UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

BIOSTATISTICS III
ASSIGNMENT 3

Final

Survival following bone marrow transplant

STUDENT NAME: Nokwanda Themba
STUDENT NO.: THMNOK003

COURSE CODE: PPH7095F

DUE DATE:
Monday 15 May 2023

THIS ASSIGNMENT INCLUDES:

1. **Software used: R**
2. **Data Set(s): transplant_demographic_data_38.csv; transplant_survival_data_38.csv**
3. **THMNOK003_BIO III_Ass3_Final**

LIST OF APPENDICES:

I. Supplementary material

Abstract

Analysis of Leukaemia Patients Undergoing Bone Marrow Transplants

Bone marrow transplantation is a successful treatment for both acute and chronic leukaemia, a type of cancer that affects the blood and bone marrow. The transplant, however, has a risk of mortality that includes dying from a disease relapse or from other causes.

The primary **aim** of this analysis is to determine if there is an association between different types of leukaemia and times to death from relapse or death from other causes following bone marrow transplantation.
Data from patients with three types of leukaemia who underwent bone marrow transplants between 1985 and 1998 was analysed using Cause-specific hazard regression models and covariate-adjusted Cumulative incidence curves.

The results of the analysis showed that there is a significant association between different types of leukaemia and time to death from relapse or from other causes. Among patients with Acute myeloid leukaemia, death due to other causes increases by HR=1.3 (95% CI, 0.8 - 2.1) and HR=1.7 (95% CI, 0.9 -3.2) for death due to relapse. For patients with Chronic myelogenous leukaemia, the Hazard of death due to other causes decreases by HR=0.3 (95% CI, 0.2 - 0.6) and the Hazard of death due to relapse increases HR=1.2 (95% CI, 0.5 -2.6) for the same group.

The results may help us gain a better understanding of the risks associated with bone marrow transplantation, leading to improved therapies and outcomes for patients with leukaemia.

Methods

In this analysis, demographic and survival data included 5410 patients with three types of leukaemia—acute lymphocytic leukaemia *(N =1153)*, acute myeloid leukaemia *(N=2165)*, and chronic myelogenous leukaemia *(N=2092)* who underwent bone marrow transplants between 1985 and 1998 —on time-to-death from relapse or death from other causes.

We observed that 32%*(n=1754*) of the T cell depletion data was missing at random (suppl. Fig.3A) due to observed variables. Given our data did not follow normal distribution, the Conditional multiple imputation approach (Conditional MI) was used to model the conditional distribution of a variable given other variables.Using MICE a separate model for each variable with missing values was fitted, using parameters estimated from other variables in the data set as it provides unbiased estimates of the data and produces more accurate results than simple imputation methods. Multiple imputations (m=100) were conducted on the binary variable dataset using the predictive mean matching (pmm) method.

Statistical analysis was conducted using cause-specific hazard models to obtain Hazard ratios (HR) and 95% confidence interval (CI) to compare hazards between patient groups. The effect of covariates age, transplant reason, same sex, t-cell depletion, as well as the interactions between age and transplant reason on cause-specific hazard were estimated with cox proportional hazard regression for each competing event, separately. Additionally, covariate-adjusted cumulative incidence curves were used to provide an accurate representation of the probability of failure at any given time due to death from relapse and death from other causes, as well as for the different leukaemia disease classifications over time.

The cause specific hazard model diagnostics were performed using cox proportional hazard assumptions, of which were not violated, as well as employing Schoenfeld residual plots (suppl. Fig.4 A-B).It is noted that the residuals follow a constant distribution across failure times, indicating PH assumptions have been met. Finally, given the complexity of our model, final model fitting was measured using AIC.

R version Version: 2023.03.1+446 was used for the analysis of data and $p < 0.05$ was considered as the level of significance for all models.

Results and Discussion

Table 1: Baseline demographic characteristics of the total sample population stratified by type of leukaemia

Characteristic	**Overall,** ***(N=5410)***	**Acute lymphocytic leukaemia,** ***N = 1153***	**Acute myeloid leukaemia,** ***N = 2165***	**Chronic myelogenous leukaemia,** ***N = 2092***
Age (years), Mean(s.d.)	30.8 (14.2)	25.2 (11)	29.7 (12.2)	35 (11.3)
Same sex donor, n(%)	4116 (76%)	827 (72%)	1693 (78%)	1596 (76%)
Time (months), Mean(s.d.)	60 (48.6)	56.6 (49.2)	58.5 (48.9)	63.3 (47.8)
T-cell depletion, n(%)				
No	2659 (73%)	613 (77%)	1160 (78%)	886 (64%)
Yes	997 (27%)	185 (23%)	323 (22%)	489 (36%)
Event status, n(%)				
Censored	3992 (74%)	798 (69%)	1581 (73%)	1613 (77%)
Death from relapse	769 (14%)	235 (20%)	361 (17%)	173 (8.3%)
Death from other causes	649 (12%)	120 (10%)	223 (10%)	306 (15%)

s.d.,standard deviation; n / N (%) - varies owing to missing values
time: time to event or time to censoring (months); Same sex donor: if the donor and recipient were the same sex;
T-cell depletion: reduction in T cells, Yes or No

