Survival Analysis

Amit Agarwal, Lorena Romeo, Nikolai Len, Quentin Camilleri

Data set: Survival of patients awaiting transplant in the Jasa heart data of the survival package.

Objective:

Assess the effect on survival of transplantation, treating the patient population as homogeneous, while in the influence of a number of covariates was investigated through pairwise correlations and explore techniques to see the simultaneous effect of several covariates and see for what values of these covariates, if any, transplantation or surgery is likely to prolong survival.

Load Libraries

```
invisible(lapply(c("dplyr", "tidyr", "ggplot2", "magrittr", "lubridate", "BSDA",
    "tidyverse","broom", "xtable", "webr", "gtsummary","modelsummary",
    "epiDisplay", "mgcv", "survival", "ggfortify", "gridExtra", "survminer",
    "epiR", "swimplot", "muhaz", "asaur", "maxLik", "survivalROC", "plyr",
    "glmnet", "tidycmprsk", "mstate", "cmprsk", "timeROC", "survAUC",
    "tidycmprsk","openxlsx","VSURF", "Hmisc","pec","riskRegression","car"),
    library, character.only = TRUE))
```

Datasets

Three dataset are available. 1. jasa, the original data with 103 observations 2. heart (the main data set) with 172 observations 3. stanford2 (Stanford Heart Transplant data in a different format) with 184 observations (ignored as the number of subjects are not the same). jasa, jasa1, heart, all have 103 subjects while stanford2 has 184 subjects. id and subject are the same. actual_age = accept.dt - birth.dt futime = fu.date - accept.dt wait.time = tx.date - accept.dt stop = fu.date - accept.dt + 1 = end of followup - acceptance into program + 1 wait.time = tx.date - accept.dt + 1 = transplant date - acceptance into program + 1 (stop - start) time between the events. event = fustat = death of patient

We use jasa data, that is the most complete, and extract relevant columns.

```
# Load the datasets
jasa_data <- jasa
colnames(jasa_data)[colnames(jasa_data) == "age"] <- "actual_age"
jasa_data <- cbind(id = 1:103, jasa_data)
jasa_data$age <- jasa_data$actual_age - 48
jasa_data$stop <- as.numeric(jasa_data$fu.date - jasa_data$accept.dt + 1)
jasa_data$age_group <- cut(jasa_data$actual_age, breaks = seq(0, 80, 10),
              labels = c("0-10", "10-20", "20-30", "30-40", "40-50",
                    "50-60", "60-70", "70-80"))
jasa_data$duration_e_1 <- as.numeric(jasa_data$fu.date - jasa_data$accept.dt) / 365.25

# Check for censoring in the dataset
jasa_data$censored <- ifelse(jasa_data$fustat == 0, TRUE, FALSE)
print(table(jasa_data$censored))

##
## FALSE  TRUE
##   75   28
```

75 observations are not censored, i.e. death has occurred within the end of the study. 28 subjects are censored, i.e. the event has not occurred by the end of the study period.

```
data_uncensored <- jasa_data %>% filter(fustat == 1)
jasa_data$event <- ifelse(jasa_data$censored, 0, 1)
```

```r
# Removing specified redundant or useless covariates
jasa_data <- dplyr::select(jasa_data, -fustat, -mismatch, -hla.a2, -mscore, -reject)
#str(jasa_data) #commented out to reduce pages
#'data.frame':  103 obs. of  16 variables:
#id, birth.dt, accept.dt, tx.date, fu.date, surgery, actual_age, futime, wait.time, transplant,
#age, stop, age_group, duration_e_1, censored, event

surv_object <- Surv(time = jasa_data$stop, event = jasa_data$event)
```

