COLUMBIA UNIVERSITY
IRVING MEDICAL CENTER

Survival Analysis Final Report

Survival Analysis in AIDS: Analyzing and Comparing the Effect of Treatment Strategies

Advisor:
Prof. Helen Li
Teaching Assistant:
Ryan Wei

Group Members:
Zhuodiao Kuang(leader zk2275)*,
Anjing Liu(al4225), Shangsi Lin(sl5232),
Wenjia Zhu(wz2631), Ziqi Liao(zl3384)

Department of Biostatistics

December 20, 2023

Contents

1 Abstract

Introduction

Acquired Immunodeficiency Syndrome (AIDS) remains a global health challenge affecting millions with HIV. This study investigates the efficacy of antiretroviral therapies in HIV/AIDS treatment, focusing on the AIDS Clinical Trials Group Study 175 (ACTG 175) dataset. It compares the efficacy of various antiretroviral therapies, including monotherapies and combination treatments, aiming to assess their impact on patient survival probability and hazard rates in the context of HIV treatment evolution.

Methods

Our study methodically analyzes the ACTG 175 dataset, employing non-parametric survival estimation techniques like Life Tables, Kaplan-Meier, and Nelson-Aalen Estimators. We utilize Log-rank and Trend Log-rank Tests for survival distribution comparisons across treatment groups. The Cox Proportional Hazard Model, including time-varying coefficients, is applied for a nuanced assessment of treatment impacts on survival, with thorough checks for model assumptions and robustness.

Results

Our study's exploratory data analysis confirmed a well-balanced distribution of essential variables across the HIV/AIDS treatment groups. Life table, Kaplan-Meier Estimators and Nelson-Aalen Estimators revealed marked differences in survival curves among these groups. The log-rank test and the log-rank Trend Test for each treatment group yielded significant p-values below 0.05, indicating that ddI monotherapy and combination therapies outperformed ZDV monotherapy in terms of patient survival. The Cox Proportional Hazards model, enhanced with time-varying coefficients, consistently demonstrated significant hazard ratios for different treatment approaches. Notably, the hazard ratio for ZDV combined with ddI was 0.64 ($p < 0.001$), for ZDV combined with Zal was 0.62 ($p < 0.001$), and for ddI monotherapy was 0.67 ($p < 0.001$). These ratios significantly lower than 1 compared to ZDV monotherapy.

Conclusion

Our study concludes that in the treatment of HIV/AIDS, combination therapies (ZDV + ddI, ZDV + Zal) and ddI monotherapy demonstrate superior efficacy compared to ZDV monotherapy. These regimens show higher survival probabilities and significantly lower hazard ratios, indicating a more effective reduction in mortality risk among HIV-infected patients. Based on these findings, we advocate for a revision of current treatment protocols to prioritize these more effective therapies, potentially improving patient outcomes in the ongoing management of HIV/AIDS.

Keywords: AIDS, Survival Analysis, Kaplan-Meier curves, Log-rank tests, Hazard ratio, Cox Proportional Hazards Model

2 Introduction

Acquired Immunodeficiency Syndrome (AIDS), caused by the human immunodeficiency virus (HIV), remains a significant global health concern, impacting millions of lives worldwide[20]. The evolution of treatment strategies for HIV/AIDS is a critical area of research, focusing on improving patient survival and quality of life[6]. Historically, Zidovudine has been known to improve survival and reduce opportunistic infections in patients with advanced HIV-1 infections and to slow disease progression in those with mild symptoms[8][9]. However, its effectiveness tends to diminish over time, particularly in asymptomatic patients undergoing prolonged therapy[19]. The potential for more effective HIV treatment lies in combination therapy. Early intervention with potent regimens, particularly those combining ZDV with other

antiretrovirals, has shown promise in enhancing and sustaining immune responses[5][15]. Nevertheless, certain combinations, like ZDV and zalcitabine, have not yielded significant clinical benefits[10]. In this context, the AIDS Clinical Trials Group Study 175 (ACTG 175) provides crucial insights[12]. This study, a randomized, double-blind, placebo-controlled trial, evaluates the effectiveness of various antiretroviral therapies in adults infected with HIV-1 and having CD4 cell counts between 200 and 500 per cubic millimeter[7]. Our research is grounded in the detailed analysis of the ACTG 175 dataset. This study focuses on comparing the effects of four therapies, zidovudine (ZDV) or didanosine (ddI) monotherapy and the combination of zidovudine plus didanosine and zidovudine plus zalcitabine. A key aspect of our investigation is understanding the differential impacts of these treatments on patient survival. This involves a focus on survival probabilities and hazard ratios across the treatment groups, crucial for assessing the efficacy of these regimens. Our analysis, employing statistical methods such as life tables, Kaplan-Meier survival curves, log-rank tests, and the Cox proportional hazards model[4], aims to dissect these complexities. By doing so, we hope to contribute meaningfully to the field of HIV/AIDS treatment, enhancing the understanding of effective therapeutic strategies.

3 Methods

3.1 Data Description

The AIDS Clinical Trials Group protocol 175 (ACTG 175) dataset is derived from a rigorous randomized controlled trial (RCT) conducted in 1996, serves as an extensive and detailed collection of medical data for patients diagnosed with HIV/AIDS. It comprises data from 2139 HIV-infected patients, each represented by a comprehensive set of 23 attributes. The good news is that the data is complete, meaning there are no missing values.

The dataset is characterized by a non-informative right censoring approach. In terms of content, the dataset includes a diverse array of variables. These encompass treatment regimens, demographic details, historical medical information, and the Karnofsky performance score – a vital measure for evaluating a patient's functional status. The dataset segregates patients into four treatment groups: Zidovudine (ZDV) monotherapy, ZDV combined with didanosine (ddI), ZDV combined with zalcitabine, and ddI monotherapy. Our research focuses on comparing the treatment effects of the four treatment groups. The primary outcome was the survival probability and hazard ratio across these treatment groups. For a detailed breakdown of the 23 variables and their descriptions, please refer to the accompanying form in Table 7 in the Appendix.

3.2 Data Exploration

3.2.1 Data Summary

This part analyzed the baseline characteristics of participants enrolled in the AIDS Clinical Trials Group Study 175. The dataset comprises a diverse range of variables, categorized into continuous and categorical types, each offering insights into the participant profiles and study conditions. See Table 1 and Table 2.

The continuous variables include age, CD4 and CD8 counts at baseline and 20 weeks, duration of pre-treatment antiretroviral therapy, time to failure or censoring, and weight at baseline. Participants ranged in age from 12 to 70 years, with a median age of 34 years. The median CD4 count at baseline was 340, increasing slightly to 353 at 20 weeks. Similarly, CD8 counts showed a decrease from a median of 893 at baseline to 865 at 20 weeks. The participants had a varied duration of pre-treatment with antiretroviral therapy, ranging from 0 to 2851 days.

Table 1: Summary statistics of baseline characteristics (Continuous Variables)

Variable	Min	Median	Mean	Max	Q1	Q3	Std. Dev.
Age (Years)	12	34.00	35.25	70.00	29.00	40.00	8.71
CD4 Count at Baseline	0	340.00	350.50	1199.00	263.00	423.00	118.57
CD4 Count at 20 Weeks	49	353.00	371.31	1119.00	269.00	460.00	144.63
CD8 Count at Baseline	40	893.00	986.63	5011.00	654.00	1208.00	480.20
CD8 Count at 20 Weeks	124	865.00	935.37	6035.00	631.00	1147.00	444.98
Pre-Treatment Duration (Days)	0	142.00	379.18	2851.00	0.00	740.00	468.66
Time to Failure/Censoring (Days)	14	997.00	879.10	1231.00	727.00	1091.00	292.27
Weight at Baseline (kg)	31.00	74.39	75.13	159.94	66.68	82.55	13.26

The median time to failure or censoring was observed at 997 days. Weight varied widely among participants, with a median of 74.39 kg.

In the study, the categorical variables included treatment types, hemophilia status, homosexual activities, history of IV drug use, Karnofsky score, prior use of non-ZDV antiretroviral therapy, use of ZDV in the 30 days prior to the study, prior use of ZDV, race, gender, antiretroviral therapy history, stratification based on antiretroviral history, symptomatic status, treatment type, and the indicator for discontinuing treatment before 96±5 weeks. The treatment groups were evenly distributed across four categories, providing a balanced overview of different treatment modalities. This also confirms the success of the Randomized controlled trial. The majority of the 2,139 participants did not have hemophilia, with 1,959 individuals (approximately 92%) indicating absence of the condition. Interestingly, a significant portion of the cohort, representing 66%, reported engaging in homosexual activities. Intravenous drug use was relatively uncommon among the participants, with 87% reporting no history of such activities. The Karnofsky score, which measures a patient's functional status, revealed that a high proportion of participants were in good health, with 59% scoring the maximum 100 points. In terms of prior antiretroviral therapy, the majority had not been on non-ZDV antiretroviral therapy before the study, and all participants had previously used ZDV. Demographically, the study population was predominantly male, with males constituting 83% of the cohort. In terms of race, the majority were white, accounting for 71% of the participants. Regarding their antiretroviral therapy history, a significant number of participants were experienced in antiretroviral therapy, suggesting a cohort with substantial prior exposure to such treatments. Most participants were asymptomatic, and when it came to discontinuing treatment before 96±5 weeks, the data showed that a larger number of participants did not go off treatment. In terms of the study's outcome measure, censoring was more prevalent than failure, as indicated by the censoring indicator, with 76% not experiencing failure (censoring) in the context of the study's endpoint.

3.3 Non-parametric Survival Estimate

3.3.1 Life Table

Life tables can provide interval-based summaries of survival data[18]. For each treatment group, we calculated the number at risk, the number of events, and the proportion surviving at each interval. The hazard function was estimated as the ratio of the number of events to the number at risk in each interval. These estimates allowed us to observe the mortality rate's pattern over time and identify any periods with unusually high or low rates.

Table 2: Summary statistics of baseline characteristics (Categorical Variables)

Variable			Treatment Group			
	Levels	**Overall, N = 2139**	**0, N = 532**	**1, N = 522**	**2, N = 524**	**3, N = 561**
hemo	0	1959 (92%)	490 (92%)	479 (92%)	478 (91%)	512 (91%)
	1	180 (8.4%)	42 (7.9%)	43 (8.2%)	46 (8.8%)	49 (8.7%)
homo	0	725 (34%)	191 (36%)	176 (34%)	176 (34%)	182 (32%)
	1	1414 (66%)	341 (64%)	346 (66%)	348 (66%)	379 (68%)
drugs	0	1858 (87%)	469 (88%)	449 (86%)	448 (85%)	492 (88%)
	1	281 (13%)	63 (12%)	73 (14%)	76 (15%)	69 (12%)
karnof	70	9 (0.4%)	4 (0.8%)	0 (0%)	3 (0.6%)	2 (0.4%)
	80	80 (3.7%)	17 (3.2%)	22 (4.2%)	18 (3.4%)	23 (4.1%)
	90	787 (37%)	197 (37%)	189 (36%)	180 (34%)	221 (39%)
	100	1263 (59%)	314 (59%)	311 (60%)	323 (62%)	315 (56%)
oprior	0	2,092 (98%)	516 (97%)	513 (98%)	511 (98%)	552 (98%)
	1	47 (2.2%)	16 (3.0%)	9 (1.7%)	13 (2.5%)	9 (1.6%)
z30	0	962 (45%)	241 (45%)	234 (45%)	230 (44%)	257 (46%)
	1	1,177 (55%)	291 (55%)	288 (55%)	294 (56%)	304 (54%)
zprior	1	2,139 (100%)	532 (100%)	522 (100%)	524 (100%)	561 (100%)
race	0	1,522 (71%)	376 (71%)	384 (74%)	374 (71%)	388 (69%)
	1	617 (29%)	156 (29%)	138 (26%)	150 (29%)	173 (31%)
gender	0	368 (17%)	100 (19%)	88 (17%)	89 (17%)	91 (16%)
	1	1,771 (83%)	432 (81%)	434 (83%)	435 (83%)	470 (84%)
str2	0	886 (41%)	223 (42%)	213 (41%)	212 (40%)	238 (42%)
	1	1,253 (59%)	309 (58%)	309 (59%)	312 (60%)	323 (58%)
strat	1	886 (41%)	223 (42%)	213 (41%)	212 (40%)	238 (42%)
	2	410 (19%)	96 (18%)	106 (20%)	106 (20%)	102 (18%)
	3	843 (39%)	213 (40%)	203 (39%)	206 (39%)	221 (39%)
symptom	0	1,769 (83%)	443 (83%)	426 (82%)	435 (83%)	465 (83%)
	1	370 (17%)	89 (17%)	96 (18%)	89 (17%)	96 (17%)
treat	0	532 (25%)	532 (100%)	0 (0%)	0 (0%)	0 (0%)
	1	1,607 (75%)	0 (0%)	522 (100%)	524 (100%)	561 (100%)
offtrt	0	1,363 (64%)	316 (59%)	348 (67%)	322 (61%)	377 (67%)
	1	776 (36%)	216 (41%)	174 (33%)	202 (39%)	184 (33%)
cid	0	1,618 (76%)	351 (66%)	419 (80%)	415 (79%)	433 (77%)
	1	521 (24%)	181 (34%)	103 (20%)	109 (21%)	128 (23%)

3.3.2 Kaplan-Meier Estimator

The Kaplan-Meier estimator is a non-parametric estimation of the survival function, S(t), which represents the probability of surviving past time t[3]. The Kaplan-Meier estimator for the survival function at time t is given by:

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

where t_i are the distinct observed event times, d_i is the number of events at t_i, and n_i is the number of individuals at risk just prior to t_i. For each group, the Kaplan-Meier curve was plotted to illustrate the survival experience of participants over the study period. This method allowed us to utilize the full timeline of each participant, taking into account right-censoring, a common assumption where individuals' end of study data may not be due to the event of interest but rather due to loss of follow-up or study termination.