The results in Table 1. showing baseline characteristics of patients in the total sample stratified by type of leukaemia showed that the study population (N=5410) had a mean age of 30.8 years (s.d 14.2) and about 76% of the total study population had a donor and recipient of the same sex. In addition, there was a higher proportion of patients without T cell depletion (73%) compared to those who had T Cell depletion (27%). In the total study population, patients had a mean time to event or time to censoring of 60 months (s.d. 48.6). Additionally, in terms of event status, 74% patients were censored, 14% had death from relapse and 12% experienced death from other causes. Among those censored 77% were patients diagnosed with CML, 73% were patients diagnosed with AML and 69% were patients diagnosed with ALM. Among those who experienced death from relapse, 20% were patients diagnosed with ALL,17% were patients with AML and 8.3% were patients with CML. Lastly, 10% of patients diagnosed with AML, 10% of patients diagnosed with ALL and 15% of patients diagnosed with CML died from other causes. The mean time to event or censoring was the highest for patients with Chronic Myelogenous Leukaemia(CML) at 63.3 months (s.d. 47.8), followed by patients with Acute Lymphocytic Leukaemia (ALL) and patients with Acute Myeloid Leukaemia (AML) at 56.6 (s.d. 49.2) and 58.5 (s.d. 48.9) months, respectively.

The percentage of patients with the same sex donor and recipient was highest (78%) in the group of 2165 patients with AML, followed by a proportion of 76% among the 2092 patients with CML, and the least (72%) among the 1153 patients with ALL. Among patients who were diagnosed with AML, there was a lower proportion of those who had T cell depletion (22%) compared to those without (78%). Similarly, the proportion of patients with T cell depletion (23%) was lower compared to those without (77%) among patients diagnosed with ALL. Lastly There was a higher proportion of individuals with CML who did not have T cell depletion (64%) compared to those who did (36%).

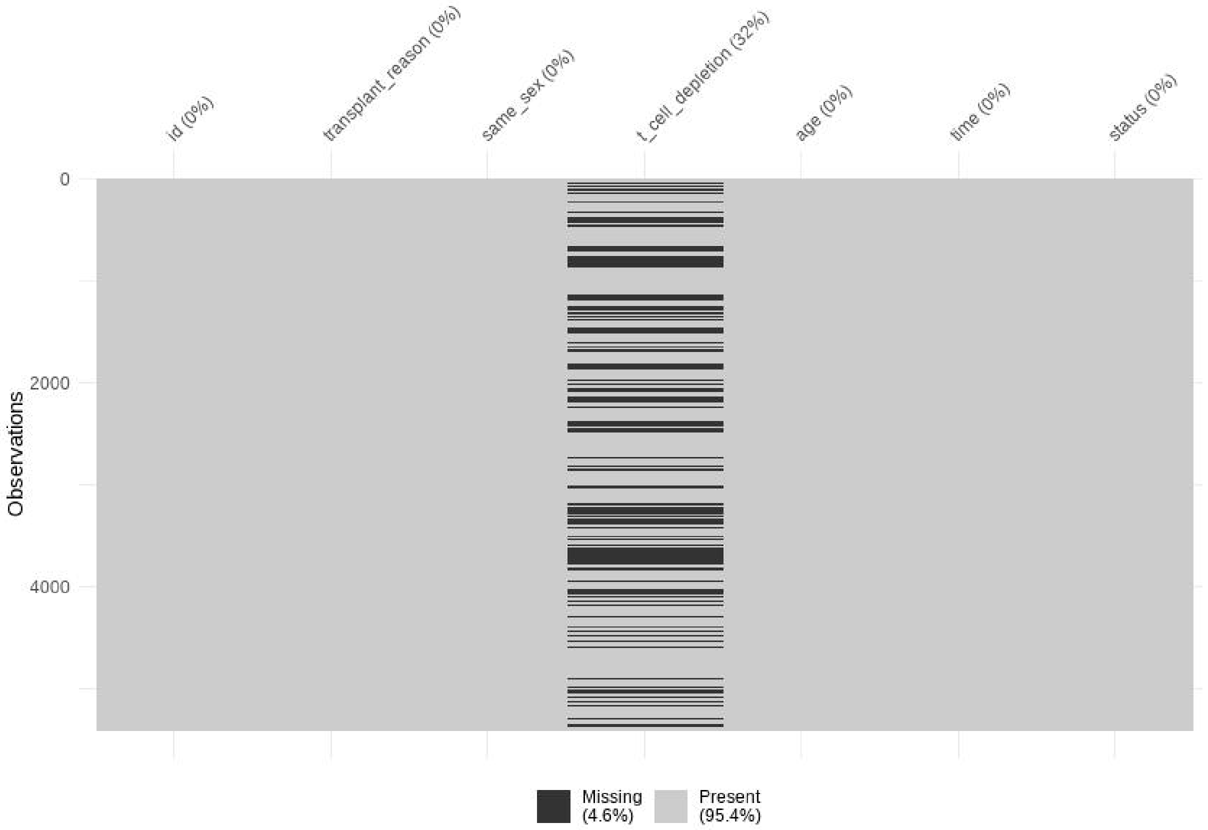

Figure 1.: Missingness map illustrating percentage of missing data in the total data set.

The map of missingness in our data set (Fig. 1.) provides a specific visualisation of the overall percentage of missing data, showing in black the location of missing values (4.6%) and present values (95.4%), also providing information on the percentage of missing values in each variable (suppl. Fig. 3A).This diagram indicates that the T cell depletion variable had most of its values missing (32%).