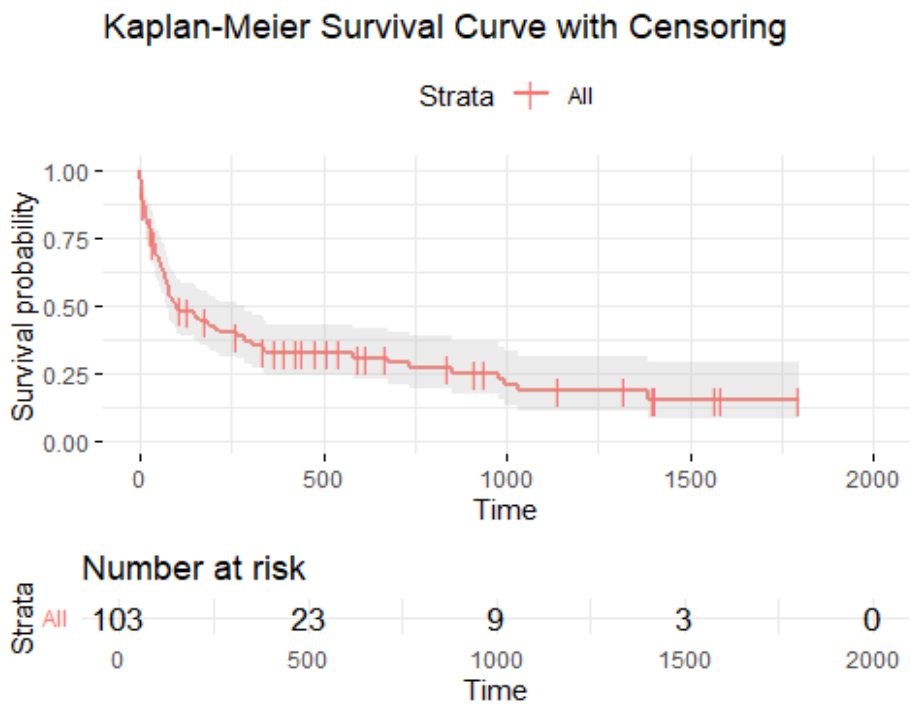
Modelling: Survival and Hazard Curves

Survival: without considering grouping or stratification based on other variables

```r
# Kaplan-Meier estimator
km_fit <- survfit(surv_object ~ 1,data=jasa_data)
km_summary <- summary(km_fit)
print(paste("survival probabilities: ", min(km_summary$surv)))

## [1] "survival probabilities:  0.151912117361313"

print(paste("survival times: ", max(km_summary$time)))

## [1] "survival times:  1387"
```

We can see that over time, the probability of survival steadily declines, in fact we see that after 1387 days of study (~3.8 years) only 15.2% of the individuals are expected to survive beyond this.

```r
ggsurvplot(km_fit, conf.int = TRUE, risk.table=TRUE, ggtheme = theme_minimal(),
    title = "Kaplan-Meier Survival Curve with Censoring", censor.shape = '|',
    censor.size = 4)
```



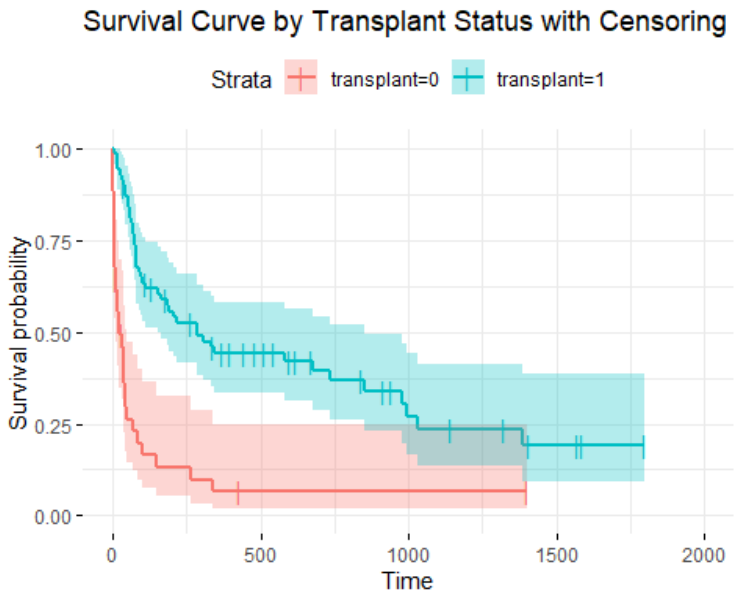
The survival plot confirms that the probability of survival decreases over time. Initially, the survival probability drops sharply and then continues to decline at a slower rate as time progresses.

