**Midterm Project Report**

**Statistical Analysis**

Descriptive statistics were used to summarize data from the "midtermdata.csv" file, reflecting a double-blind, placebo-controlled trial comparing the efficacy of a three-drug regimen (indinavir, open-label zidovudine or stavudine, and lamivudine) against a two-drug regimen (zidovudine or stavudine and lamivudine) in 1151 HIV-infected patients. Stratification was based on CD4 cell count during the screening phase. Continuous variables (e.g., age and CD4 count) were summarized using means and standard deviations, while categorical variables (e.g., sex and race/ethnicity) were presented as frequencies and percentages. Due to low participant counts, certain categorical variables, such as "raceth" categories 3, 4, 5, "ivdrug" categories 2 and 3, and "karnof" categories 80 and 70, were combined in the descriptive analysis.

For the comparison of progression-free survival (PFS) between the three-drug and two-drug regimens, Kaplan-Meier curves were generated, and log-rank tests were conducted. Wilcoxon rank-sum tests were also performed. A similar approach was taken to assess overall survival (OS). These methodologies provide a comprehensive evaluation of treatment effects on PFS and OS. Comprehensive analyses of PFS and OS included univariate Cox proportional hazards models and subsequent multivariate models adjusting for various factors. Model reduction techniques, such as the log partial likelihood ratio test and stepwise variable selection based on the AIC criteria, were applied to identify the most parsimonious model. Effect modification tests were conducted by introducing interaction terms and performing log-likelihood ratio tests.

Statistical analyses were performed using R version 4.1.1 and various packages (survival, ggsurvfit, survMisc, lmtest, MASS, tidyverse). The significance level for hypothesis tests was set at 0.05.

**Results**

The overall cohort contained 1151 participants, with 577 in the two-drug regimen group and 574 in the three-drug regimen group (**Table 1).** Notably, race/ethnicity, Karnofsky Performance Scale, and IV drug use history variables were recoded due to low patient counts. The average age at enrollment was comparable between the two groups, with a mean of 38.6 years (SD=8.82) for the two-drug regimen and 38.7 years (SD=8.81) for the three-drug regimen. Baseline CD4 counts exhibited a similar pattern, with mean values of 84.3 cells/mL (SD=70.1) and 88.6 cells/mL (SD=70.0) for the two-drug and three-drug regimens, respectively. The distribution of Karnofsky Performance Scale categories, IV drug use history, CD4 stratum at screening, sex, race/ethnicity, and months of prior zidovudine (ZDV) use were also comparable across the two groups.

In the analysis of progression-free survival (PFS) among participants in the clinical trial, the log-rank test between the "Two-drug regimen without IDV" (577 participants) and the "Three-drug regimen including IDV" (574 participants) revealed a statistically significant difference in PFS between the two groups (χ² = 10.5, df = 1, p = 0.001) . Specifically, the observed number of PFS events in the "Two-drug regimen without IDV" group was 63, while the expected events were 47.1 **(Table 1a)**. In the "Three-drug regimen including IDV" group, the observed events were 33, and the expected events were 48.9. The chi-squared statistic indicates a significant discrepancy between the observed and expected events, suggesting a notable difference in PFS outcomes between the two treatment regimens. While evaluating the overall survival, the log-rank test showed a significant difference in OS between the "Two-drug regimen without IDV" and "Three-drug regimen including IDV" groups (χ² = 4.06, df = 1, p = 0.044) **(Table 1b).** The observed number of OS events in the "Two-drug regimen without IDV" group was 18, with an expected count of 12.9. In the "Three-drug regimen including IDV" group, the observed events were 8, and the expected events were 13.1. The significant p-value indicates a notable difference in OS outcomes between the two treatment groups, exemplifying the impact of the treatment regimen on overall survival. The attached Kaplan Meier curves illustrate the overall survival and progression-free survival for the respective treatment groups in **figures 1 and 2.**

