R for Statistical Data Science Autumn 2024

Survival Analysis in HIV-Positive Patients: A Study of Treatment Regimens and Contributing Factors

Analysing the Role of Clinical and Biological Factors in Survival Outcomes

Rory O'Sullivan

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

MSc Data Science and Statistical Learning
Department of Mathematics and Statistics

December 13, 2024

Abstract: This study evaluates survival outcomes in 2,133 HIV-positive patients across different treatment regimens, considering a range of factors including CD4/CD8 cell counts, off-treatment status, symptomatic status, and other relevant variables. Kaplan-Meier survival curves reveal significant survival benefits for combination therapies and `ddI` compared to `ZDV` alone ($p < 0.0001$), while treatment discontinuation before 96 ± 5 weeks is associated with poorer outcomes. Cox Proportional Hazards models identify CD4 counts, symptoms, and off-treatment status as key predictors of survival. Despite violations of the proportional hazards assumption, this analysis provides valuable insights into survival trends and emphasises the need for methodological refinements in future statistical modelling.

Contents

1. Introduction

In this project, we analyse the "AIDS Clinical Trials Group Study 175" dataset from Kaggle [1]. This dataset originates from a clinical trial conducted in 1996 to evaluate the efficacy of different nucleoside treatments in patients with HIV-1. The trial included adults with baseline CD4 cell counts between 200 to 500 per cubic millimetre [2].

The analysis presented in this report focuses on understanding survival trends, treatment efficacy and the influence of various baseline and follow-up factors. Additionally, we evaluate the performance of statistical models, assess the proportional hazards assumption and explore areas for methodological improvement. This project aims to provide an introductory exploration of survival analysis techniques applied to clinical trial data.

2. Methods

This is a randomised, double-blind, placebo-controlled trial study. The aim is to evaluate treatment with either a single nucleoside or two nucleosides in adults infected with Human Immunodeficiency Virus type 1 (HIV-1) whose CD4 cell counts were between 200 to 500 per cubic millimetre. The primary end point of the study was defined as a >50 percent decline in the CD4 cell count, development of AIDS, or death. Patients were randomised into 4 treatment groups, Zidovudine (ZDV) 600mg, Zidovudine 600mg plus didanosine (ddI) 400mg, zidovudine 600mg plus zalcitabine (Zal) 2.25mg, didanosine 400mg.

The `survival` package in R was used to analyse the dataset. The primary aim of this project is to identify effects of treatment on survival, alongside relationships between other factors and survival.

2.1 Kaplan-Meier Survival Curves

Kaplan-Meier curves are utilised to visualise survival distributions between groups. These curves are very useful and widely applicable as they are nonparametric. Kaplan-Meier curves make use of the survival function

$$S(t) = Pr(T > t). \tag{1}$$

This function quantifies the probability of surviving past time t. The Kaplan-Meir estimator of the survival curve takes the form

$$\hat{S}(d_k) = \prod_{j=1}^{k} \left(\frac{r_j - q_j}{r_j} \right). \tag{2}$$

In this study, d_k denotes the k^{th} unique death time among non-censored patients, q_k denotes the number of patients that died at time d_k and r_k denotes the number of patients alive in the study just before d_k, i.e, the patients at risk. As there will naturally be time jumps between d_k and d_{k+1}, the curve will have a step-like shape [3].

2.2 Log-Rank Test

We use the **Log-Rank Test** to compare the survival of different groups. The log-rank test is used to compare the "equality" of two or more Kaplan-Meier survival curves. In this paper, we will consider explaining this test for just the case for two classes. For us to test $H_0 : E(X) = 0$, for some random variable X, we consider a test statistic

$$W = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}$$

where $E(X)$ and $Var(X)$ are the expectation and variance, respectively of X under H_0. Applying this to our survival analysis, the test statistic W takes the form

$$W = \frac{\sum_{k=1}^{K}(q_{1k} - \mathrm{E}(q_{1k}))}{\sqrt{\sum_{k=1}^{K} \mathrm{Var}(q_{1k})}} \tag{3}$$

