model, those three variables are significant with a level of 0.01, and the variable gender is significant with a p-value of 0.1

Model Fit

To check which of the three models is the most appropriate to explain the data, the residuals are looked at. Figure 3 shows the residuals of each fitted model in comparison with the expected theoretical distribution. The residuals of the log-logistic model do not match the distribution that would be expected. Also, the residual distribution of the lognormal model differs from the empirical distribution as moves outside of the 95%-confidence intervals of the distribution. The residuals of the Weibull distribution seem to match the theoretical distribution very well. It becomes clear, that the Weibull model is the best fit for the data.

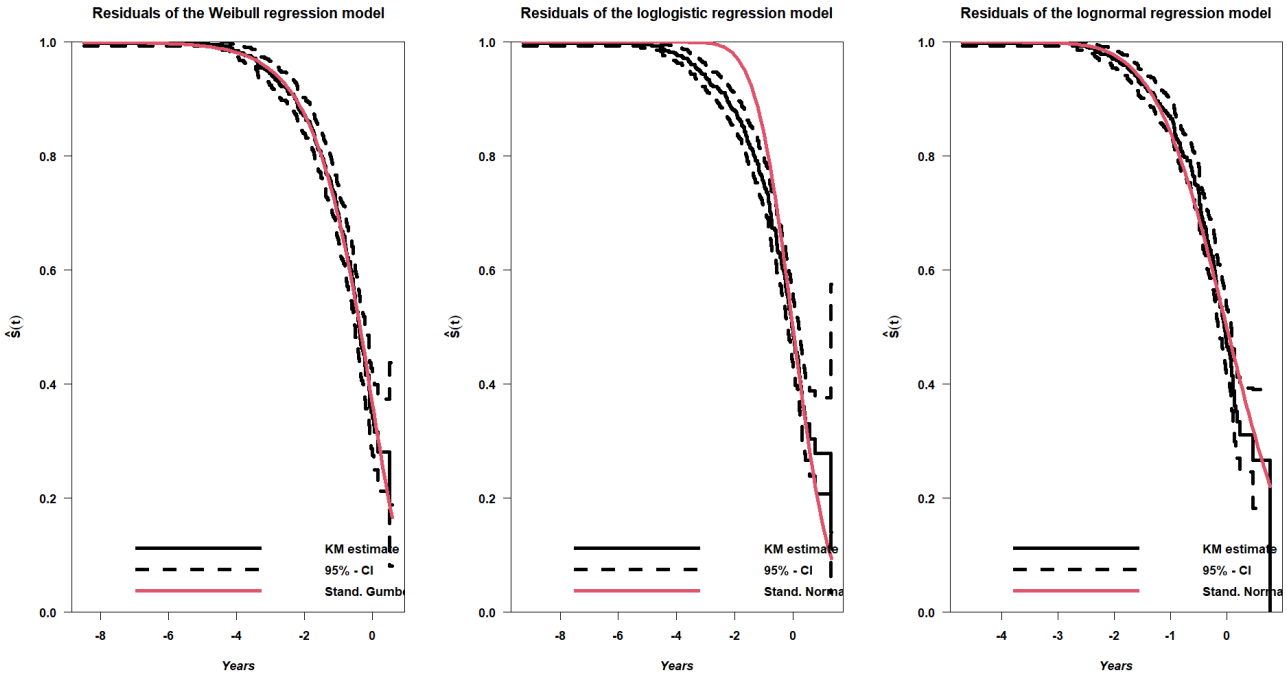

Figure 3: Residuals of models

Acceleration Factor and Relative Odds

As the Weibull model with the variables AIDS and CD4 count was chosen to explain the survival function the best, it is now used to investigate the impact of the two variables in

more detail. The acceleration factor of the variable noAIDS is 0.53. The acceleration factor indicates to which degree a covariate accelerates or decelerates the life span. An acceleration factor <1 is an indication that the covariate is extending the survival time, hence, it can be said that having no prior opportunistic infections prolongs the survival time of a patient. The odds ratio of noAIDS is 0.39, which means that is a patient without AIDShas only a 40% chance of dying compared to a person with AIDS.