For PFS, the univariate Cox analysis revealed several covariates with statistically significant associations **(Table 2)**. Lower CD4 counts (HR = 0.98, p<0.0001), a Karnofsky Performance Scale of 70 (HR = 4.76, p<0.0001), and the presence of the " CD4≤ 50" stratification in "strat2" (HR=3.84, p<0.0001) were all significantly associated with worse PFS. Additionally, the three-drug regimen demonstrated a significantly better PFS compared to the two-drug regimen (HR=0.50, p=0.001). For OS, the univariate Cox analysis similarly revealed significant associations **(Table 3).** Older age (HR=1.07, p<0.001), lower CD4 (HR=0.99, p=0.004), a Karnofsky Performance Scale of 70 (HR=13.76, p<0.001), and the CD4≤ 50" stratification in "Strat2" (HR=3.49, p=0.003) were significantly associated with worse OS. Here, the three-drug regimen also demonstrated a significantly better OS compared to the two-drug regimen (HR=0.44, p=0.050).

In the multivariate analysis for PFS, there were several key associations that showed influence on this outcome. For PFS, older age (HR=1.02, p=0.046), lower baseline CD4 counts (HR=0.99, p<0.0001), and a Karnofsky Performance Score of 70 were associated with worse PFS (HR=3.24, p<0.0001), while the three-drug regimen continued to show a favorable impact on PFS compared to the two-drug regimen (HR=0.52, p=0.002). Other variables, including sex and the "Strat2: CD4≤50" stratification, did not exhibit significant associations with PFS **(Table 4).** In the multivariate analysis for OS, old age (HR=1.08, p<0.0001), and a Karnofsky Performance Scale of 70 was associated with worse OS (HR=8.29, p=0.005), while the three-drug regimen continued to show a positive impact on OS compared to the two-drug regimen (HR=0.43, p=0.050). The “CD4 ≤ 50" stratification and baseline CD4 count did not show a significant association with OS. **(Table 5).**

The log partial likelihood ratio test and stepwise variable selection based on the AIC criterion used to reduce the models for OS and PFS resulted in insignificant findings on my selected reduced models being better than the full models. However, for the sake of simplicity and parsimony, the reduced models for PFS and OS were used as our final models. For both the reduced PFS and OSmodels, the results of the likelihood ratio tests suggest that there is no evidence to support the presence of significant effect modifiers for the relationship between treatment regimens in the context of the specified interaction terms.

**Conclusion**

These findings indicate that the three-drug regimen is associated with improved PFS and OS compared to the two-drug regimen in this HIV-infected patient population. Additionally, age, CD4 counts, Karnofsky Performance Scale, and the stratification of CD4 counts emerge as significant predictors of both PFS and OS. No significant effect modifiers were identified in either multivariate reduced model.