3.3.3 Nelson-Aalen Estimator

Similar to the Kaplan-Meier estimator, the Nelson-Aalen estimator is another non-parametric approach used in survival analysis. However, while the Kaplan-Meier estimator focuses on es-

timating the survival function, the Nelson-Aalen estimator is primarily used to estimate the cumulative hazard function, denoted as H(t).

$$H(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \quad \& \quad \hat{S}(t) = \prod_{t_i \leq t} e^{-\frac{d_i}{n_i}}$$

It is worth noting that the Nelson-Aalen estimator generally provides more conservative estimates of the cumulative hazard function compared to the Kaplan-Meier estimator[1][16].

3.4 Non-parametric test

3.4.1 Log-rank test

The log-rank test is a nonparametric statistic that we use to compare survival distributions across multiple groups. It is particularly effective in assessing differences in survival probabilities, and has optimal power when the hazard ratio is constant.

Consider $t_1, t_2, \ldots, t_D$ as the distinct event times observed across all groups. At each event time t_i, for each group k out of K total groups, let O_i be a vector representing the observed number of events, where each element d_{ki} corresponds to the observed number of events in group k. Similarly, let E_i be the vector of expected number of events for each group under the null hypothesis, calculated as $E_i = \begin{pmatrix} n_{1i} \\ \vdots \\ n_{Ki} \end{pmatrix} \frac{d_i}{n_i}$, where n_{ki} is the number of subjects at risk in group k at time t_i, d_i is the total number of events at time t_i, and n_i is the total number at risk at time t_i.

The covariance matrix V_i at each event time is defined as:

$$V_i = \begin{pmatrix} v_{11i} & v_{12i} & \cdots & v_{1Ki} \\ & v_{22i} & \cdots & v_{2Ki} \\ & & \cdots & \\ & & & v_{KKi} \end{pmatrix}$$

where $\nu_{kk'i}$ is the covariance between groups k and k' at time t_i, calculated as:

$$\nu_{kk'i} = \frac{n_{ki} d_i (n_i - d_i)}{n_i (n_i - 1)} \left(\delta_{kk'} - \frac{n_{k'i}}{n_i} \right)$$

with $\delta_{kk'}$ being the Kronecker delta function.

The overall log-rank test statistic is then computed as:

$$LM = {L_K}' V^{-1} L_K$$

where $L_K = \sum_{i=1}^{r'} (O_i - E_i)$ and $V = \sum_{i=1}^{r'} V_i$. Under the null hypothesis of no difference between the survival curves of the groups, LM follows a chi-square distribution with K degrees of freedom[17].

3.4.2 Trend Log-rank Test

The trend log-rank rest is an extension of the log-rank test, designed to evaluate survival data across multiple groups, especially when there is an ordinal relationship or trend among the groups. This test is particularly useful in scenarios where the proportional hazards assumption holds true across these groups.

For the trend log-rank rest, let's denote K groups, and for the k^{th} group, the log-rank statistics can be represented as:

$$L_k = \sum_{i=1}^{r'_k} (d_{0i} - e_{ki})$$

where d_{0i} is the observed number of events and e_{ki} is the expected number of events under the null hypothesis for the i^{th} time period in the k^{th} group.

The variance of L_k is given by:

$$\text{var}(L_k) = \sum_{i=1}^{r'_k} \frac{n_{0i} n_{ki} d_{ki} (n_{0ki} - d_{0ki})}{n_{0ki}^2 (n_{0ki} - 1)}$$

where n_{0i}, n_{ki}, d_{ki}, and n_{0ki} represent the number of subjects at risk, the number of subjects at risk in the k^{th} group, the number of events in the k^{th} group, and the total number of subjects at risk at the i^{th} time point, respectively.

The trend log-rank test statistic, LS, is then calculated as follows:

$$LS = \left(\frac{\sum_{k=1}^{K} \omega_k L_k}{\sqrt{\sum_{k=1}^{K} (\omega_k - \overline{\omega})^2 \text{var}(L_k)}} \right)^2 \quad \& \ \overline{\omega} = \frac{\sum_{k=1}^{K} \omega_k e_k}{\sum_{k=1}^{K} e_k}$$

Here, ω_k represents the weight for the k^{th} group, and $\overline{\omega}$ is the average weight. Under the null hypothesis of no trend across the groups, the test statistic LS follows a chi-square distribution with one degree of freedom (χ_1^2)[21].

3.4.3 Weighted (Trend) Log-rank Test

The weighted (trend) log-rank test extends the traditional (trend) log-rank test by incorporating weights into the analysis. This approach enhances the flexibility of the test, allowing it to cater to specific study designs or hypotheses. In this variant, the statistic $L_{k,w}$ is used, which is defined as $L_{k,w} = \sum_{i=1}^{k} \omega_i (d_{0i} - e_{0i})$. Here, ω_i denotes the weight assigned to the i^{th} event. Different weighting schemes can be applied depending on the study requirements. The choice of weights can significantly affect the sensitivity of the test to early or late differences in survival times.

Please refer to Table 8 in the appendix for more details[2][14].

3.5 Cox Proportional Hazard(PH) Model

In this study, another modeling approach used is Cox PH Model. It is built to perform survival analysis and predict the survival time of patients and predictor variables, As a semi-parametric model, it makes fewer assumptions about the underlying hazard function compared to fully parametric models. The primary assumption of the Cox PH model is that the hazard ratio is constant over time, which implies that the relative risk of an event remains constant across different time points[13].

Compared with prior approaches such as the log-rank test, Cox PH model allows analysis that considers multiple factors, the covariates. Final model was chosen based on stepwise approaches including forward, backward, and both. The model with the lowest AIC is selected. Generally, the Cox-PH model has the function as:

$$h(t|Z = z) = h_0(t) e^{\beta z}$$

where $h_0(t)$ is the baseline hazard function, Z can be a vector of p covariates, β is a vector of p coefficients.

3.6 Assumptions and Model Checking

3.6.1 Model Assumptions

The significance level for all tests in our analysis is set at 0.05. And we assumed that censoring occurs at random which is referred to as non-informative right censoring.

Given the non-parametric nature of the Kaplan-Meier estimator and the log-rank test, we do not assume proportional hazards in these methods. These approaches are robust to the violation of the proportional hazards assumption, making them suitable for a wide range of survival data analyses.

In addition to these non-parametric methods, we employed the Cox proportional hazards model. For this model, we need to assume the existence of proportional hazards. Hence, it is crucial to test the assumption of proportional hazards.

3.6.2 Model Checking Methods

1) Graphical Approach

We employed graphical methods to assess the proportional hazards (PH) assumption in our Cox model. Recall the PH model formulation, $S(t|Z = z) = S_0(t)^{e^{\beta z}}$, where $S_0(t)$ is the baseline survival function. By applying a log-log transformation to this model, we get the following relationship.

$$\log\{-\log \hat{S}(t|Z = z)\} - \log\{-\log \hat{S}_0(t)\} = \beta z$$

Under the PH assumption, this equation implies that the plot of $\log\{-\log \hat{S}(t|Z = z)\}$ against time for different values of Z should yield roughly parallel lines, as the right-hand side of the equation is constant for each group defined by Z.

2) Rao Score Test

Furthermore, we will utilize the Rao score test to evaluate the proportional hazards assumption.

The Rao score test is used to test the null hypothesis that the slope of the regression of the weighted Schoenfeld residuals on time is zero. The test statistic is based on the weighted Schoenfeld residuals, r_{ji}, for each covariate j at each failure time i. the weighted Schoenfeld residuals are an important tool for assessing the proportional hazards assumption. These residuals are defined as follows:

$$r_i = dI(\hat{\beta})^{-1} r_{Si}$$

where r_{Si} is the vector of Schoenfeld residuals for each covariate at each event time, and $I(\hat{\beta})$ is the information matrix evaluated at the estimated coefficients $\hat{\beta}$. The Schoenfeld residuals for each covariate j at each failure time i are given by:

$$r_{Sji} = \delta_i \left(Z_{ji} - \frac{\sum_{l \in R(t_i)} Z_{jl} \exp(\beta_j Z_{jl})}{\sum_{l \in R(t_i)} \exp(\beta_j Z_{jl})} \right)$$

In these equations, δ_i is an indicator that denotes whether an event (such as failure or death) occurred at time t_i. Z_{ji} represents the value of the j-th covariate for the individual who experienced the event at time t_i. The set $R(t_i)$ denotes the risk set at time t_i, which includes all individuals at risk of experiencing the event at that time. The term d represents the total number of events. The information matrix $I(\hat{\beta})$ plays a crucial role in the calculation of the weighted Schoenfeld residuals. It reflects the variability of the estimated coefficients $\hat{\beta}$ in the Cox model, and its inverse provides the necessary weighting in the calculation of r_i.

The hypothesis tested using the Rao score test is:

$$H_0 : \text{The slope of } \beta_j + r_{ji} \text{ versus } T_i = 0 \quad \text{for all } j$$

versus the alternative hypothesis that the slope is non-zero for at least one covariate. A significant result from this test indicates a violation of the proportional hazards assumption for the Cox model[11].

This methodology allows for a comprehensive evaluation of the proportional hazards assumption, ensuring the robustness and validity of our Cox model analysis.

4 Result

4.1 Exploratory Data Analysis Findings

In the exploratory data analysis of the clinical study, a thorough examination of continuous and categorical variables was conducted to assess the comparability of treatment groups within a randomized controlled trial (RCT) framework. The heatmap analysis, comprising scatter plots, histograms, and correlation coefficients, revealed a successful randomization process, indicated by the similar distribution shapes and spreads of continuous covariates like age, CD4 and CD8 counts, weight, and pre-treatment duration across four treatment regimens. This uniformity ensures each group's comparability at baseline, crucial for evaluating treatment efficacy and safety. While the CD4 and CD8 counts, key markers in HIV treatment, showed a significant increase across treatments, suggesting effective immune restoration, the analysis also revealed a notable exception in multicollinearity, primarily between CD4 and CD8 counts. These findings underscore the need to employ statistical techniques that address the interdependency of these counts, ensuring accurate coefficient estimation and maintaining the integrity of the analysis(see Figure 4 in Appendix).

As illustrated in Figure 5 and Figure 6 in the Appendix, the exploration of categorical variables through bar charts further affirmed the balanced distribution of key demographics and clinical characteristics like hemophilia status, sexual orientation, and drug use history across the treatment groups, reinforcing the effectiveness of the randomization. The boxplot visualization highlighted the relationship between various categorical variables and event time, a critical endpoint in the study. While some variables like the Karnofsky score and treatment type showed potential influences on event timing, others exhibited less variability, indicating a small impact.

4.2 Life table

The life tables stratified by the treatment groups provided each 100-day interval-specific survival information for patients receiving different AIDS treatments over time. Across all groups, they began with a survival probability of 100%. The hazard rate followed a pattern of increase until the 800 to 900-day interval, after which it showed a decline at different rates, and we need to perform further test to convince this difference. The median survival times were not attainable as the survival probability did not reach 50% within the study time for any group. Considering the 80% survival time instead, Treatment Group 0(ZDV only) showed the earliest time with 600-700 days, Treatment Group 3(ddI only) displayed intermediate time with 900-1000, and Treatment Group 1(ZDV + ddI) and 2(ZDV + Zal) showed latest time with 1000-1100 days.

4.3 Non-parametric Survival Estimate

As Figure **??** shows, we utilized the Kaplan-Meier survival analysis and Nelson-Aalen Estimator with 95% confidence interval to estimate the survival probabilities for different treatment groups over time which extended up to 1250 days, displayed in Figure 1. According to the result of Kaplan-Meier survival analysis, we observed that the magnitude of the differences between the curves varies over time. Initially, the curves start close together, indicating similar survival probabilities across all groups. However, as time progresses, the curves diverge, with Treatment Group for ZDV + ddI consistently showing the highest survival probability and Treatment Group for ZDV only showing the lowest. The curves for Treatment Groups for ZDV + Zal and ddI only display intermediate survival probabilities, with Group ZDV + Zal generally above Group ddI only. At the 80% quantile, the Kaplan-Meier analysis indicated survival times of 569 days for ZDV only, 986 days for ZDV + ddI, 972 days for ZDV + Zal, and 898 days for ddI only. The Nelson-Aalen Estimator gives a very similar result.

4.4 Non-parametric test

Table 3: Log-rank Family of Test Result

Test	Chi-Square	p-value
Log-rank (Mantel-Cox)	49.194110	< 0.0001
Gehan-Breslow-Wilcoxon	56.430494	< 0.0001
Tarone-Ware	53.333356	< 0.0001
Peto-Peto	52.969995	< 0.0001
Modified Peto-Peto	52.975187	< 0.0001
Fleming-Harrington (p=0, q=1)	18.105713	0.0004183
Fleming-Harrington (p=1, q=0)	52.964058	< 0.0001
Fleming-Harrington (p=1, q=1)	20.758158	0.0001182
Fleming-Harrington (p=0.5, q=0.5)	30.576382	< 0.0001
Fleming-Harrington (p=0.5, q=2)	9.122593	0.0277046

Table 4: Log-rank Family of Trend Test Result

Test	Z	p-value
Log-rank (Mantel-Cox)	5.438449	< 0.0001
Gehan-Breslow-Wilcoxon	5.895011	< 0.0001
Tarone-Ware	5.702505	< 0.0001
Peto-Peto	5.669496	< 0.0001
Modified Peto-Peto	5.669831	< 0.0001
Fleming-Harrington (p=0, q=1)	3.155372	0.0016029
Fleming-Harrington (p=1, q=0)	5.668639	< 0.0001
Fleming-Harrington (p=1, q=1)	3.421890	0.0006219
Fleming-Harrington (p=0.5, q=0.5)	4.227214	< 0.0001
Fleming-Harrington (p=0.5, q=2)	2.073866	0.0380918

1) Log-rank Test

A series of log-rank family tests were conducted to compare survival distributions across the four treatment groups in the study: ZDV only (0), ZDV + ddI (1), ZDV + Zal (2), and ddI only (3). Under the null hypothesis $H_0 : S_0(t) = S_1(t) = S_2(t) = S_3(t)$, where $S_i(t)$ represents the survival function of the i^{th} treatment group at time t, the tests sought to detect any differences in survival functions among the treatment groups.