By the end of the observation period, the survival probability is quite low, indicating that most participants either experienced the event or were censored.

Survival: Grouping and Stratification modelling

Kaplan-Meier estimator stratified by transplant status

```
km_fit_transplant <- survfit(surv_object ~ transplant, data = jasa_data)
#summary(km_fit_transplant) # commented out to reduce the pages
ggsurvplot(km_fit_transplant, conf.int = TRUE, ggtheme = theme_minimal(),
      title = "Survival Curve by Transplant Status with Censoring",
      censor.shape = '|', censor.size = 4)
```

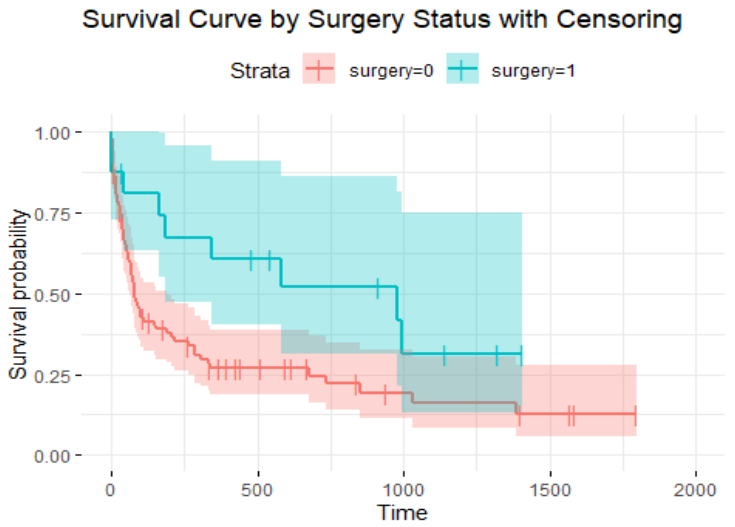


By the end of the study only 19% of individuals who received a transplant were still alive compared to only 6.5% of those who did not receive a transplant.

The survival curve for transplanted patients declines more gradually, indicating that they experienced better longevity compared to those who did not receive a transplant.

Kaplan-Meier estimator stratified by surgery

```
km_fit_surgery <- survfit(surv_object ~ surgery, data = jasa_data)
#summary(km_fit_surgery) #commented out to reduce the pages
ggsurvplot(km_fit_surgery, conf.int = TRUE, ggtheme = theme_minimal(),
      title = "Survival Curve by Surgery Status with Censoring",
      censor.shape = '|', censor.size = 4)
```



Stratifying by surgery status reveals that by the end of the study, approximately 13% of individuals who did not undergo surgery were still alive, compared to ~31% of those who had surgery.

Additionally, the plot shows that the survival curve for patients who had surgery declines more gradually and maintains a higher survival rate over time. This analysis suggests that surgery positively impacts patient survival.