Tables & Figures

**Table 1**

**A screenshot of a medical report

Description automatically generated**

**Figure 1**

A graph of a patient

Description automatically generated with medium confidence

**Figure 2**

A graph of a patient's life cycle

Description automatically generated with medium confidence

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2** | | | | | | | | |
| *Univariate Cox Analysis for PFS* | | | | | | |  | |
|  | | |  | *95% CI* | | |  | |
| Covariate | | | *HR* | *LL* | *UL* | | *P-value* | |
| Age | | | 1.02 | 1.00 | 1.04 | | 0.061 | |
| Baseline CD4 count | | | 0.98 | 0.98 | 0.99 | | **<0.0001** | |
| Hemophiliac: Yes | | | 1.02 | 0.32 | 3.22 | | 0.972 | |
| IV drug use history: Currently/Previously | | | 1.50 | 0.80 | 2.80 | | 0.209 | |
| Karnofsky Scale: 70 | | | 4.76 | 2.70 | 8.38 | | **<0.0001** | |
| Karnofsky Scale: 90 | | | 1.61 | 0.91 | 2.86 | | 0.101 | |
| Months of prior ZDV use | | | 1.00 | 0.99 | 1.01 | | 0.511 | |
| Race/Ethnicity: Black | | | 0.80 | 0.48 | 1.33 | | 0.384 | |
| Race/Ethnicity: Other/unknown | | | 1.30 | 0.80 | 2.11 | | 0.292 | |
| Sex: Female | | | 1.08 | 0.62 | 1.88 | | 0.778 | |
| CD4 stratum at screening: CD4 ≤ 50 | | | 3.84 | 2.49 | 5.94 | | **<0.0001** | |
| Treatment indicator: three-drug regimen including IDV | | | 0.50 | 0.33 | 0.77 | | **0.001** | |
| *Note:* Significant P-values are bolded. | | | | | | |  | |
| **Table 3** |  |  | |  | |  | |
| *Univariate Cox Analysis for OS* | | | | | |  | |
|  |  | *95% CI* | | | |  | |
| Covariate | *HR* | *LL* | | *UL* | | *P-value* | |
| Age | 1.07 | 1.03 | | 1.11 | | **<0.001** | |
| Baseline CD4 count | 0.99 | 0.98 | | 1.00 | | **0.004** | |
| Hemophiliac: Yes | 1.29 | 0.17 | | 9.51 | | 0.804 | |
| IV drug use history: Currently/Previously | 0.80 | 0.30 | | 2.12 | | 0.652 | |
| Karnofsky Scale: 70 | 13.76 | 3.15 | | 60.18 | | **<0.001** | |
| Karnofsky Scale: 90 | 3.22 | 0.70 | | 14.89 | | 0.135 | |
| Months of prior ZDV use | 0.99 | 0.97 | | 1.01 | | 0.211 | |
| Race/Ethnicity: Black | 1.67 | 0.69 | | 4.03 | | 0.256 | |
| Race/Ethnicity: Other/unknown | 1.49 | 0.55 | | 4.04 | | 0.429 | |
| Sex: Female | 0.82 | 0.31 | | 2.17 | | 0.686 | |
| CD4 stratum at screening: CD4 ≤ 50 | 3.49 | 1.52 | | 8.03 | | **0.003** | |
| Treatment indicator: three-drug regimen including IDV | 0.44 | 0.19 | | 1.00 | | **0.050** | |
| *Note:* Significant P-values are bolded. | | | | | |  | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4** | | | | | | | | | |
| *Multivariate Cox Analysis for PFS* | | | | | | | | | |
|  | | |  | *95% CI* | | | |  | |
| Covariate | | | *HR* | *LL* | | *UL* | | *P-value* | |
| Age | | | 1.02 | 1.00 | | 1.05 | | **0.046** | |
| Sex: Female | | | 0.91 | 0.52 | | 1.59 | | 0.751 | |
| Treatment indicator: three-drug regimen including IDV | | | 0.52 | 0.34 | | 0.79 | | **0.002** | |
| Baseline CD4 Count | | | 0.99 | 0.98 | | 0.99 | | **<0.0001** | |
| Karnofsky Scale: 70 | | | 3.24 | 1.82 | | 5.76 | | **<0.0001** | |
| Karnofsky Scale: 90 | | | 1.55 | 0.87 | | 2.74 | | 0.137 | |
| CD4 stratum at screening: CD4 ≤ 50 | | | 0.97 | 0.50 | | 1.91 | | 0.938 | |
| *Note:* Significant P-values are bolded. | | | | | | | |  | |
| **Table 5** |  |  | | |  | |  | |
| *Multivariate Cox Analysis for OS* |  |  | | |  | |  | |
|  |  | *95% CI* | | | | |  | |
| Covariate | *HR* | *LL* | | | *UL* | | *P-value* | |
| Age | 1.08 | 1.04 | | | 1.12 | | **<0.001** | |
| Sex: Female | 0.62 | 0.23 | | | 1.65 | | 0.338 | |
| Treatment indicator: three-drug regimen including IDV | 0.43 | 0.19 | | | 1.00 | | **0.050** | |
| Baseline CD4 count | 0.99 | 0.98 | | | 1.01 | | 0.285 | |
| Karnofsky Scale: 70 | 8.29 | 1.87 | | | 36.82 | | **0.005** | |
| Karnofsky Scale: 90 | 2.91 | 0.63 | | | 13.51 | | 0.172 | |
| CD4 stratum at screening: CD4 ≤ 50 | 1.90 | 0.51 | | | 7.02 | | 0.337 | |
| *Note:* Significant P-values are bolded. | |  | | |  | |  | |

