TCGA-pancancer

Code from the Science Translational Medicine article about predicting patients' prognosis based on their transcriptomic phenotype

This R Markdown document is part of a series containing the code that was used to perform some of the analyses from the article "Loss of TGFB signaling increases alternative end-joining DNA repair that sensitizes to genotoxic therapies across cancer types", published on Science Translational Medicine on 2021 (link: https://stm.sciencemag.org/content/13/580/eabc4465).

Specifically, this is the code that was used to analyze the pancancer dataset from The Cancer Genome Atlas (TCGA). Many of the results from this analysis can be seen in Figure 6 of the article.

PART 1: PREPARE THE ENVIRONMENT

#Step 1: Load (+/- install) the necessary packages.

```r
library("GSVA")
library("GSEABase")
library("methods")
library("edgeR")
library("geneplotter")
library("genefilter")
library("BiocGenerics")
library("Biobase")
library("graph")
library("XML")
library("lattice")
library("limma")
library("shinythemes")
library("shiny")
library("RColorBrewer")
library("parallel")
library("cluster")
library("Matrix")
library("locfit")
library("snow")
library("dplyr")
library("ggplot2")
library ("remotes")
library("OIsurv")
library("survival")
library("KMsurv")
library("splines")
library("survminer")
library("readxl")
library("reshape2")
library("data.table")
```

PART 2: IMPORT THE FILES FOR THE ANALYSIS

#Step 1: Import the files with the ssGSEA scores.

```
sBLCA = fread("Input/ssgsea scores/BLCA.ssgseas.tsv", data.table=FALSE)
sBRCA = fread("Input/ssgsea scores/BRCA.ssgseas.tsv", data.table=FALSE)
sCOAD = fread("Input/ssgsea scores/COAD.ssgseas.tsv", data.table=FALSE)
sESCA = fread("Input/ssgsea scores/ESCA.ssgseas.tsv", data.table=FALSE)
sGBM = fread("Input/ssgsea scores/GBM.ssgseas.tsv", data.table=FALSE)
sHNSC = fread("Input/ssgsea scores/HNSC.ssgseas.tsv", data.table=FALSE)
sKIRC = fread("Input/ssgsea scores/KIRC.ssgseas.tsv", data.table=FALSE)
sLIHC = fread("Input/ssgsea scores/LIHC.ssgseas.tsv", data.table=FALSE)
sLUAD = fread("Input/ssgsea scores/LUAD.ssgseas.tsv", data.table=FALSE)
sLUSC = fread("Input/ssgsea scores/LUSC.ssgseas.tsv", data.table=FALSE)
sOV = fread("Input/ssgsea scores/OV.ssgseas.tsv", data.table=FALSE)
sPAAD = fread("Input/ssgsea scores/PAAD.ssgseas.tsv", data.table=FALSE)
sPRAD = fread("Input/ssgsea scores/PRAD.ssgseas.tsv", data.table=FALSE)
sSKCM = fread("Input/ssgsea scores/SKCM.ssgseas.tsv", data.table=FALSE)
sTGCT = fread("Input/ssgsea scores/TGCT.ssgseas.tsv", data.table=FALSE)
sTHCA = fread("Input/ssgsea scores/THCA.ssgseas.tsv", data.table=FALSE)
sUCEC = fread("Input/ssgsea scores/UCEC.ssgseas.tsv", data.table=FALSE)
##These files were sent by collaborators (Miquel Angel, Roderic and Luis) on March 2020.
##Contain the ssSGSEA scores from TCGA-pancancer patients calculated by them.
```

#Step 2: Merge all the ssGSEA scores in one file.

```
ssgsea <- rbind(sBLCA,sBRCA,sCOAD,sESCA,sGBM,sHNSC,sKIRC,sLIHC,sLUAD,sLUSC,sOV,sPAAD,sPRAD,sSKCM,sTGCT,
ssgsea$sampleID <- ssgsea$V1
ssgsea$sampleID <- chartr(".", "-", ssgsea$sampleID) #turn "."s into "-"s
ssgsea$sampleID <- substring(ssgsea$sampleID,1,15) #keep only characters 1-15
##Dimensions: 7115 samples x 5 variables.
```

#Step 3: Import the file with clinical information.

```
PanCancer_ClinicalInformation <- read_excel("Input/PanCancer_ClinicalInformation.xlsx")
##This file contatins clinical information from TCGA-pancancer patients with primary solid tumors.
##It was sent by a collaborator (Mao) on March 2020, who generated it from the from the integrated TCGA
##Dimensions: 10967 samples x 13 variables.
```

PART 3: PUT ALL THE INFORMATION THAT WE WILL NEED FOR FURTHER ANALYSES IN ONE SINGLE DATAFRAME

#Step 1: Merge the "ssgsea" and the "PanCancer_ClinicalInformation" dataframes.

```
ssgsea$ID2 <- substr(ssgsea$sampleID, 0, 15) #Keep only the first 15 characters of sample IDs.
ALL <- merge(ssgsea, PanCancer_ClinicalInformation,
             by.x = "ID2", by.y = "Sample ID",
             all.x=FALSE, all.y=FALSE) #7115 patients with ssGSEA scores, 6949 of them with clinical da
dim(ALL) #[1] 6949  18
```

```
## [1] 6949   18
```

#Step 2: Prepare the survival variables so that they are in the appropriate format.

```
table(ALL$`Overall Survival Status`)
ALL$OSstatus <- ifelse(ALL$`Overall Survival Status` == "DECEASED", 1,
                       ifelse(ALL$`Overall Survival Status` == "LIVING", 0, NA))
class(ALL$OSstatus) #[1] "numeric"
class(ALL$`Overall Survival (Months)`) #[1] "character"
ALL$`Overall Survival (Months)` <- as.numeric(ALL$`Overall Survival (Months)`)
```

PART 4: ELIMINATE RECURRENT AND NORMAL TISSUE SAMPLES

#Step 1: Create a variable that shows the sample type.

```
ALL$Sample.type.2 <- substr(ALL$ID2, 13, 15) #Keep only sample type numbers.
table(ALL$Sample.type.2) #6949 primary; 0 normal; 0 recurrent.
```

```
##
##  -01
## 6949
```