where q_{1k} is the total number of patients in group 1 who died at time d_k and

$$E(q_{1k}) = \frac{r_{1k}}{r_k} q_k$$

where q_k denotes the total number of patients in both groups who died at d_k, r_k is the total number of patients at risk at time d_k and r_{1k} is the total number of patients in group 1 at risk at time d_k. Intuitively, the test statistic is simply being formed as the observed number of patients who died in group 1 vs the number expected, all over $\mathrm{Var}(X)$. If $\mathrm{E}(X)$ is much different to the observed value then the test statistic will yield a significant p-value and thus the groups are different. The log-rank test will play a pivotal role in deciding which variables to use for modelling later. Initially, variables will be individually ran with the log rank test and their significance will be noted [3].

2.3 Cox's Proportional Hazards Model

In the modelling part of this project, we will use Cox's Proportional Hazards Model. First, the **hazard function** is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{4}$$

This can be intuitively understood as the death rate instantly after time t, given survival past that time. The **proportional hazards assumption** states that

$$h(t|x_i) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right), \tag{5}$$

where $h_0(t)$ is known as the baseline hazard. This is the hazard for an individual with all features $x_{i1} = \cdots = x_{ip} = 0$. The equation gets the "proportional" part from the fact that the hazard function for a patient with feature vector x_i is the unknown baseline hazard $h_0(t)$ times the factor $\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$. The latter is known as the relative risk for feature vector x_i.

Since $h_0(t)$ is unknown, direct estimation of $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ using a full likelihood function is not feasible. Instead, Cox's model employs a **partial likelihood** approach, which relies on comparing the hazard for an individual at the event time y_i to the total hazard among all individuals still at risk at that time. This approach effectively cancels

out $h_0(t)$ in the likelihood, allowing for the estimation of β without specifying the baseline hazard. In terms of the linear predictor, the model can be written as:

$$\log h(t \mid x_i) = \log h_0(t) + \sum_{j=1}^{p} x_{ij}\beta_j, \tag{6}$$

where $\log h(t|x_i)$ represents the log hazard, expressed as the sum of the log baseline hazard $\log h_0(t)$ and a linear combination of the covariates x_{ij} weighted by their respective coefficients β_j.

A critical assumption of the Cox model is the **proportional hazards assumption**. This assumes that the effect of a covariate on the hazard remains constant over time and does not vary with t. Violations of the proportional hazards assumption may lead to biased estimates and invalid inferences, requiring further investigation or adjustments to the model, such as time-dependent covariates. We may examine if this assumption holds by examining the models' Schoenfeld residuals [4].

2.4 Model Evaluation Techniques

Previously, we introduced the Log-Rank Test to evaluate predictors individually against the target variable. This approach effectively shortlists features for modelling. For continuous variables, we fit a Cox model with the feature of interest and assess its significance using the Score test, which aligns with the log-rank test. Here, we introduce two additional methods for feature selection and model comparison: the **Likelihood-Ratio Test** and the **Akaike Information Criterion**.

2.4.1 Likelihood-Ratio Test

The Likelihood-Ratio Test is used to compare nested models by evaluating their log-likelihoods. The likelihood ratio statistic is defined as:

$$2\left(l(\hat{\beta}_{\text{full}}) - l(\hat{\beta}_{\text{reduced}})\right), \tag{7}$$

where $l(\hat{\beta}_{\text{full}})$ and $l(\hat{\beta}_{\text{reduced}})$ represent the log-likelihoods of the full and reduced models, respectively. This statistic follows a χ^2 distribution, with degrees of freedom equal to the difference in the number of parameters between the models. A significant result indicates that the full model provides a better fit than the reduced model. This test can be implemented via an ANOVA comparison between the models in R [4].

2.4.2 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a statistical measure that balances model fit with complexity. We use it to compare models that aren't necessarily nested. It is defined as:

$$\text{AIC} = -2\log(L(\hat{\theta}|y)) + 2K,$$

where $\hat{\theta}$ are the estimated model parameters, y is the observed data, and K is the number of parameters in the model.

