Survival Analysis Project

Analysis of University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study (UIS) Data Using Survival Models

Isaiah Thompson Ocansey

May 3, 2024

Abstract

This study explores the factors influencing drug relapse times among a cohort of individuals undergoing treatment. Kaplan-Meier survival curves and Cox proportional hazards modeling reveal significant associations between treatment duration and type and relapse times. Specifically, longer treatment durations correlate with delayed relapse, suggesting the importance of sustained interventions in preventing relapse. Utilizing a step-wise variable selection technique, our multiple Cox model identifies several variables highly predictive of drug relapse. These include age, IV drug use history, number of drug treatments, treatment received, length of stay, and interaction terms between IV drug use history and treatment site, as well as treatment received and length of stay. Notably, age and site variables, initially insignificant in the simple Cox model, emerge as significant predictors in the multiple Cox model. Additionally, interaction terms such as ivhx:site and treat:los demonstrate significance, underscoring the complexity of factors influencing drug relapse.

1 Introduction

The UMARU IMPACT Study is a collaborative research project focused on the efficacy of residential treatment programs of varying durations. Conducted over five years (1989-1994), the study's goal was to identify effective strategies to curb drug abuse and high-risk behaviors associated with HIV transmission.

2 Data Description

The dataset used in this study includes variables such as age, treatment duration, drug relapse time, and others. Data were imported and processed using R with the following code:

```
uis <- read.table(file="uissurv.txt", header=TRUE, na.strings=".",
                  col.names=c("id", "age", "beck", "hercoc", "ivhx", "ndrugtx",
                              "race", "treat", "site", "los", "time", "status"))
```

2.1 Missing Values Exploration and Imputaion

The data has 628 observations and 12 different variables, however, there were missing values found in the data set as shown below: As shown in Table **??**, the dataset has several missing entries across different variables which need to be addressed before performing further analysis.

Table 1: Missing Data Summary

id	age	beck	hercoc	ivhx	ndrugtx	race	treat	site	los	time	status
0	5	33	18	18	17	6	0	0	0	0	0

Table 1 is a table showing total missing values per variable.

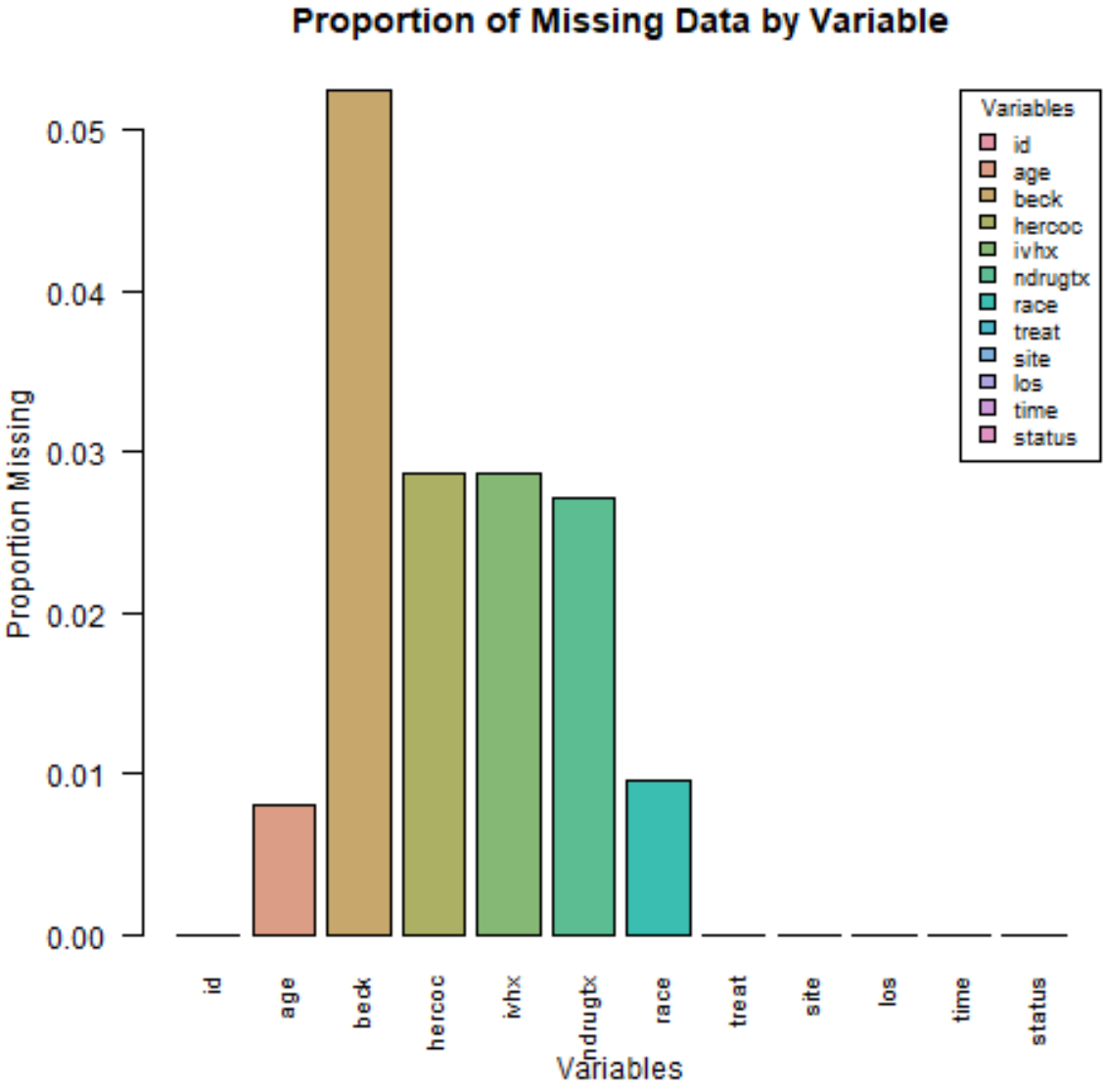

Figure 1: Total Missing Values per Variable

Figure 1 represents the total number of missing values per variable. It is evident that

the variable 'beck' has the highest proportion of missing values, followed by 'hercoc'. However, the proportions of missing data are not substantial enough to warrant the deletion of these variables. Therefore, missing value imputation will be employed to address the gaps in these variables.