PART 5: CALCULATE THE BALT SCORE OF EACH SAMPLE

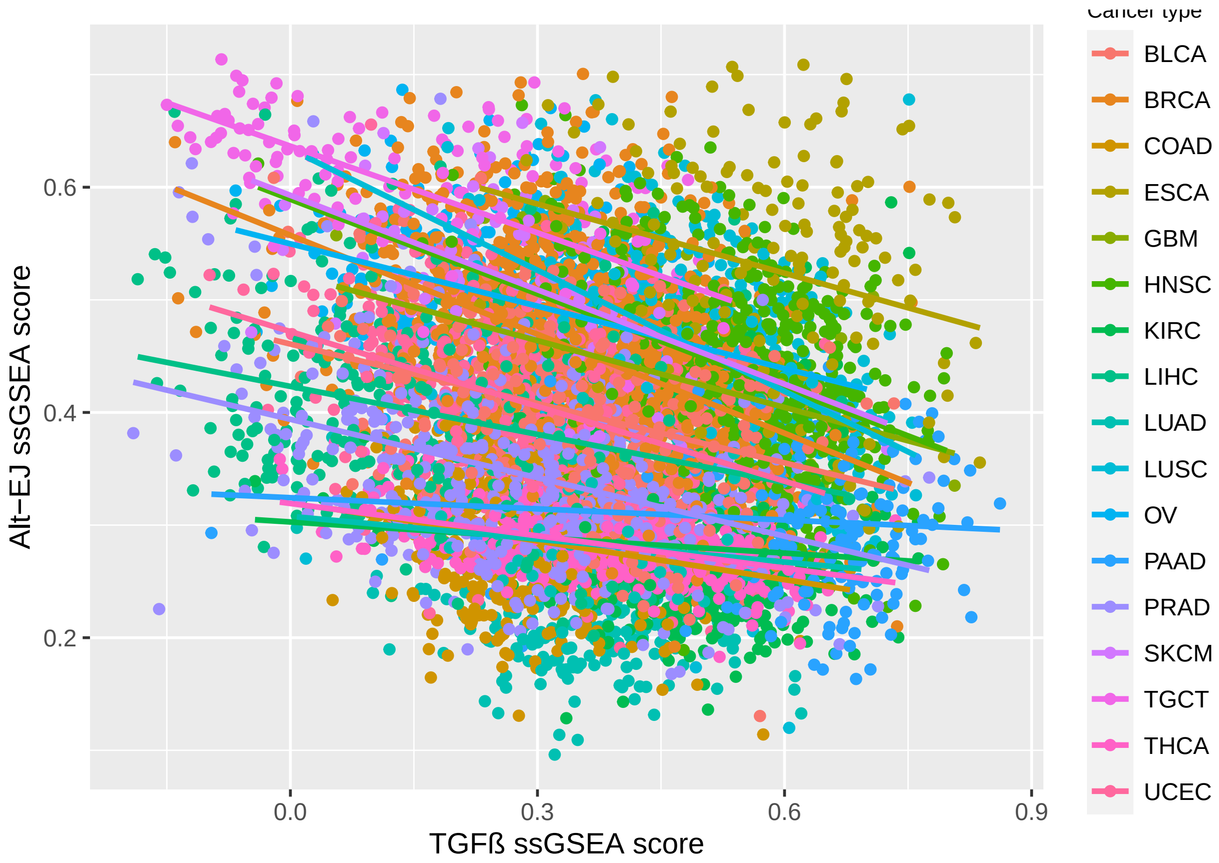
#Step 1: Create a new variable that is be the Balt score.

```
ALL$balt <- sqrt((max(ALL$ALT_EJ_repair)-ALL$ALT_EJ_repair)^2+
                     (min(ALL$Upregulated_TGF_beta)-ALL$Upregulated_TGF_beta)^2) -
            sqrt((min(ALL$ALT_EJ_repair)-ALL$ALT_EJ_repair)^2+
                     (max(ALL$Upregulated_TGF_beta)-ALL$Upregulated_TGF_beta)^2)
ALL$balt <- ALL$balt * -1
```

PART 6: ANALYSIS OF THE CORRELATION BETWEEN TGFB AND ALTEJ SSGSEA SCORES

#Step 1: Create a scatterplot of TGFB versus ALTEJ ssGSEA scores, coloring the cancer type.

```
ggplot(ALL, aes(x=ALL$Upregulated_TGF_beta, y=ALL$ALT_EJ_repair, col=ALL$`TCGA PanCanAtlas Cancer Type
  geom_point() +
  geom_smooth(method = "lm", fill = NA) +
  labs(x = "TGF  ssGSEA score", y = "Alt-EJ ssGSEA score", color = "Cancer type") +
  theme(legend.title=element_text(size=8))
```



#Step 2: Calculate Pearson correlation coefficient.

```
cor.test(ALL$Upregulated_TGF_beta, ALL$ALT_EJ_repair, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  ALL$Upregulated_TGF_beta and ALL$ALT_EJ_repair
## t = -17.213, df = 6947, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2246959 -0.1795933
## sample estimates:
##        cor
## -0.2022518
```

#Step 3: Calculate the Pearson correlation coefficient BY CANCER TYPE.

```
library(broom)
TumCor <- ALL %>%
  group_by(`TCGA PanCanAtlas Cancer Type Acronym`) %>%
  do(tidy(cor.test(.$Upregulated_TGF_beta, .$ALT_EJ_repair, method = "pearson"))) #All anticorrelated.
```

PART 7: ANALYSIS OF THE CORRELATION OF "AGE VS BALT SCORE"