The results of the log-rank family of tests are given in Table 3. The Log-rank (Mantel-Cox), Gehan-Breslow-Wilcoxon, Tarone-Ware, Peto-Peto, and Modified Peto-Peto tests all yielded highly significant p-values below the 0.0001 threshold, firmly rejecting the null hypothesis at a conventional significance level of 0.05. This indicates that there are statistically significant differences in the survival functions across the treatment groups.

The Fleming-Harrington family of tests, which gives different weights to events at different times, also supported the rejection of the null hypothesis for most configurations of p and q parameters. However, for the FH test with $p = 0.5, q = 2$, the p-value was 0.0277046, which is above the more stringent significance level of 0.01 but still indicates significant differences at the 0.05 level.

2) Trend Log-rank Test

A weighted log-rank family trend test was applied to assess the survival functions across four treatment groups, informed by the initial findings from Kaplan-Meier estimations. The Kaplan-Meier curves suggested a notably lower survival probability for the treatment group 0

(ZDV only) compared to the other groups, which exhibited similar survival probabilities. Consequently, we state the null hypothesis $H_0 : S_0(t) = S_1(t) = S_2(t) = S_3(t)$ and the alternative $H_1 : S_0(t) \leq S_1(t) = S_2(t) = S_3(t)$, with weights (1,3,3,3) assigned to reflect the observed KM trends.

The results of the trend log-rank family of tests are summarized in Table 4. the Log-rank (Mantel-Cox) test, Gehan-Breslow-Wilcoxon test, and Tarone-Ware test all indicated significant differences in survival experiences among the treatment groups, with p-values well below the 0.05 threshold. The Peto-Peto and Modified Peto-Peto tests substantiated these findings with similarly significant p-values. Fleming-Harrington tests across various parameter settings revealed significant trends, notably for $p = 0, q = 1$ (p = 0.0016029), indicating significant differences at the 0.05 significance level. The results were consistent for other parameter configurations, with the exception of $p = 0.5, q = 2$, which presented a p-value of 0.0380918, signifying a weaker trend that is significant at the 0.05 level.

4.5 Cox PH model

4.5.1 Model selection

Based on the three different step selection directions, three models are generated. The models used backward selection and bidirection selection generated the same model, variables selected as statistically significant at the 0.05 alpha level are trt1(ZDV+ddI), trt2(ZDV+Zal), trt3(ddI), age, drugs1(history of IV drug use), karnof, preanti, symptom1, offtrt1, cd40(cd4 at baseline), cd420(CD4 at 20 +/- 5 weeks), and cd820(CD8 at 20 +/- 5 weeks). The model used forward selection selected variables that are statistically significant at the 0.05 alpha level also includes the same variables, however it includes much more variables that are not statistically significant at the 0.05 alpha level in the final model. Since the models produced by bidirection and forward selection has the lowest AIC, and it is also simpler, it is being selected as our Cox PH model for further discussion.

4.5.2 The Diagnostic Process

1) Graphical Diagnosis of PH Assumptions in Treatment Group Comparisons

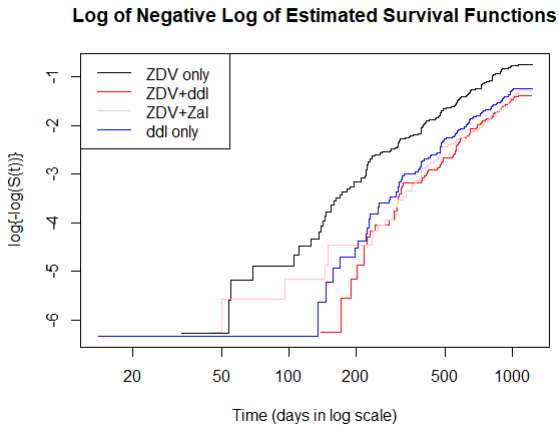

Figure 1: Log-Log Plots of Estimated Survival Functions for Four Treatment Groups

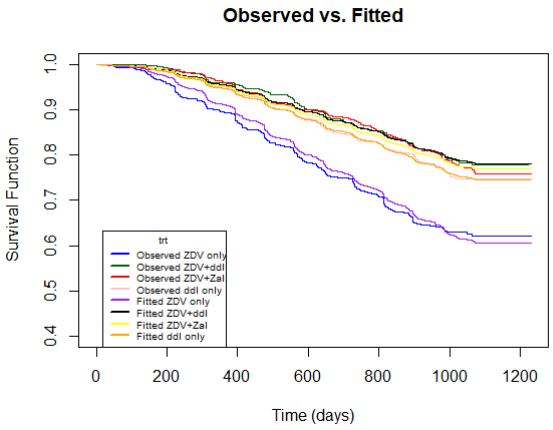

Figure 2: Observed and Fitted Survival Functions for Four Treatment Groups

As shown in Figure 1 and Figure 2, except for a few deviations at the early stages, the log-log survival plots for the four treatment groups exhibited approximately parallel lines, and

the fits of the Cox and KM models are in general agreement. These graphical observations suggest a reasonable adherence to the proportional risks relationship among the treatment groups. The parallelism and consistency in these plots support the validity of the proportional hazards assumption in our model, particularly given that our model includes only one indicator variable.

2) Rao Score Test with Covariate

Following the graphical analysis, we extended our assessment of the proportional hazards (PH) assumption to include more covariates using the Rao score test.

In our analysis, we applied the Rao score test to various covariates including treatment type, pre-antiretroviral therapy status, symptomatic status, CD8 count at 20 weeks, Karnofsky score (karnof), age, and history of drug use. The test results were indicative of the PH assumption holding for these variables. Specifically, the p-values for these tests ranged from 0.06140 to 0.71101, suggesting no significant deviation from the proportional hazards assumption for each of these covariates.The comprehensive results of the Rao score test are presented in Table 5.

Table 5: Rao Score Test Results

Test	**Chi-squared (χ^2)**	**df**	**p-value**
trt	7.355	3	0.06140
preanti	0.137	1	0.71101
symptom	0.689	1	0.40636
offtrt	17.437	1	3.0×10^{-5}
cd420	75.515	1	$< 2 \times 10^{-16}$
cd820	0.255	1	0.61385
karnof	0.488	1	0.48477
age	1.176	1	0.27809
drugs	0.632	1	0.42658
cd40	12.240	1	0.00047
GLOBAL	96.034	12	3.3×10^{-15}

However, at the 0.05 confidence level, the test revealed significant evidence against the proportional hazards assumption for the variables off-treatment before 96±5 weeks (offtrt, $p = 3.0 \times 10^{-5}$), CD4 count at 20 weeks (cd420, $p < 2 \times 10^{-16}$), and CD4 count at baseline (cd40, $p = 0.00047$).

To further investigate these deviations, we plotted Schoenfeld residuals for these variables, providing a visual assessment of how their relationship with the hazard changes over time.

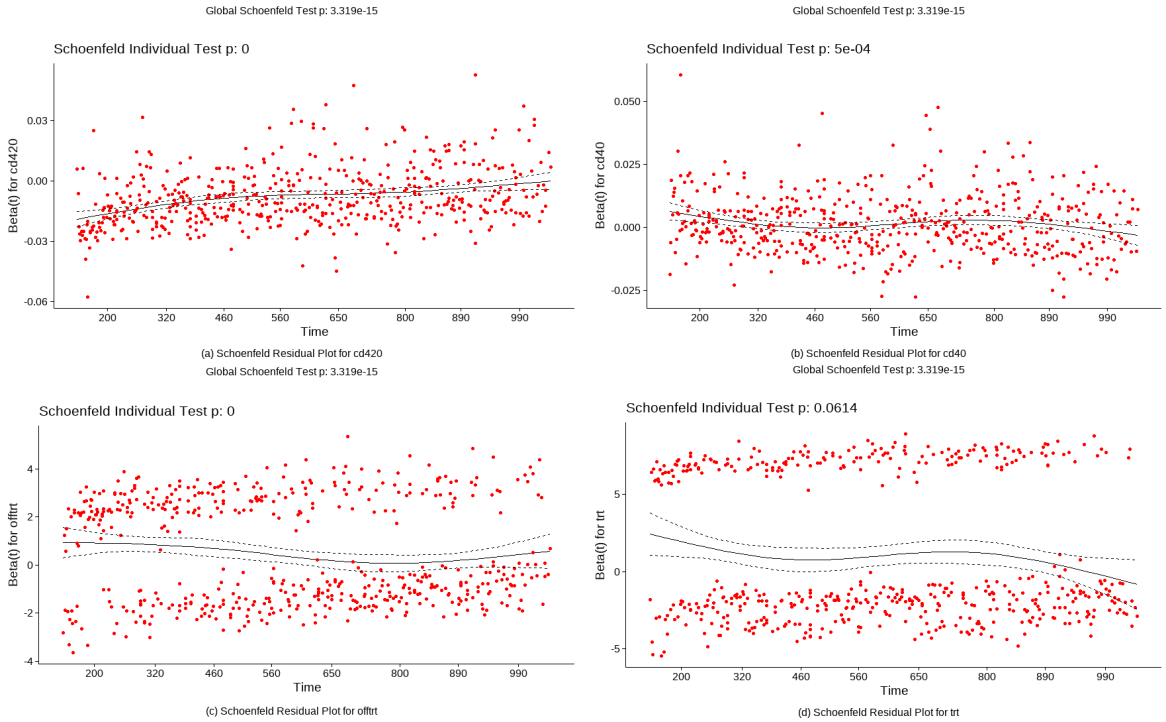

Figure 3: Schoenfeld Residual Plot of Three Target Covariate(a,b,c) and Treatment Indicator(d)

The Schoenfeld residuals analysis revealed significant deviations from the proportional hazards assumption for several covariates. Specifically, the residuals for the CD4 count at 20 weeks suggested a potential increase in the hazard ratio over time, with a clear upward trajectory, indicating a violation of the proportional hazards assumption (see Figure 3). For the baseline CD4 count (cd40), a subtle undulating pattern was observed in the residuals, which also suggests a departure from the assumption, albeit to a lesser extent than the other variables. The plot for off-treatment before 96±5 weeks demonstrated a non-random dispersion of residuals, further corroborating the non-proportionality.

Regarding treatment indicator, unusual patterns were observed, suggesting a potential variability in the hazard ratio over time. Despite these irregularities, the Schoenfeld individual test yielded a p-value of 0.0614, which is above the selected alpha level of 0.05. Consequently, we opted to regard the proportional hazards assumption for the treatment effect as tenable.

4.5.3 Time-varying Cox PH Model

To address the observed deviations from the proportional hazards assumption for the variables off-treatment time, CD4 count at 20 weeks, and CD4 count at baseline, we incorporated a time-linear interaction term for each of these covariates in our Cox model. This approach allows the hazard ratio to change linearly over time, thereby accommodating the non-proportional effects indicated by the Schoenfeld residuals.

Table 6: Comparison Table of Cox Regression With and Without Time-varying Coefficient

	Time-varying		Origin	
Predictors	**Hazard Ratio**	**p**	**Hazard Ratio**	**p**
trt [1]	0.64 ***	< 0.001	0.64 ***	< 0.001
	(0.50 – 0.82)		(0.50 – 0.82)	
trt [2]	0.62 ***	< 0.001	0.60 ***	< 0.001
	(0.49 – 0.79)		(0.47 – 0.77)	
trt [3]	0.67 ***	< 0.001	0.66 ***	< 0.001
	(0.53 – 0.84)		(0.53 – 0.83)	
preanti	1.00 **	0.003	1.00 **	0.002
	(1.00 – 1.00)		(1.00 – 1.00)	
symptom	1.38 **	0.002	1.42 ***	0.001
	(1.13 – 1.69)		(1.16 – 1.74)	
offtrt	2.86 ***	< 0.001	1.62 ***	< 0.001
	(1.79 – 4.58)		(1.35 – 1.94)	
cd420	0.98 ***	< 0.001	0.99 ***	< 0.001
	(0.98 – 0.98)		(0.99 – 0.99)	
cd820	1.00 ***	< 0.001	1.00 ***	< 0.001
	(1.00 – 1.00)		(1.00 – 1.00)	
karnof	0.99	0.092	0.99	0.057
	(0.97 – 1.00)		(0.97 – 1.00)	
age	1.01 *	0.023	1.01 *	0.036
	(1.00 – 1.02)		(1.00 – 1.02)	
drugs	0.73 *	0.030	0.72 *	0.027
	(0.54 – 0.97)		(0.54 – 0.96)	
cd40	1.00 **	0.003	1.00 **	0.003
	(1.00 – 1.01)		(1.00 – 1.00)	
cd420 · time	1.00 ***	< 0.001		
	(1.00 – 1.00)			
cd40 · time	1.00	0.059		
	(1.00 – 1.00)			
offtrt · time	1.00 **	0.007		
	(1.00 – 1.00)			
Observations: 2139				
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$				

In our study, we compared Cox regression models with and without time-varying coefficients to assess treatment effects and other covariates' impact on survival. The Table 6 presents hazard ratios (HRs) and associated p-values for each predictor in both models. Once again, we use a confidence level of 0.05.