For the acceleration factor of the variable CD4 count, the higher cell count categories are compared to the lowest cell count category. The second-lowest gives an acceleration factor of 0.74, the third-lowest one of 0.51, and the highest category of CD4 cell count is 0.23. Compared to the lowest category, any increase in cell count extends the survival time of a patient. The higher the cell count, the longer the extension of the survival time. The odds ratio confirms this statement. The categories give an odds ratio of 0.65, 0.38, and 0.13, from the second-lowest to the highest category. With a cell count of above 109, a patient only has a 12% chance of dying compared to someone with a cell count of below 11.

Semi-parametric survival model

1. A well-known example of a semi-parametric model is the Cox proportional hazards model. If we are interested in studying the time T to an event such as death due to cancer or failure of a light bulb, the Cox model specifies the following distribution function for T:

$$F(t) = 1 - \exp\left(-\int_0^t \lambda_0(u)e^{\beta x}du\right) \tag{1}$$

where x is the covariate vector, and β and $\lambda_0(u)$ are unknown parameters. $\theta = (\beta, \lambda_0(u))$. Here β is finite-dimensional and is of interest; $\lambda_0(u)$ is an unknown non-negative function of time (known as the baseline hazard function) and is often a nuisance parameter. The set of possible candidates for $\lambda_0(u)$ is infinite-dimensional.

```
coxph(formula = svfc ~ CD4_cells + treatment + gender + AIDS +
    AZT, data = aids)

  n= 467, number of events= 188

                   coef exp(coef)  se(coef)      z Pr(>|z|)
CD4_cells      -0.008669  0.991368  0.001659 -5.227 1.72e-07 ***
treatmentddI    0.241020  1.272547  0.146617  1.644 0.100202
gendermale     -0.301401  0.739781  0.243971 -1.235 0.216682
AIDSnoAIDS     -0.802456  0.448227  0.234814 -3.417 0.000632 ***
AZTintolerance -0.178681  0.836372  0.162039 -1.103 0.270156
---
```

```
Signif. codes:  0     ***     0.001     **     0.01     *     0.05

               exp(coef) exp(-coef) lower .95 upper .95
CD4_cells         0.9914     1.0087     0.9882     0.9946
treatmentddI      1.2725     0.7858     0.9547     1.6962
gendermale        0.7398     1.3518     0.4586     1.1934
AIDSnoAIDS        0.4482     2.2310     0.2829     0.7102
AZTintolerance    0.8364     1.1956     0.6088     1.1490

Concordance= 0.712  (se = 0.019 )
Likelihood ratio test= 102.9  on 5 df,   p=<2e-16
Wald test            = 68.5  on 5 df,   p=2e-13
Score (logrank) test = 81.73  on 5 df,   p=4e-16
```

2. We know the higher the concordance, we have a better fit in our model with a value of over 0.71, close to 1(the perfect fit), we have a good fit. We observe the two variables CD4 cell count and the AIDS or no AIDS indicator to be statistically significant in our model, unlike other variables (namely treatment, gender, and AZT) that are not statistically significant in our model. Finally, we also see all three tests have lower p values close to zero, which proves the global statistical significance of the model.

3. The relative hazard or hazard ratio function is an exponential function of Z:

$$HR(t|Z) = \frac{\lambda(t|Z)}{\lambda_0(t)} = \exp\{\beta_1 Z_1 + \cdots + \beta_p Z_p\}. \tag{2}$$

And the logarithm of HR is the linear function of x:

$$= \ln \frac{\lambda(t|Z)}{\lambda_0(t)} = \beta_1 Z_1 + \ + \beta_p Z_p \tag{3}$$

We have estimated a Cox proportional hazards regression model and related an indicator of each variable to time to death.