Hazard: based on risk groups

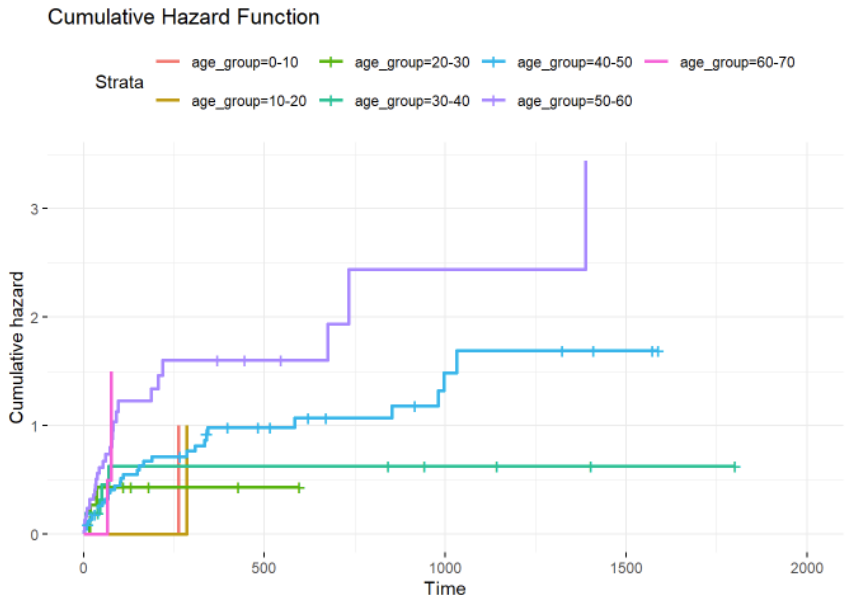
Cumulative hazard function using Nelson-Aalen estimator by age-group

```
na_fit_age <- survfit(Surv(stop, event) ~ age_group,
            data = jasa_data, type = "fleming-harrington")

#summary(na_fit_age) # commented out to reduce the pages
summary(na_fit_age)$table # See results for breakdown by age groups
```

```
##              records n.max n.start events     rmean se(rmean) median 0.95LCL
## age_group=0-10       1     1     1      1  828.4307   0.00000    263     263
## age_group=10-20      1     1     1      1  842.3374   0.00000    285     285
## age_group=20-30      8     8     8      3 1173.9715 314.99539     NA      36
## age_group=30-40     11    11    11      5  979.7527 290.85939     NA      50
## age_group=40-50     48    48    48     34  574.7142 108.57838    188     100
## age_group=50-60     32    32    32     29  260.1087  95.43696     61      32
## age_group=60-70      2     2     2      2  457.1251 276.56721     77      66
##              0.95UCL
## age_group=0-10      NA
## age_group=10-20     NA
## age_group=20-30     NA
## age_group=30-40     NA
## age_group=40-50    852
## age_group=50-60    186
## age_group=60-70     NA
```

In the above summary we note that the age-group 50-60 has the highest risk of experiencing the event, in fact the survival rate is only ~3% during the study period. While the age-group 20-30 has the lowest risk, with a survival rate of ~65%.

```
ggsurvplot(na_fit_age, fun = "cumhaz", ggtheme = theme_minimal(),
      title = "Cumulative Hazard Function")
```



The plot shows that Age group 50-60 (purple) exhibits the highest risk. Age group 20-30 (green) has the lowest cumulative hazard even if we do not have complete data on the long run. The cumulative hazard increases over time for most age groups, but at different rates. This suggests that the risk of the event increases as time progresses, but it is not uniform across age groups. Therefore we conclude that the hazard risk is likely to be impacted by the age of the patient.

Tests to compare Survival distribution between groups

Stratified Log-rank test on transplant+surgery

```
stratified_logrank <- survdiff(surv_object ~ transplant + strata(surgery),
                               data = jasa_data)
print(paste("Log-rank test for transplant + surgery : ", stratified_logrank))
```

```
## [1] "Log-rank test for transplant + surgery :  c(`transplant=0` = 34, `transplant=1` = 69)"
## [2] "Log-rank test for transplant + surgery :  c(27, 39, 3, 6)"
## [3] "Log-rank test for transplant + surgery :  c(12.9687343286493, 53.0312656713507, 0.39775641025641, 8.60224358974359)"
## [4] "Log-rank test for transplant + surgery :  c(10.3597418346794, -10.3597418346794, -10.3597418346794, 10.3597418346794)"
## [5] "Log-rank test for transplant + surgery :  26.7066143880863"
## [6] "Log-rank test for transplant + surgery :  2.36809494670142e-07"
## [7] "Log-rank test for transplant + surgery :  c(`surgery=0` = 87, `surgery=1` = 16)"
## [8] "Log-rank test for transplant + surgery :  survdiff(formula = surv_object ~ transplant + strata(surgery), data = jasa_data)"
```

The chi-square [5] is quite high, therefore suggesting a stronger difference in survival between the groups. The p-value, indicator [6], associated to the chi-square test statistic is very small, indicating that the observed differences in survival between the groups are highly statistically significant. This suggests strong evidence against the null hypothesis of no difference in survival.