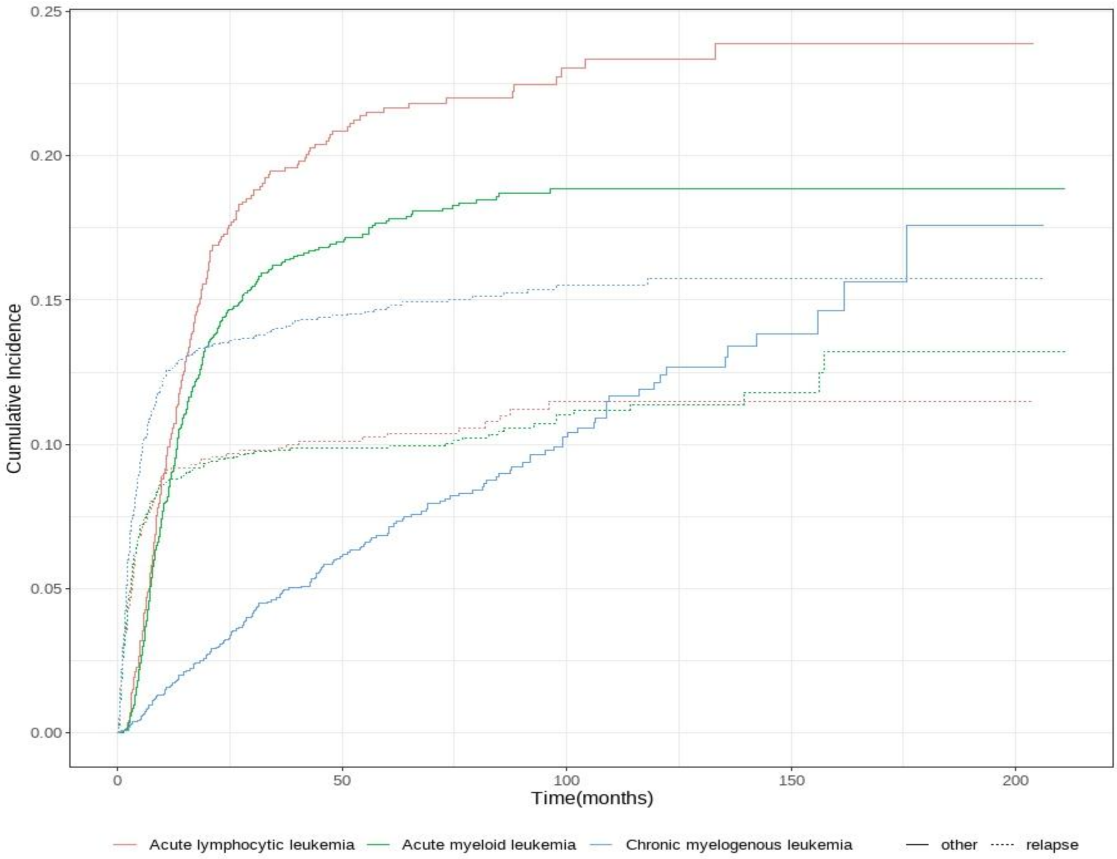

Figure 2.: Cumulative Incidence curves of death due to relapse and other causes stratified by the different leukaemia disease classifications

The predicted Cumulative Incidence curves for the probability of failure due to death from relapse and death from other cases stratified by the different leukaemia disease classifications as competing risks (Fig 2.) indicates that at any given time after the transplant, a patient diagnosed with ALL is more likely to die due to other causes than a patient diagnosed with AML, and a patient diagnosed with AML is more likely to die due to other causes than a patient diagnosed with CML. Alternatively, we can see that at any given time after the transplant, a patient diagnosed with CML is more likely to relapse than a patient diagnosed with AML, and a patient diagnosed with AML is more likely to relapse than a patient diagnosed with CML.

Table 2. Cause-specific hazard ratios and 95% CIs for death due to relapse and other causes in the total sample including an interaction between patient age and transplant reason.

Variables	Cause-specific hazard models	
	Death due to other causes	Death due to relapse
	HR (95% CI)	HR (95% CI)
Age (years)	1.0 (1.0 - 1.0)	1.0 (1.0-1.0)
T cell depletion		
No TCD	-	-
TCD	1.5 (1.3 - 1.8)	1.4 (1.2 - 1.8)
Type of leukaemia		
Acute lymphocytic leukaemia	-	-
Acute myeloid leukaemia	1.3 (0.8 - 2.1)	1.7 (0.9 -3.2)
Chronic myelogenous leukaemia	0.3 (0.2 - 0.6)	1.2 (0.5 - 2.6)
Same sex donor		
No	-	-
Yes	0.9 (0.7 - 1.1)	0.8 (0.6 - 1.0)
Type of leukaemia * age		
Acute myeloid leukaemia * age	1 (1.0 - 1.0)	1 (1.0 - 1.0)
Chronic myelogenous leukaemia * age	1 (1.0 - 1.0)	1.0 (1.0 - 1.0)

1 HR-hazard ratio; CI-confidence interval;
Same sex donor: if the donor and recipient were the same sex;
Type of leukaemia i.e.transplant reason; T-cell depletion: reduction in T cells, Yes or No

In Table 2. holding all other covaries constant, the Hazard of death due to other causes increases by 1.3 (95% CI, 0.8 - 2.1) for patients diagnosed with AML, similarly, the hazard of death due to relapse increases 1.7 (95% CI, 0.9 -3.2) for the same group, indicating the adverse effect of this variable on survival time. For patients with CML, the Hazard of death due to other causes decreases by 0.3 (95% CI, 0.2 - 0.6) and the hazard of death due to relapse increases 1.2 (95% CI, 0.5 -2.6) for the same group.For same sex donors, the hazard of death due to other causes 0.9 (95% CI, 0.7 - 1.1) and death due to relapse 0.8 (95% CI, 0.6 - 1.0) are both reduced, indicating that the variable is beneficial. For those who had a T cell reduction, the hazard of death due to other causes increases 1.5 (95% CI, 1.3 - 1.8) and death due to relapse 1.4 (95% CI, 1.2 - 1.8) both increase, indicating a rather harmful effect on survival. However, the hazard of death due to relapse and death due to other causes was estimated to be 1.0 (95% CI,1.0 - 1.0) for every unit increase in age, indicating that age does not show a significant contribution to death due to relapse and other causes. Similarly, there seems to be no significant contribution to death due to relapse and other causes in interaction between age and AML (1.0 (95% CI,1.0 - 1.0) and age and CML 1.0 (95% CI,1.0 - 1.0).