The AIC operates on an interval scale, meaning comparisons are based on absolute differences rather than percentages. Lower AIC values indicate a better balance between goodness of fit and parsimony, making the model more suitable for selection [5].

2.4.3 Harrell's Concordance Index

Harrell's Concordance Index, or the C-index, is an adaptation of the Area Under the Curve (AUC) for survival data. It evaluates a model's ability to correctly rank predicted risks. We define an estimated risk score for patient i as $\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$, for $i = 1, \ldots, n$, using the Cox model coefficients. If $\hat{\eta}_{i'} > \hat{\eta}_i$, the Cox model predicts that the i'th observation has a greater hazard risk than the ith observation, implying that the survival time t_i will be longer than $t_{i'}$.

In survival analysis, due to censoring, we observe the possibly censored survival times $y_1, \ldots, y_n$ along with censoring indicators $\delta_1, \ldots, \delta_n$, where $\delta_i = 1$ indicates that the event was observed (uncensored) and $\delta_i = 0$ indicates that it was censored. The C-index is calculated as the proportion of comparable observation pairs for which the model correctly ranks the risk scores. It is given by:

$$C = \frac{\sum_{i,i':y_i>y_{i'}} I(\hat{\eta}_{i'} > \hat{\eta}_i)\delta_{i'}}{\sum_{i,i':y_i>y_{i'}} \delta_{i'}},$$

where $I(\hat{\eta}_{i'} > \hat{\eta}_i)$ is an indicator function equal to 1 if $\hat{\eta}_{i'} > \hat{\eta}_i$, and 0 otherwise. The numerator counts the number of concordant pairs—those where the model correctly predicts that i survives longer than i' given that the event occurred for i'. The denominator represents the total number of comparable pairs in the data. In essence, the C-index measures the proportion of observation pairs for which the model's predicted risk scores correctly reflect the observed survival time ordering [3].

3. Exploratory Data Analysis

The full dataset contains 2139 samples and 24 features. Six patients[1] were excluded due to missing values and outliers. Table 1 provides a summary of all variables.

The treatment groups are balanced, each with similar sample sizes. The majority of patients are male (82.8%) and white (71.2%). Nearly a quarter of patients (24.4%) reached

[1]Patients 217, 218, and 724 were excluded due to having a CD40 cell count of 0 (missing). Patient 2125 was excluded due to extremely high CD820 levels (over 6000), which were much higher than the mean of 935. Patient 1145 had an exceptionally high CD40 value (over 1199), significantly higher than the next highest value of 918. Patient 601's CD80 level was 40, with a corresponding CD40 value of 400. This is possibly due to imputed values where a '0' was missed. Additionally, the next lowest CD80 value is 105, representing a large jump from 40, which further justifies the removal of this patient.

Variable	Description	Summary*(N=2133)
time	Time to event (days)	878.73 ± 292.45
label	Censoring Indicator (0=censoring, 1=failure)	521 (24.4%)
age	Age in years at baseline	35.24 ± 8.70
wtkg	Weight in kilograms at baseline	75.14 ± 13.27
hemo	Hemophilia (0=no, 1=yes)	179 (8.4%)
homo	Homosexual activity (0=no, 1=yes)	1410 (66.1%)
drugs	History of IV drug use (0=no, 1=yes)	280 (13.1%)
karnof	Karnofsky score (70, 80, 90, 100)	(9, 80, 784, 1260)
oprior	Non-ZDV antiretroviral therapy prior to study (0=no, 1=yes)	47 (2.2%)
z30	ZDV use 30 days prior to treatment (0=no, 1=yes)	1173 (55.0%)
zprior	ZDV use prior to treatment (0=no, 1=yes)	2133 (100%)
preanti	Number of days of pre-study antiretroviral therapy	377.95 ± 466.74
race	Race (0=white, 1=non-white)	614 (28.8%)
gender	Gender (0=female, 1=male)	1766 (82.8%)
str2	Antiretroviral history (0=naive, 1=experienced)	1249 (58.6%)
strat	Antiretroviral history stratification	(884, 409, 840)
symptom	Symptomatic indicator (0=asymptomatic, 1=symptomatic)	370 (17.3%)
trt	Treatment indicator (0=ZDV only, 1=ZDV+ddl, 2=ZDV+Zal, 3=ddl only)	(532, 517, 524, 560)
treat	Binary treatment indicator (0=ZDV only, 1=others)	1601 (75.1%)
offtrt	Off treatment before 96±5 weeks (0=no, 1=yes)	773 (36.2%)
cd40	CD4 cell count at baseline	350.48 ± 116.49
cd420	CD4 cell count at 20±5 weeks	370.80 ± 144.41
cd80	CD8 cell count at baseline	985.60 ± 472.21
cd820	CD8 cell count at 20±5 weeks	933.09 ± 431.06