Log-rank test on transplant vs. no transplant

```
logrank_test_transplant <- survdiff(surv_object ~ transplant, data = jasa_data)
print(paste("Log-rank test for transplant: ", logrank_test_transplant))
```

```
## [1] "Log-rank test for transplant:  c(`transplant=0` = 34, `transplant=1` = 69)"
## [2] "Log-rank test for transplant:  c(30, 45)"
## [3] "Log-rank test for transplant:  c(12.091021143621, 62.908978856379)"
## [4] "Log-rank test for transplant:  c(9.64838436230081, -9.64838436230081, -9.64838436230081, 9.64838436230081)"
## [5] "Log-rank test for transplant:  33.2419928181372"
## [6] "Log-rank test for transplant:  8.13741373144396e-09"
## [7] "Log-rank test for transplant:  survdiff(formula = surv_object ~ transplant, data = jasa_data)"
```

The small p-value indicates that the transplant has a highly significant impact on survival. Thus, receiving a transplant affects survival times. The high chi-square value (33.24) further supports the significant effect.

Log-rank test on surgery vs. no surgery

```
logrank_test_surgery <- survdiff(surv_object ~ surgery, data = jasa_data)
print(paste("Log-rank test for surgery: ",logrank_test_surgery))
```

```
## [1] "Log-rank test for surgery:  c(`surgery=0` = 87, `surgery=1` = 16)"
## [2] "Log-rank test for surgery:  c(66, 9)"
## [3] "Log-rank test for surgery:  c(58.5875940157749, 16.4124059842251)"
## [4] "Log-rank test for surgery:  c(12.3658501352826, -12.3658501352826, -12.3658501352826, 12.3658501352826)"
## [5] "Log-rank test for surgery:  4.4431852136238"
## [6] "Log-rank test for surgery:  0.0350408140477288"
## [7] "Log-rank test for surgery:  survdiff(formula = surv_object ~ surgery, data = jasa_data)"
```

The chi-square statistic (4.443) and the p-value (0.035) suggest that there is a statistically significant difference between the survival curves of the two groups surgery and not surgery. The low p-value indicates that the difference in survival between surgery and non-surgery groups is significant, but less significant than the transplant group.

Log-rank test on age_groups

```
logrank_test_age <- survdiff(surv_object ~ age_group, data = jasa_data)
print(paste("Log-rank test for age groups: ",logrank_test_age))
```

```
## [1] "Log-rank test for age groups:  c(`age_group=0-10` = 1, `age_group=10-20` = 1, `age_group=20-30` = 8, `age_group=30-40` = 11,
`age_group=40-50` = 48, `age_group=50-60` = 32, `age_group=60-70` = 2)"
## [2] "Log-rank test for age groups:  c(1, 1, 3, 5, 34, 29, 2)"
## [3] "Log-rank test for age groups:  c(0.936259415948052, 0.993402273090909, 5.2401479103105, 9.75279429838162,
39.5726408080367, 17.4736596990747, 1.03109559515754)"
## [4] "Log-rank test for age groups:  c(0.916346052775715, -0.0156823759628872, -0.0723331922130203, -0.0993992897026912, -
0.490725775493497, -0.224949401710104, -0.0132560176935162, -0.0156823759628872, 0.970223603796124, -0.0755024599201032, -
```

```
0.107322458970398, -0.524003086417867, -0.234457204831353, -0.0132560176935162, -0.0723331922130203, -0.0755024599201032,
4.79565752612084, -0.569917870003763, -2.70850376064568, -1.28836542419102, -0.081034819147249, -0.0993992897026912, -
0.107322458970398, -0.569917870003763, 8.16893934692461, \n-5.20832063365711, -2.07651372612089, -0.107465368469759, -
0.490725775493497, -0.524003086417867, -2.70850376064568, -5.20832063365711, 18.4386628573733, -8.99863570111134, -
0.508473900047858, -0.224949401710104, -0.234457204831353, -1.28836542419102, -2.07651372612089, -8.99863570111134,
13.0983321628675, -0.275410704902819, -0.0132560176935162, -0.0132560176935162, -0.081034819147249, -0.107465368469759, -
0.508473900047858, -0.275410704902819, 0.998896827954718)"
## [5] "Log-rank test for age groups:  12.9962514916591"
## [6] "Log-rank test for age groups:  0.0430955121328494"
## [7] "Log-rank test for age groups:  survdiff(formula = surv_object ~ age_group, data = jasa_data)"
```