Conclusion and Limitations

In conclusion, there seems to be a considerable association between different types of leukaemia and times to death from relapse or death from other causes following bone marrow transplantation.Patients diagnosed with Chronic myelogenous leukaemia have an improved survival amongst other leukaemia type groups for death due to other causes. According to death due to other causes and relapse, patients diagnosed with acute myeloid leukaemia show an overall harmful effect. The interaction between age and leukaemia classification seems to show no considerable significance.

For MAR data, multiple imputation is acceptable and can result in accurate estimates, but there is no way to guarantee that bias has been eliminated.The imputed values may or may not correspond to "real-life" values. The Cox regression model assumes independence between the event of interest and other possible events, while the cause-specific hazard model assumes that the risk of failure depends on the cause of failure and may be dependent on each other.

Cause-specific hazard models are used in competing risk analysis to estimate the effect of covariates on the hazard of a specific cause of failure. However, there are some limitations, such as ignoring the dependence between competing risks which may result in biased estimates of the cause-specific hazard as the assumption might not be true in reality,making it difficult to assess cause-specific hazard ratios, and not directly addressing the question of whether patients with a particular attribute(e.g. Older patients) have a lower risk **(1)**. These two assumptions can sometimes clash when analysing competing risks, as the hazard function of one cause of failure can affect the hazard function of another cause of failure.

Supplementary material

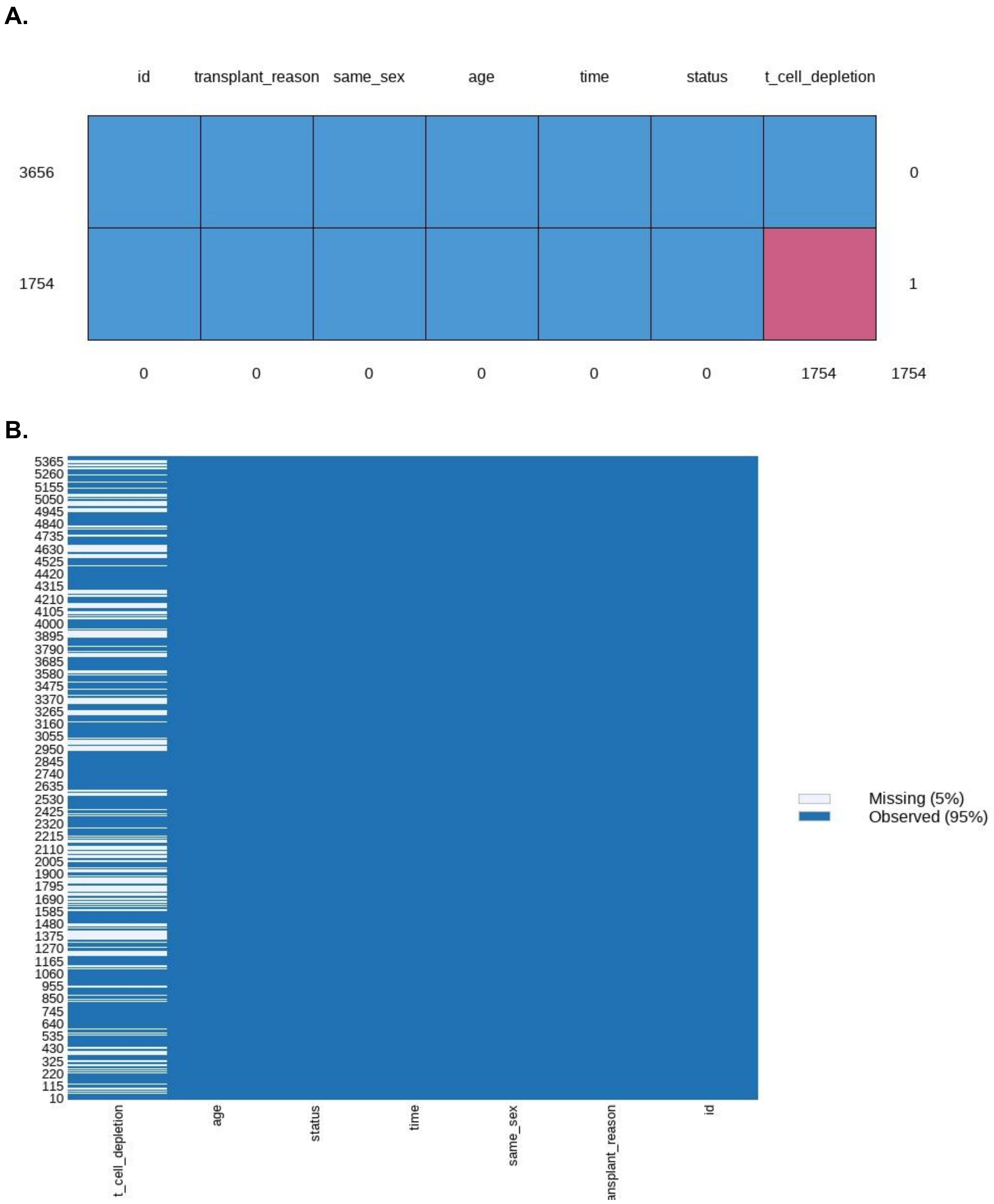

Figure 3.: (A) Missing values pattern present in each variable in a data set **(B)** Missingness map illustrating percentage of missing data in the total data.