**Appendix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1a.** |  |  |  |  |  |
| *PFS Log-Rank Test* |  |  |  |  |  |
|  | *N* | Observed | Expected | (O-E)^2/E | (O-E)^2/V |
| tx=Two-drug regimen without IDV | 577 | 63 | 47.1 | 5.37 | 10.5 |
| tx=Three-drug regimen including IDV | 574 | 33 | 48.9 | 5.17 | 10.5 |
| Chisq = 10.5 on 1 degrees of freedom, **p= 0.001** | | |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1b.** |  |  |  |  |  |
| *OS Log-Rank Test* |  |  |  |  |  |
|  | *N* | Observed | Expected | (O-E)^2/E | (O-E)^2/V |
| tx=Two-drug regimen without IDV | 577 | 18 | 12.9 | 2.05 | 4.06 |
| tx=Three-drug regimen including IDV | 574 | 8 | 13.1 | 2.01 | 4.06 |
| Chisq = 4.1 on 1 degrees of freedom, **p= 0.04** | | |  |  |  |

**R Code:**

---

title: "Midterm Project"

author: "Camille Okonkwo"

date: "2024-02-26"

output: github\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

library(tidyverse)

library(broom)

library(table1)

library(survival)

library(ggsurvfit)

library(survival)

library(lmtest)

library(gt)

library(survMisc)

library(MASS)

```

### Step 0: Data Preparation

```{r step\_0}

# loading data

midtermdata = read\_csv("data/midtermdata.csv")

# combining categories, creating factor variables, and labelling

# Recode the variables

midtermdata$tx <- factor(

midtermdata$tx,

levels = c("0", "1"))

midtermdata$strat2 <- factor(

midtermdata$strat2,

levels = c("1", "0"))

midtermdata$sex <- factor(

midtermdata$sex,

levels = c("2", "1"))

midtermdata$raceth <- factor(

ifelse(midtermdata$raceth %in% c(3, 4, 5, 6), 6, midtermdata$raceth),

levels = c("1", "2", "6"))

midtermdata$ivdrug <- factor(

ifelse(midtermdata$ivdrug %in% c(2, 3), 2, midtermdata$ivdrug),

levels = c("2", "1"))

midtermdata$hemophil <-

factor(midtermdata$hemophil,

levels = c("0", "1"))

midtermdata$karnof <- factor(

ifelse(midtermdata$karnof %in% c(80, 70), 70, midtermdata$karnof),

levels = c("100", "90", "70"))

```

# Descriptive statistics

```{r}

# data exploration

summary(midtermdata)

# Summarize Baseline Characteristics

library(table1)

# Define variable labels

table1::label(midtermdata$tx) <- "Treatment"

table1::label(midtermdata$age) <- "Age at Enrollment"

table1::label(midtermdata$cd4) <- "Baseline CD4 Count (cells/mL)"

table1::label(midtermdata$karnof) <- "Karnofsky Performance Scale\*"

table1::label(midtermdata$ivdrug) <- "IV Drug Use History\*"

table1::label(midtermdata$strat2) <- "CD4 Stratum at Screening"

table1::label(midtermdata$sex) <- "Sex"

table1::label(midtermdata$raceth) <- "Race/Ethnicity\*"

table1::label(midtermdata$hemophil) <- "Hemophiliac"

table1::label(midtermdata$priorzdv) <- "Months of prior ZDV use"

# Create a summary table with variable labels

summary\_table <- table1(

~ age + cd4 + karnof + ivdrug + strat2 + sex + raceth + priorzdv | tx,

data = midtermdata,

footnote = "\*Race/Ethnicity, Karnofsky Performance Scale, and IV drug use history variables were recoded due to low patient counts.",

caption = "Descriptive Statistics of Baseline Characteristics in HIV Clinical Trial Participants")

print(summary\_table)

```

is there sufficient evidence that the three-drug regimen has better PFS compared to the two-drug regimen?