Table 1: Variable Descriptions and Updated Summaries.
* Continuous variables display mean ± standard deviation.
Binary variables display count for positive class along with percentage.
Categorical variables display count for each category.

the primary endpoint. Over a third of patients (36.2%) were off treatment before 96±5 weeks. Baseline characteristics show an average age of 35.2 years and weight of 75.1 kg.

All patients had prior ZDV use, with 58.6% having an experienced antiretroviral history. Symptomatic patients constitute 17.3% of the cohort. Baseline CD4 counts average 350.48 cells/mm^3, with a slight increase at 20 weeks. Conversely, baseline CD8 counts average 985.60 cells/mm^3, with a slight decrease at 20 weeks. Time-to-event data averages at 878.7 days.

Many variables in the table are related, such as `str2`, `strat`, `preanti`, and `z30`. We prioritised using `str2` because `strat` did not significantly enhance the information gained. Additionally, using a binary variable simplifies interpretation. To avoid multicollinearity, only one of these four variables was included in the model, with `str2` being chosen for its clarity and explanatory power.

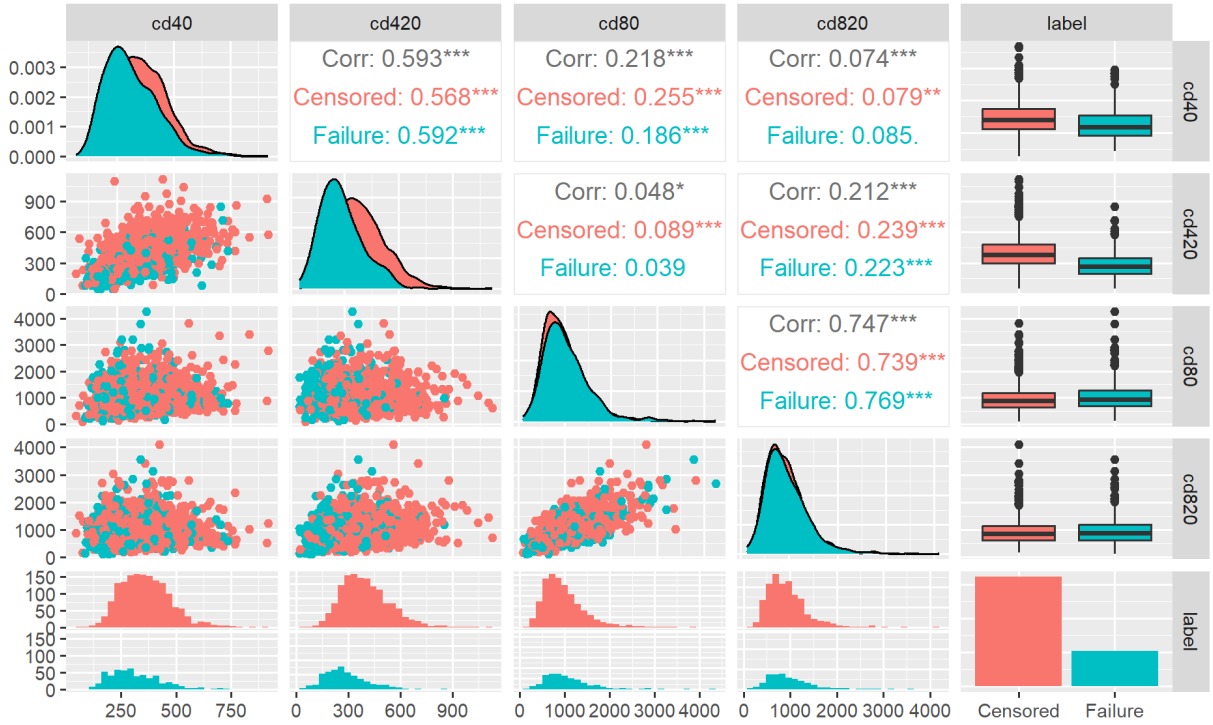

Figure 1: Distribution, Scatter and Box plots paired with correlation statistics between the four CD cell variables.