For treatment indicators (trt), the hazard ratios remained consistent across both models. In the time-varying model, the HR for trt [1] (ZDV + ddI) was 0.64 ($p < 0.001$), trt [2] (ZDV + Zal) was 0.62 ($p < 0.001$), and trt [3] (ddI) was 0.67 ($p < 0.001$), all indicating a substantial reduction in hazard compared to the baseline treatment (ZDV only).

Several covariates showed different HRs between the two models. Notably, 'offtrt' (off-treatment before 96±5 weeks) had an HR of 2.86 in the time-varying model ($p < 0.001$) compared to 1.62 in the original model ($p < 0.001$), indicating a significant change in the hazard ratio over time.

Covariates with HRs significantly different from 1 in were 'preanti', 'symptom', 'cd40', 'cd420', 'cd820', 'age', and 'drugs', indicating a substantial impact on survival. For instance, 'symptom' had an HR of 1.38 in the time-varying model ($p = 0.002$) and 1.42 in the original model ($p = 0.001$). Covariate 'karnof' did not show a significant difference in HR between the two models. The HR for 'karnof' was 0.99 in both models, with p-values of 0.092 and 0.057, respectively.

4.5.4 Final Cox Models

The final selected Time-varying Cox model is:

$$\log\left(\frac{h(t|Z)}{h_0(t)}\right) = \begin{array}{llllll} -0.4475 & \times \textbf{trt1} & -0.4794 & \times \textbf{trt2} & -0.4075 & \times \textbf{trt3} \\ +0.000274 & \times \textbf{preanti} & +0.3234 & \times \textbf{symptom1} & +1.0530 & \times \textbf{offtrt1} \\ -0.01805 & \times \textbf{cd420} & +0.000464 & \times \textbf{cd820} & -0.01199 & \times \textbf{karnof} \\ +0.01155 & \times \textbf{age} & -0.3196 & \times \textbf{drugs1} & +0.00351 & \times \textbf{cd40} \\ -0.001033 & \times \textbf{offtrt} \cdot \textbf{t} & +0.000016 & \times \textbf{cd420} \cdot \textbf{t} & -0.0000035 & \times \textbf{cd40} \cdot \textbf{t} \end{array}$$

The final selected non-Time-varying Cox model is:

$$\log\left(\frac{h(t|Z)}{h_0(t)}\right) = \begin{array}{llllll} -0.4416 & \times \textbf{trt1} & -0.5060 & \times \textbf{trt2} & -0.4117 & \times \textbf{trt3} \\ +0.0003 & \times \textbf{preanti} & +0.3531 & \times \textbf{symptom1} & +0.4809 & \times \textbf{offtrt1} \\ -0.0815 & \times \textbf{cd420} & +0.0005 & \times \textbf{cd820} & -0.0135 & \times \textbf{karnof} \\ +0.0108 & \times \textbf{age} & -0.3256 & \times \textbf{drugs1} & +0.0015 & \times \textbf{cd40} \end{array}$$

For a more detailed summary of the two models, see Table 9 and Table 10 in the Appendix.

5 Discussion

5.1 Interpretation and Findings

Our study undertook a detailed examination of the AIDS Clinical Trials Group protocol 175 (ACTG 175) dataset, which encompassed data from 2139 HIV-infected patients, characterized by 23 diverse attributes. The principal aim was to assess and contrast the efficacy of four distinct AIDS treatment regimens: Zidovudine (ZDV) monotherapy, ZDV combined with didanosine (ddI), ZDV combined with zalcitabine, and ddI monotherapy. The primary ountcome was the survival probability and hazard ratio across these treatment groups over a span of 1250 days.

Our comprehensive analysis indicated marked disparities in survival probabilities between the treatment groups. Utilizing approaches such as the Life-table method, Kaplan-Meier Estimator, and the Nelson-Aalen Estimator, we consistently observed that ZDV monotherapy lagged in terms of survival probability. In contrast, the combination treatment of ZDV and ddI, along with other groupings, exhibited higher survival probabilities. These observations were robustly supported by a series of log-rank family tests, which unequivocally refuted the hypothesis of identical survival functions across the treatment groups.

Recognizing the apparent order in survival probabilities suggested by Kaplan-Meier estimates, we employed the trend log-rank family of tests. These tests were weighted to reflect the observed survival trends among the treatment groups. The results further substantiated our initial findings, demonstrating significant differences in survival experiences with a particular emphasis on the inferiority of ZDV monotherapy compared to the other treatments.

Delving deeper, our investigation employed the Cox Proportional Hazards (PH) model, carefully chosen through stepwise selection methods to include pivotal covariates. Our initial diagnostic evaluations suggested compliance with the PH assumption for the treatment variable. However, the behavior of certain covariates deviated from this assumption, leading to the integration of time-varying coefficients into the model.

One of the most striking findings from our Cox model analysis was the consistency of hazard ratios (HRs) for treatment indicators across both the standard and time-varying model. This consistency underscores the robustness of our treatment effect findings. Specifically, the hazard ratios for combination therapies (ZDV + ddI, ZDV + Zal) and ddI monotherapy consistently indicated a significant reduction in hazard compared to ZDV monotherapy. For instance, in the time-varying model, ZDV + ddI treatment had an HR of 0.64, ZDV + Zal had an HR of 0.62, and ddI monotherapy had an HR of 0.67 compared to ZDV monotherapy. These HRs, being less than 1 and statistically significant ($p < 0.05$), strongly suggest the superior efficacy of these treatments over ZDV monotherapy.

In summary, our study yields significant insights for HIV/AIDS treatment strategies. The consistently lower survival probability linked with ZDV monotherapy underlines its relative ineffectiveness. On the other hand, the combination therapies involving ZDV and either ddI or zalcitabine, as well as ddI monotherapy, demonstrate more favorable outcomes. These results suggest a preference for combination therapies or ddI monotherapy over ZDV monotherapy in the treatment of HIV/AIDS, underscoring the fact that we should favor the use of combination therapies or ddI monotherapy.

5.2 Limitations

While our study provides significant insights into treatment efficacy for HIV/AIDS, there are limitations to consider.

In our study, significant deviations from the proportional hazards assumption were observed for specific covariates, prompting the introduction of linear time interaction terms in our Cox model. While this adjustment addresses time-varying effects, it is limited by its assumption of linear changes over time, possibly leading to an oversimplified interpretation of their effects.

Another point of consideration is the treatment indicator variable, which, despite showing unusual patterns in the Schoenfeld residuals analysis, passed the Rao score test with a p-value of 0.0614. This value, being slightly above our selected alpha level of 0.05, indicates a borderline adherence to the PH assumption. While we regarded the proportional hazards assumption for the treatment effect as tenable, it's important to acknowledge that this marginal p-value introduces a degree of uncertainty, potentially impacting the robustness and reliability of the model's findings regarding treatment effects.

6 Conclusions

From our extensive analysis, we conclude that in the context of HIV/AIDS treatment, combination therapies (ZDV + ddI, ZDV + Zal) and ddI monotherapy are more effective than ZDV monotherapy. These treatments not only demonstrate higher survival probabilities but also significantly lower hazard ratios, indicating their effectiveness in reducing the risk of mortality among HIV-infected patients. Given these findings, we recommend a reconsideration of treatment protocols in HIV/AIDS management, favoring combination therapies or ddI monotherapy over ZDV monotherapy. This shift could potentially improve patient outcomes and align treatment strategies with the evolving landscape of HIV/AIDS care.

It is important, however, to recognize the limitations of our study, including potential oversimplifications in the Cox model and the borderline adherence to the proportional hazards assumption for the treatment effect. These factors underline the necessity for ongoing research and the continuous adaptation of HIV/AIDS treatment strategies to ensure they reflect the most current and comprehensive understanding of the disease and its management.

In conclusion, our study contributes significantly to the body of knowledge in HIV/AIDS treatment, offering evidence-based guidance for enhancing patient care and outcomes. As the medical community continues to combat HIV/AIDS, these findings provide a strong foundation for optimizing treatment strategies and improving the quality of life for those affected by this chronic condition.

References

[1] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 701–726.

[2] Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., & Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models* (Vol. 4). Springer.

[3] Cleves, M. (2008). *An introduction to survival analysis using stata*. Stata press.

[4] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

[5] D'Aquila, R. T., Hughes, M. D., Johnson, V. A., & et al. (1996). Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with hiv-1 infection: A randomized, double-blind, placebo-controlled trial. *Annals of Internal Medicine*, *124*, 1019–1030.

[6] Eisinger, R. W., & Fauci, A. S. (2018). Ending the hiv/aids pandemic. *Emerging infectious diseases*, *24*(3), 413.

[7] Farhadian, M., Mohammadi, Y., Mirzaei, M., & Shirmohammadi-Khorram, N. (2021). Factors related to baseline cd4 cell counts in hiv/aids patients: Comparison of poisson, generalized poisson and negative binomial regression models. *BMC Research Notes*, *14*, 1–7.

[8] Fischl, M. A., Richman, D. D., Grieco, M. H., & et al. (1987). The efficacy of azidothymidine (azt) in the treatment of patients with aids and aids-related complex: A double-blind, placebo-controlled trial. *New England Journal of Medicine*, *317*, 185–191.

[9] Fischl, M. A., Richman, D. D., Hansen, N., & et al. (1990). The safety and efficacy of zidovudine (azt) in the treatment of subjects with mildly symptomatic human immunodeficiency virus type 1 (hiv) infection: A double-blind, placebo-controlled trial. *Annals of Internal Medicine*, *112*, 727–737.

[10] Fischl, M. A., Stanley, K., Collier, A. C., & et al. (1995). Combination and monotherapy with zidovudine and zalcitabine in patients with advanced hiv disease. *Annals of Internal Medicine*, *122*, 24–32.

[11] Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*, 515–526.

[12] Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081–1090.

[13] Kleinbaum, D. G., Klein, M., Kleinbaum, D. G., & Klein, M. (2012). The cox proportional hazards model and its characteristics. *Survival analysis: a self-learning text*, 97–159.

[14] Leissen, S., Ligges, U., Neuhäuser, M., & Hothorn, L. A. (2009). Nonparametric trend tests for right-censored survival times. In *Statistical inference, econometric analysis and matrix algebra: Festschrift in honour of götz trenkler* (pp. 41–61). Springer.

[15] Meng, T. C., Fischl, M. A., Boota, A. M., & et al. (1992). Combination therapy with zidovudine and dideoxycytidine in patients with advanced human immunodeficiency virus infection: A phase i/ii study. *Annals of Internal Medicine*, *116*, 13–20.

[16] Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, *1*(1), 27–52.

[17] Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, *135*(2), 185–198.

[18] Shepard, J. M., & Greene, R. W. (2003). Sociology and you. *(No Title)*.

[19] Volberding, P. A., Lagakos, S. W., Grimes, J. M., & et al. (1995). A comparison of immediate with deferred zidovudine therapy for asymptomatic hiv-infected adults with cd4 cell counts of 500 or more per cubic millimeter. *New England Journal of Medicine*, *333*, 401–407.

[20] Whiteside, A., & Wilson, D. (2018). Health and aids in 2019 and beyond.

[21] Yang, S., & Prentice, R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, *66*(1), 30–38.

7 Appendix

7.1 Variables and Description in the Dataset

Table 7: Variables and Description in the Dataset

Variables	Description
pidnum	Patient ID
cid	censoring indicator (1 = failure, 0 = censoring)
time	time to failure or censoring
age	age (yrs) at baseline
wtkg	weight (kg) at baseline
hemo	hemophilia (0=no, 1=yes)
homo	homosexual activity (0=no, 1=yes)
drugs	history of IV drug use (0=no, 1=yes)
karnof	Karnofsky score (on a scale of 0-100)
oprior	Non-ZDV antiretroviral therapy pre-175 (0=no, 1=yes)
z30	ZDV in the 30 days prior to 175 (0=no, 1=yes)
zprior	ZDV prior to 175 (0=no, 1=yes)
preanti	number of days pre-175 anti-retroviral therapy
race	race (0=White, 1=non-white)
gender	gender (0=Female, 1=Male)
str2	antiretroviral history (0=naive, 1=experienced)
strat	antiretroviral history stratification
symptom	symptomatic indicator (0=asymp, 1=symp)
treat	treatment indicator (0=ZDV only, 1=others)
offtrt	indicator of off-trt before 96+/-5 weeks (0=no,1=yes)
cd40	CD4 at baseline
cd420	CD4 at 20+/-5 weeks
cd80	CD8 at baseline
cd820	CD8 at 20+/-5 weeks
trt	treatment indicator (0 = ZDV; 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI)