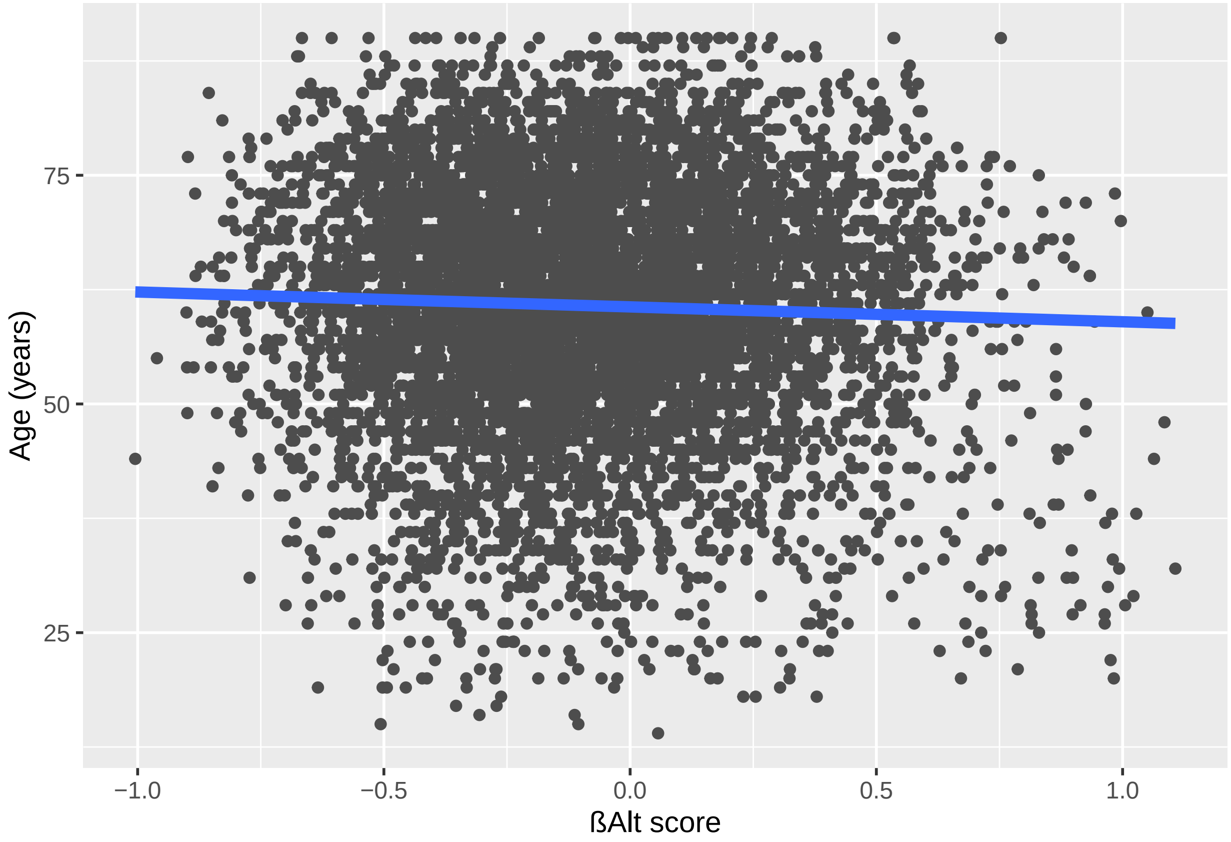
#Step 1: Calculate Pearson correlation coefficient of age vs BAlt score.

```
cor.test(ALL$balt, ALL$`Diagnosis Age`, method = "pearson") #PCC=-0.0396
```

```
##
##  Pearson's product-moment correlation
##
## data:  ALL$balt and ALL$`Diagnosis Age`
## t = -3.267, df = 6793, p-value = 0.001092
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06332505 -0.01584447
## sample estimates:
##         cor
## -0.03960712
```

#Step 2: Create a scatterplot of Age vs BAlt score.

```
ggplot(ALL, aes(x=ALL$balt, y=ALL$`Diagnosis Age`)) + geom_point(col="grey30") +
  labs(x = " Alt score", y = "Age (years)") +
  theme(legend.title=element_text(size=8))  +
  geom_smooth(method = "lm", fill = NA, size=2)
```



PART 8: SELECT PATIENTS WHO PROBABLY RECEIVED GENOTOXIC TREATMENT

#Step 1-A: Create a subdataframe with patients treated with RT.

```
table(ALL$`Radiation Therapy`)
ALL_RT <- ALL[which(ALL$`Radiation Therapy`=="Yes"),] #6.949-->1.737 patients.
```

```
dim(ALL_RT)
```

```
## [1] 1737   21
```

#Step 2-A: Create a new variable that represents the tertile that each patient belongs to, according to the value of the BAlt score.

```
ALL_RT <- ALL_RT %>% mutate(tertile = ntile(balt,3))
table(ALL_RT$`TCGA PanCanAtlas Cancer Type Acronym`, ALL_RT$tertile)
ALL_RT$Tertile <- ifelse(ALL_RT$tertile == 2, NA, ALL_RT$tertile)
```

#Step 1-B: Create a subdataframe with patients who, based on their cancer type and stage, their standard of care treatment includes RT and/or genotoxic ChT.

```
table(ALL$`Cancer Type`, ALL$`Neoplasm Disease Stage American Joint Committee on Cancer Code`)
table(ALL$`TCGA PanCanAtlas Cancer Type Acronym`, ALL$`Neoplasm Disease Stage American Joint Committee
ALL_RTChT <- ALL[which(ALL$`Radiation Therapy` == "Yes" |
                            (ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "BLCA" & ALL$`Neoplasm Diseas
                            (ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "COAD" & !ALL$`Neoplasm Disea
                            (ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "ESCA" & !ALL$`Neoplasm Disea
                            ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "GBM" |
                            (ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "HNSC" & !ALL$`Neoplasm Disea
                            (ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "LUAD" & !ALL$`Neoplasm Disea
                            (ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "LUSC" & !ALL$`Neoplasm Disea
                            ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "OV" & !ALL$`Neoplasm Disease
                            ALL$`TCGA PanCanAtlas Cancer Type Acronym` == "PAAD" |
                            ALL$`Cancer Type`== "Non-Seminomatous Germ Cell Tumor"
),] #6.949-->3.577 patients.
```

```
dim(ALL_RTChT)
```

```
## [1] 3577   21
```

```
##Inclusion criteria. RT yes or:
##BLCA: Stage 4.
##BRCA: RT.
##COAD: Stage not 1.
##ESCA (Esophageal): Stage not 1 nor 2a.
##GBM: All.
##HNSC: Stage not 1.
##KIRC (Renal): RT.
##LIHC (Liver): RT.
##LUAD and LUSC: Stage not 1.
##OV: Stage not 1a.
##PAAD (Pancreas): All.
##PRAD (Prostate): RT.
##SKCM: RT.
##TGCT seminoma: RT.
```

```
##TGCT non-seminoma: All.
##THCA (Thyroid): RT.
##UCEC (Endometrial): RT.
```

```
##Reasoning behind the inclusion criteria:
##BLCA: Stage 4 usually includes genotoxic ChT with platin agents and earlier stages are frequently tre
##BRCA: Usually treated with surgery + RT +/- HT +/- Trastuzumab +/- ChT. ChT frequently consists of an
##COAD: Stage I may be treated with surgery alone. In other stages ChT usually includes genotoxic Plati
##ESCA: Stages T1-2N0M0 (stages <IIB) may be treated with surgery alone. Otherwise RT, ChT or both are
##GBM: Standard treatment is usually surgery + RT + ChT. ChT usually consists of genotoxic Temozolamide
##HNSC: Stage I may be treated with surgery alone. Otherwise treatment usually includes RT, ChT or both
##KIRC: RT is rarely indicated. Many times treated with surgery alone and the systemic treatment, when
##LIHC: Many treated with surgery alone. Systemic treatment rarely includes genotoxic drugs, as the mos
##LUAD and LUSC: Stage I may be treated with surgery alone. Otherwise treatment usually includes RT, Ch
##OV: Stage Ia may be treated with surgery alone and stage>Ia usually receive ChT. ChT usually includes
##PAAD: Rarely treated with surgery alone. RT and/or ChT are frequently added. ChT most usually include
##PRAD: Usually treated with HT combined with surgery, RT or both. ChT is given only in some stage IV p
##SKCM: Many treated with surgery alone. Systemic treatment rarely includes genotoxic drugs.
##TGCT: Treated with surgery. In seminomas, RT is usually added (and/or ChT if stage >I) and, in non-se
##THCA: Most of them are treated with surgery alone. Few exceptions are treated with both RT and ChT, s
##UCEC: Stage I may be treated with surgery alone. Otherwise RT is usually given. ChT is used mainly in
```