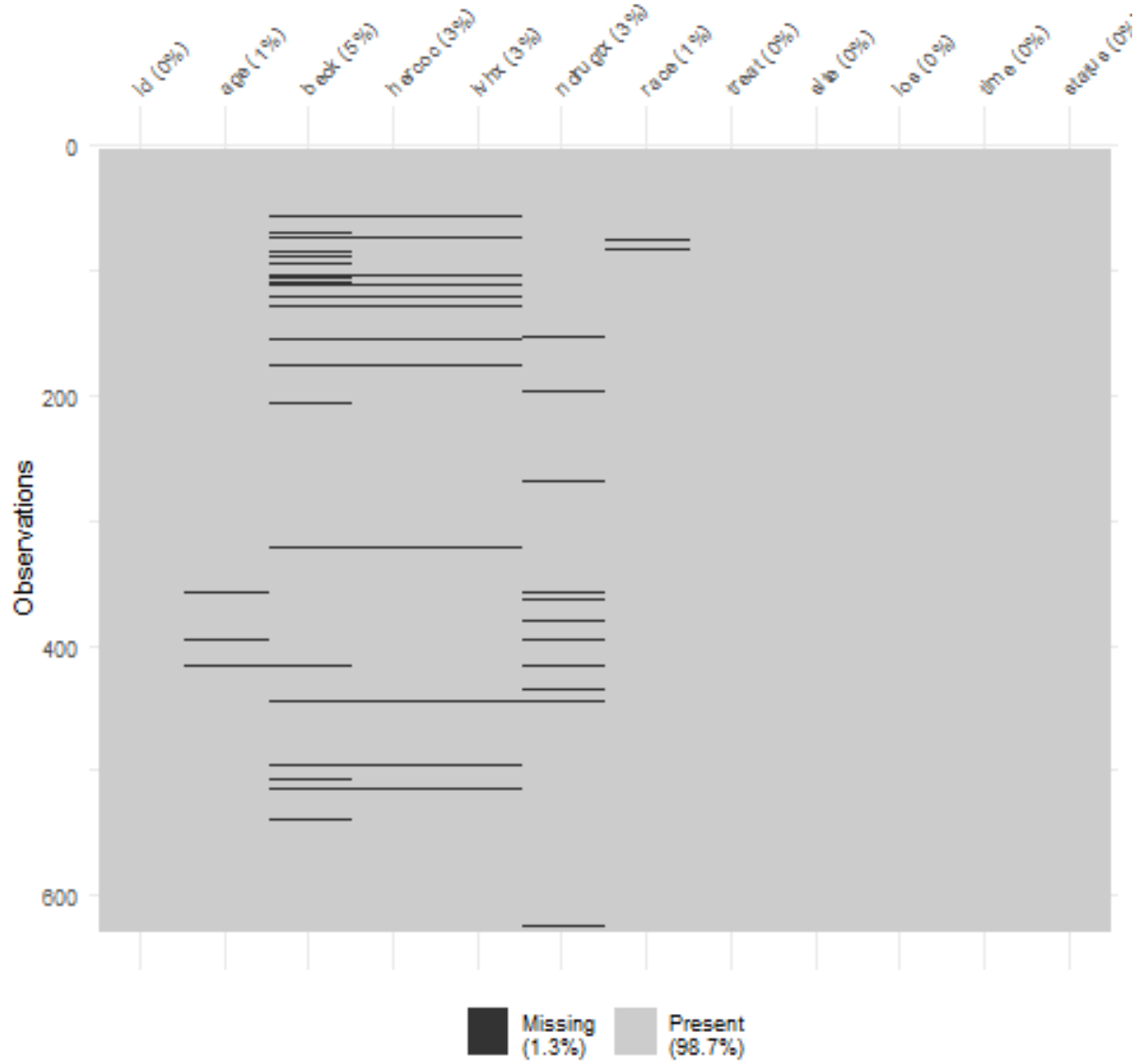

Figure 2: Total Missing Percentage

Figure 2 displays the percentage of total missing values across the entire dataset. It is noted that approximately 1.3% of the data is missing, while about 98.7% of the data is present. Given that the percentage of missing data is relatively small, we will proceed with imputing the missing values. The missing value imputation was done using the following R code:

```
    data.imputed<-preProcess(uis,method = c("medianImpute"))
data.imputed<-predict(data.imputed,uis)
data.imputed;colSums(is.na(data.imputed))
```

Table 2: Total Missing After Missing Value Imputation

id	age	beck	hercoc	ivhx	ndrugtx	race	treat	site	los	time	status
0	0	0	0	0	0	0	0	0	0	0	0

Table 2 shows that there are no missing values in the variables after imputation.

3 Exploratory Data Analysis

We begin exploring the data to uderstand the distributions of the variables in the dataset that will aid in further analysis