	logHR	HR	Lower 95%	Upper 95%
HR — CD4_cells	-0.009	0.991	0.988	0.995
HR — treatment ddI	0.241	1.273	0.955	1.696
HR — gender male	-0.301	0.740	0.459	1.193
HR — noAIDS	-0.802	0.448	0.283	0.710
HR — AZT intolerance	-0.179	0.836	0.609	1.149

We have different variables and their effect on the expected hazard function in the above table. For interpretability, we compute hazard ratios by exponentiating the parameter

estimates. For CD4_cells, $\exp(-0.009) = 0.991$. There is an $0.991 - 1 = -0.009 \Rightarrow$ 0.9% decrease in the expected hazard relative to a one-count increase in CD4_cells (or the expected hazard is 0.99 times less in a person who is one count higher in CD4_cells than another), holding AIDS factor constant. Similarly, exp(-0.802) = 0.448. The expected hazard is 0.55% higher in the case of noAIDS as compared to the case of AIDS, holding CD4_cells constant. This is extended to other variables as well.

Since All of the parameter estimates are estimated taking the other predictors into account. After accounting for CD4_cells and AIDS diagnosis there are no other statistically significant associations between variables and survival. This does not say that these risk factors are not associated with survival. Their lack of significance is likely due to interactions between variables (interrelationships among the risk factors considered). Notice that for the statistically significant risk factors (i.e., CD4_cells and AIDS variable), the 95% CIs of the hazard ratios do not contain 1 (the null value). In contrast, the 95% CIs for the non-significant risk factors (AZT, gender, and treatment type) include the null value.

4. The Cox proportional hazards model makes several assumptions. Thus, it is necessary to assess whether a fitted Cox regression model adequately describes the data. Here, we'll discuss three types of diagnostics for the Cox model: Detecting non-linearity in the relationship between the log-hazard and the covariates. Examining the influential observations (or outliers). Testing the proportional hazards assumption. For checking these model assumptions, the usual residual methods of the Cox model include: Martingale residuals to assess non-linearity Deviance residual (symmetric transformation of the Martingale residuals), to examine influential observations. Schoenfeld residuals to check the proportional hazards assumption.

   The proportional hazards (PH) assumption is verified using statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals. For each covariate included in a Cox regression model fit, the function cox.zph() is a convenient solution to test the proportional hazards assumption. From the output of the function, we observe that the test produces high p values for each variable, and the global test is also not statistically significant, and we cannot reject our null hypothesis of the proportional hazards assumption. Therefore our proportional hazards assumption is reasonable. Also, we plot the results of the residuals to see if the line of the mean is shifting approximately along with zero line in each variable, and we see that they are, for the most part, averaging around zero there is no pattern with time.

