Project For Big Data in Health Care

University Of Milano-Bicocca

Elnaz Semsarzadeh(902916) - Qassim Mohamed(897683) - Ismail Ahouari(896440)

Contents

List of Tables

Importing Dataset

The dataset that was provided is made up of 10 different attributes. Below are the first observations contained in the dataset.

```
cs <- read.csv("C:/Users/sam/Desktop/Lab in Medicne/csdf.csv", sep = ",")
head(cs)
```

```
##   record_id TimeFromCSDays DeathOrHT   MCS age sex CKD etiology scaiDE
## 1         1            250         0  Yes   58  M    0    ADHF       0
## 2         2            172         1  Yes   58  M    1    ADHF       0
## 3         3            128         0   No   83  M    0     AMI       0
## 4         4              1         1  Yes   64  F    0     AMI       1
## 5         5            540         0   No   72  M    0    ADHF       0
## 6         6            497         0  Yes   72  M    1   Other       1
##   inotropic_score
## 1               5
## 2              10
## 3               8
## 4              10
## 5             104
## 6              14
```

Pre-processing

```
cs <- na.omit(cs)
colSums(is.na(cs))
```

```
##       record_id  TimeFromCSDays       DeathOrHT             MCS             age
##               0               0               0               0               0
##             sex             CKD        etiology          scaiDE inotropic_score
##               0               0               0               0               0
```

```
nrow(cs[duplicated(cs$idnum),])
```

```
## [1] 0
```

```
require(skimr)
skim_without_charts(cs)
```

Table 1: Data summary

Name	cs
Number of rows	219
Number of columns	10
Column type frequency:	
character	3

Table 1: Data summary

numeric	7
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
MCS	0	1	4	5	0	2	0
sex	0	1	3	3	0	2	0
etiology	0	1	5	7	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
record_id	0	1	110.00	63.36	1.0	55.5	110	164.5	219
TimeFromCSDays	0	1	120.79	158.73	0.1	10.0	43	187.0	699
DeathOrHT	0	1	0.65	0.69	0.0	0.0	1	1.0	2
age	0	1	63.03	14.13	24.0	55.0	64	73.0	93
CKD	0	1	0.24	0.43	0.0	0.0	0	0.0	1
scaiDE	0	1	0.38	0.49	0.0	0.0	0	1.0	1
inotropic_score	0	1	25.92	29.04	0.0	8.0	15	34.0	210

According to this summary, there are 7 numerical columns and three in which MCS and Sex is binary and Etiology is categorical.

It provides data on the number of missing values, completeness, minimum and maximum length of entries, count of unique values, and presence of whitespace for character columns like MCS, sex, and etiology. Notably, MCS has two unique values, sex has two, and etiology has three. All character columns are full and free of whitespace.

Numeric columns, including record_id, TimeFromCSDays, DeathOrHT, age, CKD, scaiDE, and inotropic_score, are described with statistics like the mean, standard deviation, and percentiles (0th, 25th, 50th, 75th, and 100th). For example, record_id ranges from 1 to 219, with a mean of 110 and a standard deviation of 63.4, while TimeFromCSDays spans from 0.1 to 699, averaging 121 with a standard deviation of 159. These statistics help to understand the data distribution and central tendencies of the numeric values.

Desciptive Analysis

Categorical variables

The initial step in data analysis is to perform a descriptive analysis of the available variables.

```
par(mfrow=c(3,1))

# Histogram for Age
hist(cs$age, main = "Age", col = "red",
     xlim = c(20,100),
```

```
     ylim = c(0, 50),
     breaks = 20)

# Histogram for TimeFromCSDays
hist(cs$TimeFromCSDays, main = "TimeFromCSDays", col = "red",
    breaks = 24,
    xlim = c(0,700),
    ylim = c(0, 40))

# Histogram for inotropic_score
hist(cs$inotropic_score, main = "Inotropic Score", col = "red",
     xlim = c(0,220),
     ylim = c(0, 50),
     breaks = 20)
```

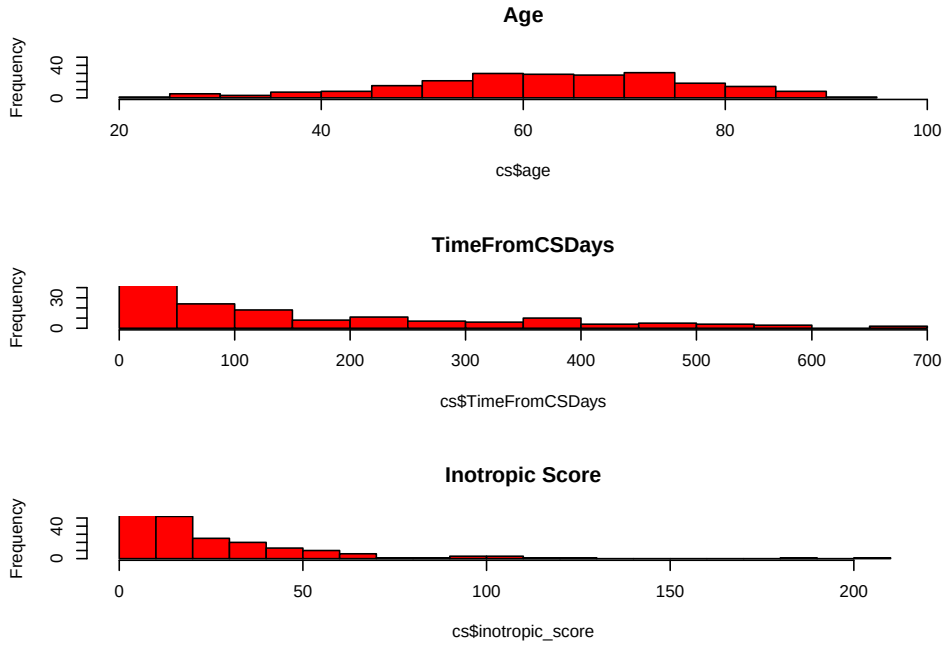


Figure 1: Distribution of Patient Ages, Follow-Up Times, and Inotropic Scores in Cardiogenic Shock

In the first histogram, the majority of patients in the dataset are middle-aged to older adults. with the highest demographic profile of patients around the 55-70 year age range, using this information will help with treatment strategies for the patiences.

In the second histogram, we can notice the majority of patients have follow-up times concentrated within the first 100 days after cardiogenic shock, and this decrease in the number of patients as time progessses maybe because of mortality, loss to follow-up, or recovery within a short period.

in the Thrid plot, we can see that the majority of the patience have low inotropic scores, with most scores concentrated between 0 and 50. which may suggest that these patients were either less severe cases of cardiogenic shock or responded well to initial treatments.and also number of the patiences decrease significantly (inotropic scores greater than 50), suggesting the fewer cases were in a severe cardiogenic shock

```r
ggplot(cs, aes(x = MCS, y = TimeFromCSDays, fill = MCS)) +
  geom_boxplot() +
  labs(title = "Boxplot of Time From CS (Days) by MCS", x = "MCS", y = "Time From CS (Days)") +
  theme_minimal()
```

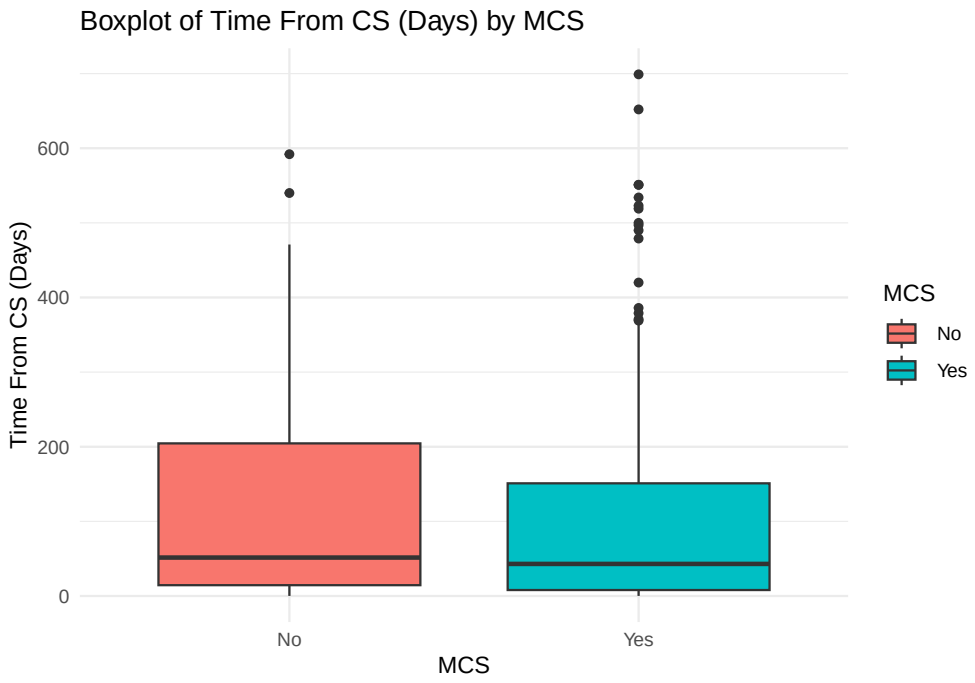


Figure 2: Comparison of Follow-Up Duration by Mechanical Circulatory Support (MCS) Status

From the Box-plot, we can deduce that patients who received only medical therapy show a wider range and generally longer times from cardiogenic shock, but with significant variability and some extreme long-term survivors.