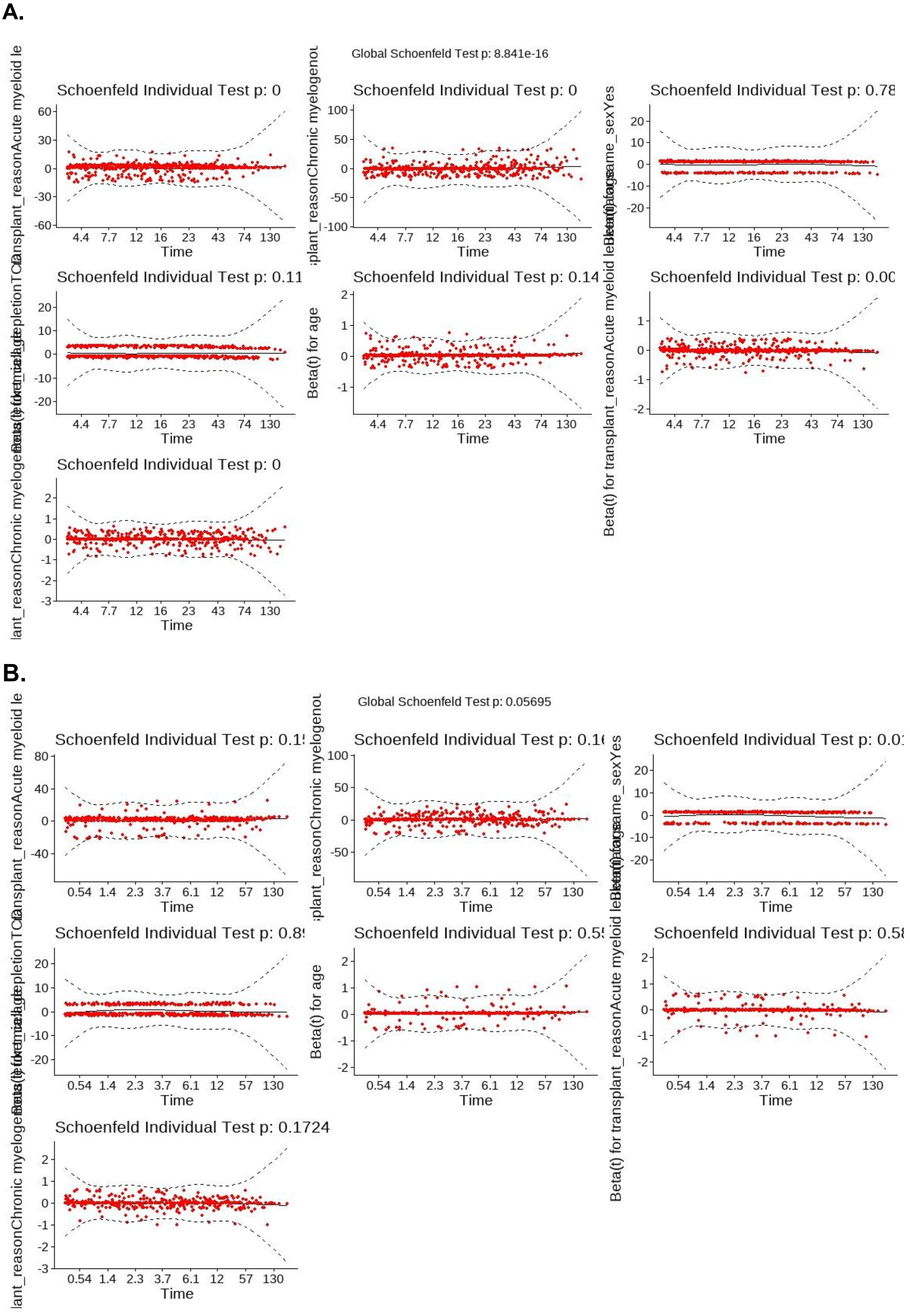

Figure 4.: **(A).**Schoenfeld residuals model diagnostics to check the proportional hazards assumption for death due to other causes **(B)**. Schoenfeld residuals model diagnostics to check the proportional hazards assumption for death due to relapse.

References:

1. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: Possibilities and pitfalls. International Journal of Epidemiology. 2012;41(3):861–70. doi:10.1093/ije/dyr213