```{r PFS\_step1}

library(survival)

# PFS K-M table

km\_pfs = survfit(Surv(time, censor) ~ tx, data = midtermdata, conf.type = "log-log")

summary(km\_pfs)

# log-rank test

log\_rank\_PFS <- survdiff(Surv(time, censor) ~ tx, data = midtermdata)

# Extract relevant information

chi\_squared <- log\_rank\_PFS$chisq

degrees\_of\_freedom <- length(log\_rank\_PFS$n) - 1

p\_value <- pchisq(chi\_squared, degrees\_of\_freedom, lower.tail = FALSE)

# Create a custom table

PFS\_logrank <- data.frame(

"Chi-Squared" = chi\_squared,

"Degrees of Freedom" = degrees\_of\_freedom,

"P-Value" = p\_value

)

# Display the custom table

knitr::kable(PFS\_logrank, caption = "Log-Rank Test Results for PFS", digits = 5)

library(ggsurvfit)

# PFS K-M plot

survfit2(Surv(time, censor) ~ tx, data = midtermdata) |>

ggsurvfit() +

add\_pvalue(location = "annotation",

caption = "Log-rank {p.value}") +

add\_confidence\_interval() +

add\_risktable() +

add\_censor\_mark() +

add\_legend\_title(title = "PFS Curve")

```

#### Our hypotheses:

( where s1(t) is tx = 0 and s2(t) is tx = 1)

\* H0 :S1(t) = S2(t), for all t ≤ τ

\* Hα :One of the Sk (t) is different for some t ≤ τ

#### Test statistic:

\* Q\_log-rank = 10.5

#### Degree of freedom:

\* df = 1

#### P-value:

\* Pr(χ21 ≥ 10.5) = 0.00117 < 0.05

#### Conclusion:

We reject H0 at the significance level 0.05. The survival curves for patients in two hormone therapy groups are significantly different, and there is sufficient evidence to conclude that the three-drug regimen has better PFS compared to the two-drug regimen.

```{r pfs\_wilx}

# repeat with wilxocon

library(survMisc)

pfs\_wilx = ten(survfit(Surv(time, censor) ~ tx, data = midtermdata))

comp(pfs\_wilx)

knitr::kable(attributes(pfs\_wilx)$lrt[, c(1, 6:8)], "simple", digits = 4)

# Same conclusion with Wilxocon Rank Test

```

are there any other variables significantly associated with PFS?

```{r pfs\_step2}

library(survival)

library(lmtest)

library(gt)

# age

age\_pfs = coxph(Surv(time, censor) ~ age,

data = midtermdata,

ties = "efron")

summary(age\_pfs)

# cd4

cd4\_pfs = coxph(Surv(time, censor) ~ cd4,

data = midtermdata,

ties = "efron")

summary(cd4\_pfs)

# hemophil

hemophil\_pfs = coxph(Surv(time, censor) ~ hemophil,

data = midtermdata,

ties = "efron")

summary(hemophil\_pfs)

# ivdrug

ivdrug\_pfs = coxph(Surv(time, censor) ~ ivdrug,

data = midtermdata,

ties = "efron")

summary(ivdrug\_pfs)

# karnof

karnof\_pfs = coxph(Surv(time, censor) ~ karnof,

data = midtermdata,

ties = "efron")

summary(karnof\_pfs)