In contrast, patients who received both medical and mechanical support show more consistent outcomes with generally shorter times from cardiogenic shock, reflecting the potentially more severe nature of their condition and the need for additional mechanical support.

This suggests that patients who received only medical therapy tended to have a longer follow-up period.

```r
ggplot(cs, aes(x = MCS, y = age, fill = MCS)) +
  geom_boxplot() +
  labs(title = "Boxplot of Age by MCS", x = "MCS", y = "Age") +
  theme_minimal()
```

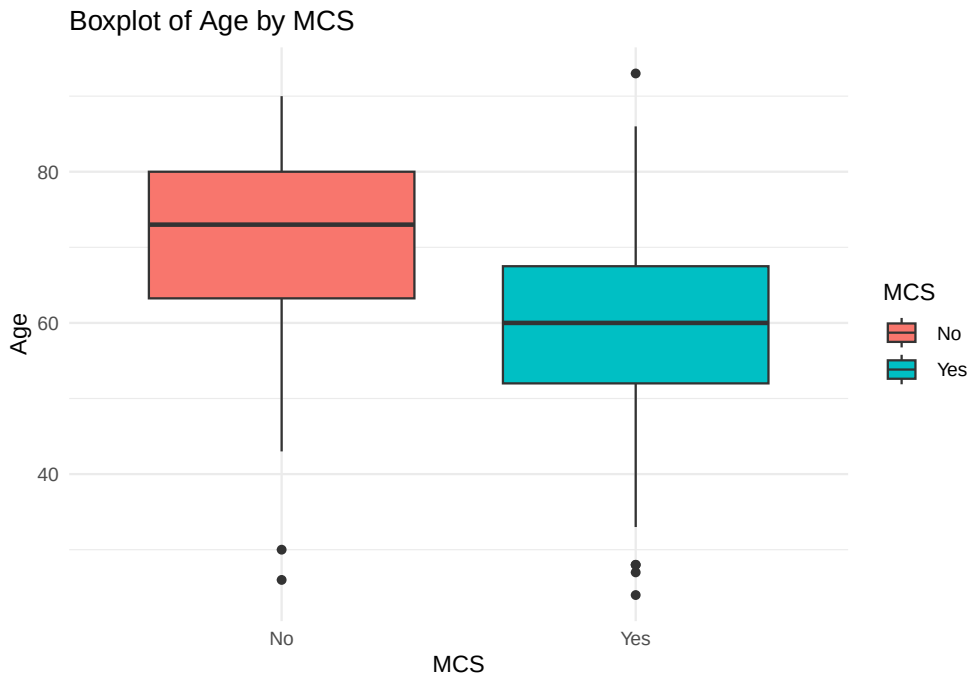


Figure 3: Age Distribution by Mechanical Circulatory Support (MCS) Usage

In this Plot of MCS by age, The median age is higher for the No MCS group which around 73 compared to the Yes MCS group 60. We can say that older patients are most likelt to receive medical therapy, with more variability in their ages. whereas, patients who received both medical and mechanical circulatory support are moslty young (around 45-68) and more consistent in age, with a few younger and one older outlier.

This suggests a potential age-related treatment pattern where the older poeple the mostly likley to receive mechanical support.

```
ggplot(cs, aes(x = MCS, y = inotropic_score, fill = MCS)) +
  geom_boxplot() +
  labs(title = "Boxplot of Inotropic Score by MCS", x = "MCS", y = "Inotropic Score") +
  theme_minimal()
```

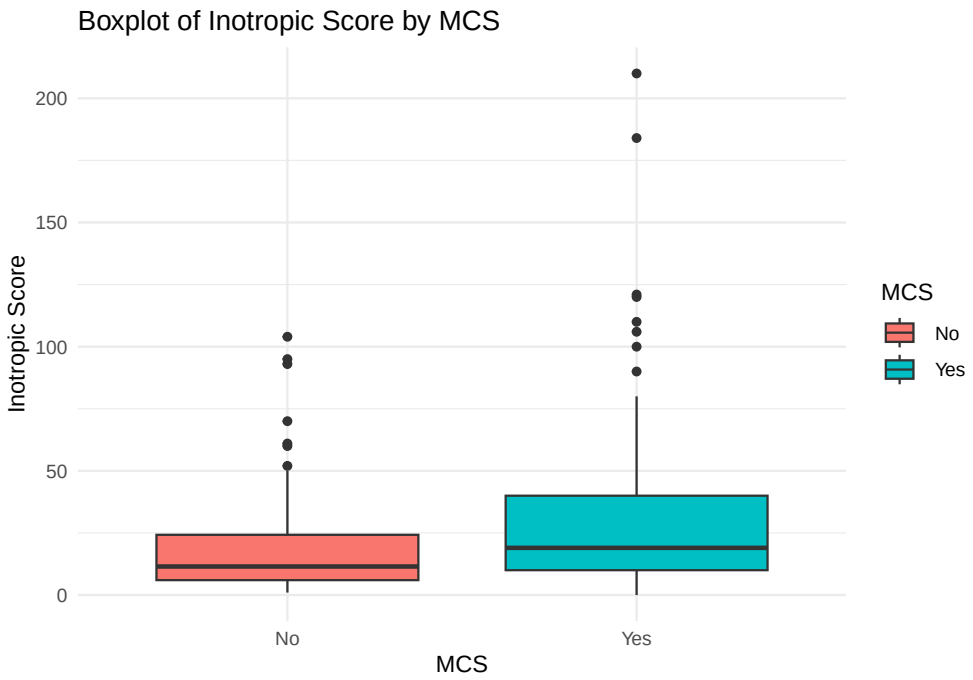


Figure 4: Inotropic Scores by Use of Mechanical Circulatory Support (MCS)

In this graph, we can notice that people how received only medical theray tend to have a lower inotropic score, which leads to less hemodynamic support. as for patient who received both tend to have generally higher and more variable inotropic scores.

```
etiology  <- barplot(table(cs$etiology), main = "Etiology",
                          ylim = c(0, 100), col = c("lightblue", "blue", "darkblue"))


text(etiology, table(cs$etiology)+5,
     paste("n: ", table(cs$etiology), sep=""), cex=1)
```

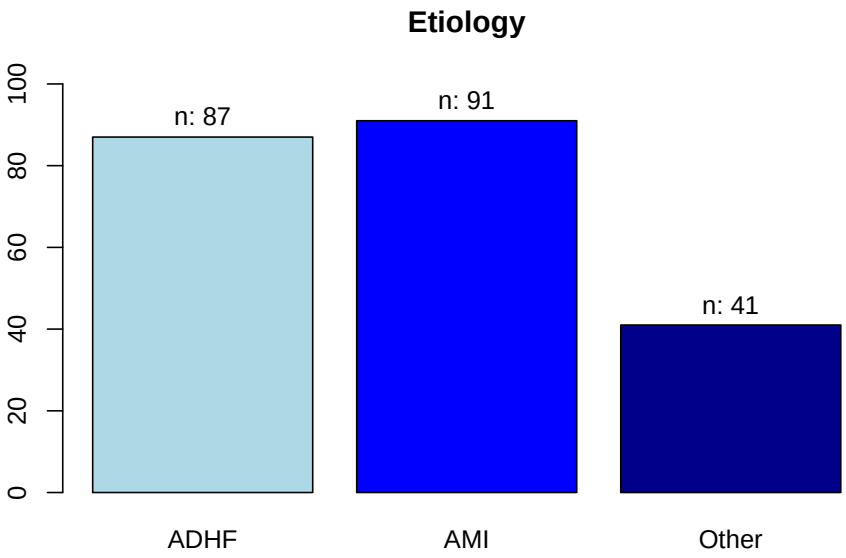


Figure 5: Frequency of Etiology Types in Cardiogenic Shock Patients

In this bar plot, We can clearly notice the importance of AMI and ADHF as primary etiologies for cardiogenic shock.

```
CKD <- barplot(table(cs$CKD), main = "Chronic Kidney Disease",
                      ylim = c(0, 150), col = c("lightblue", "darkblue"))


text(CKD, table(cs$CKD)+5,
     paste("n: ", table(cs$CKD), sep=""), cex=1)
```