R Coding Script

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("skimr")
install.packages("readr")
install.packages("lattice")
install.packages("tidyr")
install.packages("here")
install.packages("tidyverse")
install.packages("ggubr")
install.packages("describer")
install.packages("psych")
install.packages("gtsummary")
install.packages("janitor")
install.packages("stringr")
install.packages("gmodels")
install.packages("summarytools")
install.packages("epiDisplay")
install.packages("epiR")
install.packages("survival")
install.packages("tab")
install.packages("gtools")
install.packages("gtable")
install.packages("stringi")
install.packages("MASS")
install.packages("glue")
install.packages("broom")
install.packages("palmerpenguins")
install.packages("Hmisc")
install.packages("car")
install.packages("caret")
install.packages("ROCR")
install.packages("labelled")
install.packages("modelr")
install.packages("comprehenr")
install.packages("epitools")
install.packages("broom")
install.packages("aod")
install.packages("effectsize")
install.packages("modeldata")
install.packages("haven")
install.packages("interplot")
install.packages("sjPlot")
install.packages("epiR")
install.packages("readr")
install.packages("dplyr")
install.packages("stringr")
install.packages("skimr")
install.packages("psych")
install.packages("hmisc")
install.packages("gmodels")
install.packages("summarytools")
install.packages("epiDisplay")
install.packages("ggpubr")
install.packages("car")
install.packages("MASS")
install.packages("rcompanion")
install.packages("tidyverse")
install.packages("moments")
install.packages("dlookr")
install.packages("ISLR")
install.packages("ggstatsplot")
install.packages("bestNormalize")
install.packages("forecast")
install.packages("geepack")
install.packages("gee")
install.packages("broom")
install.packages("nlme")
install.packages("lme4")
install.packages("broom.mixed")
install.packages("GLMMadaptive")
install.packages("missForest")
install.packages("corrplot")
install.packages("ggeffects")
install.packages("glmmTMB")
install.packages("stargazer")
install.packages("naniar")
install.packages("lubridate")
installed.packages("janitor")
installed.packages("glue")
install.packages("survival")
install.packages("VIM package")
install.packages("missmap")
install.packages("visdat")
install.packages("Amelia")
install.packages("gmodels")
install.packages("summarytools")
install.packages("epiDisplay")
install.packages("tidycmprsk")
install.packages("ggsurvfit")
install.packages("survival")
install.packages("survminer")
install.packages("cmprsk")
install.packages("devtools")
install.packages("MASS")
install.packages("mice")
install.packages("nlme")
install.packages("lme4")
library("survival")
library("glue")
library("lubridate")
library("janitor")
library("naniar")
library("stargazer")
library("glmmTMB")
library("corrplot")
library("gee")
library("geepack") # for GEE modelling
library("broom") # for obtaining confidence intervals from GEE models
library("nlme") # for LMM modelling
library("lme4") # for GLMM modelling
library("broom.mixed") # for obtaining confidence intervals from GLMM models
library("GLMMadaptive") # more GLMM modelling (extensions)
library("missForest") # to generate some missing data at random (extensions)
library("forecast")
library("ggpubr")
library("gmodels")
library("summarytools")
library("epiDisplay")
library("Hmisc")
library("psych")
library("readr")
library("epiR")
library("dplyr")
library("stringr")
library("skimr")
library("gtsummary")
library("gtools")
library("car")
library("MASS")
library("rcompanion")
library("tidyverse")
library("moments")
library("dlookr")
library("ISLR")
library("ggstatsplot")
library("bestNormalize")
library("ggeffects")
library("sjPlot")
library("here")
library("dplyr")
library("skimr")
library("ggplot2")
library("readr")
library("lattice")
library("tidyr")
library("tidyverse")
library("describe")
library("psych")
library("gtsummary")
library("janitor")
library("stringr")
library("gmodels")
library("summarytools")
library("epiDisplay")
library("epiR")
library("survival")
library("tab")
library("gtools")
library("gtable")
library("stringi")
library("MASS")
library("glue")
library("broom")
library("palmerpenguins")
library("Hmisc")
library("car")
library("caret")
library("ROCR")
library("labelled")
library("modelr")
library("comprehenr")
library("epitools")
library("broom")
library("aod")
library("effectsize")
library("modeldata")
library("haven")
library("interplot")
library("nlme")
library("lme4")
library("mice")
library("MASS")
library("devtools")
library("cmprsk")
library("tidycmprsk")
library("ggsurvfit")
library("survival")
library("survminer")
library("gmodels")
library("summarytools")
library("VIM package")
library("missmap")
library("visdat")
library("Amelia")

##-----------------------------------------------------------------------------------------------------------
##Data Analysis
#Set working directory #
setwd("C:/Users/Student/Downloads")
# find what directory you are in
getwd()
#### Read in the data ####
transplant_demographic_data_38 <- read_csv("transplant_survival_data_38")
transplant_demographic_data_38 <-
  read_csv("C:/Users/Student/Downloads/transplant_survival_data_38")
##----------------------------------------------------------------------------------------------------------
```

```
##Exploring Data
#Getting an idea of the DEMOGRAPHIC data
View(transplant_demographic_data_38)##view the whole dataset
print(transplant_demographic_data_38)
head(transplant_demographic_data_38)# first 6 rows
tail(transplant_demographic_data_38)##Last rows
tail(transplant_demographic_data_38, n=10)##Last 10 rows
str(transplant_demographic_data_38)##Provides the structure of the data set
glimpse(transplant_demographic_data_38)
summary(transplant_demographic_data_38)##Provides basic descriptive statistics and frequencies
names(transplant_demographic_data_38) ##Lists variables in the dataset
describe(transplant_demographic_data_38)
skim(transplant_demographic_data_38)
dim(transplant_demographic_data_38)
nrow(transplant_demographic_data_38)
ncol(transplant_demographic_data_38)
##Descriptive Statistics-Age Variable
mean(transplant_demographic_data_38$age) # Mean of all numeric variables
median(transplant_demographic_data_38$age)
sd(transplant_demographic_data_38$age)# Standard deviation
var(transplant_demographic_data_38$age)# Variance
max(transplant_demographic_data_38$age) # Max value
min(transplant_demographic_data_38$age) # Min value
range(transplant_demographic_data_38$age) # Range
IQR(transplant_demographic_data_38$age)
quantile(transplant_demographic_data_38$age)
by(transplant_demographic_data_38$age)
fivenum(transplant_demographic_data_38$age)
length(transplant_demographic_data_38$age)
which.max(transplant_demographic_data_38$age)#Determines the location of the (first) maximum of a numeric vector
which.min(transplant_demographic_data_38$age)#Determines the location of the (first) minimum of a numeric vector
table(transplant_demographic_data_38$age) # Mode by frequencies
prop.table(transplant_demographic_data_38$age)
hist(transplant_demographic_data_38$age, main="Distribution of Age",
    xlab="Age", lwd=3, col="pink")
boxplot(transplant_demographic_data_38$age, horizontal = T, col="pink", main="Distribution of Age", xlab="Age")
#Getting an idea of the SURVIVAL data
View(transplant_survival_data_38)##view the whole dataset
print(transplant_survival_data_38)
head(transplant_survival_data_38)# first 6 rows
tail(transplant_survival_data_38)##Lastsd(transplant_demographic_data_38)
tail(transplant_survival_data_38, n=10)##Last 10 rows
str(transplant_survival_data_38)##Provides the structure of the data set
glimpse(transplant_survival_data_38)
summary(transplant_survival_data_38)##Provides basic descriptive statistics andfrequencies
names(transplant_survival_data_38) ##Lists variables in the dataset
describe(transplant_survival_data_38)
skim(transplant_survival_data_38)
dim(transplant_survival_data_38)
```