The CD4 and CD8 cell counts prove important indicators of AIDS progression in patients. In Figure 1, we observe that the variables `cd40`, `cd420`, `cd80`, and `cd820` exhibit distinct distributions, with higher variability in `cd40` and `cd420` compared to `cd80` and `cd820`. The density plots show some overlap between the censored and failure groups, with notable separation for `cd40` and `cd420`, indicating their potential relevance in distinguishing between these groups.

The pairwise correlations reveal varying degrees of association between the CD variables. Notably, `cd80` and `cd820` are strongly correlated (0.747^{***}). We saw earlier the little change in mean between these two and this may be why there is such a high correlation. `cd40` and `cd420` are also moderately correlated (0.593^{***}), though it is not as high, indicating a more stark difference between CD4 cell counts at the two timepoints.

Boxplots provide further insights, showing that CD4 cell counts at both time points visually differ between groups, whereas for CD8 this difference is not as obvious. Additionally, the distribution tails of these plots are also skewed right, indicating a log transformation of the variables may be required for modelling.

4. Results

Kaplan-Meier survival curves were fitted to compare survival times among the four treatment groups. Figure 2 illustrates the comparison of survival probabilities across these groups. It is evident that both combination therapy approaches and **ddI** alone provide better survival probabilities compared to **ZDV** alone. The log-rank test yields a p-value < 0.0001, emphasizing the statistically significant differences in survival between the treatment groups.