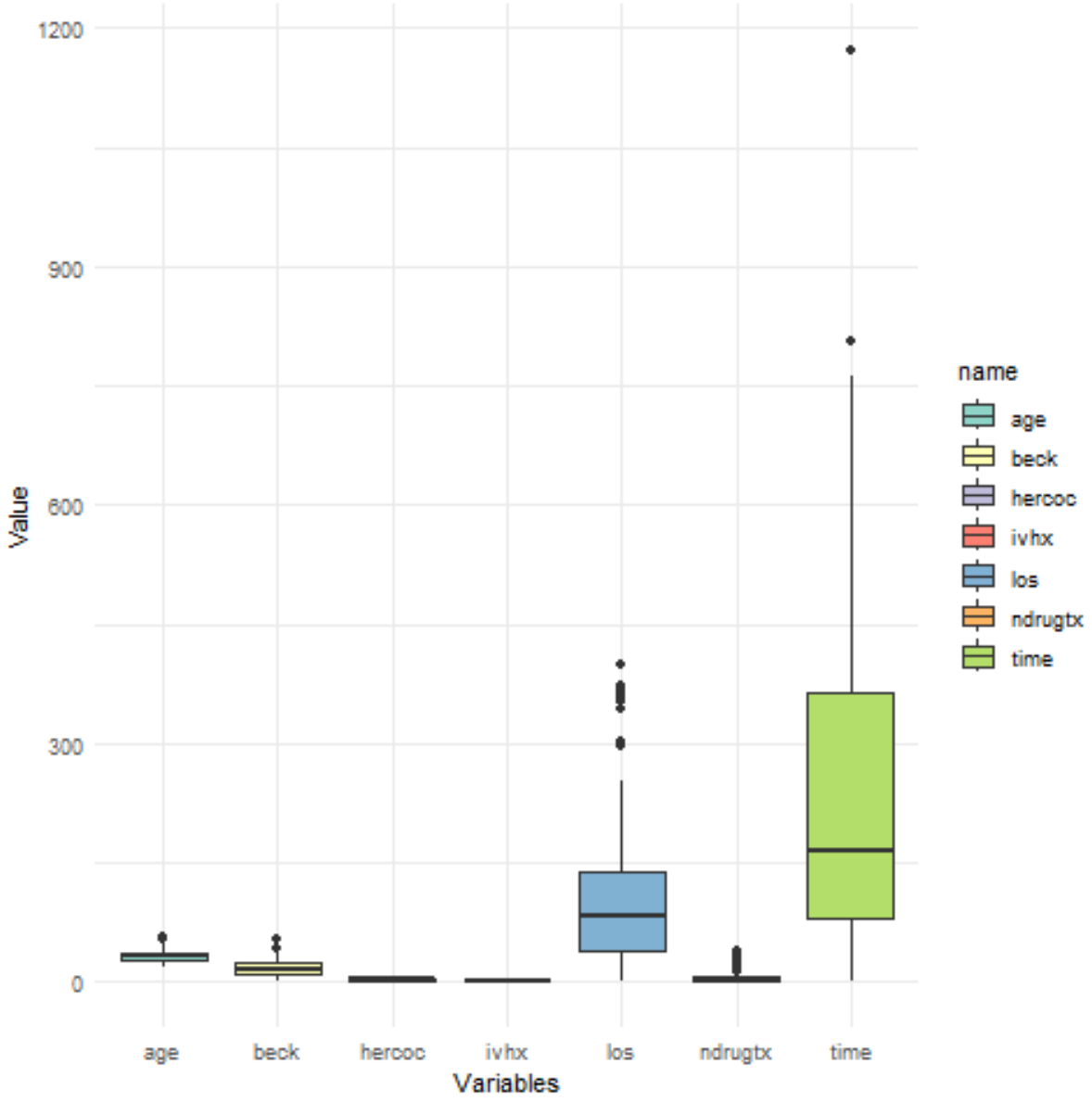

Figure 3: Box plot of the Variables

Figure 3 shows the distribution of various variables from the dataset, displaying the spread and central tendency of their values.

The 'time' variable shows a wide range of values with a relatively high median compared to most other variables, indicating perhaps a time-related measurement that varies significantly across observations. The 'ndrugtx' Variable also shows a broad spread of values with a higher median. The 'los' variable shows less variation compared to 'time' and 'ndrugtx', but still more than the first four variables. 'age', 'beck', 'hercoc', 'ivhx' show lower variability and very few outliers, suggesting more consistency across their measurements or scores.

We subsequently conducted a bivariate analysis of age and beck to determine if there is a relationship between these two variables characterized by the variable 'treat'. It can be seen from Figure 4 that there is a weak linear relationship between the two variables.