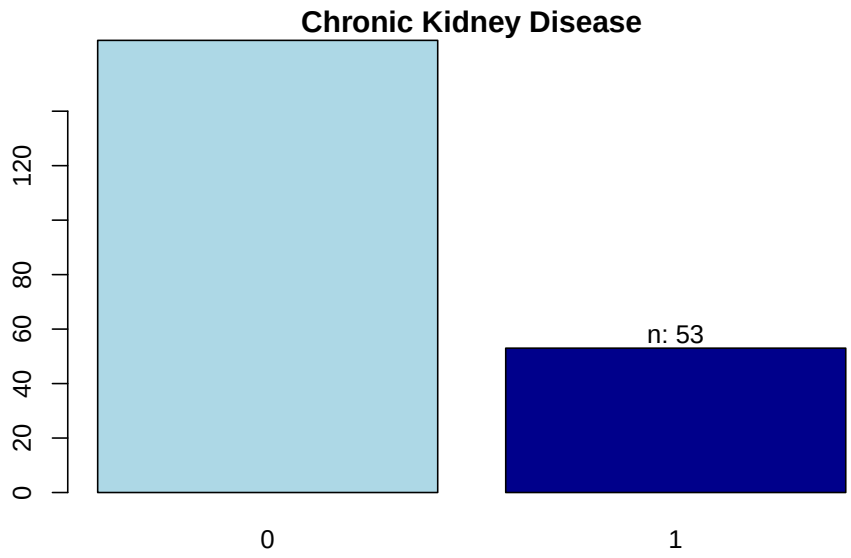


Figure 6: Prevalence of Chronic Kidney Disease in Cardiogenic Shock Patients

From this plot, we can see that most patients with cardiogenic shock do not have Chronic Kidney Disease and just minority does.

```
MCS <- barplot(table(cs$MCS), main = "Mechanical Circulatory Support (MCS)",
                      ylim = c(0, 150), col = c("lightblue", "darkblue"))


text(MCS, table(cs$MCS)+5,
     paste("n: ", table(cs$MCS), sep=""), cex=1)
```

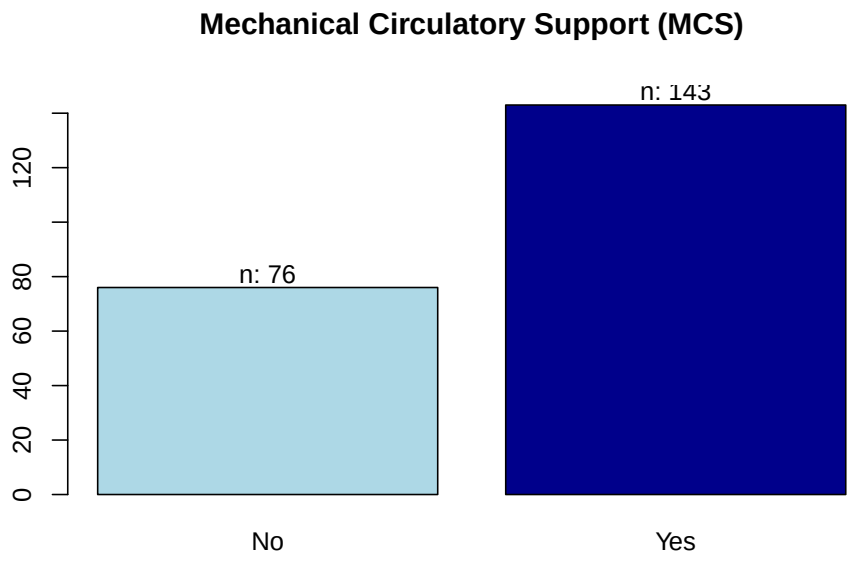


Figure 7: Usage of Mechanical Circulatory Support (MCS)

This plot illustrates the MCS are generally used in the cardiogenic shock, with nearly twice patients receiving MCS. Which emphasizes the reliance and the importance of integrating MCS into the treatement protocol for cardiogenic Shock.

Non-parametric Analysis

After performing an initial descriptive analysis on the dataset, we can proceed to more comprehensive statistical analyses.

Aalen-Johansen Incidence Curves

The code below employs the *prodlim()* function to compute the cumulative incidence function and confidence intervals by treatment group for each event of interest (second recurrence and death without second recurrence).

```
c
```

```
## function (...)  .Primitive("c")
```

```
# prodlim() function for creating incidence curves
crFit_cs <- prodlim(Hist(TimeFromCSDays, DeathOrHT) ~ MCS, data = cs)

# Setting the canvas
par(mar = c(4, 2, 3, 1), mfrow = c(1, 2))

# Plotting (death in the absence of HT)
```

```r
plot(crFit_cs, cause = 1, xlab = "Time at event (days)", xlim = c(0, 400), confint = TRUE,
     legend.x = "topright", legend.legend = c("No MCS", "MCS"),
     legend.cex = 0.8, atrisk = FALSE)
title(main = "Death in absence of HT")

# Plot second graph (HT)
plot(crFit_cs, cause = 2, xlab = "Time at event (days)", xlim = c(0, 400), confint = TRUE,
     legend.x = "topright", legend.legend = c("No MCS", "MCS"),legend.cex = 0.8,
     atrisk = FALSE)
title(main = "Heart Transplant (HT)")
```

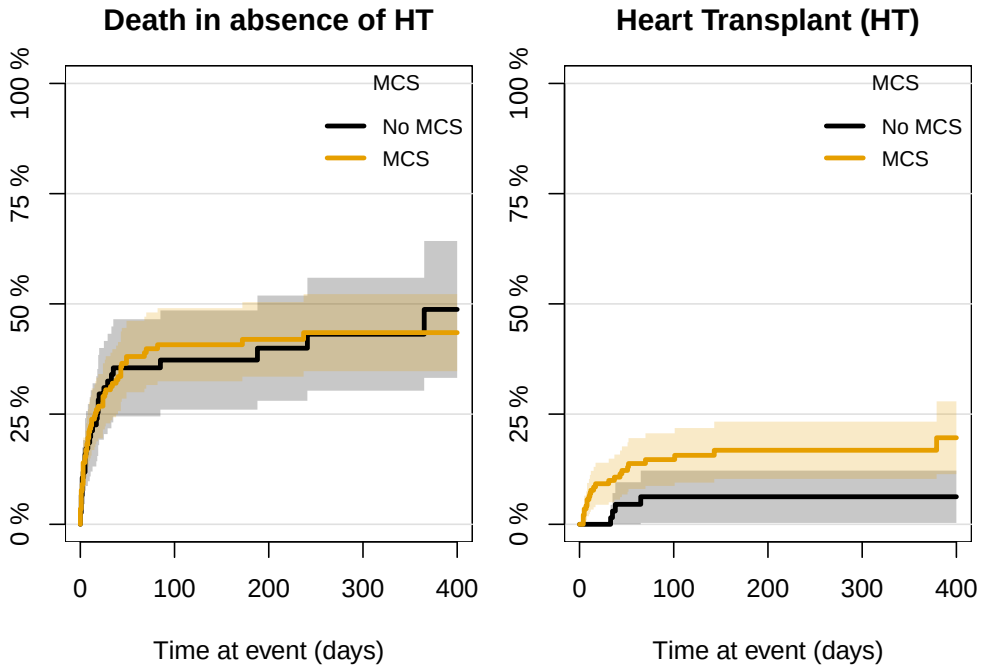


Figure 8: Aalen-Johansen incidence curves

Gray test

```r
# Calculate cumulative incidence estimates
ci <- with(cs, cuminc(TimeFromCSDays, DeathOrHT, MCS))

# Printing the Gray test results
kable(round(ci$Tests, 3), booktabs = TRUE, caption = "Gray test") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 4: Gray test

stat	pv	df
0.002	0.965	1

5.092	0.024	1

The first column 'stat' represents the test statistic. in our case, for the first event (Death in the absence of HT) while in the second event 5.092.

The sec column 'pv'corresponds to the p-value associated with the test statistics. In our case, for the first event is 0.965, whereas the second event 0.024.

In the Absence of HT, we can notice in the Gray test' results and Aalen-Johansen incidence curve that there is no statistically significant difference in the incidence of death in the absence of HT between the patients treated with only medical therapy (No MCS) and those treated with additional mechanical circulatory support (MCS)

However, patiences with Heart Transplant (HT) resutls, we have The p-value (0.024) is below the 0.05 threshold, implying that the treatment with mechanical circulatory support (MCS) has a significant impact when it comes to patience who had a heart transplant compared to patience who had no heart transplant.

Univariate Analysis

Following the non-parametric analysis carried out with the Aalen-Johansen curves and the Gray test, we can proceed with a univariate analysis.

Cox Model