```
nrow(transplant_survival_data_38)
ncol(transplant_survival_data_38)
##Descriptive Statistics- Time Variable
mean(transplant_survival_data_38$time) # Mean of all numeric variables
median(transplant_survival_data_38$time)
sd(transplant_survival_data_38$time)# Standard deviation
var(transplant_survival_data_38$time)# Variance
max(transplant_survival_data_38$time) # Max value
min(transplant_survival_data_38$time) # Min value
range(transplant_survival_data_38$time) # Range
IQR(transplant_survival_data_38$time)
quantile(transplant_survival_data_38$time)
by(transplant_survival_data_38$time)
fivenum(transplant_survival_data_38$time)
length(transplant_survival_data_38$time)
which.max(transplant_survival_data_38$time)#Determines the location of the (first)maximum of a
numeric vector
which.min(transplant_survival_data_38$time)#Determines the location of the (first) minimum of a
numeric vector
table(transplant_survival_data_38$time) # Mode by frequencies
freq(transplant_survival_data_38$time, order = "freq")
prop.table(transplant_survival_data_38$time)
hist(transplant_survival_data_38$time, main="Distribution of Time",
   xlab="Age", lwd=3, col="cyan")
boxplot(transplant_survival_data_38$time, horizontal = T, col="cyan", main="Distribution of
Time", xlab="Age")
#------------------------------------------------------------------------------------------------------------
##Data Cleaning
#DEMOGRAPHIC DATA
#Fix structural errors
names(transplant_demographic_data_38)#check names of variables in the data frame;-Lists variables
in the dataset
clean_names(transplant_demographic_data_38)##gives better names to data set
transplant_demographic_data_38 = clean_names(transplant_demographic_data_38)
names(transplant_demographic_data_38)
View(transplant_demographic_data_38)
#Remove duplicates
duplicated(transplant_demographic_data_38)#Identify Duplicates
sum(duplicated(transplant_demographic_data_38))
view(transplant_demographic_data_38)
#Remove irrelevant observations
##Implausible values
#Handle unwanted outliers
#Handle missing data
is.na(transplant_demographic_data_38)
sum(is.na(transplant_demographic_data_38))##find the sum of non-missing values # any missing
data
rowSums(is.na(transplant_demographic_data_38))##Number of missing per variable
sum(rowSums(is.na(transplant_demographic_data_38)))
#=================================================================
##Data Cleaning
#SURVIVIAL DATA
```

```
#Fix structural errors
names(transplant_survival_data_38)#check names of variables in the data frame;-Lists variables in the dataset
clean_names(transplant_survival_data_38)##gives better names to data set
transplant_survival_data_38$status<-factor(transplant_survival_data_38$status,levels = c(0,1,2),labels = c("censored","other","relapse"))
transplant_survival_data_38 = clean_names(transplant_survival_data_38)
names(transplant_survival_data_38)
View(transplant_survival_data_38)
#Remove duplicates
duplicated(transplant_survival_data_38)#Identify Duplicates
sum(duplicated(transplant_survival_data_38))
distinct(transplant_survival_data_38)
#Remove irrelevant observations
##Implausible values
#Handle unwanted outliers
#Handle missing data
is.na(transplant_survival_data_38)
sum(is.na(transplant_survival_data_38))##find the sum of non-missing values # any missing data
rowSums(is.na(transplant_survival_data_38))##Number of missing per variable
sum(rowSums(is.na(transplant_survival_data_38)))
###========================================================================
########Joining Data########
#perform an inner join
inner_join(transplant_demographic_data_38, transplant_survival_data_38, by = "id")
join_inner <- inner_join(transplant_demographic_data_38, transplant_survival_data_38, by ="id")
view(join_inner)
dim(join_inner)#get dimensions
head(join_inner)#preview the new object
glimpse(join_inner)##Now preview the dataset
##Combine Tables
transplant_merged_data_38<-bind_cols(join_inner)##Returns tables placed side by side as a single table
View(transplant_merged_data_38)
##Changing days to months
transplant_merged_data_38$time <- c(transplant_merged_data_38$time/30.4375)
data.frame(transplant_merged_data_38$time)
View(transplant_merged_data_38$time)
##Exploratory data analysis
##baseline demographic data stratified by type of leukaemia
transplant_merged_data_38 %>% tbl_summary(by = transplant_reason,statistic = list(
  all_continuous() ~ "{mean} ({sd})",
  all_categorical() ~ "{n} / {N} ({p}%)"
),
digits = all_continuous() ~ 2 ,missing = "no") %>%
  add_overall() %>%
  modify_caption("Baseline demographic data stratified by type of leukaemia",) %>%
  bold_labels()
##extent of missing data
missmap(transplant_merged_data_38)
vis_miss(transplant_merged_data_38)
vis_dat(transplant_merged_data_38)
```