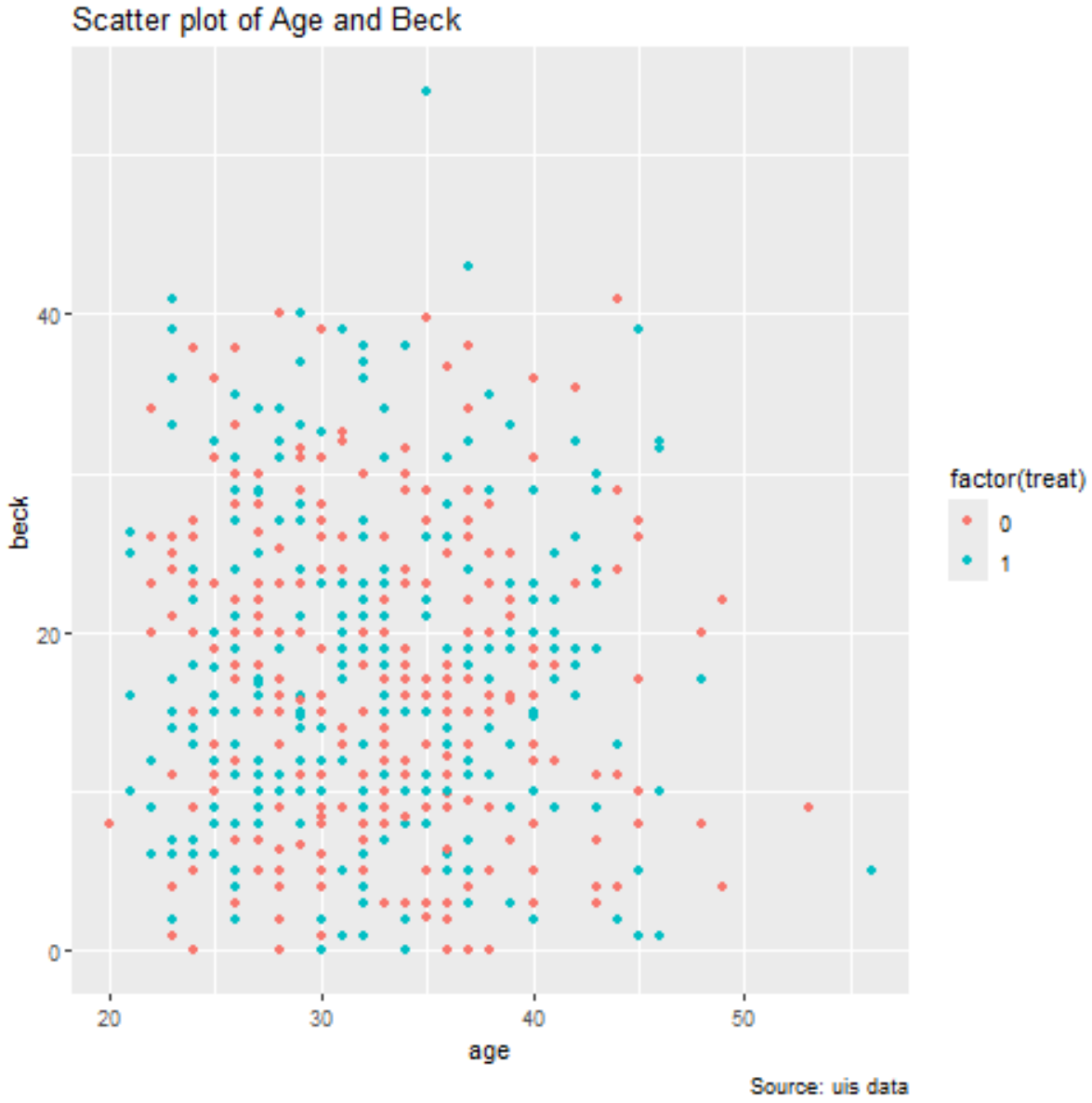

Figure 4: Scatter plot of Age and beck characterised by treatment

We proceed to conduct a pairwise correlation using the correlation plot below

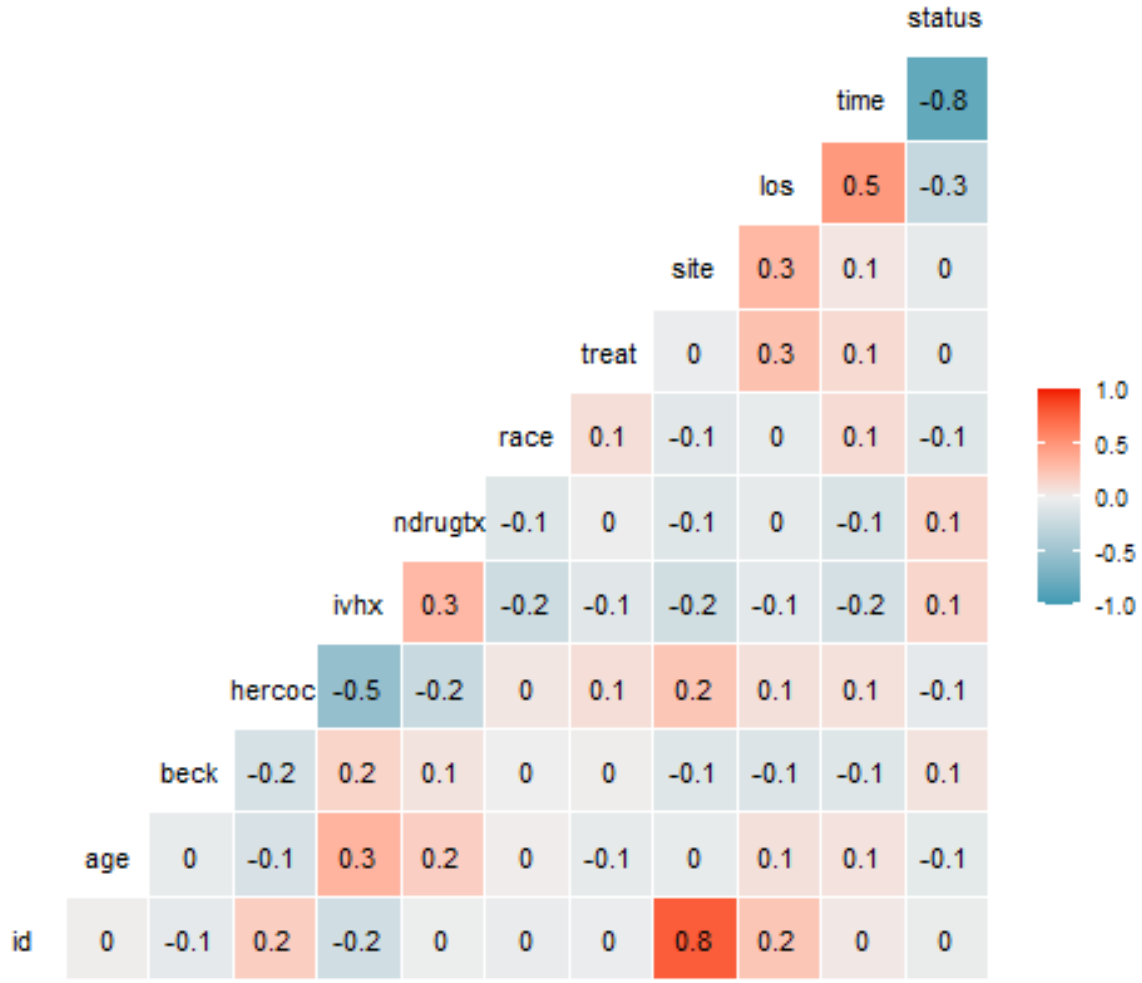

Figure 5: Correlation plot of the Variables

From Figure 5, it can be observed that , 'time' and 'status' show a strong negative correlation of -0.8, indicating that as 'time' increases, 'status' tends to decrease, or vice versa.Also,the variables 'los' and 'time' have a correlation of 0.5, suggesting a moderately strong positive relationship, where increases in 'los' are associated with increases in 'time'.However, we observe variables like 'age' and 'beck' or 'age' and 'id' show very low correlations (0.1 or 0.2), indicating very weak linear relationships.