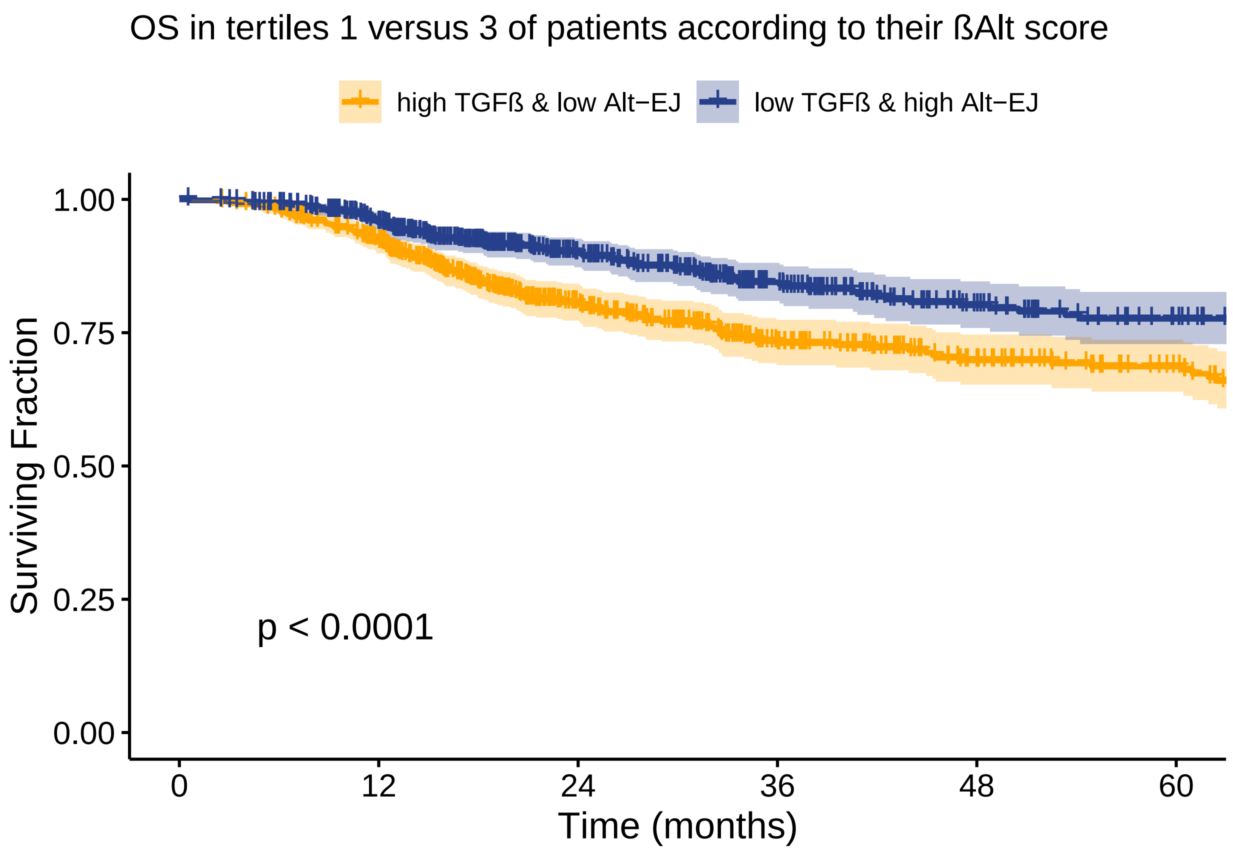
#Step 2-B: Create a new variable that represents the tertile that each patient belongs to, according to the value of the BAlt score.

```
ALL_RTChT <- ALL_RTChT %>% mutate(tertile = ntile(balt,3))
table(ALL_RTChT$`TCGA PanCanAtlas Cancer Type Acronym`, ALL_RTChT$tertile)
ALL_RTChT$Tertile <- ifelse(ALL_RTChT$tertile == 2, NA, ALL_RTChT$tertile)
```