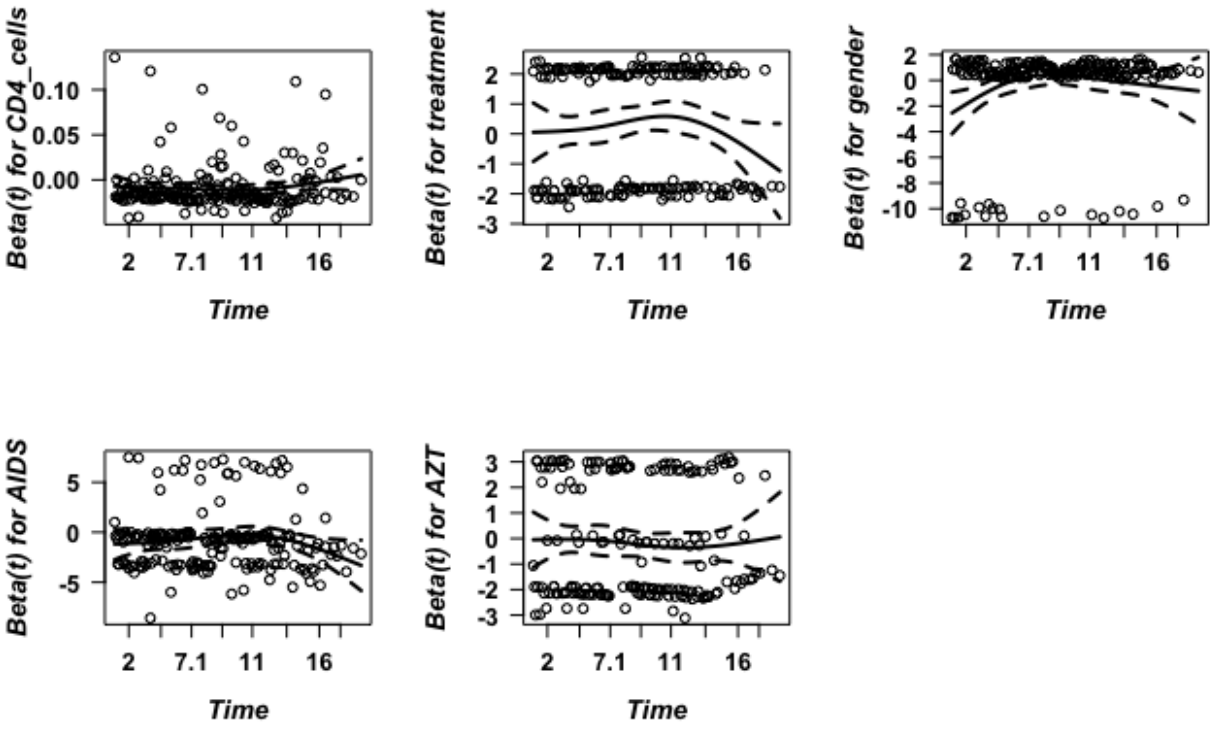

Figure 4: Schoenfeld Residuals

To test influential observations or outliers, we use the dfbeta values. By comparing the coefficients of the variables to dfbeta values, we conclude that none of them individually are very influential in our model.

Plotting the Martingale residuals against continuous covariates is a common approach used to detect non-linearity or, in other words, to assess the functional form of a covariate. For a given continuous covariate, patterns in the plot may suggest that the variable is not properly fit. Non-linearity is not an issue for categorical variables, so we only examine plots of martingale residuals and partial residuals against a continuous variable. Martingale residuals may present any value in the range (,1)(,1). A value of martingale residuals near 1 represents individuals that died too soon, and large negative values correspond to individuals that lived too long. We deduce by plotting this residual that the mean is zero, thus the linearity assumption is acceptable.

Conclusion

In conclusion, it can be said that mainly two variables can explain the behavior of the survival function of the patients, namely CD4 cell count and the AIDS indicator. There are several observations in this analysis that lead to this conclusion. In the explanatory analysis, those were the variables that showed the strongest correlation with survival and censoring. According to the log-rank test, one can with certainty say that different cell counts show very differently behaving survival functions. The same holds for the AIDS variable.
Using a Weibull model seems the best way to describe the given data compared to the other models tested. The acceleration factor and ratio odds put the previous statements into numbers by comparing the groups of CD4 cell counts and the AIDS indicator.

The semi-parametric analysis strengthens these findings. The two variables, CD4 cells, and AIDS indicators have low p-values in the models. The model fit is acceptable, and we can deduce how much it will change the hazard rate by increasing one unit in the variable and keeping every other variable constant. Finally, we diagnose the model using residuals for non-linearity, outliers, and whether it holds the cox proportionality assumption. After diagnoses, it can be concluded, the model can be used with the available data and that it achieves to explain the behavior of the survival function.

These different tools of analysis combined lead to the conclusion that with HIV patients it is crucial to look at previous opportunistic infections, and the CD4 cell count as they have shown to be a good indicator for the disease progression.

References

[1] R. Pebody, "Cd4 cell counts — aidsmap," 2021, may 2021. [Online]. Available: https://www.aidsmap.com/about-hiv/cd4-cell-counts

[2] U. D. of Health Human Services, "Opportunistic infections — living with hiv — hiv basics — hiv/aids — cdc," 2021, may 20, 2021. [Online]. Available: https://www.cdc.gov/hiv/basics/livingwithhiv/opportunisticinfections.html