7.2 Exploratory Data Analysis

7.2.1 Continuous Variable Exploration

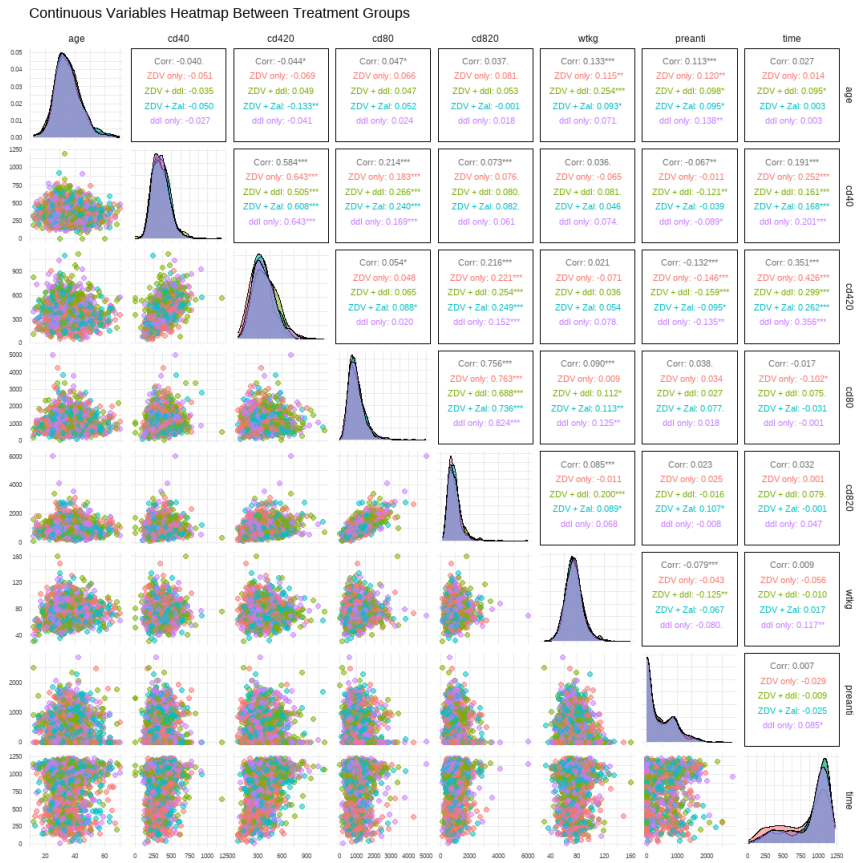

Figure 4: Continuous Variables Heatmap Between Treatment Groups

7.2.2 Categorical Variable Exploration

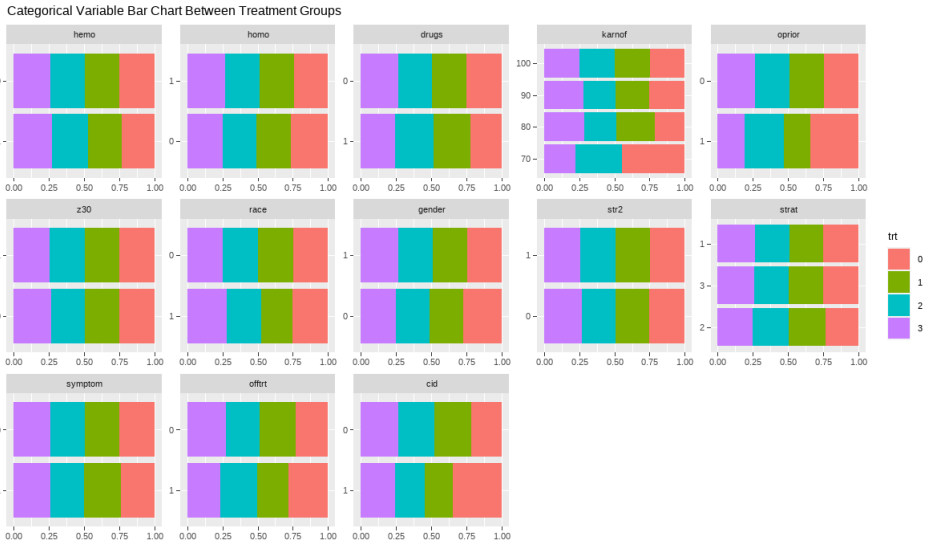

Figure 5: Categorical Variable Bar Chart Between Treatment Groups

7.2.3 Event Time Exploration

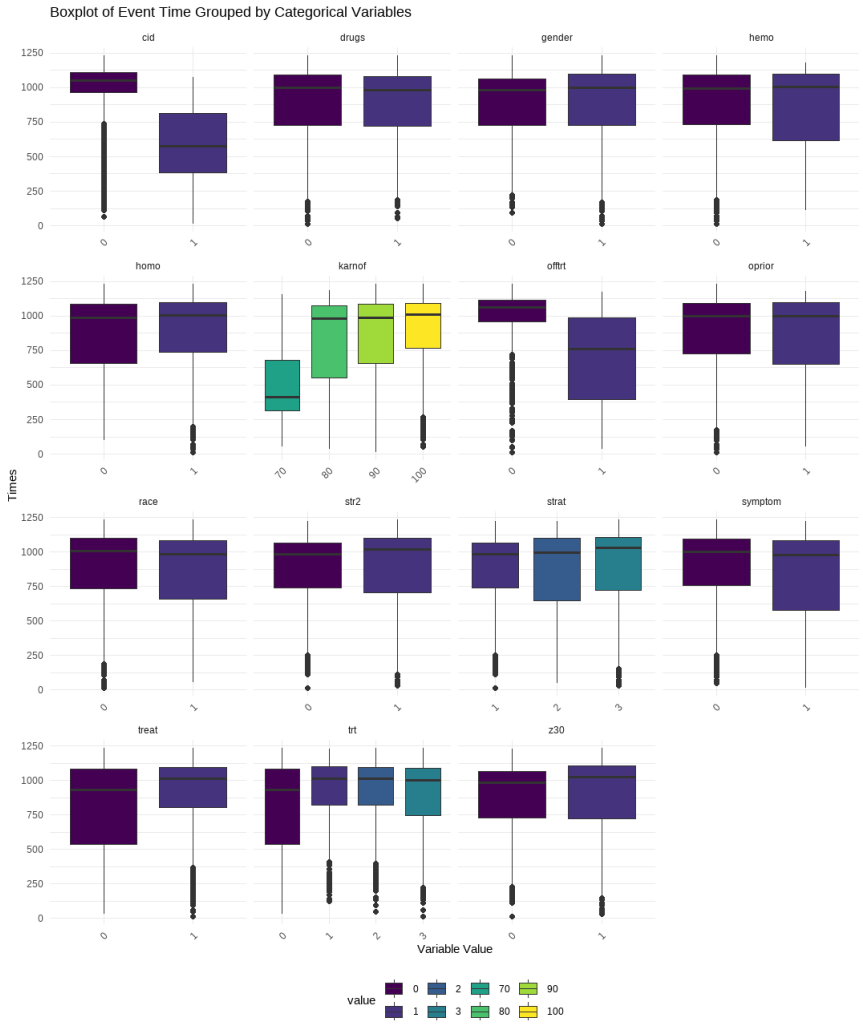

Figure 6: Boxplot of Event Time Grouped by Categorical Variables

7.3 Weighted Log-rank Test Table

Table 8: Weighting Schemes for the Weighted Trend Log-rank Test

Weighting Scheme	**Formula for Weight** ω_i
Log-rank (Mantel-Cox) test	$\omega_i = 1$
Gehan-Breslow-Wilcoxon test	$\omega_i = n_i$
Peto-Peto test	$\omega_i = S(t_i)$
Fleming-Harrington test	$\omega_i = S(t_{i-1})^p(1 - S(t_{i-1}))^q$, $p, q \geq 0$
Tarone-Ware test	$\omega_i = \sqrt{n_i}$
Modified Peto-Peto test	$\omega_i = \frac{S(t_i)n_i}{n_i+1}$

7.4 Cox Regression With and Without Time-varying Coefficient

Note: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

tt = function(x, t, ...) as.numeric(x) * t

Table 9: Cox Proportional Hazards Model Without Time-varying Coefficient Summary

	coef	exp(coef)	se(coef)	z	Pr(> |z|)	95% CI
trt1	-0.4416	0.6430	0.1245	-3.546	0.000391 ***	(0.5038, 0.8208)
trt2	-0.5060	0.6029	0.1219	-4.151	3.31e-05 ***	(0.4748, 0.7656)
trt3	-0.4117	0.6626	0.1166	-3.531	0.000414 ***	(0.5272, 0.8326)
preanti	0.0003	1.0003	0.0001	3.158	0.001589 **	(1.0001, 1.0005)
symptom1	0.3531	1.4235	0.1029	3.430	0.000603 ***	(1.1634, 1.7418)
offtrt1	0.4809	1.6175	0.0929	5.175	2.28e-07 ***	(1.3482, 1.9406)
cd420	-0.0815	0.9919	0.0005	-14.925	< 2e-16 ***	(0.9908, 0.9929)
cd820	0.0005	1.0005	0.0001	4.936	7.99e-07 ***	(1.0003, 1.0007)
karnof	-0.0135	0.9866	0.0071	-1.906	0.056699 .	(0.9731, 1.0004)
age	0.0108	1.0108	0.0051	2.092	0.036404 *	(1.0007, 1.0210)
drugs1	-0.3256	0.7221	0.1472	-2.213	0.026922 *	(0.5411, 0.9635)
cd40	0.0015	1.0015	0.0005	2.959	0.003089 **	(1.0005, 1.0025)
Statistic	**Value**	**Details**				
Concordance	0.772	se = 0.01				
Likelihood Ratio Test	506.3	df = 12, p < 2e-16				
Wald Test	432.9	df = 12, p < 2e-16				
Score (log-rank) Test	446.8	df = 12, p < 2e-16				

Table 10: Cox Proportional Hazards Model With Time-varying Coefficient Summary

	coef	exp(coef)	se(coef)	z	Pr(> |z|)	95% CI
trt1	-0.4475	0.6392	0.1249	-3.584	0.000339 ***	(0.5005, 0.8165)
trt2	-0.4794	0.6192	0.1221	-3.927	8.60e-05 ***	(0.4874, 0.7865)
trt3	-0.4075	0.6653	0.1167	-3.491	0.000481 ***	(0.5293, 0.8363)
preanti	0.000274	1.0003	0.000092	2.981	0.002871 **	(1.0001, 1.0005)
symptom1	0.3234	1.3818	0.1031	3.138	0.001702 **	(1.1291, 1.6911)
offtrt1	1.0530	2.8649	0.2391	4.401	1.08e-05 ***	(1.7928, 4.5779)
cd420	-0.01805	0.9821	0.001373	-13.146	¡ 2e-16 ***	(0.9795, 0.9848)
cd820	0.000464	1.0005	0.000092	5.028	4.96e-07 ***	(1.0003, 1.0006)
karnof	-0.01199	0.9881	0.007122	-1.684	0.092215 .	(0.9744, 1.0020)
tt(offtrt)	-0.001033	0.9990	0.000386	-2.675	0.007478 **	(0.9982, 0.9997)
age	0.01155	1.0116	0.005082	2.272	0.023088 *	(1.0016, 1.0217)
drugs1	-0.3196	0.7264	0.1470	-2.175	0.029663 *	(0.5446, 0.9689)
cd40	0.00351	1.0035	0.001176	2.984	0.002849 **	(1.0012, 1.0058)
tt(cd420)	0.000016	1.0000	0.000002	8.106	5.24e-16 ***	(1.0000, 1.0000)
tt(cd40)	-0.0000035	1.0000	0.0000019	-1.886	0.059362 .	(1.0000, 1.0000)
Statistic	**Value**	**Details**				
Concordance	0.771	se = 0.01				
Likelihood Ratio Test	595.3	df = 15, p < 2e-16				
Wald Test	480.7	df = 15, p < 2e-16				
Score (log-rank) Test	497	df = 15, p < 2e-16				