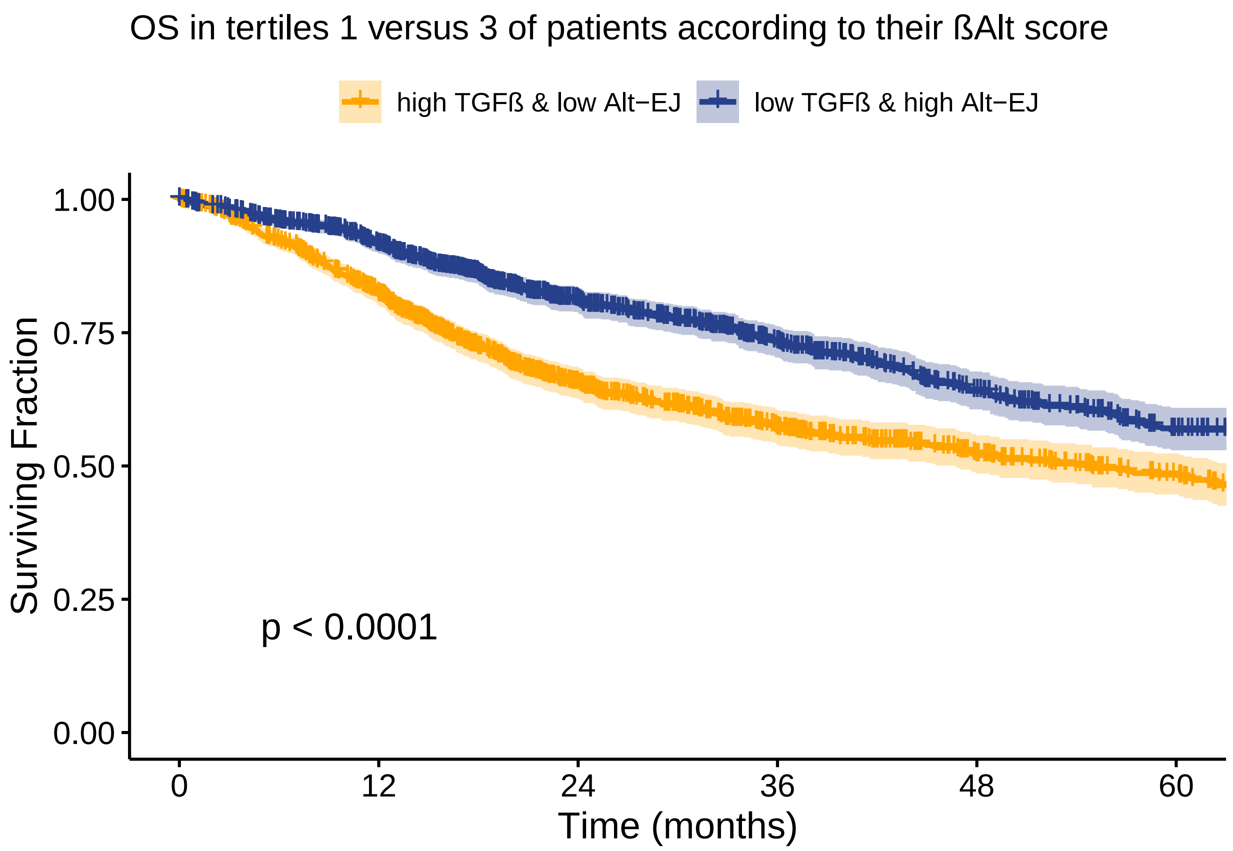
PART 9: OVERALL SURVIVAL CURVES OF THE BALT SCORE GROUPS

#Step 1: Plot and compare the OS curves between Balt score top and bottom tertiles.

```
##In patients treated with RT
my.surv.object <- Surv(time=ALL_RT$`Overall Survival (Months)`, event=ALL_RT$OSstatus)
my.fit<-survfit(my.surv.object~ALL_RT$Tertile)
my.fit
ggsurvplot(my.fit, data=ALL_RT, pval = TRUE, conf.int = TRUE, break.time.by = 12,
           xlab="Time (months)", ylab="Surviving Fraction", xlim=c(0, 60),
           legend.labs=c("high TGF  & low Alt-EJ", "low TGF  & high Alt-EJ"),
           legend.title=" ", title="OS in tertiles 1 versus 3 of patients according to their  Alt score"
           font.main=13, palette=c("orange", "royalblue4"))
```



```
##In patients treated with RT and/or probably genotoxic ChT
my.surv.object <- Surv(time=ALL_RTChT$`Overall Survival (Months)`, event=ALL_RTChT$OSstatus)
my.fit<-survfit(my.surv.object~ALL_RTChT$Tertile)
my.fit
ggsurvplot(my.fit, data=ALL_RTChT, pval = TRUE, conf.int = TRUE, break.time.by = 12,
           xlab="Time (months)", ylab="Surviving Fraction", xlim=c(0, 60),
           legend.labs=c("high TGF  & low Alt-EJ", "low TGF  & high Alt-EJ"),
           legend.title=" ", title="OS in tertiles 1 versus 3 of patients according to their  Alt score"
           font.main=13, palette=c("orange", "royalblue4"))
```



PART 10: CALCULATE THE HAZARD RATIOS OF THE SURVIVAL CURVES FROM PART 9

#Step 1: In the variable tertile, make "high TGFB and low ALTEJ" the reference level so that the Hazard Ratios are expressed as 0.x instead of 1.x.

```
##In patients treated with RT
ALL_RT$Tertile <- ifelse(ALL_RT$Tertile == "1", "high TGFB and low ALTEJ",
                         ifelse(ALL_RT$Tertile == "3", "low TGFB and high ALTEJ", NA))
ALL_RT$Tertile <- factor(ALL_RT$Tertile, levels = c("high TGFB and low ALTEJ", "low TGFB and high ALTEJ
table(ALL_RT$Tertile)
```

```
##In patients treated with RT and/or genotoxic ChT
ALL_RTChT$Tertile <- ifelse(ALL_RTChT$Tertile == "1", "high TGFB and low ALTEJ",
                            ifelse(ALL_RTChT$Tertile == "3", "low TGFB and high ALTEJ", NA))
ALL_RTChT$Tertile <- factor(ALL_RTChT$Tertile, levels = c("high TGFB and low ALTEJ", "low TGFB and high
table(ALL_RTChT$Tertile)
```