```
#Amount of missing data
sum(is.na(transplant_merged_data_38$t_cell_depletion))
md.pattern(transplant_merged_data_38)
##--------------------------------------------------------------------------------
#Multiple Imputation
# Traditional mice pipeline goes mice() -> with() -> pool()
##Now, let’s impute the missing values
# impute missing values using mean imputation
##Create an imputation model using mice
set.seed(400)
imputed_data = mice(transplant_merged_data_38, m = 5, maxit = 10, method = "pmm", seed = 400)
summary(imputed_data)
head(imputed_data)
tail(imputed_data)
##check imputed values
imputed_data$imp$t_cell_depletion
sum(is.na(imputed_data$imp$t_cell_depletion))
#check the imputation method used for each variable
imputed_data$meth
# Generate an imputed data set
imputed_complete = mice::complete(imputed_data, action = "long")
summary(imputed_complete)
head(imputed_complete)
tail(imputed_complete)
sapply(imputed_complete, function (x) sum(is.na(x)))
imputed_complete <- complete(imputed_data,1)
#show that we have all of our imputed datasets in this "complete" dataset
table(imputed_complete$.imp)
#visual
densityplot(imputed_data)
stripplot(imputed_data, time ~ status + transplant_reason + same_sex + t_cell_depletion +  age |
.imp, pch = 20)

###----------------------------------------------------------------------------------------
##comparing the probability of failure due to death from relapse and death from other cases
##Plotting Cumulative Incidence Curves
##Fit the CI curve-Google
CI_curve <-survfit(Surv(time = time, event = status) ~ transplant_reason, data =
                    transplant_merged_data_38)
cuminc(Surv(time, status) ~ transplant_reason, data = transplant_merged_data_38) %>%
  ggcuminc(outcome = c("relapse", "other")) + labs(x = "Time(months)")

#using the survival multi-state model
CI_fit<-survfit(Surv(time, status) ~ transplant_reason, data = transplant_merged_data_38) %>%
  ggcuminc(outcome = c("relapse", "other")) + labs(x = "Time(months)")

##Cox Proportional-Hazards Model
##Using the imputed data set -imputed_complete
#Main effects model
##Other
transother.cox <- coxph(Surv(time, status =="other")~ transplant_reason + same_sex +
t_cell_depletion +  age, data = imputed_complete)
```

```r
transother.cox
summary(transother.cox)
#Tabulate-Other
transother.cox %>%
  tbl_regression(
    exp = TRUE)
other_tab <-tbl_regression(transother.cox, exponentiate = TRUE,conf.int = TRUE)
other_tab

##With the interaction term
inter_other.cox <- coxph(Surv(time, status =="other") ~ transplant_reason + same_sex +
t_cell_depletion +  age + age*transplant_reason, data = imputed_complete)
inter_other.cox
summary(inter_other.cox)
#Tabulate-Other x interaction
inter_other.cox %>%
  tbl_regression(
    exp = TRUE)
otherinter_tab <-tbl_regression(inter_other.cox, exponentiate = TRUE,conf.int = TRUE)
otherinter_tab

##Variable selection
#AIC
AIC(transother.cox,inter_other.cox)

#Model Diagnosis
##To test for the proportional-hazards (PH) assumption
#Other
testother.ph <- cox.zph(inter_other.cox, terms = F)
testother.ph
ggcoxzph(testother.ph)
ggcoxdiagnostics(testother.ph,'schoenfeld',ox.scale = 'time')
ggcoxdiagnostics(testother.ph, type = "schoenfeld", linear.predictions = FALSE)

##Relapse
transrelapse.cox <- coxph(Surv(time, status == "relapse") ~ transplant_reason + same_sex +
t_cell_depletion +  age, data = imputed_complete)
transrelapse.cox
summary(transrelapse.cox)
#Tabulate-relapse
transrelapse.cox %>%
  tbl_regression(
    exp = TRUE)
relapse_tab <-tbl_regression(transrelapse.cox, exponentiate = TRUE,conf.int = TRUE)
relapse_tab
##With the interaction term
inter_relapse.cox <- coxph(Surv(time, status =="relapse") ~ transplant_reason + same_sex +
t_cell_depletion +  age + age*transplant_reason, data = imputed_complete)
inter_relapse.cox
summary(inter_relapse.cox)

#Tabulate-Relapse x interaction
```

```
inter_relapse.cox %>%
  tbl_regression(
    exp = TRUE)
relapseinter_tab <-tbl_regression(inter_relapse.cox, exponentiate = TRUE,conf.int = TRUE)
relapseinter_tab

##Variable selection
#AIC
AIC(transrelapse.cox,inter_relapse.cox)
#Model Diagnosis
##To test for the proportional-hazards (PH) assumption
#Relapse
testrelapse.ph <- cox.zph(inter_relapse.cox, terms = F)
testrelapse.ph
ggcoxzph(testrelapse.ph)
ggcoxdiagnostics(testrelapse.ph,'schoenfeld',ox.scale = 'time')
ggcoxdiagnostics(testrelapse.ph, type = "schoenfeld", linear.predictions = FALSE)

##Tabulating-final MODEL
##combine OTHER and RELAPSE x interaction models in a single table
tbl_merge(
  tbls = list(otherinter_tab, relapseinter_tab),
  tab_spanner = c("**otherinter_Model**", "**relapseinter_Model**"))
#AIC
AIC(inter_other.cox,inter_relapse.cox)
```

-THE END-