7.5 Code

```
ggcoxzph <- function (fit, resid = TRUE, se = TRUE, df = 4, nsmo = 40, var,
                      point.col = "red", point.size = 1, point.shape = 19,
    point.alpha = 1,
                      caption = NULL,
                      ggtheme = theme_survminer(), ...){

  x <- fit
  if(!methods::is(x, "cox.zph"))
    stop("Can't handle an object of class ", class(x))

  xx <- x$x
  yy <- x$y
  d <- nrow(yy)
  df <- max(df)
  nvar <- ncol(yy)
  pred.x <- seq(from = min(xx), to = max(xx), length = nsmo)
  temp <- c(pred.x, xx)
  lmat <- splines::ns(temp, df = df, intercept = TRUE)
  pmat <- lmat[1:nsmo, ]
  xmat <- lmat[-(1:nsmo), ]
  qmat <- qr(xmat)
  if (qmat$rank < df)
    stop("Spline fit is singular, try a smaller degrees of freedom")
  if (se) {
    bk <- backsolve(qmat$qr[1:df, 1:df], diag(df))
    xtx <- bk %*% t(bk)
    seval <- d * ((pmat %*% xtx) * pmat) %*% rep(1, df)
  }
  ylab <- paste("Beta(t) for", dimnames(yy)[[2]])
  if (missing(var))
    var <- 1:nvar
  else {
    if (is.character(var))
      var <- match(var, dimnames(yy)[[2]])
    if (any(is.na(var)) || max(var) > nvar || min(var) <
        1)
      stop("Invalid variable requested")
  }
  if (x$transform == "log") {
    xx <- exp(xx)
    pred.x <- exp(pred.x)
  }
  else if (x$transform != "identity") {
    xtime <- as.numeric(dimnames(yy)[[1]])
    indx <- !duplicated(xx)
    apr1 <- approx(xx[indx], xtime[indx], seq(min(xx), max(xx),
                                              length = 17)[2 * (1:8)])
    temp <- signif(apr1$y, 2)
    apr2 <- approx(xtime[indx], xx[indx], temp)
    xaxisval <- apr2$y
    xaxislab <- rep("", 8)
    for (i in 1:8) xaxislab[i] <- format(temp[i])
  }
  plots <- list()
  lapply(var, function(i) {
    invisible(round(x$table[i, 3],4) -> pval)
    ggplot() + labs(title = paste0('Schoenfeld Individual Test p: ', pval))
   + ggtheme -> gplot
```

```
    y <- yy[, i]
    yhat <- as.vector(pmat %*% qr.coef(qmat, y))
    if (resid)
      yr <- range(yhat, y)
    else yr <- range(yhat)
    if (se) {
      bk <- backsolve(qmat$qr[1:df, 1:df], diag(df))
      xtx <- bk %*% t(bk)
      seval <- ((pmat %*% xtx) * pmat) %*% rep(1, df)
      temp <- as.vector(2 * sqrt(x$var[i, i] * seval))
      yup <- yhat + temp
      ylow <- yhat - temp
      yr <- range(yr, yup, ylow)
    }
    if (x$transform == "identity") {
      gplot + geom_line(aes(x=pred.x, y=yhat)) +
        xlab("Time") +
        ylab(ylab[i]) +
        ylim(yr) -> gplot
    } else if (x$transform == "log") {
      gplot + geom_line(aes(x=log(pred.x), y=yhat)) +
        xlab("Time") +
        ylab(ylab[i]) +
        ylim(yr)  -> gplot
    } else {
      gplot + geom_line(aes(x=pred.x, y=yhat)) +
        xlab("Time") +
        ylab(ylab[i]) +
        scale_x_continuous(breaks = xaxisval,
                           labels = xaxislab) +
        ylim(yr)-> gplot
    }

    if (resid)
      gplot <- gplot + geom_point(aes(x = xx, y =y),
                                  col = point.col, shape = point.shape, size
    = point.size, alpha = point.alpha)

    if (se) {
      gplot <- gplot + geom_line(aes(x=pred.x, y=yup), lty = "dashed") +
        geom_line(aes( x = pred.x, y = ylow), lty = "dashed")
    }

    ggpubr::ggpar(gplot, ...)


  }) -> plots
  names(plots) <- var
  class(plots) <- c("ggcoxzph", "ggsurv", "list")

  if("GLOBAL" %in% rownames(x$table)) # case of multivariate Cox
    global_p <- x$table["GLOBAL", 3]
  else global_p <- NULL # Univariate Cox
  attr(plots, "global_pval") <- global_p
  attr(plots, "caption") <- caption
  plots

}
# rewrite functions
ggcoxzph <- function (fit, resid = TRUE, se = TRUE, df = 4, nsmo = 40, var,
```

```
                      point.col = "red", point.size = 1, point.shape = 19,
 point.alpha = 1,
                      caption = NULL,
                      ggtheme = theme_survminer(), ...){

x <- fit
if(!methods::is(x, "cox.zph"))
  stop("Can't handle an object of class ", class(x))

xx <- x$x
yy <- x$y
d <- nrow(yy)
df <- max(df)
nvar <- ncol(yy)
pred.x <- seq(from = min(xx), to = max(xx), length = nsmo)
temp <- c(pred.x, xx)
lmat <- splines::ns(temp, df = df, intercept = TRUE)
pmat <- lmat[1:nsmo, ]
xmat <- lmat[-(1:nsmo), ]
qmat <- qr(xmat)
if (qmat$rank < df)
  stop("Spline fit is singular, try a smaller degrees of freedom")
if (se) {
  bk <- backsolve(qmat$qr[1:df, 1:df], diag(df))
  xtx <- bk %*% t(bk)
  seval <- d * ((pmat %*% xtx) * pmat) %*% rep(1, df)
}
ylab <- paste("Beta(t) for", dimnames(yy)[[2]])
if (missing(var))
  var <- 1:nvar
else {
  if (is.character(var))
    var <- match(var, dimnames(yy)[[2]])
  if (any(is.na(var)) || max(var) > nvar || min(var) <
      1)
    stop("Invalid variable requested")
}
if (x$transform == "log") {
  xx <- exp(xx)
  pred.x <- exp(pred.x)
}
else if (x$transform != "identity") {
  xtime <- as.numeric(dimnames(yy)[[1]])
  indx <- !duplicated(xx)
  apr1 <- approx(xx[indx], xtime[indx], seq(min(xx), max(xx),
                                            length = 17)[2 * (1:8)])
  temp <- signif(apr1$y, 2)
  apr2 <- approx(xtime[indx], xx[indx], temp)
  xaxisval <- apr2$y
  xaxislab <- rep("", 8)
  for (i in 1:8) xaxislab[i] <- format(temp[i])
}
plots <- list()
lapply(var, function(i) {
  invisible(round(x$table[i, 3],4) -> pval)
  ggplot() + labs(title = paste0('Schoenfeld Individual Test p: ', pval))
 + ggtheme -> gplot
  y <- yy[, i]
  yhat <- as.vector(pmat %*% qr.coef(qmat, y))
  if (resid)
```

```
      yr <- range(yhat, y)
    else yr <- range(yhat)
    if (se) {
      bk <- backsolve(qmat$qr[1:df, 1:df], diag(df))
      xtx <- bk %*% t(bk)
      seval <- ((pmat %*% xtx) * pmat) %*% rep(1, df)
      temp <- as.vector(2 * sqrt(x$var[i, i] * seval))
      yup <- yhat + temp
      ylow <- yhat - temp
      yr <- range(yr, yup, ylow)
    }
    if (x$transform == "identity") {
      gplot + geom_line(aes(x=pred.x, y=yhat)) +
        xlab("Time") +
        ylab(ylab[i]) +
        ylim(yr) -> gplot
    } else if (x$transform == "log") {
      gplot + geom_line(aes(x=log(pred.x), y=yhat)) +
        xlab("Time") +
        ylab(ylab[i]) +
        ylim(yr)  -> gplot
    } else {
      gplot + geom_line(aes(x=pred.x, y=yhat)) +
        xlab("Time") +
        ylab(ylab[i]) +
        scale_x_continuous(breaks = xaxisval,
                           labels = xaxislab) +
        ylim(yr)-> gplot
    }

    if (resid)
      gplot <- gplot + geom_point(aes(x = xx, y =y),
                                  col = point.col, shape = point.shape, size
    = point.size, alpha = point.alpha)

    if (se) {
      gplot <- gplot + geom_line(aes(x=pred.x, y=yup), lty = "dashed") +
        geom_line(aes( x = pred.x, y = ylow), lty = "dashed")
    }

    ggpubr::ggpar(gplot, ...)


  }) -> plots
  names(plots) <- var
  class(plots) <- c("ggcoxzph", "ggsurv", "list")

  if("GLOBAL" %in% rownames(x$table)) # case of multivariate Cox
    global_p <- x$table["GLOBAL", 3]
  else global_p <- NULL # Univariate Cox
  attr(plots, "global_pval") <- global_p
  attr(plots, "caption") <- caption
  plots

}


library(tidyverse)
library(knitr)
library(kableExtra)
```

```r
library(summarytools)
library(corrplot)
library(survminer)
library(ggplot2)
library(survMisc)
library(flexsurv)
library(dplyr)
# Decide which columns to include in each table
cols_part1 <- names(data_2)[1:(ncol(data_2)/2+1)]
cols_part2 <- names(data_2)[(ncol(data_2)/2 + 2):ncol(data_2)]

# Create two separate tables
data_2 %>%
  select(cols_part1) %>%
  head(n = 10) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

data_2 %>%
  select(cols_part2) %>%
  head(n = 10) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

# Summary
data_2 %>%
  mutate(across(everything(), as.character)) %>%  # Convert all columns to
    character
  pivot_longer(cols = colnames(.)) %>%
  group_by(name) %>%
  summarize(unique_values = n_distinct(value)) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

# Describe
data_2 %>%
  descr(transpose=TRUE, stats=c("min", "med", "mean", "max", "q1", "q3", "sd
    ")) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

# Split data into continuous and discrete variables
continuous_vars <- data_2 %>% select_if(~is.numeric(.))
discrete_vars <- data_2 %>% select_if(~is.factor(.))

# Summary for continuous variables
summary_continuous <- continuous_vars %>%
  summarise(across(everything(), list(mean = ~mean(.),
                                      sd = ~sd(.),
                                      min = ~min(.),
                                      q25 = ~quantile(., 0.25),
                                      median = ~median(.),
                                      q75 = ~quantile(., 0.75),
                                      max = ~max(.))))

# Summary for discrete variables
summary_discrete <- discrete_vars %>%
  map(~table(.)) %>%
  enframe(name = "variable", value = "counts")

```

```r
# Print summaries
print(summary_continuous)
print(summary_discrete)

# Histogram for continuous variable
tmp = data_2 %>%
  select(age, cd40, cd420, cd80, cd820, wtkg) %>%
  pivot_longer(everything())
ggplot(tmp, aes(x=value)) +
  geom_histogram(aes(y=..density..), alpha=0.5) +
  geom_density() +
  facet_wrap(. ~ name, scales="free") +
  theme(text = element_text(size=10))

# install.packages("GGally")
# Plot continuous heapmap between treatment groups
data_2 %>%
  select(age, cd40, cd420, cd80, cd820, wtkg, preanti, time, trt) %>%
  ggpairs(
    columns = 1:8,
    aes(color = factor(trt, labels = c("ZDV only", "ZDV + ddI","ZDV + Zal","
   ddI only")), alpha = 0.5),
    title = "Continuous Variables Heatmap Between Treatment Groups",
    upper = list(continuous = wrap("cor", size = 3))
  ) +
  theme(axis.text = element_text(size = 6))

data_for_cor <- data_2 %>%
  select(-c(age, cd40, cd420, cd80, cd820, wtkg, preanti, time, zprior)) %>%
  fastDummies::dummy_cols(remove_selected_columns = TRUE, remove_first_dummy
    = TRUE)

corrs = cor(data_for_cor)
high_corr <- abs(corrs) > 0.5
diag(high_corr) <- FALSE

# Filter higher corr variable
keep_vars <- apply(high_corr, 1, any)
data_filtered <- data_for_cor[, keep_vars]

corrs_filtered <- cor(data_filtered)

g <- data_filtered %>%
  mutate(trt = data_2$trt) %>%
  ggpairs(
    col = 1:7,
    aes(color = factor(trt, labels = c("ZDV only", "ZDV + ddI","ZDV + Zal","
   ddI only")), alpha = 0.5),
    title = "Categorical Variables Heatmap Between Treatment Groups",
    upper = list(continuous = wrap("cor", size = 3)),
    lower = list(continuous = "blank"),
    diag = list(continuous = "blankDiag")
  ) +
  theme(axis.text = element_text(size = 6))

gpairs_upper <- function(g) {
  # Remove the last row
  g$plots <- g$plots[-((g$nrow * (g$ncol - 1) + 1):(g$nrow * g$ncol))]
  g$yAxisLabels <- g$yAxisLabels[-g$nrow]
  g$nrow <- g$nrow - 1
```

```
348
349   # Remove the first column
350   g$plots <- g$plots[-(seq(1, length(g$plots), by = g$ncol))]
351   g$xAxisLabels <- g$xAxisLabels[-1]
352   g$ncol <- g$ncol - 1
353
354   g
355 }
356 gpairs_upper(g)
357
358
359
360
361 # bar charts of discrete features
362 plot_bar(data_2 %>%
363             select(-c(age, cd40, cd420, cd80, cd820, wtkg, preanti, time,
    zprior)),
364           theme_config=list(text = element_text(size = 10)),
365           order_bar=TRUE)
366
367
368
369 # bar charts of discrete features
370 plot_bar(data_2 %>%
371             select(-c(age, cd40, cd420, cd80, cd820, wtkg, preanti, time,
    zprior, treat)), nrow = 3L, ncol = 5L,
372           theme_config=list(text = element_text(size = 10)),
373           order_bar=TRUE, by="trt",title = "Categorical Variable Bar Chart
    Between Treatment Groups")
374
375
376
377
378
379
380 # Outlier summary
381 outlier_summary = data_2 %>% diagnose_outlier() %>% filter(outliers_cnt > 0)
382 outlier_var_iqr = data.frame(t(apply(data_2[, outlier_summary$variables], 2,
383                                      quantile, c(0.25, 0.75), na.rm=TRUE)))
384 outlier_var_iqr %>%
385   rename(Q25="X25.", Q75="X75.") %>%
386   rownames_to_column("variables") %>%
387   left_join(outlier_summary, by="variables") %>%
388   select(variables, Q25, Q75, outliers_mean, outliers_cnt, with_mean,
     without_mean) %>%
389   mutate(across(where(is.numeric), round, 2))
390
391
392 # use dlookr to examine outliers
393 plot_outlier(data_2,
394              diagnose_outlier(data_2) %>%
395                filter(outliers_cnt >= 0) %>%
396                select(variables) %>%
397                unlist())
398
399 # bar plot remain categorical variables
400 plot_outlier(data_2,
401              diagnose_outlier(data_2) %>%
402                filter(outliers_cnt < 0) %>%
403                select(variables) %>%
```

```
                unlist())

# Normality distribution of each variables
normality(data_2)  #shapiro.test
plot_normality(data_2)




# Compare censored and uncensored observations
tmp = data_2 %>%
  select(-c(age, cd40, cd420, cd80, cd820, wtkg, preanti, time, zprior)) %>%
  pivot_longer(-one_of("trt")) %>%
  group_by(trt, name, value) %>%
  summarise(count=n())


ggplot(tmp, aes(x=value, y=count)) +
  geom_bar(aes(fill=trt), position="dodge", stat="identity") +
  facet_wrap(. ~ name, scales="free") +
  theme(text = element_text(size=12),
        legend.title = element_blank(),
        legend.position = "top",
        axis.text.x = element_text(angle = 45, hjust = 1))

# Calculating percentages
tmp <- tmp %>%
  group_by(name, trt) %>%
  mutate(total = sum(count),             # Total count per group
         percent = count / total * 100)  # Percentage calculation

# Plotting with percentages
ggplot(tmp, aes(x=value, y=percent, fill=trt)) +
  geom_bar(position="dodge", stat="identity") +
  facet_wrap(. ~ name, scales="free") +
  theme(text = element_text(size=12),
        legend.title = element_blank(),
        legend.position = "top",
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  ylab("Percentage")  # Updating y-axis label

# Survival Time Different
# Reshape data_2 from wide to long format
long_data <- data_2 %>%
  select(-c(age, cd40, cd420, cd80, cd820, wtkg, preanti, zprior)) %>%
  pivot_longer(
    cols = -time, # Exclude the survival_months column
    names_to = "variable",
    values_to = "value"
  )

# Create the boxplots
ggplot(long_data, aes(x=value, y=time, fill = value)) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Variable Value", y = "Times") +
  ggtitle("Boxplot of Event Time Grouped by Categorical Variables")