#Step 2: Calculate the OS hazard ratio of BAlt tertile 1 versus 3.

```
##In patients treated with RT
my.surv.object <- Surv(time=ALL_RT$`Overall Survival (Months)`, event=ALL_RT$OSstatus)
cox<-coxph(my.surv.object ~  ALL_RT$Tertile)
summary(cox) #HR=0.56
```

```
## Call:
## coxph(formula = my.surv.object ~ ALL_RT$Tertile)
##
##   n= 1157, number of events= 243
##    (580 observations deleted due to missingness)
##
##                                             coef exp(coef) se(coef)      z
## ALL_RT$Tertilelow TGFB and high ALTEJ -0.5788    0.5606   0.1332 -4.345
##                                          Pr(>|z|)
## ALL_RT$Tertilelow TGFB and high ALTEJ  1.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                          exp(coef) exp(-coef) lower .95 upper .95
## ALL_RT$Tertilelow TGFB and high ALTEJ    0.5606      1.784    0.4318    0.7278
##
## Concordance= 0.575  (se = 0.017 )
## Likelihood ratio test= 19.62  on 1 df,   p=9e-06
## Wald test            = 18.88  on 1 df,   p=1e-05
## Score (logrank) test = 19.41  on 1 df,   p=1e-05
```

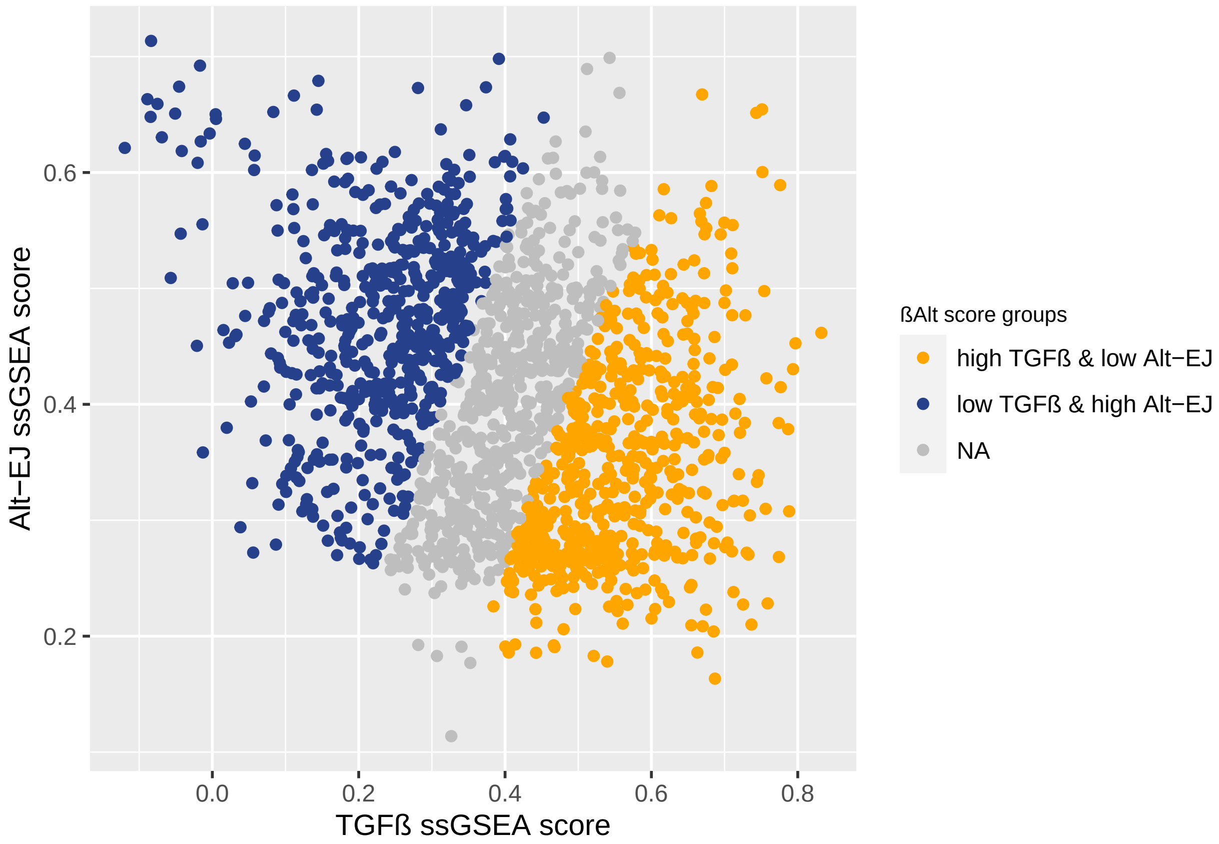
```
##In patients treated with RT and/or genotoxic ChT
my.surv.object <- Surv(time=ALL_RTChT$`Overall Survival (Months)`, event=ALL_RTChT$OSstatus)
cox<-coxph(my.surv.object ~  ALL_RTChT$Tertile)
summary(cox) #HR=0.60
```

```
## Call:
## coxph(formula = my.surv.object ~ ALL_RTChT$Tertile)
##
##   n= 2370, number of events= 861
##    (1207 observations deleted due to missingness)
##
##                                                 coef exp(coef) se(coef)      z
## ALL_RTChT$Tertilelow TGFB and high ALTEJ -0.50528   0.60333  0.06925 -7.296
##                                             Pr(>|z|)
## ALL_RTChT$Tertilelow TGFB and high ALTEJ 2.96e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                             exp(coef) exp(-coef) lower .95
## ALL_RTChT$Tertilelow TGFB and high ALTEJ    0.6033      1.657    0.5268
##                                             upper .95
## ALL_RTChT$Tertilelow TGFB and high ALTEJ     0.691
##
## Concordance= 0.577  (se = 0.009 )
## Likelihood ratio test= 54.04  on 1 df,   p=2e-13
## Wald test            = 53.23  on 1 df,   p=3e-13
## Score (logrank) test = 54.35  on 1 df,   p=2e-13
```

PART 11: SCATTERPLOTS SHOWING THE TWO BALT SCORE GROUPS COMPARED IN PART 9

#Step 1: Create a scatterplot of TGFB versus ALTEJ ssGSEA scores, coloring the BAlt tertiles.

```
##In patients treated with RT
ggplot(ALL_RT, aes(x=ALL_RT$Upregulated_TGF_beta, y=ALL_RT$ALT_EJ_repair, col=ALL_RT$Tertile)) +
  geom_point() +
  scale_color_manual(values=c("orange", "royalblue4"),
                     labels = c("high TGF  & low Alt-EJ", "low TGF  & high Alt-EJ"),
                     na.translate=TRUE, na.value="grey") +
  labs(x = "TGF  ssGSEA score", y = "Alt-EJ ssGSEA score", color = " Alt score groups") +
  theme(legend.title=element_text(size=8))
```