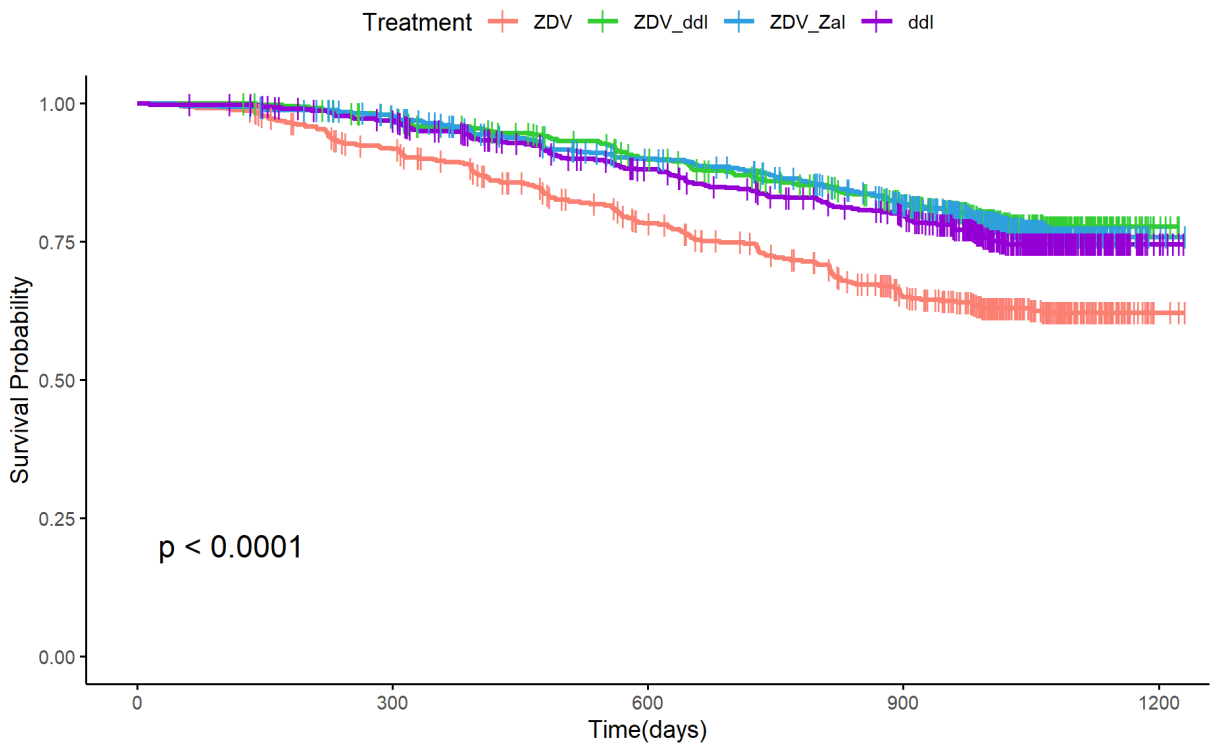

Figure 2: Kaplan-Meier survival curves for the four treatment groups.

A Kaplan-Meier survival curve comparison was also performed to assess survival times

between individuals who did and did not discontinue treatment before 96±5 weeks (`offtrt`). Figure 3 illustrates the comparison of survival probabilities across these two groups.

The red curve represents individuals who remained on treatment, while the blue curve represents those who discontinued treatment before 96±5 weeks. The survival probabilities for these two groups diverge significantly, as indicated by the log-rank test ($p < 0.0001$). Patients who remained on treatment had consistently higher survival probabilities over time compared to those who discontinued treatment. Notably, there are no censoring marks before 96±5 weeks in the `offtrt = No` group, as these individuals were defined as remaining on treatment until at least 96±5 weeks. In contrast, the `offtrt = Yes` group shows numerous censoring events prior to 96±5 weeks, reflecting the earlier discontinuation of treatment in this group.

This analysis underscores the critical role of treatment adherence in survival outcomes. However, it is important to acknowledge that discontinuation of treatment may have been influenced by factors such as medical advice, disease progression or other underlying conditions. Patients who continued treatment demonstrated markedly better survival outcomes compared to those who discontinued, regardless of the reason for discontinuation.