```

```r
aids_data <- read_csv(file = "aids_clinical_trials_group_study_175.csv")



# First, decide which columns to include in each table
cols_part1 <- names(aids_data)[1:(ncol(aids_data)/2)]
cols_part2 <- names(aids_data)[(ncol(aids_data)/2 + 1):ncol(aids_data)]

# Create two separate tables
aids_data %>%
  select(cols_part1) %>%
  head(n = 10) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

aids_data %>%
  select(cols_part2) %>%
  head(n = 10) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))


aids_data %>%
  pivot_longer(cols=colnames(.)) %>%
  group_by(name) %>%
  summarize(unique_values=n_distinct(value)) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

aids_data %>%
  descr(transpose=TRUE, stats=c("min", "med", "mean", "max", "q1", "q3", "sd
    ")) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

tmp = aids_data %>%
  select(age, cd40, cd420, cd80, cd820, wtkg) %>%
  pivot_longer(everything())
ggplot(tmp, aes(x=value)) +
  geom_histogram(aes(y=..density..), alpha=0.5) +
  geom_density() +
  facet_wrap(. ~ name, scales="free") +
  theme(text = element_text(size=10))

tmp = aids_data %>%
  select(preanti, time) %>%
  pivot_longer(everything())
ggplot(tmp, aes(x=value)) +
  geom_histogram(aes(y=..density..), alpha=0.5) +
  geom_density() +
  facet_wrap(. ~ name, scales="free") +
  theme(text = element_text(size=10))

# compare censored and uncensored observations
tmp = aids_data %>%
  select(cid, age, cd40, cd420, cd80, cd820, wtkg) %>%
  pivot_longer(-one_of("cid")) %>%
  mutate(cid=ifelse(cid==0,"Censored", "Died"))
ggplot(tmp, aes(x=value)) +
  geom_density(aes(color=cid)) +
```

```
  facet_wrap(. ~ name, scales="free") +
  theme(text = element_text(size=10),
        legend.title = element_blank(),
        legend.position = "top")

# compare censored and uncensored observations
tmp = aids_data %>%
  select(cid, preanti, time) %>%
  pivot_longer(-one_of("cid")) %>%
  mutate(cid=ifelse(cid==0,"Censored", "Died"))
ggplot(tmp, aes(x=value)) +
  geom_density(aes(color=cid)) +
  facet_wrap(. ~ name, scales="free") +
  theme(text = element_text(size=10),
        legend.title = element_blank(),
        legend.position = "top")

# dlookr: correlation heatmap
data_for_cor <- aids_data %>%
  select(age, cd40, cd420, cd80, cd820, wtkg, preanti, time)

corrs = cor(data_for_cor)
corrplot(corrs, type="upper", method="color", addCoef.col = "black", order="hclust", hclust.method = 'ward.D2',
         tl.col="black", na.label = "")


# dlookr: correlation heatmap
data_for_cor <- aids_data %>%
  select(-age, -cd40, -cd420, -cd80, -cd820, -wtkg, -preanti, -time, -zprior, -treat)

corrs = cor(data_for_cor)
high_corr <- abs(corrs) > 0.3
diag(high_corr) <- FALSE

# Filter higher corr variable
keep_vars <- apply(high_corr, 1, any)
data_filtered <- data_for_cor[, keep_vars]

corrs_filtered <- cor(data_filtered)

corrplot(corrs_filtered, type="upper", method="color", addCoef.col = "black", order="hclust", hclust.method = 'ward.D2',
         tl.col="black", na.label = "")

# Time effect
time_df <- aids_data %>%
  select(age, cd40, cd420, cd80, cd820, wtkg, preanti, time) %>%
  pivot_longer(cols = c(age, cd40, cd420, cd80, cd820, wtkg, preanti), names_to="variable")

ggplot(time_df, aes(x=time, y=value)) +
  geom_jitter() +
  geom_smooth(method="lm", se=TRUE) +
  facet_wrap(vars(variable), nrow=2, scales="free")



lm_data <- aids_data %>%
```

```r
  select(age, cd40, cd420, cd80, cd820, wtkg, preanti, time, trt)

summary(lm(time ~ ., data = lm_data))




# Time effect
time_df <- aids_data %>%
  select(-age, -cd40, -cd420, -cd80, -cd820, -wtkg, -preanti, -zprior, -treat, -cid) %>%
  pivot_longer(cols = c(-time), names_to="variable")

ggplot(time_df, aes(x=time, y=value)) +
  geom_jitter() +
  geom_smooth(method="lm", se=TRUE) +
  facet_wrap(vars(variable), nrow=3, scales="free")



lm_data <- aids_data %>%
  select(-zprior, -treat, -cid)

summary(lm(time ~ ., data = lm_data))

fit_overall = survfit(Surv(time, event=cid) ~ 1, data=aids_data)
print(fit_overall)

# creates the survival table
f <- summary(fit_overall)
df_overall_fit <- data.frame(f$time, f$n.risk, f$n.event, f$n.censor, f$surv, f$lower, f$upper)
names(df_overall_fit) <- c("time", "n.risk", "n.event", "n.censor", "survival", "ci_95_lower", "ci_95_upper")
head(df_overall_fit, n=10)

ggsurvplot(fit_overall,
           title = "Overall K-M Survival Estimation",
           xlab="Days",
           ylab="Overall survival probability",
           ylim=c(0.6,1),
           conf.int=TRUE)

km_trt = survfit(Surv(time, event=cid) ~ strata(trt), data=aids_data)
print(km_trt)

ggsurvplot(
  survfit(Surv(time, event=cid) ~ trt, data=aids_data),
   #survival model we want to plot
  pval = TRUE,              #displays p-value of log-rank test, if p-value < 0.05, then the difference between the two curves are statistically significant
  conf.int = TRUE,          #plots a confidence interval for each curve
  xlab = "Time in days",
  break.time.by = 150,      # break X axis in time intervals by 100.
  ggtheme = theme_light(),  # customize theme with a grid for better readability
  risk.table = "abs_pct",   # absolute number and percentage at risk
  risk.table.y.text.col = T,# colour risk table text annotations
  risk.table.y.text = FALSE,# show bars instead of names in text annotations
```

```
  fontsize = 2.5,
  ylim=c(0.55,1),
  ncensor.plot = TRUE,      # plot the number of censored subjects at time t
  legend.labs=c("ZDV only", "ZDV + ddI","ZDV + Zal","ddI only"), legend.
    title="trt",
  palette=c("dodgerblue2", "orchid2","grey","green"),
  title="Kaplan-Meier Curve by treatment",
  risk.table.height=.3)

# Plotting the FH
ggsurvplot(
  survfit(Surv(time, event=cid) ~ trt, data=aids_data, type="fh"),
              #survival model we want to plot
  pval = TRUE,              #displays p-value of log-rank test, if p-value <
    0.05, then the difference between the two curves are statistically
    significant
  conf.int = TRUE,          #plots a confidence interval for each curve
  xlab = "Time in days",
  break.time.by = 150,      # break X axis in time intervals by 100.
  ggtheme = theme_light(),  # customize theme with a grid for better
    readability
  risk.table = "abs_pct",   # absolute number and percentage at risk
  risk.table.y.text.col = T,# colour risk table text annotations
  risk.table.y.text = FALSE,# show bars instead of names in text annotations
  fontsize = 2.5,
  ylim=c(0.55,1),
  ncensor.plot = TRUE,      # plot the number of censored subjects at time t
  legend.labs=c("ZDV only", "ZDV + ddI","ZDV + Zal","ddI only"), legend.
    title="trt",
  palette=c("dodgerblue2", "orchid2","grey","green"),
  title="Fleming-Harrington Curve by Treatment",
  risk.table.height=.3
)

b=ten(Surv(time, event=cid) ~ factor(trt), data=aids_data)
comp(b,p=c(0,1,1,0.5,0.5),q=c(1,0,1,0.5,2),scores=c(1,3.5,3,2.5))



# "lrt" - the long-rank family of tests
vanilla_test <-attr(b,"lrt")
vanilla_dataframe <- data.frame(
  W = numeric(),
  chiSq = numeric(),
  df = integer(),
  pChisq = numeric()
)
vanilla_dataframe <- data.frame(
  W = vanilla_test$W,
  chiSq = as.numeric(vanilla_test$chiSq),
  df = as.integer(vanilla_test$df),
  pChisq = as.numeric(vanilla_test$pChisq)
) %>%
  mutate(W = case_when(
    W == "1" ~ "Log-rank (Mantel-Cox) test",
    W == "n" ~ "Gehan-Breslow-Wilcoxon test",
    W == "sqrtN" ~ "Tarone-Ware test",
    W %in% c("S1", "S2") ~ ifelse(W == "S1", "Peto-Peto test", "Modified
    Peto-Peto test"),
    grepl("FH_p=", W) ~ paste("Fleming-Harrington test", W)
```

```
  ))

knitr::kable(vanilla_dataframe) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

#"tft" - test for trend
trend_test=attr(b,"tft")$tft
trend_dataframe <- data.frame(
  W = numeric(),
  Q = numeric(),
  Var = numeric(),
  Z = numeric(),
  pNorm = numeric()
)
trend_dataframe <- data.frame(
  W = trend_test$W,
  Q = as.numeric(trend_test$Q),
  Var = as.integer(trend_test$Var),
  Z = as.numeric(trend_test$Z),
  pNorm = as.numeric(trend_test$pNorm)
) %>%
  mutate(W = case_when(
    W == "1" ~ "Log-rank (Mantel-Cox) test",
    W == "n" ~ "Gehan-Breslow-Wilcoxon test",
    W == "sqrtN" ~ "Tarone-Ware test",
    W %in% c("S1", "S2") ~ ifelse(W == "S1", "Peto-Peto test", "Modified Peto-Peto test"),
    grepl("FH_p=", W) ~ paste("Fleming-Harrington test", W)
  ))

knitr::kable(trend_dataframe) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

# Create a function to transform survfit object for ggplot
transform_survfit_km <- function(survfit_obj) {
  data.frame(source = rep("KM", length(survfit_obj$time)),
             time = survfit_obj$time,
             surv = survfit_obj$surv,
             strata = rep(names(survfit_obj$strata), survfit_obj$strata))
}

# Transform survfit object
km_data <- transform_survfit_km(km_trt)


# Plotting the log(-log) survival curves
ggplot(km_data, aes(x = log(time), y = log(-log(surv)), color = strata)) +
  geom_step() +
  scale_color_manual(values = c("dodgerblue2", "orchid2", "grey", "green"),
                     name = "Treatment",
                     labels = c("ZDV only", "ZDV + ddI", "ZDV + Zal", "ddI only")) +
  labs(title = "Log(-log) Survival Curves by Treatment",
       x = "Log(Time)",
       y = "Log(-log Survival)") +
  theme_bw() +
  theme_minimal()
```

```r
ggsurvplot(survfit(Surv(time, cid) ~ trt, data=aids_data),
           ggtheme = theme_bw(),
           fun = "event",
           legend.labs=c("ZDV only", "ZDV + ddI","ZDV + Zal","ddI only"),
   legend.title="trt",
           palette=c("dodgerblue2", "orchid2","grey","green"),
           title="Cumulative Events by treatment", )

ggsurvplot(survfit(Surv(time, cid) ~ trt, data=aids_data),
           ggtheme = theme_bw(),
           fun = "cumhaz",
           legend.labs=c("ZDV only", "ZDV + ddI","ZDV + Zal","ddI only"),
   legend.title="trt",
           palette=c("dodgerblue2", "orchid2","grey","green"),
           title="Cumulative Hazard Function by treatment")

fit.weibull <- flexsurvreg(Surv(time, cid) ~ factor(trt), data=aids_data,
   dist = "weibull")
fit.ggama <- flexsurvreg(Surv(time, cid) ~ factor(trt), data=aids_data, dist
    = "gengamma")
fit.lnorm <- flexsurvreg(Surv(time, cid) ~ factor(trt), data=aids_data, dist
    = "lognormal")

# Plot the survival curves for the Weibull model.
plot(fit.weibull, ylim=c(0.6,1), xlab="Time", ylab="Survival Probability",
   main="Weibull Model",col=c("dodgerblue2", "orchid2", "grey", "green"))
legend("bottomleft", legend=c("ZDV only", "ZDV + ddI", "ZDV + Zal", "ddI
   only"),
       col=c("dodgerblue2", "orchid2", "grey", "green"), lty=1)

# Plot the survival curves for the generalized gamma model.
plot(fit.ggama, ylim=c(0.6,1), xlab="Time", ylab="Survival Probability",
   main="Generalized Gamma Model",col=c("dodgerblue2", "orchid2", "grey", "
   green"))
legend("bottomleft", legend=c("ZDV only", "ZDV + ddI", "ZDV + Zal", "ddI
   only"),
       col=c("dodgerblue2", "orchid2", "grey", "green"), lty=1)

# Plot the survival curves for the log-normal model.
plot(fit.lnorm, ylim=c(0.6,1), xlab="Time", ylab="Survival Probability",
   main="Log-normal Model",col=c("dodgerblue2", "orchid2", "grey", "green"))
legend("bottomleft", legend=c("ZDV only", "ZDV + ddI", "ZDV + Zal", "ddI
   only"),
       col=c("dodgerblue2", "orchid2", "grey", "green"), lty=1)


# model selection(coxph)

# modify data for further modeling, the binary column zprior has only 1
   level in the dataset, thus exclude it
df = read.csv(file = "aids_clinical_trials_group_study_175.csv") %>%
  dplyr::select(-zprior) %>%
  mutate(trt = as.factor(trt),
         hemo = as.factor(hemo),
         homo = as.factor(homo),
         drugs = as.factor(drugs),
         oprior = as.factor(oprior),
         z30 = as.factor(z30),
         race = as.factor(race),
         gender = as.factor(gender),
```

```
          str2 = as.factor(str2),
          strat = as.factor(strat),
          symptom = as.factor(symptom),
          treat = as.factor(treat),
          offtrt = as.factor(offtrt))

summary(df)

cox.mod = coxph(Surv(time, cid) ~., data = df)
final_model = step(cox.mod, direction = "backward", trace = TRUE)
step(cox.mod, direction = "forward", trace = TRUE)
stepwise_cox_mod = stepAIC(cox.mod, direction = "both")
# Generates summary on the final model
summary(final_model)

# nonparametric methods-test









library(gridExtra)
data$trt <- factor(data$trt, labels = c("ZDV only", "ZDV + ddI", "ZDV + Zal"
    , "ddI only"))
km_fit <- survfit(Surv(time, cid) ~ trt, data)
km_plot = km_fit %>% autoplot() +
  labs(x = "Time (Days)", y = "Estimated Survival Probability",
       title = "(a) Kaplan-Meier Survival Estimate") +
  theme_bw()


data$trt <- factor(data$trt, labels = c("ZDV only", "ZDV + ddI", "ZDV + Zal"
    , "ddI only"))
fh_fit <- survfit(Surv(time, cid) ~ trt, data, type = "fh")
fh_plot = fh_fit %>% autoplot() +
  labs(x = "Time (Days)", y = "Estimated Survival Probability",
       title = "(b) Nelson-Aalen Survival Estimate") + theme_bw()

grid.arrange(km_plot, fh_plot, ncol = 2)


print(km_fit)

lifetab_0 <- lifetab2(Surv(time, cid) ~ 1, data[data$trt == 0,], breaks =
    seq(0, 1300, by = 100))

lifetab_1 <- lifetab2(Surv(time, cid) ~ 1, data[data$trt == 1,], breaks =
    seq(0, 1300, by = 100))

lifetab_2 <- lifetab2(Surv(time, cid) ~ 1, data[data$trt == 2,], breaks =
    seq(0, 1300, by = 100))

lifetab_3 <- lifetab2(Surv(time, cid) ~ 1, data[data$trt == 3,], breaks =
    seq(0, 1300, by = 100))

print(lifetab_0)
print(lifetab_1)
```

```
print(lifetab_2)
print(lifetab_3)
```


```
# tx group 0
lifetab_0 %>%
  mutate(time = tstart + (tstop - tstart)/2) %>%
  ggplot(aes(x = time, y = hazard)) +
  geom_point() + geom_line() + theme_bw() +
  labs(x = "Time (Days)", y = "Hazard Rate",
       title = "Hazard Function for Treatment Group 0 based on life-table
    estimate")

# tx group 1
lifetab_1 %>%
  mutate(time = tstart + (tstop - tstart)/2) %>%
  ggplot(aes(x = time, y = hazard)) +
  geom_point() + geom_line() + theme_bw() +
  labs(x = "Time (Days)", y = "Hazard Rate",
       title = "Hazard Function for Treatment Group 1 based on life-table
    estimate")

# tx group 2
lifetab_0 %>%
  mutate(time = tstart + (tstop - tstart)/2) %>%
  ggplot(aes(x = time, y = hazard)) +
  geom_point() + geom_line() + theme_bw() +
  labs(x = "Time (Days)", y = "Hazard Rate",
       title = "Hazard Function for Treatment Group 2 based on life-table
    estimate")

# tx group 3
lifetab_3 %>%
  mutate(time = tstart + (tstop - tstart)/2) %>%
  ggplot(aes(x = time, y = hazard)) +
  geom_point() + geom_line() + theme_bw() +
  labs(x = "Time (Days)", y = "Hazard Rate",
       title = "Hazard Function for Treatment Group 3 based on life-table
    estimate")


This variable has four levels, and the test is performed across these four
    treatment groups.
The log-rank test results in a Chi-squared value of 49.2 with 3 degrees of
    freedom and a p-value of 1e-10. This is highly significant, indicating
    that there are significant differences in survival among the four
    treatment groups.

# Ensure 'trt' is a factor
data$trt <- as.factor(data$trt)

# Create the survival object
surv_obj <- Surv(data$time, data$cid)

# Perform the log-rank test across multiple trt groups
log_rank_test_trt <- survdiff(surv_obj ~ trt, data = data)
```
log-rank score test of drugs

```
p=0.06, not diff

In this case, with a p-value of 0.06, the result is not statistically
    significant at the 0.05 level, meaning there is not enough evidence to
    conclude that the survival experiences of the two groups (those with and
    without a history of intravenous drug use) are different. However, the p-
    value is close to the threshold, suggesting that there might be a trend
    worth exploring with a larger sample size or additional research.

# Ensure 'drugs' is a factor if it's not already
data$drugs <- as.factor(data$drugs)

# Create the survival object
surv_obj <- Surv(data$time, data$cid)

# Perform the log-rank test
log_rank_test_drugs <- survdiff(surv_obj ~ drugs, data = data)



### log-rank score test of genders

p=0.09, not diff

# Ensure 'gender' is a factor if it's not already
data$gender <- as.factor(data$gender)

# Create the survival object
surv_obj <- Surv(data$time, data$cid)

# Perform the log-rank test
log_rank_test_gender <- survdiff(surv_obj ~ gender, data = data)


### log-rank score test of homo

p=0.06, not diff

# Ensure 'homo' is a factor
data$homo <- as.factor(data$homo)

# Create the survival object
surv_obj <- Surv(data$time, data$cid)

# Perform the log-rank test
log_rank_test_homo <- survdiff(surv_obj ~ homo, data = data)




# diagnosis

# Model selection with tests


aids<-read_csv("aids_clinical_trials_group_study_175.csv")
n <- nrow(aids)
data<- aids |>
  janitor::clean_names()|>
  dplyr::select(time, cid, trt, everything())|>
```

```
  mutate_at(c(3), .funs = ~as.factor(.))
fit1<-coxph(Surv(time, cid) ~ . , data = data)
stepwise <- stepAIC(fit1, direction = "both",trace = F) # BIC


summary(stepwise)
broom::tidy(stepwise) |>kbl()



fit2 <- coxph(Surv(time,cid)~trt+preanti+karnof+age+drugs+symptom+offtrt+
    cd40+cd420+ cd820,data=data)
summary(fit2)




# Graphical Methods



library(ggfortify)
library(StepReg)
data_fit <- data |>
  dplyr::select(time, cid, trt, preanti, symptom, offtrt, cd420, cd820,
     karnof, age, drugs, cd40) |>
  mutate_at(c(3, 5, 6, 11), .funs = ~as.factor(.))


# --- treat ---
# km plot
fit_km_treat <- survfit(Surv(time, cid) ~ trt, data_fit)
autoplot(fit_km_treat) + theme_bw() +
  labs(x = "Time (days)", y = "Survival Function",
       title = "Kaplan-Meier Survival Estimate")

# loglog vs. log time
# png("ph_checking_1.png", width = 500, height = 400)
plot(fit_km_treat, fun = "cloglog", col = c("black", "red","pink","blue"),
     xlab = "Time (days in log scale)", ylab = "log{-log(S(t))}",
     main = "Log of Negative Log of Estimated Survival Functions")
legend("topleft", legend = c("ZDV only", "ZDV+ddl", "ZDV+Zal", "ddl only"),
    col = c("black", "red","pink","blue"),
       lty = 1, cex = 1)

# observed vs. fitted
fit_ph_treat <- coxph(Surv(time, cid) ~ trt, data_fit)


# png("ph_checking_2.png", width = 500, height = 400)
plot(fit_km_treat, col = c("blue", "darkgreen", "red","pink"),
     xlab = "Time (days)", ylab = "Survival Function",
     ylim = c(0.4,1),
     main = "Observed vs. Fitted")
lines(survfit(fit_ph_treat, newdata = data.frame(trt = as.factor(0))), # 0
      col = "purple", conf.int = FALSE)
lines(survfit(fit_ph_treat, newdata = data.frame(trt = as.factor(1))), # 1
      col = "black", conf.int = FALSE)
lines(survfit(fit_ph_treat, newdata = data.frame(trt = as.factor(2))), # 2
      col = "yellow", conf.int = FALSE)
```

```r
lines(survfit(fit_ph_treat, newdata = data.frame(trt = as.factor(3))), # 3
      col = "orange", conf.int = FALSE)
legend("bottomleft", legend = c("Observed ZDV only", "Observed ZDV+ddl",
                                "Observed ZDV+Zal", "Observed ddl only",
                                "Fitted ZDV only", "Fitted ZDV+ddl",
                                "Fitted ZDV+Zal", "Fitted ddl only"),
       col = c("blue", "darkgreen", "red","pink",
               "purple", "black", "yellow", "orange"), lty = 1, cex = 0.6,
    lwd = 2,
       inset=c(0.05, 0), title="trt", xpd = TRUE)

library(survminer)
# --- to be updated ---

data_fit <- data |>
  dplyr::select(time, cid, trt, preanti, symptom, offtrt, cd420, cd820,
    karnof, age, drugs, cd40) |>
  mutate_at(c(3, 5, 6, 11), .funs = ~as.factor(.))

aids_fit <- coxph(Surv(time, cid == 1) ~ trt+ preanti+ symptom+ offtrt +
    cd420+ cd820 + karnof+ age+drugs+cd40, data_fit)

cox.zph(aids_fit)

plot(cox.zph(aids_fit))


# interaction
aids_interaction_fit <- coxph(Surv(time, cid == 1) ~ trt+ preanti+ symptom+
    offtrt + cd420+ cd820 + karnof+ age+drugs+cd40+tt(cd40)+ tt(cd420), data_
    fit, tt = function(x, t, ...) x*t)

summary(aids_interaction_fit)$coefficients |>
  kable("latex",
        digits = 4,
        escape = F,
        booktabs = T,
        caption = "Regresion Coefficients Estimates of the Cox Model with
    Time Interactions") |>
  kable_styling(position = "center",
                latex_options = "hold_position")

sjPlot::plot_models(aids_fit,aids_interaction_fit, show.values=T, grid=T,
    value.size = 2.5, m.labels = c("Original Cox", "Time-varying Cox"))

aids_fit<-coxph(Surv(time, cid) ~ ., data_fit)
# residual
ggcoxzph(cox.zph(aids_fit), var = c("cd420"), df = 2, nsmo = 1000)

# ggsave("ph_checking_4.png", width = 6, height = 4)
ggcoxzph(cox.zph(aids_fit), var = c("cd820"), df = 2, nsmo = 1000)

# ggsave("ph_checking_4.png", width = 6, height = 4)
ggcoxzph(cox.zph(aids_fit), var = c("cd40"), df = 2, nsmo = 1000)




# Parametric Analysis