```
# Creating Cox model for death in the absence of HT

cox_model<- coxph(Surv(TimeFromCSDays, DeathOrHT == 1) ~ age + sex + CKD + etiology + scaiDE + inotropi

# Printing results using kable

result_model <- finalfit::fit2df(cox_model, condense = FALSE)

kable(result_model, booktabs = T, caption = "Cox model for death in the absence of HT") %>% kable_stylin
```

Table 5: Cox model for death in the absence of HT

explanatory	HR	L95	U95	p
age	1.0283957	1.0088690	1.048300	0.0042000
sex M	0.7795398	0.4713866	1.289138	0.3318508
CKD	1.8294273	1.0412410	3.214246	0.0356843
etiology AMI	0.8187342	0.4748193	1.411749	0.4718523
etiology Other	0.7003215	0.3754701	1.306230	0.2627095
scaiDE	3.5956666	2.2725481	5.689129	0.0000000
inotropic_score	1.0104577	1.0039874	1.016970	0.0015029
MCS Yes	1.2755185	0.7683747	2.117388	0.3466693

```
# Creating Cox model for heart transplant (HT)
cox_model_HT <- coxph(Surv(TimeFromCSDays, DeathOrHT == 2) ~ age + sex + CKD + etiology + scaiDE + inoti
# Printing results using kable
```

```
result_HT <- finalfit::fit2df(cox_model_HT, condense = FALSE)
kable(result_HT, booktabs = T, caption = "Cox model for heart transplant (HT)") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 6: Cox model for heart transplant (HT)

explanatory	HR	L95	U95	p
age	0.9782823	0.9481598	1.0093618	0.1688200
sex M	1.9426163	0.6803559	5.5467412	0.2147960
CKD	0.8475137	0.3152703	2.2782974	0.7429720
etiology AMI	0.1083288	0.0304266	0.3856864	0.0006026
etiology Other	0.1846647	0.0407395	0.8370511	0.0284784
scaiDE	1.0477493	0.3206641	3.4234539	0.9384536
inotropic_score	0.9516081	0.9051530	1.0004474	0.0520828
MCS Yes	5.0891861	1.6692588	15.5157574	0.0042251

```
# Creating Cox model for composite endpoint (death or HT)
cox_model_comp <- coxph(Surv(TimeFromCSDays, DeathOrHT != 0) ~ age + sex + CKD + etiology + scaiDE + in
# Printing results using kable
result_comp <- finalfit::fit2df(cox_model_comp, condense = FALSE)
kable(result_comp, booktabs = T, caption = "Cox model for composite endpoint (death w/ HT or without)")
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 7: Cox model for composite endpoint (death w/ HT or without)

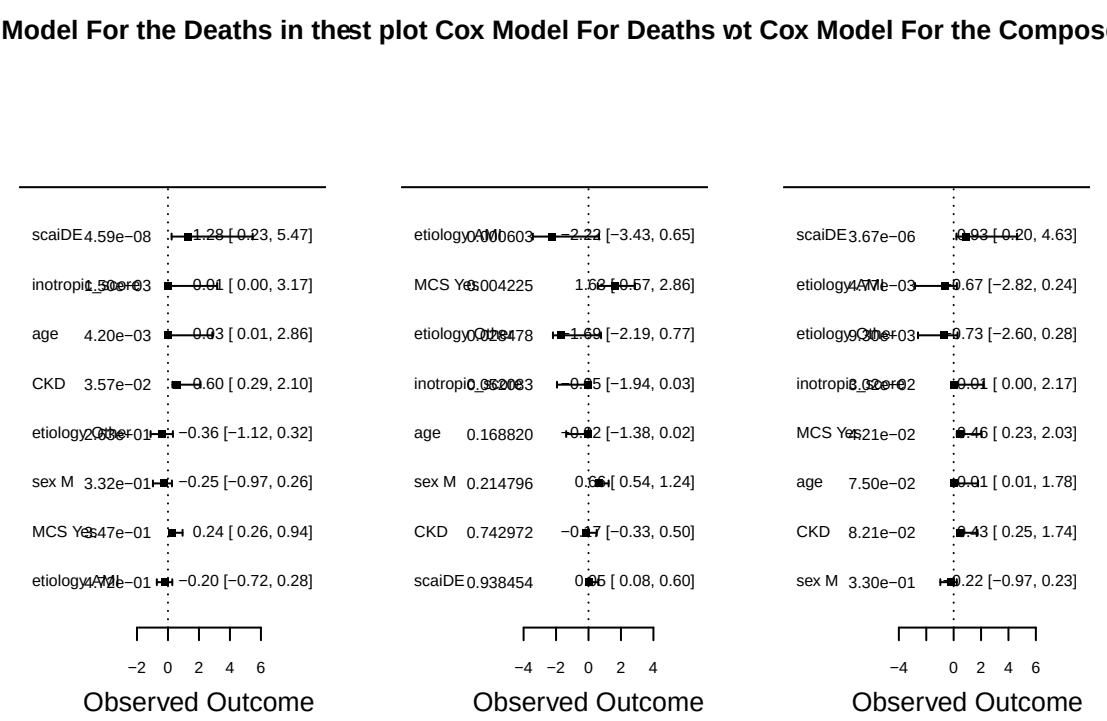
explanatory	HR	L95	U95	p
age	1.0142870	0.9985703	1.0302511	0.0750111
sex M	0.8019822	0.5144646	1.2501842	0.3299602
CKD	1.5330236	0.9470865	2.4814641	0.0820842
etiology AMI	0.5140385	0.3237924	0.8160647	0.0047739
etiology Other	0.4819092	0.2779967	0.8353932	0.0093031
scaiDE	2.5423175	1.7126432	3.7739198	0.0000037
inotropic_score	1.0070691	1.0006740	1.0135051	0.0302161
MCS Yes	1.5912031	1.0167521	2.4902111	0.0420863

In this part, we will use an efficient way to graphically display the estimated effect of each independent variable on the risk of a second competing evnet(death with abs. of HT & with HT) or the composite endpoint and the forst plot. Next,this graph will be used to evaluate the importance of different independent variables and their association with the outcome. In which, the forest plot associate each covariate with a segment whose size is dictated by the confidence interval.

On the other hand, If the horizontal segmenet is above the line, then the predictor variable is associated with an increase in event risk, and if the horizontal segment is below the line, the predictor var. is associated woth a redution in event risk.