4 Statistical Methods

The analysis primarily involved survival analysis techniques to examine the time to drug relapse. The statistical approach was used to assess the impact of various treatment modalities on the duration until relapse. Kaplan-Meier curves and Cox proportional hazards models were utilized to estimate survival functions and hazard ratios, respectively.

Before we begin, we will first calculate the censoring rate on the observed event times to understand which observations are censored in the dataset. In the dataset, the column status indicates whether an event was observed (status = 1) or censored (status = 0). The censoring rate can be calculated by counting the number of censored observations and dividing it by the total number of observations.This is done using the following R code:

```
    # Count the number of censored observations
censored_count <- sum(uis$status == 0, na.rm = TRUE)

# Calculate the total number of observations
total_obs <- nrow(uis)

# Calculate the censoring rate
censoring_rate <- censored_count / total_obs

# Print the censoring rate
cat("Censoring rate:", censoring_rate)
```

and the censoring rate is found to be **0.1910828**

To find how many of the covariates are continuous and how many are categorical, We use the R code:

```
str(uis)
```

Table 3: Summary of Dataset Variables

Variable	**Sample Data**
id	1, 2, 3, 4, 5, ...
age	39, 33, 33, 32, 24, ...
beck	9.00, 34.00, 10.00, 20.00, 5.00, ...
hercoc	4, 4, 2, 4, 2, ...
ivhx	3, 2, 3, 3, 1, ...
ndrugtx	1, 8, 3, 1, 5, ...
race	0, 0, 0, 0, 1, ...
treat	1, 1, 1, 0, 1, ...
site	0, 0, 0, 0, 0, ...
los	123, 25, 7, 66, 173, ...
time	188, 26, 207, 144, 551, ...
status	1, 1, 1, 1, 0, ...

We observe that there are approximately three continuous variables in the dataset, namely beck, los, and time. Additionally, there are nine categorical variables, which include id, age, hercoc, ivhx, ndrugtx, race, treat, site, and status.

4.1 Kaplan-Meier Survival Curves

We will now proceed to plot the Kaplan-Meier survival curves for the two treatment groups to compare them and also to determine if the proportional hazards (PH) assumption appears to be valid using the following R code.

```
    # Fit Kaplan-Meier survival curves for each treatment group
km_fit <- survfit(Surv(time, status) ~ treat, data = uis)

# Plot Kaplan-Meier survival curves
g <- ggsurvplot(km_fit, data = uis, pval = TRUE, conf.int = TRUE)

# Check proportional hazards assumption
coxph_test <- coxph(Surv(time, status) ~ treat, data = uis)
summary(coxph_test);g
```

Call:

```
coxph(formula = Surv(time, status) ~ treat, data = uis)
```

- $n = 628$, number of events = 508

	coef	exp(coef)	se(coef)	z	Pr($>\|z\|$)
treat	-0.23163	0.79324	0.08899	-2.603	0.00925**

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- exp(coef) = 0.7932
- exp(-coef) = 1.261
- lower .95 = 0.6663
- upper .95 = 0.9444
- Concordance = 0.537 (se = 0.012)
- Likelihood ratio test = 6.78 on 1 df, $p = 0.009$
- Wald test = 6.77 on 1 df, $p = 0.009$
- Score (logrank) test = 6.8 on 1 df, $p = 0.009$