```

```r
### Fit exponential and Weilbull distributions


library(flexsurv)
#parametric survival function
fit_exp_others = flexsurvreg(Surv(time, cid == 1) ~ 1,
                             data = subset(data_fit, treat == 1), dist = "exp")
fit_exp_ZDV = flexsurvreg(Surv(time, cid == 1) ~ 1,
                          data = subset(data_fit, treat == 0), dist = "exp")
fit_weib_others = flexsurvreg(Surv(time, cid == 1) ~ 1,
                              data = subset(data_fit, treat == 1), dist = "weibull")
fit_weib_ZDV = flexsurvreg(Surv(time, cid == 1) ~ 1,
                           data = subset(data_fit, treat == 0), dist = "weibull")


#plot km, exp fitted and weib fitted
plot(fit_exp_ZDV, conf.int = FALSE, ci = FALSE, col = "red", col.obs = "pink",
     lty = "longdash", xlim = c(0,1300),
     xlab = "Days", ylab = "Survival Probability",
     main = "KM and Parametric Est")
par(new = TRUE)
plot(fit_exp_others, conf.int = FALSE, ci = FALSE, col = "blue", col.obs = "skyblue", lty = "longdash", xlim = c(0,1000), xaxt = "n")
plot(fit_weib_ZDV, add = TRUE, ci = FALSE, col = "brown4")
plot(fit_weib_others, add = TRUE, ci = FALSE, col = "blue4")
legend("bottomleft", legend = c("Obs ZDV", "Obs others", "Exp ZDV", "Exp others", "Weib ZDV", "Weib others"),
       col = c("pink", "skyblue", "red", "blue", "brown4", "blue4"),
       lty = c("solid", "solid", "longdash", "longdash", "solid", "solid"),
       lwd = c(2,2,2,2,2,2))


## Parametric Regression Models

### Parametric PH Models


library(eha)
#backward selection, significance level = 0.05
fit_ph1 = eha::phreg(Surv(time, cid==1) ~ .,
                     data = data_fit, dist = "weibull")
summary(fit_ph1)|>
  kable("latex",
        digits = 4,
        escape = F,
        booktabs = T,
        caption = "Weibull (Parametric) PH Model Fitting") |>
  kable_styling(position = "center",
                latex_options = "hold_position")
# it can be our final model


#compare the estimated baseline hazards with a non-parametric ph model
fit_cox =eha::coxreg(Surv(time, cid==1) ~ ., data = data_fit)
eha::check.dist(fit_ph1, fit_cox)
```