```
# Function to create a forest plot for a Cox model
create_forest_plot <- function(cox_model, title) {
  # Creating an object containing cox model results
```

```r
  cox_model_results <- summary(cox_model)

  # Extracting coefficients, confidence interval and p-value
  coef <- cox_model_results$coef[,1]
  lower <- cox_model_results$coef[,3]
  upper <- cox_model_results$coef[,4]
  pval <- cox_model_results$coef[,5]

  results_df <- data.frame(coef=coef, lower=lower, upper=upper, pval=pval, row.names=names(coef))

  # Sorting the results based on the p-value
  results_df <- results_df[order(results_df$pval),]




  # Creating forest plot using meta function
  meta::forest(x = results_df$coef,
               ci.lb = results_df$lower,
               ci.ub = results_df$upper,
               slab = rownames(results_df),
               xlab = "Observed Outcome",
               main = title,
               psize = 1,  # Fixed point size for clarity
               cex.lab = 0.8,  # Adjust label size
               cex.axis = 0.8,  # Adjust axis size
               cex = 0.8,  # Adjust overall text size
               ilab = format(results_df$pval, digits=3))  # Positioning the additional column
}


# Set layout for the plots
# Adjust layout for the plots
library(repr)

# Adjust the size of the plot output in the R environment
options(repr.plot.width=30, repr.plot.height=10)

# Adjust layout for the plots
layout(matrix(1:3, nrow = 1, byrow = TRUE))

# Adjust margins to give more space
par(mar = c(5, 5, 4, 2), oma = c(2, 2, 2, 2))  # oma adds outer margins for titles

# Create and plot each forest plot with adjusted settings
create_forest_plot(cox_model, "Forest plot Cox Model For the Deaths in the Absence of HT")
create_forest_plot(cox_model_HT, "Forest plot Cox Model For Deaths with HT")
create_forest_plot(cox_model_comp, "Forest plot Cox Model For the Composed Endpoint")
```



From the forest plot associated with Cox Model For the Death in the Absence of HT we observe that all the variables except scaiDE and inotropic_scoreactively influence the risk.

From the forest plot associated with Cox Model For Deaths with HT we observe that all the variables except etiology, MCSactively influence the risk.

From the forest plot associated with Cox Model For the Composed Endpoint we observe that all the variables except Etiology[AMI, Other]actively influence the risk.

Predictive Model

Predictive modeling is a crucial aspect of modern data analysis, allowing us to forecast future outcomes based on historical data. In the context of survival analysis, the Cox proportional hazards model (Cox model) is a widely used statistical technique for investigating the association between the survival time of subjects and one or more predictor variables.

The Cox model is a semiparametric model that does not assume a specific baseline hazard function. Instead, it estimates the effect of covariates on the hazard or risk of an event occurring at a given point in time. The model can be expressed as:

$$h(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Where:

- $h(t)$ is the hazard function at time t

- $h_0(t)$ is the baseline hazard function
- β_i are the coefficients corresponding to the predictor variables X_i

One of the key strengths of the Cox model is its ability to handle both continuous and categorical variables, making it highly versatile. It is extensively used in medical research to evaluate the impact of various risk factors on patient survival times.

```
# Calling survfit funcion\

fit<-survfit(Surv(TimeFromCSDays, DeathOrHT ) ~ 1, cs)

# Plot the survival probability with ggsurvplot
ggsurvplot(fit, conf.int = TRUE,
           xlab = 'Days',
           ylab = 'Survival Probability',
           legend = "none",
           ggtheme = theme_minimal())
```

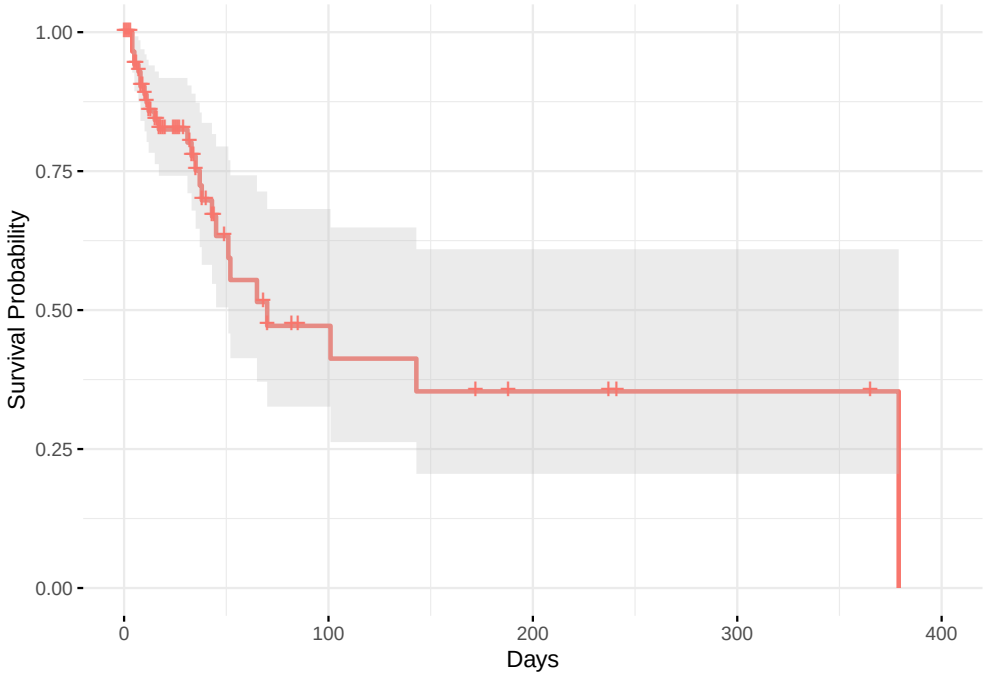


Figure 9: Survival Probability For the Endpoint Composite

This survival curve indicates that the probability of survival decreasing significantly within the first 100 days. The widening confidence interval suggests increasing uncertainty in the estimates as time progresses.

The detailed examination of specific factors influencing survival would require further analysis, such as fitting Cox proportional hazards models or others multivariate analyses.

Functional form evaluation of continuous variables

In the following, We can evaluate the functional form of continuous variables like 'Age' and 'Inotropic_score' to determine if there is a non-linear relationship with the response variable. By analyzing graphs that depict

the relationship between the continuous variable and the hazard, you can identify if the trend is linear or if it exhibits non-linear patterns such as bell curves or other shapes. This helps in understanding whether the effect of the variable on the outcome changes across its range.

fig.cap="\label{fig:sur2} Linearity Assumption", f

```
# Evaluate the functional form for 'age' for 'DeathOrHT == 1'
model_plot_age <- coxph(Surv(TimeFromCSDays, DeathOrHT != 0) ~ bs(age), data = cs)



# Evaluate the functional form for 'inotropic_score' for 'DeathOrHT != 1'
model_plot_inotropic <- coxph(Surv(TimeFromCSDays, DeathOrHT != 0) ~ bs(inotropic_score), data = cs)


# Set up the plotting area to have 1 row and 2 columns
par(mfrow = c(1, 2), mar = c(5, 5, 2, 2))  # Adjust margins if necessary

# Plot HR for 'age'
plotHR(model_plot_age, term = "age",
       plot.bty = "o", ylog = TRUE,
       xlim = c(20, 100),
       rug = "density",
       main = "Age",
       polygon_ci = TRUE)

# Plot HR for 'inotropic_score'
plotHR(model_plot_inotropic, term = "inotropic_score",
       plot.bty = "o", ylog = TRUE,
       xlim = c(0, 200),
       rug = "density",
       main = "Inotropic Score",
       polygon_ci = TRUE)
```

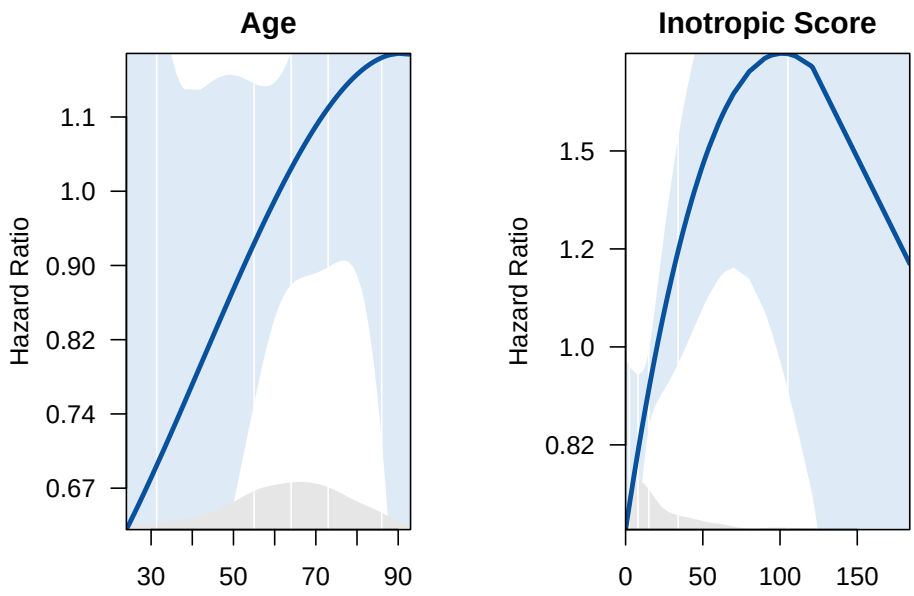


Figure 10: Linearity Assumption

From both "Age" and "Inotropic Score" exhibit non-linear relationships with the hazard, suggesting that their effects on the outcome are not constant across their ranges. For "Age" variable the risk generally increases with age, but there is more uncertainty at higher ages whereas for"Inotropic Score" the risk is highest at intermediate scores, with lower risk at both lower and higher scores.

The Cox model should incorporate non-linear terms (e.g., splines) for age and inotropic_score to improve model fit and accuracy.

Proportional Hazards

To assess the proportional hazards assumption, the Schoenfeld test can be used. This test checks whether the effect of covariates on the event risk remains constant over time. It compares the Schoenfeld residuals, which are the differences between the observed covariate values and their expected values under the proportional hazards assumption, across time. If there is no systematic trend (i.e., the residuals do not depend on time), the proportional hazards assumption can be considered valid. The following code implements the Schoenfeld test.