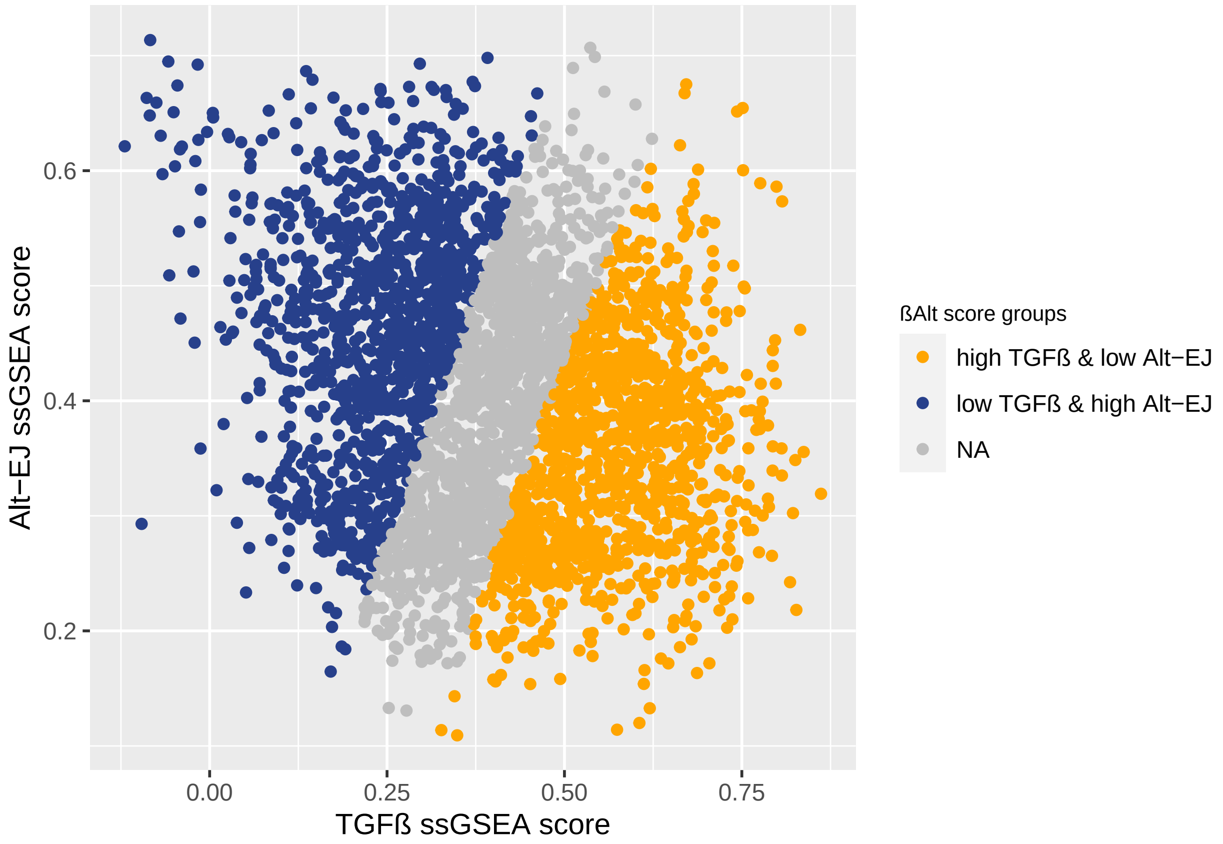


```
cor.test(ALL_RT$Upregulated_TGF_beta, ALL_RT$ALT_EJ_repair, method = "pearson") #PCC=-0.23
```

```
##
##  Pearson's product-moment correlation
##
## data:  ALL_RT$Upregulated_TGF_beta and ALL_RT$ALT_EJ_repair
## t = -10.048, df = 1735, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2784559 -0.1895514
## sample estimates:
##        cor
## -0.2344939
```

##In patients treated with RT and/or genotoxic ChT

```r
ggplot(ALL_RTChT, aes(x=ALL_RTChT$Upregulated_TGF_beta, y=ALL_RTChT$ALT_EJ_repair, col=ALL_RTChT$Tertil
  geom_point() +
  scale_color_manual(values=c("orange", "royalblue4"),
                     labels = c("high TGF  & low Alt-EJ", "low TGF  & high Alt-EJ"),
                     na.translate=TRUE, na.value="grey") +
  labs(x = "TGF  ssGSEA score", y = "Alt-EJ ssGSEA score", color = " Alt score groups") +
  theme(legend.title=element_text(size=8))
```



```r
cor.test(ALL_RTChT$Upregulated_TGF_beta, ALL_RTChT$ALT_EJ_repair, method = "pearson") #PCC=-0.16
```

```
##
##  Pearson's product-moment correlation
##
## data:  ALL_RTChT$Upregulated_TGF_beta and ALL_RTChT$ALT_EJ_repair
## t = -9.6255, df = 3575, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1907181 -0.1268263
## sample estimates:
##        cor
## -0.1589386
```