The chi-square statistic (12.996) and the p-value (0.043) suggest statistically significant difference between the survival curves of the age-groups. The low p-value indicates that the difference in survival between the age-groups is significant, but less significant than the transplant group.

Our conclusion is that all groups are significantly impacting the survival rate but the transplant is the most impactful.

Cox proportional hazards model

To understand how various covariates influence survival time

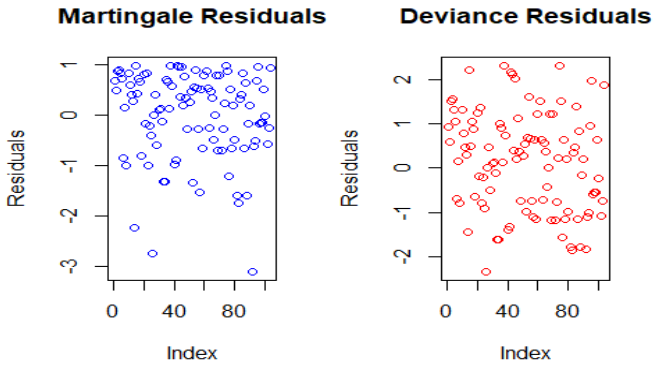
```
cox_model <- coxph(surv_object ~ age_group+surgery+transplant,data = jasa_data)
tbl_regression(cox_model)
```

Characteristic	log(HR)[1]	95% CI[1]	p-value
age_group			
0-10	—	—	
10-20	1.8	-1.0, 4.7	0.2
20-30	0.32	-2.0, 2.6	0.8
30-40	0.07	-2.1, 2.2	>0.9
40-50	1.2	-0.81, 3.2	0.2
50-60	2.1	0.06, 4.2	0.044
60-70	2.7	0.17, 5.2	0.036
70-80			
surgery	-0.28	-1.0, 0.46	0.5
transplant	-1.9	-2.5, -1.3	<0.001

[1]HR = Hazard Ratio, CI = Confidence Interval

Transplant is likely to highly impact positively the survival rate (due to negative correlation) while age_groups 50-60 and 60-70 highly impact the risk of the death event (due to positive correlation)

```
par(mfrow = c(1, 2))
plot(residuals(cox_model, type = "martingale"), main = "Martingale Residuals",
   ylab = "Residuals", xlab = "Index", col = "blue")
plot(residuals(cox_model, type = "deviance"), main = "Deviance Residuals",
   ylab = "Residuals", xlab = "Index", col = "red")
```