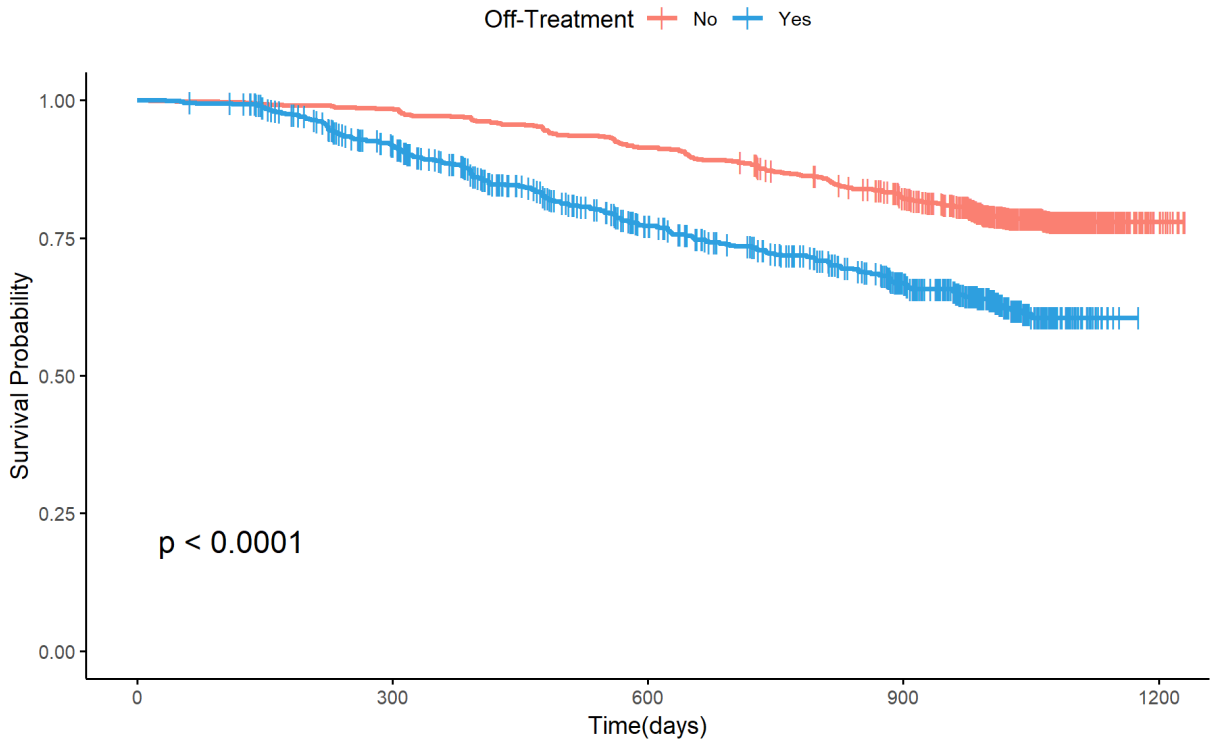

Figure 3: Kaplan-Meier survival curves comparing individuals who did and did not discontinue treatment before 96±5 weeks.

Following the Kaplan-Meier analysis, we fitted three Cox Proportional Hazards models to assess the factors influencing survival. The first model, the **Baseline Model**, in-

cluded covariates such as treatment group, previous IV drug use, symptoms, antiretroviral experience, and baseline CD4/CD8 levels. The second model, the **Updated Model**, incorporated both follow-up CD4/CD8 levels as well as off-treatment status. The CD cell count variables in both models were transformed using the log operation due to skewness in their distributions. Additionally, a third model, denoted the **Final Model**, combined both baseline and updated covariates to predict the best model.

Covariate	**Baseline Model**	**Updated Model**	**Final Model**
`trt.ZDV_ddI`	-0.79† (0.45)	-0.51† (0.60)	-0.47† (0.62)
`trt.ZDV_Zal`	-0.67† (0.51)	-0.51† (0.60)	-0.49† (0.61)
`trt.ddI`	-0.56† (0.57)	-0.44† (0.64)	-0.41† (0.66)
`drugs.yes`	-0.30** (0.74)	-0.31** (0.73)	NA
`symptom.symptomatic`	0.38*** (1.46)	0.31*** (1.37)	0.37*** (1.45)
`log_cd40`	-1.40† (0.25)	NA	0.41** (1.51)
`log_cd80`	0.54† (1.71)	NA	NA
`log_cd420`	NA	-2.18† (0.11)	-2.42† (0.09)
`log_cd820`	NA	0.57† (1.76)	0.56† (1.75)
`str2.experienced`	0.39† (1.47)	0.13 (1.14)	NA
`karnof.70`	0.87* (2.39)	0.83* (2.28)	NA
`karnof.80`	0.39** (1.48)	0.12 (1.13)	NA
`karnof.90`	0.30*** (1.34)	0.18* (1.19)	NA
`offtrt.yes`	NA	0.40† (1.49)	0.40† (1.49)
AIC	7509.12	7225.87	7221.56
C-index	0.70	0.77	0.77

Table 2: Comparison of Hazard Ratios and p-values for Baseline, Updated, and Final Cox Models.
Coefficients for the models are presented along with their corresponding hazard ratios (exponentiated coefficients).
p-values: * $=<.1$, ** $=<.05$, *** $=<.01$, † $=<.001$