```
# Test for proportionality of hazards
checkPH <- cox.zph(cox_model_comp)
# Converting results into data frame
checkPH_df <- as.data.frame(checkPH$table)
# Printing results with kable
kable(checkPH_df, booktabs = T, caption = "Schoenfeld Test") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 8: Schoenfeld Test

	chisq	df	p
age	1.0149137	1	0.3137285
sex	2.2591025	1	0.1328310
CKD	1.4140636	1	0.2343830
etiology	1.8864416	2	0.3893717
scaiDE	9.8368592	1	0.0017105
inotropic_score	5.7501105	1	0.0164876
MCS	0.0270504	1	0.8693611
GLOBAL	15.8791549	8	0.0441428

From the table above, the global p-value is 0.0196537, which is below 0.05. This indicates that there is some evidence of a violation of the proportional hazards assumption means that the effect of one or more covariates on the hazard function is not constant over time when considering all the covariates.and also This suggests that the model may need to be adjusted to account for the time-varying effects of the covariates.

Individually, most covariates (age, sex, CKD, etiology, inotropic_score, and MCS) do not show significant evidence of violating the proportional hazards assumption, though sex and CKD on the line.

evaluation of proportional hazards assumption

```
ggcoxzph(checkPH[1:3])
```

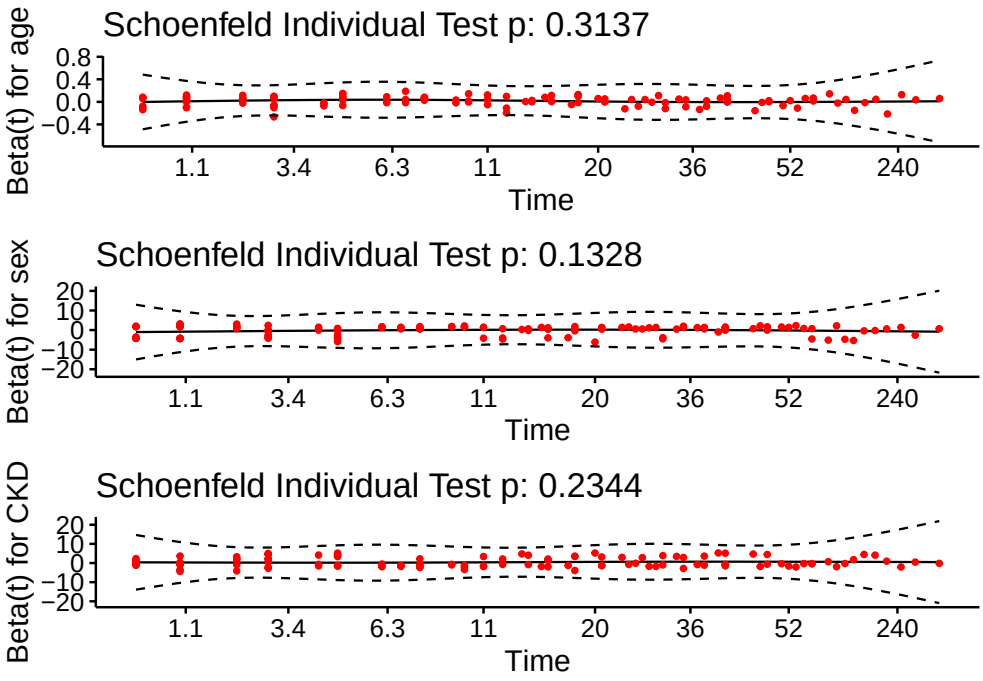


Figure 11: Evaluation of Proportional Hazards Assumption-1

```
ggcoxzph(checkPH[1:3])
```

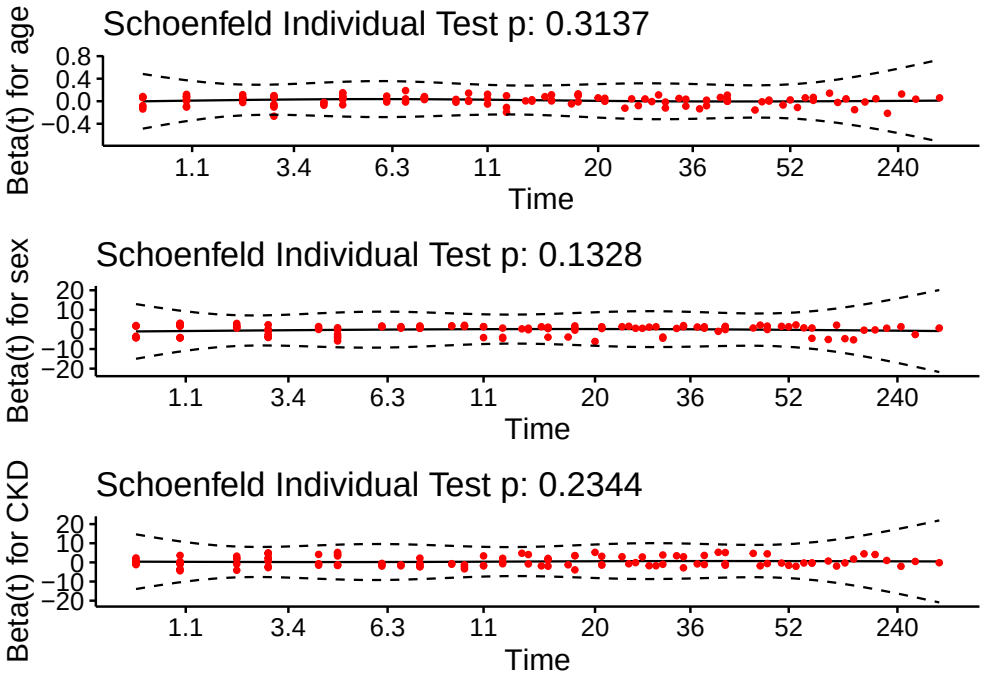


Figure 12: Evaluation of Proportional Hazards Assumption-2

```
ggcoxzph(checkPH[4:7])
```

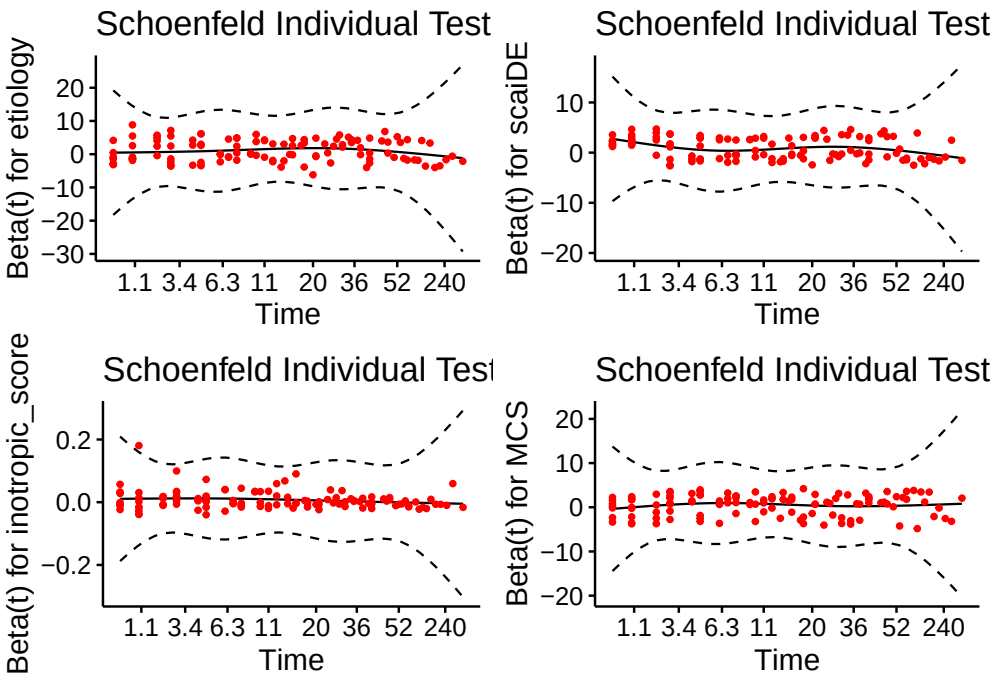


Figure 13: Evaluation of Proportional Hazards Assumption-2

Overall, the plots do not show any strong evidence against the proportional hazards assumption for each

covariate. In this case, none of the covariates show a clear trend suggesting that there is a violation of the proportional hazards assumption.