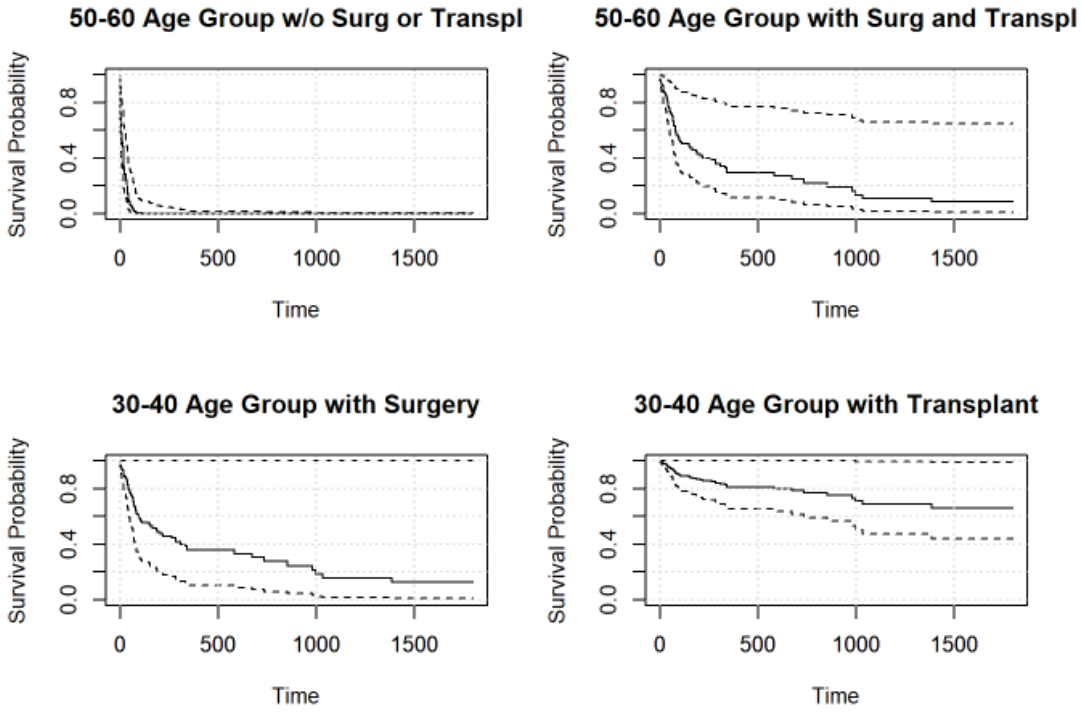


The Martingale plot shows that the residuals are scattered fairly randomly across the index range, with no obvious pattern or systematic structure. The deviance plot shows residuals that are symmetrically distributed around 0. Residuals are randomly scattered across the plot without showing any systematic patterns or trends.

Conclusion: the residuals plot suggest a good fit of the model using covariates: Transplant, Surgery and Age_Group.

Making predictions on new data based on age_group, surgery, and transplant.

```
par(mfrow = c(2, 2))
new_data1 <- data.frame(age_group = c("50-60"), surgery = c(0), transplant = c(0))
new_data2 <- data.frame(age_group = c("50-60"), surgery = c(1), transplant = c(1))
new_data3 <- data.frame(age_group = c("30-40"),surgery = c(1),transplant = c(0))
new_data4 <- data.frame(age_group = c("30-40"),surgery = c(0),transplant = c(1))
surv_fit_specific1 <- survfit(cox_model, newdata = new_data1)
surv_fit_specific2 <- survfit(cox_model, newdata = new_data2)
surv_fit_specific3 <- survfit(cox_model, newdata = new_data3)
surv_fit_specific4 <- survfit(cox_model, newdata = new_data4)
plot(surv_fit_specific1, xlab = "Time", ylab = "Survival Probability", conf.int = TRUE,
   main = "50-60 Age Group w/o Surg or Transpl")
axis(1, at = seq(0, 1500, by = 500)); grid();
plot(surv_fit_specific2, xlab = "Time", ylab = "Survival Probability", conf.int = TRUE,
   main = "50-60 Age Group with Surg and Transpl")
axis(1, at = seq(0, 1500, by = 500)); grid();
plot(surv_fit_specific3,xlab="Time",ylab="Survival Probability",conf.int=TRUE,
   main = "30-40 Age Group with Surgery")
axis(1, at = seq(0,1500, by = 500)); grid();
plot(surv_fit_specific4,xlab="Time",ylab="Survival Probability",conf.int=TRUE,
   main = "30-40 Age Group with Transplant")
axis(1, at = seq(0,1500, by = 500)); grid()
```



Conclusion

From the prediction plots we can see that also for new data, transplant and surgery are highly positively impacting the survival rate of the patients. Between surgery and transplant the results from the model are suggesting that Transplant is the most influencing the survival rate.

While transplant is highly correlated to the survival rate of the patient, the age is highly correlated to the hazard risk to encounter the event of death. This is visible from the prediction plots, that even with transplant or surgery the survival rate is less high with increase of age.