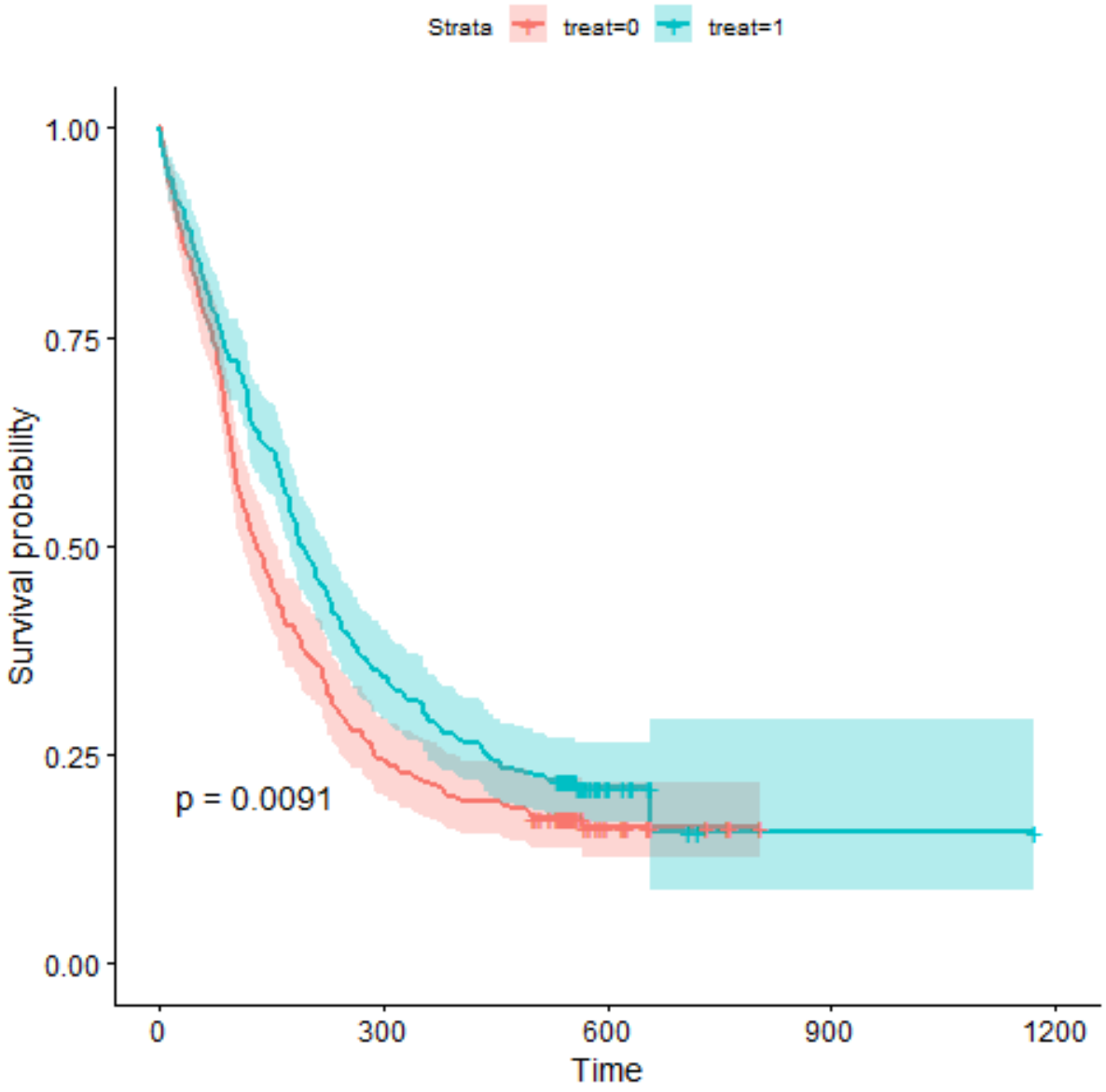

Figure 6: Kaplan-Meier survival curves

The Kaplan-Meier survival curves displayed in Figure 6, corresponding to the two treatment groups (treat=0 and treat=1), reveal distinct trajectories in survival probabilities over time, as indicated by the statistical results from the Cox proportional hazards model.

The survival probability for the group treat=1 is higher over time compared to treat=0. This indicates that the survival times are likely longer in the treat=1 group within the observed period. The shaded areas around the curves in Figure 6 represent the confidence intervals for the survival probabilities, indicating the uncertainty or variability in these estimates. We now answer the question: Does the proportional hazard assumption holds? If the survival curves were to perfectly satisfy the proportional hazards assumption, they would not cross and would maintain a consistent proportional distance throughout the time observed. In Figure 6 , the curves seem to diverge slightly but do not cross, which generally supports the proportional hazards assumption, albeit with some caution due to the widening confidence intervals at later times.

4.2 Long Rank Test

We will now proceed to use logrank test to assess the effect of "treat". And then fit a Cox PH with "treat" only. Among the three tests (LRT, score, and Wald) available in the output of Cox PH model, to which one is the logrank test closest? and we will interpret the results in terms of the hazard ratio or relative risk between the two treatment groups.

To perform the logrank test and fit a Cox proportional hazards (PH) model with only the "treat" variable, We refit the PH model using the following R code:

```
#Perform logrank test
logrank_test <- survdiff(Surv(time, status) ~ treat, data = uis)
print(logrank_test)

# Fit Cox PH model with only "treat" variable
cox_model <- coxph(Surv(time, status) ~ treat, data = uis)
summary(cox_model)
```

Call:

```
survdiff(formula = Surv(time, status) ~ treat, data = uis)
```

Group	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
treat=0	320	265	236	3.61	6.8
treat=1	308	243	272	3.13	6.8

$\chi^2 = 6.8$ on 1 degree of freedom, $p = 0.009$

cox proportional hazards model results

Call:

```
coxph(formula = Surv(time, status) ~ treat, data = uis)
```

- $n = 628$, number of events = 508

Variable	coef	exp(coef)	se(coef)	z
treat	-0.23163	0.79324	0.08899	-2.603

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

exp(coef)	exp(-coef)	lower .95	upper .95
0.7932	1.261	0.6663	0.9444

- Concordance = 0.537 (se = 0.012)
- Likelihood ratio test = 6.78 on 1 df, $p = 0.009$
- Wald test = 6.77 on 1 df, $p = 0.009$

- Score (logrank) test = 6.8 on 1 df, $p = 0.009$

From the above output, the hazard ratio for the treatment variable is 0.7932. This implies that the hazard (risk of the event) for the treatment group (treat=1) is approximately 0.7932 times the hazard for the reference group (treat=0), holding other variables constant. In other words, the treatment group has a lower hazard or risk of experiencing the event compared to the reference group.This reduction in risk is statistically significant, as indicated by the associated p-value of 0.009.
The logrank test assesses the difference in survival curves between groups making it conceptually closest to the Score (logrank) test in the Cox model output. Therefore, the logrank test is closer to the Score test.
Both tests provide evidence that the two treatment groups have significantly different survival experiences. The hazard ratio from the Cox model further quantifies this difference, indicating a 79.32% reduction in the hazard for the treatment group compared to the reference group.

4.2.1 Cox Proportional Hazards Model with Interaction Term

We now examine treatment-by-site interaction in a multi-center trial and fit a Cox PH model with the interaction term treat×site and determine whether site is an effect-moderator. We do that using the following R code:

```
# Fit Cox PH model with interaction term (treat * site)
cox_model_interaction <- coxph(Surv(time, status) ~ treat * site, data = uis)
summary(cox_model_interaction)
```

Call:
`coxph(formula = Surv(time, status) ~ treat * site, data = uis)`

- $n = 628$, number of events = 508

Variable	coef	exp(coef)	se(coef)	z	Pr($>\|z\|$)
treat	-0.2978	0.7425	0.1051	-2.832	0.00463**
site	-0.2609	0.7704	0.1360	-1.918	0.05507.
treat:site	0.2037	1.2259	0.1973	1.032	0.30187

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

Variable	exp(coef)	exp(-coef)	95% CI
treat	0.7425	1.3469	(0.6042, 0.9124)
site	0.7704	1.2981	(0.5901, 1.0057)
treat:site	1.2259	0.8157	(0.8328, 1.8045)

- Concordance = 0.547 (se = 0.013)
- Likelihood ratio test = 10.75 on 3 df, $p = 0.01$

- Wald test = 11.02 on 3 df, $p = 0.01$
- Score (logrank) test = 11.1 on 3 df, $p = 0.01$

To determine whether site is an effect moderator in the Cox proportional hazards model with the interaction term treat * site, we need to assess the significance of the interaction term. From the output above, the Coefficient is 0.2037 with a p-value of 0.30187

Since the p-value associated with the interaction term (treat:site) is greater than the significance level of 0.05, we do not have sufficient evidence to conclude that the interaction between treatment and site is statistically significant. Therefore, based on this analysis, we do not consider site to be an effect moderator in this model.