Performance evaluation

At this point it is necessary to evaluate the performance of the model. The evaluation can be carried out by exploiting the creation of three different graphs that evaluate: *discrimination*, *calibration* and *net benefit*.

```
# Computing model for performance evaluation
model_eval <- coxph(formula = Surv(TimeFromCSDays, DeathOrHT != 0) ~ age + sex + CKD + etiology + scaiD

# Calling survfit function
fit <- survfit(model_eval, newdata = cs)
cs$risk <- 1 - as.numeric(summary(fit, times = 180)$surv)

# Calling score function from riskRegression library
score <- Score(list("model1" = model_eval),
               formula = Surv(TimeFromCSDays, DeathOrHT != 0) ~ 1,
               data = cs, conf.int = TRUE,
               times = 180,
               plots = c("calibration", "ROC"))
```

Calibration plot

```
par(mfrow = c(1, 1))

# Plotting calibration plot for 180 days
plotCalibration(score, cens.method = "local", method = "quantile", q = 10)
title(main = "Calibration at 180 days")
```

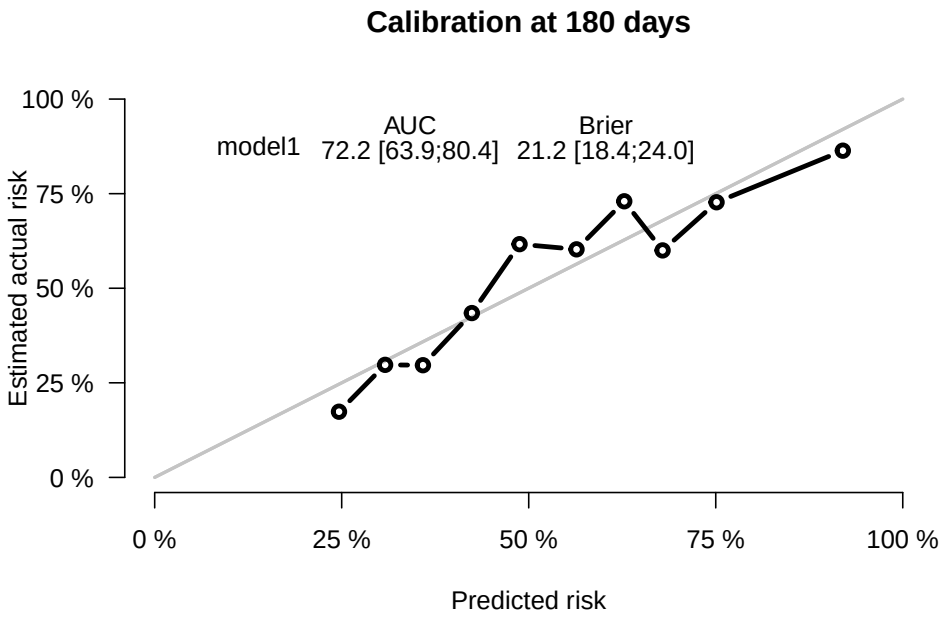


Figure 14: Calibration Plot

The calibration plot at 180 days shows the relationship between predicted and observed risks from the Cox model over the 180 day's period.

The x-axis shows the predicted probability of death or heart transplant within 180 days, and the y-axis shows the actual observed risk. Ideally, points should fall on the diagonal line, indicating perfect calibration. The black points represent model calibration across different data quantiles, with points close to the diagonal line indicating good calibration.

and also when looking at the AUC score, we have AUC score 72.2% with a 95% confidence interval of [63.9, 80.4], Indicating moderate discrimination ability, showing the model can reasonably distinguish between patients who will experience the event and those who will not. on the other hand, Brier Score is at around 21.2 with a 95% confidence interval of [18.4, 24.0].

Overall, the Cox model shows good calibration and moderate discrimination ability for predicting the risk of death or heart transplant within 180 days, with some overestimation at higher predicted risk levels.

```
# Clear previous plotting layout
par(mfrow = c(1, 1))


# Plotting ROC curve for 180 days
plotROC(score, cens.method = "local")
title(main = "Time-dependent ROC at 180 days")
```

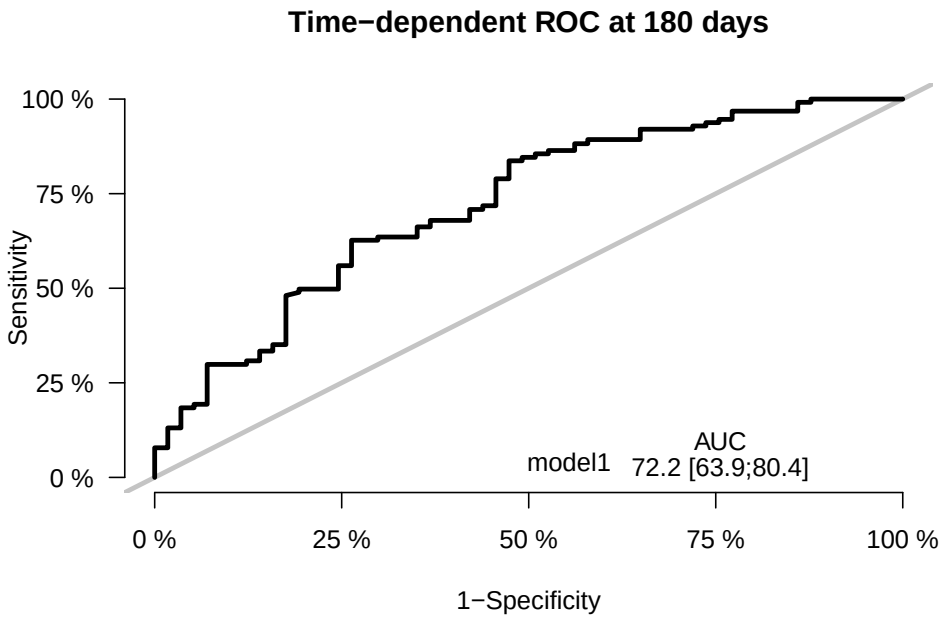


Figure 15: ROC Curve

In this plot, we have Sensitivity on (y-axis) which measures the proportion of actual positives correctly identified by the model and the Specificity (1 - x-axis) measures the proportion of actual negatives correctly identified. in our case, the curve indicates that the model performs better than random guessing, with reasonable trade-offs between sensitivity and specificity over the 180 days period.