# priorzdv

priorzdv\_pfs = coxph(Surv(time, censor) ~ priorzdv,

data = midtermdata,

ties = "efron")

summary(priorzdv\_pfs)

# raceth

raceth\_pfs = coxph(Surv(time, censor) ~ raceth,

data = midtermdata,

ties = "efron")

summary(raceth\_pfs)

# sex

sex\_pfs = coxph(Surv(time, censor) ~ sex,

data = midtermdata,

ties = "efron")

summary(sex\_pfs)

# strat2

strat2\_pfs = coxph(Surv(time, censor) ~ strat2,

data = midtermdata,

ties = "efron")

summary(strat2\_pfs)

# tx

tx\_pfs = coxph(Surv(time, censor) ~ tx,

data = midtermdata,

ties = "efron")

summary(tx\_pfs)

# multivariate Cox model

multivar\_pfs = coxph(Surv(time, censor) ~ tx + sex + age + karnof + cd4 + strat2,

data = midtermdata,

ties = "efron")

summary(multivar\_pfs)

```

reducing multivariable model using log partial likelihood ratio test

```{r}

# LRT using lmtest package

library(lmtest)

r\_model <- coxph(Surv(time, censor) ~ tx + sex + age + karnof + cd4, data = midtermdata, ties = "efron")

f\_model <- coxph(Surv(time, censor) ~ tx + sex + age + karnof + cd4 + strat2 + raceth + ivdrug + hemophil + priorzdv, data = midtermdata, ties = "efron")

lrtest(f\_model, r\_model)

# since there is insufficient evidence to conclude the reduced model is better than the full, I will try another method.

# selection tests

library(MASS)

stepwise\_cox\_model <- stepAIC(f\_model, direction = "both")

summary(stepwise\_cox\_model)

lrtest(f\_model, stepwise\_cox\_model)

# There is insufficient evidence to conclude that the stepwise model is better than the full model.

lrtest(stepwise\_cox\_model, r\_model)

# Since the step wise model and reduced model are not significantly better than the full model, for the sake of parsimony I will choose the reduced model.

```

Are there any significant effect modifiers for the relation between the treatment regimens and PFS?

```{r}

# test interactions (pfs 2)

multivar\_pfs2 = coxph(Surv(time, censor) ~ tx + sex + age + karnof + cd4 + cd4\*karnof,

data = midtermdata,

ties = "efron")

summary(multivar\_pfs2)

lrtest(r\_model, multivar\_pfs2)

```

There are no significant effect modifiers for the relation between the treatment regimens and PFS.

```{r}

# final model

final\_pfs <- coxph(Surv(time, censor) ~ tx + sex + age + karnof + cd4, data = midtermdata, ties = "efron")

```

Is there sufficient evidence that the three-drug regimen has better OS compared to the two-drug regimen?

```{r OS\_step1}

# OS K-M table

km\_os = survfit(Surv(time\_d, censor\_d) ~ tx, data = midtermdata, conf.type = "log-log")

summary(km\_os)

# log-rank test OS

log\_rank\_OS <- survdiff(Surv(time\_d, censor\_d) ~ tx, data = midtermdata)

library(ggsurvfit)

# K-M plot

survfit2(Surv(time\_d, censor\_d) ~ tx, data = midtermdata) |>

ggsurvfit() +

add\_pvalue(location = "annotation",

caption = "Log-rank {p.value}") +

add\_risktable() +

add\_confidence\_interval() +

add\_censor\_mark() +

add\_legend\_title(title = "OS Curve")

```

#### Our hypotheses:

( where s1(t) is tx = 0 and s2(t) is tx = 1)

\* H0 :S1(t) = S2(t), for all t ≤ τ

\* Hα :One of the Sk (t) is different for some t ≤ τ

#### Test statistic:

\* Q\_log-rank = 4.1

#### Degree of freedom:

\* df = 1

#### P-value:

\* Pr(χ21 ≥ 4.1) = 0.0438 < 0.05

#### Conclusion:

We reject H0 at the significance level 0.05. The survival curves for patients in two hormone therapy groups are significantly different, and there is sufficient evidence to conclude that the three-drug regimen has better OS compared to the two-drug regimen.

are there any other variables significantly associated with OS?