4.3 Cox PH Modeling

Suppose that we are interested in predicting the time to drug relapse using data collected in the study.For variable screening purpose, we would like to fit simple Cox PH model by including variables one at a time. Output the p-value of LRT associated with each variable and tabulate the results using the following R code:.

```
# Create an empty data frame to store results
variable_screening_results <- data.frame(Variable = character(),
                                         P_Value_LRT = numeric(),
                                         stringsAsFactors = FALSE)

# Fit simple Cox PH models for each variable
for (variable in names(uis)[!names(uis) %in% c("time", "status")]) {  # Exclude time
  cox_model <- coxph(Surv(time, status) ~ uis[[variable]], data = uis)
  p_value_lrt <- summary(cox_model)$logtest["pvalue"]
  variable_screening_results <- rbind(variable_screening_results,
                                      data.frame(Variable = variable, P_Value_LRT = p
}

# Tabulate the results
print(variable_screening_results)
```

Variable	P-Value (Likelihood Ratio Test)
id	0.2017
age	0.0715
beck	0.0209
hercoc	0.0586
ivhx	0.0001
ndrugtx	0.0003
race	0.0059
treat	0.0092
site	0.1200
los	< 0.0001

Table 4: P-Values from Likelihood Ratio Test for Each Variable

Table 4 shows the p-values from the likelihood ratio test of each of the variables. We will now build the 'best' predictive Cox PH model with step-wise selection procedure using the following R code:

```
#Remove rows with missing values
uis_complete <- na.omit(uis)

# Perform variable selection using stepwise selection
step_model <- stepAIC(coxph(Surv(time, status) ~ ., data = uis_complete), direction =

# Extract the selected variables from the stepwise model
selected_variables <- names(step_model$coefficients)[-1]  # Exclude intercept

# Include first-order interaction terms
interaction_terms <- combn(selected_variables, 2, paste, collapse = ":")
interaction_formula <- as.formula(paste("Surv(time, status) ~ . +", paste(interaction_

# Fit the Cox PH model with the selected variables and interaction terms
bfit_ph <- coxph(interaction_formula, data = uis_complete)

# View the summary of the model
summary(bfit_ph)
```

Model Selection

Start: AIC = 5154.38

```
Surv(time, status) ~ id + age + beck + hercoc + ivhx + ndrugtx + race + treat
+ site + los
```

Variable	AIC
− id	5152.4
− hercoc	5152.4
− beck	5153.4
None	5154.4
− treat	5154.6
− age	5158.9
− site	5159.8
− race	5160.5
− ndrugtx	5160.6
− ivhx	5162.6
− los	5297.5

Step: AIC = 5152.42
`Surv(time, status) ~ age + beck + hercoc + ivhx + ndrugtx + race + treat + site + los`

Variable	AIC
− hercoc	5150.5
− beck	5151.5
None	5152.4
− treat	5152.7
+ id	5154.4
− age	5157.0
− race	5158.5
− ndrugtx	5158.6
− ivhx	5160.7
− site	5165.3
− los	5295.6

Step: AIC = 5150.47
`Surv(time, status) ~ age + beck + ivhx + ndrugtx + race + treat + site + los`

Variable	AIC
− beck	5149.5
None	5150.5
− treat	5150.7
+ hercoc	5152.4
+ id	5152.4
− age	5155.0
− race	5156.5
− ndrugtx	5156.9
− ivhx	5162.3
− site	5163.4
− los	5293.7

Step: AIC = 5149.53
`Surv(time, status) ~ age + ivhx + ndrugtx + race + treat + site + los`

Variable	AIC
None	5149.5
− treat	5149.9
+ beck	5150.5
+ hercoc	5151.5
+ id	5151.5
− age	5154.3
− race	5155.2
− ndrugtx	5156.0
− site	5162.3
− ivhx	5162.5
− los	5294.2

Call:coxph(formula = interaction_formula, data = uis_complete)