PART 12: MULTIVARIATE COX OF THE ASSOCIATION OF THE BALT SCORE WITH OS, ADJUSTED FOR AGE AND STAGE

#Step 1: Group stages into 4 groups (1, 2, 3 and 4)

```
##In patients treated with RT
table(ALL_RT$`Neoplasm Disease Stage American Joint Committee on Cancer Code`)
ALL_RT$stage <- ALL_RT$`Neoplasm Disease Stage American Joint Committee on Cancer Code`
ALL_RT$stage <- ifelse(ALL_RT$stage == "STAGE I", "1",
                          ifelse(ALL_RT$stage == "STAGE IA", "1",
                          ifelse(ALL_RT$stage == "STAGE IB", "1",
                          ifelse(ALL_RT$stage == "STAGE II", "2",
                          ifelse(ALL_RT$stage == "STAGE IIA", "2",
                          ifelse(ALL_RT$stage == "STAGE IIB", "2",
                          ifelse(ALL_RT$stage == "STAGE IIC", "2",
                          ifelse(ALL_RT$stage == "STAGE III", "3",
                          ifelse(ALL_RT$stage == "STAGE IIIA", "3",
                          ifelse(ALL_RT$stage == "STAGE IIIB", "3",
                          ifelse(ALL_RT$stage == "STAGE IIIC", "3",
                          ifelse(ALL_RT$stage == "STAGE IS", "1",
                          ifelse(ALL_RT$stage == "STAGE IV", "4",
                          ifelse(ALL_RT$stage == "STAGE IVA", "4",
                          ifelse(ALL_RT$stage == "STAGE IVB", "4",
                          ifelse(ALL_RT$stage == "STAGE IVC", "4", NA))))))))))))))))
table(ALL_RT$stage)
```

```
##In patients treated with RT and/or genotoxic ChT
table(ALL_RTChT$`Neoplasm Disease Stage American Joint Committee on Cancer Code`)
ALL_RTChT$stage <- ALL_RTChT$`Neoplasm Disease Stage American Joint Committee on Cancer Code`
ALL_RTChT$stage <- ifelse(ALL_RTChT$stage == "STAGE I", "1",
                            ifelse(ALL_RTChT$stage == "STAGE IA", "1",
                            ifelse(ALL_RTChT$stage == "STAGE IB", "1",
                            ifelse(ALL_RTChT$stage == "STAGE II", "2",
                            ifelse(ALL_RTChT$stage == "STAGE IIA", "2",
                            ifelse(ALL_RTChT$stage == "STAGE IIB", "2",
                            ifelse(ALL_RTChT$stage == "STAGE IIC", "2",
                            ifelse(ALL_RTChT$stage == "STAGE III", "3",
                            ifelse(ALL_RTChT$stage == "STAGE IIIA", "3",
                            ifelse(ALL_RTChT$stage == "STAGE IIIB", "3",
                            ifelse(ALL_RTChT$stage == "STAGE IIIC", "3",
                            ifelse(ALL_RTChT$stage == "STAGE IS", "1",
                            ifelse(ALL_RTChT$stage == "STAGE IV", "4",
                            ifelse(ALL_RTChT$stage == "STAGE IVA", "4",
                            ifelse(ALL_RTChT$stage == "STAGE IVB", "4",
                            ifelse(ALL_RTChT$stage == "STAGE IVC", "4", NA))))))))))))))))
table(ALL_RTChT$stage)
```

#Step 2: Calculate survival multivariate Cox regression adjusted for age and stage of the BAlt score continuous variable.

```
##In patients treated with RT
my.surv.object <- Surv(time=ALL_RT$`Overall Survival (Months)`, event=ALL_RT$OSstatus)
cox<-coxph(my.surv.object ~  ALL_RT$balt + ALL_RT$`Diagnosis Age` + ALL_RT$stage)
summary(cox) #P=0.003
```

```
## Call:
## coxph(formula = my.surv.object ~ ALL_RT$balt + ALL_RT$`Diagnosis Age` +
##     ALL_RT$stage)
##
##   n= 1311, number of events= 235
##    (426 observations deleted due to missingness)
##
##                             coef exp(coef)  se(coef)      z Pr(>|z|)
## ALL_RT$balt            -0.733073  0.480430  0.243386 -3.012  0.00260 **
## ALL_RT$`Diagnosis Age`  0.045386  1.046432  0.005654  8.027 9.97e-16 ***
## ALL_RT$stage2           0.745358  2.107196  0.272021  2.740  0.00614 **
## ALL_RT$stage3           1.168182  3.216141  0.264604  4.415 1.01e-05 ***
## ALL_RT$stage4           1.666491  5.293561  0.265038  6.288 3.22e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## ALL_RT$balt               0.4804     2.0815    0.2982    0.7741
## ALL_RT$`Diagnosis Age`    1.0464     0.9556    1.0349    1.0581
## ALL_RT$stage2             2.1072     0.4746    1.2364    3.5913
## ALL_RT$stage3             3.2161     0.3109    1.9147    5.4022
## ALL_RT$stage4             5.2936     0.1889    3.1488    8.8992
##
## Concordance= 0.744  (se = 0.015 )
## Likelihood ratio test= 183.2  on 5 df,   p=<2e-16
## Wald test            = 152.8  on 5 df,   p=<2e-16
## Score (logrank) test = 173.9  on 5 df,   p=<2e-16
```

##In patients treated with RT and/or genotoxic ChT

```
my.surv.object <- Surv(time=ALL_RTChT$`Overall Survival (Months)`, event=ALL_RTChT$OSstatus)
cox<-coxph(my.surv.object ~  ALL_RTChT$balt + ALL_RTChT$`Diagnosis Age` + ALL_RTChT$stage)
summary(cox) #P<0.001
```

```
## Call:
## coxph(formula = my.surv.object ~ ALL_RTChT$balt + ALL_RTChT$`Diagnosis Age` +
##     ALL_RTChT$stage)
##
##   n= 2626, number of events= 797
##    (951 observations deleted due to missingness)
##
##                                coef exp(coef)  se(coef)      z Pr(>|z|)
## ALL_RTChT$balt            -1.111831  0.328956  0.129779 -8.567  < 2e-16 ***
## ALL_RTChT$`Diagnosis Age`  0.037804  1.038527  0.003046 12.410  < 2e-16 ***
## ALL_RTChT$stage2           1.136611  3.116189  0.211973  5.362 8.23e-08 ***
## ALL_RTChT$stage3           1.301994  3.676621  0.214494  6.070 1.28e-09 ***
## ALL_RTChT$stage4           1.899419  6.682014  0.211929  8.963  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## ALL_RTChT$balt                0.329     3.0399    0.2551    0.4242
## ALL_RTChT$`Diagnosis Age`     1.039     0.9629    1.0323    1.0447
## ALL_RTChT$stage2              3.116     0.3209    2.0568    4.7212
## ALL_RTChT$stage3              3.677     0.2720    2.4147    5.5979
## ALL_RTChT$stage4              6.682     0.1497    4.4108   10.1228
##
## Concordance= 0.717  (se = 0.009 )
## Likelihood ratio test= 529.6  on 5 df,   p=<2e-16
## Wald test            = 409.4  on 5 df,   p=<2e-16
## Score (logrank) test = 465.2  on 5 df,   p=<2e-16
```