The results of the Cox models are summarised in Table 2. Holding all other variables constant, we may interpret some of the variables from our models. All models highlight that both combination therapy groups and the ddI group have a lesser hazard risk than ZDV only, that is, their probability of meeting the primary end-point is lesser. Interestingly, in our first two models, patients with previous IV drug use experience a 26/27% less hazard than those who have not, though this result is marginally significant and classes are heavily imbalanced as we saw in Table 1. Patients displaying symptoms prove to be a significant predictor of meeting the end point with symptomatic patients having a 45% higher hazard risk in the final model than asymptomatic patients. Patients CD cell counts prove important in all models, though notably `log_cd80` is not in the final model. The CD-4 cells are weighted heavier than the CD-8 cells. Interestingly, a higher baseline CD4-cell count in the baseline model is associated with a lesser hazard though in the final model,

when the updated CD-4 cell count is introduced, the baseline CD-4 cell count now implies a greater hazard risk. Those with an experienced antiretroviral history exhibited a 47% increased hazard than those with a naive history, though this was only significant in the baseline model. Those with a Karnofsky Index less than 100 experienced an increased hazard than those with a score of 100. Patients being off treatment before the 96±5 week mark proved to be an important predictor of survival in the final model with those coming off treatment before this mark having a 49% higher hazard.

We observe that both the updated model and the final model achieve identical C-index values of 0.77, indicating a notable improvement in predictive accuracy compared to the baseline model, which has a C-index of 0.70. This demonstrates the updated and final models' superior ability to correctly rank survival risks. Similarly, the AIC values confirm these improvements. The updated and final models significantly outperform the baseline model, with a reduction of approximately 300 points in AIC, reflecting a better balance between model fit and complexity. While the difference in AIC between the updated and final models is smaller, at just 4.31 points, it still suggests that the final model provides a slightly better fit to the data.

As discussed in an earlier section (Cox's Proportional Hazards Model), the primary assumption of the Cox model is the proportional hazards assumption. This can be evaluated by examining the Schoenfeld residuals of the model. In this project, all three models assessed were found to violate this assumption. Specifically:

- The baseline model violates the assumption with `trt` and `log_cd40`.
- The updated model violates it with `log_cd420` and `offtrt`.
- The final model violates it with `log_cd420`, `offtrt`, and `log_cd40`.

While meeting the proportional hazards assumption is crucial, this project is intended as an introductory analysis. Addressing these violations through more advanced methods such as implementing models with time-varying coefficients is identified as an area for future exploration.

5. Conclusion and Discussion

This project provides an introductory analysis of factors influencing survival in a cohort of patients, with a focus on the impact of treatment regimens, CD4/CD8 cell counts and other baseline characteristics. The Kaplan-Meier survival analyses highlighted significant differences in survival probabilities across treatment groups and between individuals who discontinued or remained on treatment before 96±5 weeks. Notably, combination therapies and `ddI` alone demonstrated superior survival probabilities compared to `ZDV` alone, reinforcing the benefits of these treatment options.

The Cox Proportional Hazards models further highlighted key factors affecting survival. Symptomatic status emerged as a consistent and significant predictor of higher hazard risk across models. Similarly, CD4 cell counts, both at baseline and follow-up, were critical

determinants of survival, underscoring their importance in monitoring disease progression. The findings also suggest that discontinuation of treatment before the designated time frame has a marked adverse impact on survival.

However, this analysis revealed some limitations, including violations of the proportional hazards assumption in all three Cox models. These violations suggest that the relationship between covariates and hazard risk may vary over time, warranting further exploration with advanced modelling techniques such as introducing time-varying coefficients.

In conclusion, this project highlights the critical role of treatment regimens, baseline characteristics and disease markers in determining survival outcomes. It underscores the importance of adherence to therapy and regular monitoring of CD4/CD8 counts in improving patient prognosis. While the findings offer valuable insights, further work is required to refine the models and address underlying assumptions to achieve a more comprehensive understanding of survival dynamics in this patient population.

References

[1] Aids clinical trials group study 175 dataset. Kaggle, 2023. Accessed: November 23, 2024. URL: `https://www.kaggle.com/datasets/tanshihjen/aids-clinical-trials`.

[2] S. M. Hammer, D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, M. S. Hirsch, and T. C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *The New England journal of medicine*, 335(15):1081–1090, 1996. `doi:10.1056/NEJM199610103351501`.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 2nd edition, 2021.

[4] D.F. Moore. *Applied Survival Analysis Using R*. Springer, Switzerland, 2nd edition, 2016.

[5] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag Inc., New York, 2nd edition, 2002.