```{r OS\_step2}

# univariate Cox models

# age

age\_os = coxph(Surv(time\_d, censor\_d) ~ age,

data = midtermdata,

ties = "efron")

summary(age\_os)

# cd4

cd4\_os = coxph(Surv(time\_d, censor\_d) ~ cd4,

data = midtermdata,

ties = "efron")

summary(cd4\_os)

# hemophil

hemophil\_os = coxph(Surv(time\_d, censor\_d) ~ hemophil,

data = midtermdata,

ties = "efron")

summary(hemophil\_os)

# ivdrug

ivdrug\_os = coxph(Surv(time\_d, censor\_d) ~ ivdrug,

data = midtermdata,

ties = "efron")

summary(ivdrug\_os)

# karnof

karnof\_os = coxph(Surv(time\_d, censor\_d) ~ karnof,

data = midtermdata,

ties = "efron")

summary(karnof\_os)

# priorzdv

priorzdv\_os = coxph(Surv(time\_d, censor\_d) ~ priorzdv,

data = midtermdata,

ties = "efron")

summary(priorzdv\_os)

# raceth

raceth\_os = coxph(Surv(time\_d, censor\_d) ~ raceth,

data = midtermdata,

ties = "efron")

summary(raceth\_os)

# sex

sex\_os = coxph(Surv(time\_d, censor\_d) ~ sex,

data = midtermdata,

ties = "efron")

summary(sex\_os)

# strat2

strat2\_os = coxph(Surv(time\_d, censor\_d) ~ strat2,

data = midtermdata,

ties = "efron")

summary(strat2\_os)

# tx

tx\_os = coxph(Surv(time\_d, censor\_d) ~ tx,

data = midtermdata,

ties = "efron")

summary(tx\_os)

# multivariate Cox model

multivar\_os = coxph(Surv(time\_d, censor\_d) ~ tx + sex + age + karnof + cd4 + strat2,

data = midtermdata,

ties = "efron")

summary(multivar\_os)

```

reducing multivariable model using log partial likelihood ratio test

```{r}

full\_model <- coxph(Surv(time\_d, censor\_d) ~ tx + age + sex + cd4 + strat2 + karnof + raceth + ivdrug + hemophil + priorzdv, data = midtermdata, ties = "efron")

reduced\_model <- coxph(Surv(time\_d, censor\_d) ~ tx + sex + age + karnof, data = midtermdata, ties = "efron")

summary(reduced\_model)

# LRT using lmtest package

lrtest(full\_model, reduced\_model)

# since there is insufficient evidence to conclude the reduced model is better than the full, I will try another method.

# selection tests

library(MASS)

stepwise\_cox\_model2 <- stepAIC(full\_model, direction = "both")

summary(stepwise\_cox\_model2)

lrtest(full\_model, stepwise\_cox\_model2)

# There is insufficient evidence to conclude that the stepwise model is better than the full model.

# is the stepwise model better than the reduced model?

lrtest(reduced\_model, stepwise\_cox\_model2)

```

For the sake of parsimony, I will choose my initial reduced model over the stepwise model.

Are there any significant effect modifiers for the relation between the treatment regimens and OS?

```{r}

# test interactions (OS 2)

multivar\_os2 = coxph(Surv(time\_d, censor\_d) ~ tx + sex + age + karnof + karnof\*tx,

data = midtermdata,

ties = "efron")

summary(multivar\_os2)

summary(multivar\_os)

```

There are no significant effect modifiers for the relation between the treatment regimens and OS.

```{r}

#final model

final\_os <- coxph(Surv(time\_d, censor\_d) ~ tx + sex + age + karnof, data = midtermdata, ties = "efron")

summary(final\_os)

```