To sum up, the ROC curve and the AUC value suggest that the model has a good discriminative ability, with reasonable accuracy in predicting the risk of death in the absence of HT & with HT within 180 days. There is still room for improvement, particularly in enhancing the model's sensitivity and specificity.

```
# Computing Net Benefit
dca(Surv(TimeFromCSDays, DeathOrHT) ~ risk,
    data = cs, time = 180)
```

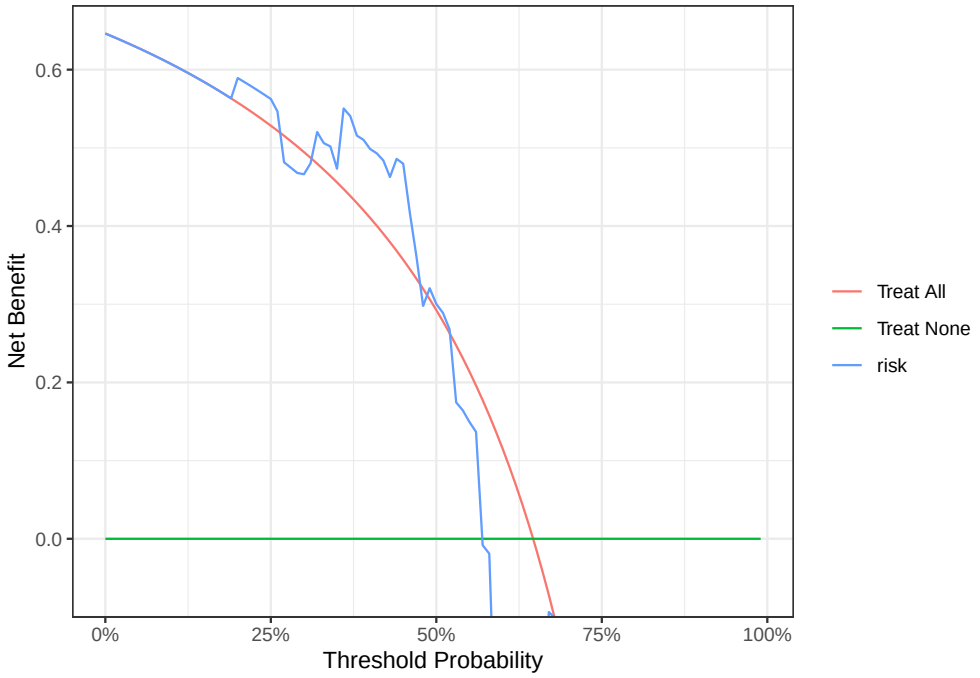


Figure 16: Net Benefit

This plot is Net Benefit graph which illustrates the efficiency of the predictive model across various threshold probabilities. We can say that this predictive model is efficient as long as the net benefit (blue line) remains above both the Treat None (green line) and Treat All (red line) strategies

we can see that in the case of high thresholds probabilities (around 50% and above), the net benefit of the predictive model (blue line) flattens and approaches the Treat None line (green). we can say that at high thresholds, the model's benefit diminishes, aligning with the strategy of not treating any patients. Whereas for low threshold probabilities (at around 0%), the blue curve follows a trend similar to the Treat All line (red). This means at very low thresholds, the model's recommendations are almost equivalent to treating all patients, capturing most true positives.

on the other hand, The small dips below Treat All suggest occasional inefficiencies, and overall the model performs well across a range of threshold probabilities.

Risk prediction

```
cs$MCS <- as.factor(cs$MCS)

cs$CKD <- as.factor(cs$CKD)

cs$sex <- as.factor(cs$sex)

cs$scaiDE <- as.factor(cs$scaiDE)

cs$etiology <- as.factor(cs$etiology)
```

```r
# Creating Cox model for composite endpoint (death or HT)
cox_model_comp <- coxph(Surv(TimeFromCSDays, DeathOrHT != 0) ~ age + sex + CKD + etiology + scaiDE + ino
# Printing results using kable
result_comp <- finalfit::fit2df(cox_model_comp, condense = FALSE)
```

```r
str(cs)
```

```
## 'data.frame':    219 obs. of  11 variables:
##  $ record_id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ TimeFromCSDays : num  250 172 128 1 540 497 420 24 52 4 ...
##  $ DeathOrHT      : int  0 1 0 1 0 0 0 1 2 2 ...
##  $ MCS            : Factor w/ 2 levels " No "," Yes ": 2 2 1 2 1 2 2 2 2 2 ...
##  $ age            : int  58 58 83 64 72 72 36 58 58 56 ...
##  $ sex            : Factor w/ 2 levels " F "," M ": 2 2 2 1 2 2 2 2 2 2 ...
##  $ CKD            : Factor w/ 2 levels "0","1": 1 2 1 1 1 2 2 1 1 1 ...
##  $ etiology       : Factor w/ 3 levels " ADHF "," AMI ",..: 1 1 2 2 1 3 3 2 1 3 ...
##  $ scaiDE         : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 2 1 1 ...
##  $ inotropic_score: int  5 10 8 10 104 14 22 8 6 4 ...
##  $ risk           : num  0.474 0.64 0.261 0.693 0.628 ...
```

```r
# Create the new patient data with consistent factor levels and no extra spaces
young_patient <- data.frame(age = 16,
                            sex = factor( ' M ' , levels = levels(cs$sex)),
                            CKD = factor(1 , levels = levels(cs$CKD)),
                            etiology = factor(" AMI ", levels = levels(cs$etiology)),
                            scaiDE = factor(1  , levels = levels(cs$scaiDE)),
                            inotropic_score = 200,
                            MCS = factor(" Yes ", levels = levels(cs$MCS)))

# Ensure there are no missing values in the new data

#View(young_patient)
# Fit the survival model
fit_young <- survfit(cox_model_comp, newdata = young_patient)
```

```r
elderly_patient  <- data.frame(age = 86,
                            sex = factor(" F ", levels = levels(cs$sex)),
                            CKD = factor( 0 , levels = levels(cs$CKD)),
                            etiology = factor(" Other ", levels = levels(cs$etiology)),
                            scaiDE = factor( 0 , levels = levels(cs$scaiDE)),
                            inotropic_score = 0,
                            MCS = factor(" No ", levels = levels(cs$MCS)))
```

```r
fit_elderly <- survfit(cox_model_comp, newdata = elderly_patient)
#View(elderly_patient)
```

```r
set.seed(123)  # For reproducibility
numero <- sample(1:nrow(cs), 1)
random_patient <- cs[numero, c("age", "sex", "CKD", "etiology", "scaiDE", "inotropic_score", "MCS")]
```

```r
# Combine the hypothetical and randomly selected patients
patients <- rbind(young_patient, elderly_patient, random_patient)


# Refit the Cox model with the correct factor levels
cox_model_comp <- coxph(Surv(TimeFromCSDays, DeathOrHT != 0) ~ age + sex + CKD + etiology + scaiDE + in

# Predicting survival probability for each patient at 180 days
fit_young <- survfit(cox_model_comp, newdata = young_patient)
fit_elderly <- survfit(cox_model_comp, newdata = elderly_patient)
fit_random <- survfit(cox_model_comp, newdata = random_patient)

# Calculate probabilities
young_prob <- 1 - summary(fit_young, times = 180)$surv
elderly_prob <- 1 - summary(fit_elderly, times = 180)$surv
random_prob <- 1 - summary(fit_random, times = 180)$surv

# Create a table with the results
tabella_prob <- data.frame(
  Patient = c("Young", "Elderly", "Random"),
  Probability = c(young_prob, elderly_prob, random_prob)
)

# Print the results with kable
kable(tabella_prob, booktabs = T, caption = "Event Probability for Three Patients at 180 Days") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 9: Event Probability for Three Patients at 180 Days

Patient	Probability
Young	0.9393684
Elderly	0.2943214
Random	0.6564873

Summary

The initial exploratory data analysis (EDA) of a dataset that included 219 patients and variables such as age, gender, chronic kidney disease (CKD), etiology, scaiDE, inotropic score, and mechanical circulatory support (MCS) showed useful results. Histograms and boxplots showed the distribution and correlations between the variables. However, the Schoenfeld residuals test revealed certain violations of the proportional hazards assumption, especially when all factors were considered combined. In addition, the nonlinearity of continuous variables such as age and inotropic score impacted model accuracy. The model's performance was examined using the ROC curve, which revealed an Area Under the Curve (AUC) of 72.2%, showing moderate discriminative capacity. The calibration plot after 180 days showed adequate calibration but

revealed some overestimation at higher anticipated risk levels. The Brier score is 21.2 suggested moderate predictive accuracy.

Using the fitted Cox model, we calculated survival probabilities at 180 days for three hypothetical patients: a young patient, an elderly patient, and a randomly selected patient from the dataset. The event probabilities for these patients were 93.94%, 29.43%, and 65.65%, respectively and the decision curve analysis further shows that the predictive model provided net benefits over treating all or none scenarios, especially at lower threshold probabilities.

To improve the model's predictive accuracy, Maybe increasing the sample size and including a variety of clinical characteristics, such as prior diseases, lifestyle factors, and exposure to common risk factors. Using non-linear modeling approaches or converting continuous variables may help the model conform with the proportional risks assumption.and also investigating alternative modeling methods, such as machine learning techniques, may yield higher predicted accuracy.