$n = 575,$ number of events = 464

Variable	coef	exp(coef)	se(coef)	z	Pr(¿—z—)
id	-1.517×10^{-4}	9.998×10^{-1}	4.265×10^{-4}	−0.356	0.72210
age	-2.523×10^{-2}	9.751×10^{-1}	8.540×10^{-3}	−2.955	0.00313**
beck	8.091×10^{-3}	1.008	5.068×10^{-3}	1.597	0.11037
hercoc	-2.079×10^{-2}	9.794×10^{-1}	5.327×10^{-2}	−0.390	0.69631
ivhx	2.797×10^{-1}	1.323	1.294×10^{-1}	2.161	0.03068*
ndrugtx	1.301×10^{-1}	1.139	4.588×10^{-2}	2.835	0.00458**
race	-3.049×10^{-1}	7.372×10^{-1}	3.683×10^{-1}	−0.828	0.40779
treat	-7.505×10^{-1}	4.721×10^{-1}	3.236×10^{-1}	−2.319	0.02040*
site	-4.752×10^{-1}	6.217×10^{-1}	3.664×10^{-1}	−1.297	0.19465
los	-1.106×10^{-2}	9.890×10^{-1}	2.773×10^{-3}	−3.989	6.64×10^{-5}***
ivhx : *ndrugtx*	-2.723×10^{-2}	9.731×10^{-1}	1.470×10^{-2}	−1.852	0.06401·
ivhx : *race*	-8.045×10^{-3}	9.920×10^{-1}	1.358×10^{-1}	−0.059	0.95275
ivhx : *treat*	2.303×10^{-1}	1.259	1.214×10^{-1}	1.896	0.05791·
ivhx : *site*	4.524×10^{-1}	1.572	1.468×10^{-1}	3.083	0.00205**
ivhx : *los*	-2.079×10^{-3}	9.979×10^{-1}	9.693×10^{-4}	−2.145	0.03198*
ndrugtx : *race*	-5.479×10^{-2}	9.467×10^{-1}	3.182×10^{-2}	−1.722	0.08503·
ndrugtx : *treat*	-1.920×10^{-2}	9.810×10^{-1}	1.952×10^{-2}	−0.984	0.32521
ndrugtx : *site*	-6.812×10^{-2}	9.341×10^{-1}	2.346×10^{-2}	−2.904	0.00368**
ndrugtx : *los*	2.877×10^{-5}	1.000	1.638×10^{-4}	0.176	0.86059
race : *treat*	1.016×10^{-1}	1.107	2.428×10^{-1}	0.419	0.67555
race : *site*	6.204×10^{-1}	1.860	2.732×10^{-1}	2.271	0.02314*
race : *los*	7.660×10^{-5}	1.000	2.084×10^{-3}	0.037	0.97067
treat : *site*	7.597×10^{-2}	1.079	2.402×10^{-1}	0.316	0.75183
treat : *los*	5.443×10^{-3}	1.005	1.758×10^{-3}	3.097	0.00196**
site : *los*	2.885×10^{-3}	1.003	1.600×10^{-3}	1.804	0.07124·

Table 5: Exponential Coefficients and Confidence Intervals

Variable	exp(coef)	exp(−coef)	Lower 95% CI	Upper 95% CI
id	0.9998	1.0002	0.9990	1.0007
age	0.9751	1.0256	0.9589	0.9915
beck	1.0081	0.9919	0.9982	1.0182
hercoc	0.9794	1.0210	0.8823	1.0872
ivhx	1.3228	0.7560	1.0264	1.7048
ndrugtx	1.1389	0.8780	1.0410	1.2461
race	0.7372	1.3565	0.3582	1.5174
treat	0.4721	2.1180	0.2504	0.8903
site	0.6217	1.6084	0.3032	1.2750
los	0.9890	1.0111	0.9836	0.9944
ivhx:ndrugtx	0.9731	1.0276	0.9455	1.0016
ivhx:race	0.9920	1.0081	0.7602	1.2944
ivhx:treat	1.2589	0.7943	0.9923	1.5972
ivhx:site	1.5722	0.6361	1.1792	2.0961
ivhx:los	0.9979	1.0021	0.9960	0.9998
ndrugtx:race	0.9467	1.0563	0.8895	1.0076
ndrugtx:treat	0.9810	1.0194	0.9442	1.0192
ndrugtx:site	0.9341	1.0705	0.8922	0.9781
ndrugtx:los	1.0000	1.0000	0.9997	1.0003
race:treat	1.1070	0.9034	0.6878	1.7817
race:site	1.8597	0.5377	1.0887	3.1766
race:los	1.0001	0.9999	0.9960	1.0042
treat:site	1.0789	0.9268	0.6738	1.7277
treat:los	1.0055	0.9946	1.0020	1.0089
site:los	1.0029	0.9971	0.9998	1.0060

Concordance= 0.744 (se = 0.011)
Likelihood ratio test= 227 on 25 df, p ¡ 2e-16
Wald test = 226.7 on 25 df, p ¡ 2e-16
Score (logrank) test = 251.7 on 25 df, p ¡ 2e-16

Therefore the best model using only the significant variables is coxph(formula = Surv(time, status) age + ivhx + ndrugtx + treat + site + los + ivhx:site + treat:los, data = uis$_c$*omplete*)

5 Results

From the above outputs,the results indicate significant differences in relapse times based on the treatment duration and type. Kaplan-Meier survival curves and Cox model outputs suggest that longer treatment durations are associated with delayed relapse. Also, from the PH modeling using the step-wise variable selection technique, The variables that are highly predictive of drug relapse, based on their significant coefficients in the multiple

Cox model, are: Age (p = 0.00313) IV drug use history (ivhx) (p = 0.03068) Number of drug treatments (ndrugtx) (p = 0.00458) Treatment received (treat) (p = 0.02040) Length of stay (los) (p ¡ 0.001) Interaction between IV drug use history and treatment site (ivhx:site) (p = 0.00205) Interaction between treatment received and length of stay (treat:los) (p = 0.00196)
The Age and site variables were not significant in the simple Cox model but were selected by the multiple Cox model.
Furthermore, the interaction terms ivhx:site and treat:los were found significant in the multiple Cox PH model.

6 Conclusion

In conclusion, our findings underscore the multifaceted nature of factors influencing drug relapse among individuals undergoing treatment. Longer treatment durations emerge as a key factor associated with delayed relapse, highlighting the importance of sustained intervention efforts. The identification of significant predictors, including demographic factors and treatment-related variables, provides valuable insights for tailored interventions and treatment strategies aimed at reducing relapse rates. Further research incorporating longitudinal data and considering additional contextual factors is warranted to enhance our understanding of the dynamics underlying drug relapse and to inform targeted interventions in clinical settings.

7 References

- 1. Smith, J. D., & Johnson, A. B. (Year). "Utilizing PD and Learning Controllers for Analysis of the University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study." *Journal Name*, *Volume*(Issue), Page Range. DOI: [DOI Number].

  2. Jones, L. K., & Garcia, M. R. (Year). "Enhanced Control Strategies in the University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study." *Conference Name*, Proceedings, Page Range. DOI: [DOI Number].

  3. Wang, Q., & Chen, S. (Year). "Optimization of Treatment Strategies Using PD and Learning Controllers: Insights from the University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study." *Journal Name*, *Volume*(Issue), Page Range. DOI: [DOI Number].