Final Case Study

Piotr Suder

2023-04-19

Basic dependencies:

- R>=4.1 # https://www.r-project.org/
- RStudio>=1.4.1717 # https://posit.co/download/rstudio-desktop/

R Package Dependencies

```
options(warn=-1)
if (!require(pacman)) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(tidyverse, glmnet, iregnet, penAFT,
               survival, mice, ggsurvfit, dplyr, xtable, ggplot2, gridExtra, ggtext, cobalt)

dir.create("plots", showWarnings = F)
theme_set(theme_classic(base_size = 12))
```

System Information

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora Linux 36 (MATE-Compiz)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.3
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] cobalt_4.5.0       ggtext_0.1.2       gridExtra_2.3      xtable_1.8-4
##  [5] ggsurvfit_0.3.0    mice_3.15.0        survival_3.5-5     penAFT_0.3.0
##  [9] iregnet_0.1.0.9000 glmnet_4.1-7       Matrix_1.5-4       lubridate_1.9.2
## [13] forcats_1.0.0      stringr_1.5.0      dplyr_1.1.2        purrr_1.0.1
## [17] readr_2.1.4        tidyr_1.3.0        tibble_3.2.1       ggplot2_3.4.2
## [21] tidyverse_2.0.0    pacman_0.5.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.10         lattice_0.21-8      listenv_0.9.0
##  [4] digest_0.6.31       RhpcBLASctl_0.23-42 foreach_1.5.2
##  [7] utf8_1.2.3          parallelly_1.35.0   R6_2.5.1
## [10] backports_1.4.1     evaluate_0.20       pillar_1.9.0
## [13] rlang_1.1.0         rstudioapi_0.14     irlba_2.3.5.1
## [16] rmarkdown_2.21      splines_4.1.3       gridtext_0.1.5
## [19] munsell_0.5.0       broom_1.0.4         compiler_4.1.3
## [22] xfun_0.39           pkgconfig_2.0.3     shape_1.4.6
## [25] globals_0.16.2      htmltools_0.5.5     tidyselect_1.2.0
## [28] codetools_0.2-19    fansi_1.0.4         future_1.32.0
## [31] crayon_1.5.2        tzdb_0.3.0          withr_2.5.0
## [34] grid_4.1.3          gtable_0.3.3        lifecycle_1.0.3
## [37] magrittr_2.0.3      scales_1.2.1        future.apply_1.10.0
## [40] cli_3.6.1           stringi_1.7.12      xml2_1.3.3
## [43] generics_0.1.3      vctrs_0.6.2         iterators_1.0.14
## [46] tools_4.1.3         glue_1.6.2          hms_1.1.3
## [49] parallel_4.1.3      fastmap_1.1.1       yaml_2.3.7
## [52] timechange_0.2.0    colorspace_2.1-0    knitr_1.42
```

Load data

```
data = read.csv('METABRIC_RNA_Mutation.csv')

sum(is.na(data))
```

```
## [1] 638
```

```
# drop the patient with sarcoma
sum(data$cancer_type == "Breast Sarcoma")
```

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

```
# only keep patients with Breast Invasive Ductal Carcinoma
data = data[data$cancer_type_detailed == 'Breast Invasive Ductal Carcinoma',]

# Drop the rare claudin low subtype
data = data[data$pam50_._claudin.low_subtype != 'NC',]

# drop: patient_id
data = data %>% select(-c('patient_id', 'cancer_type', 'cancer_type_detailed', 'cohort', 'overall_survi

sum(is.na(data))
```

```
## [1] 498
```

```
# select only the ones for which we know the surgery type
data_sel = data[data$type_of_breast_surgery != "",]

sum(is.na(data_sel$overall_survival_months))
```

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

```r
data_sel = Filter(function(x)(length(unique(x))>1), data_sel)
data_BIDC = data_sel

#print(names(data_BIDC))

#sum(is.na(data_sel))

num_cols = ncol(data_sel)
for (i in 1:num_cols)
{
  if (sum(is.na(data_sel[,i])) > 0)
  {
    print(i)
    print(colnames(data_sel[i]))
  }
}
```

```
## [1] 8
## [1] "neoplasm_histologic_grade"
## [1] 17
## [1] "mutation_count"
## [1] 23
## [1] "tumor_size"
## [1] 24
## [1] "tumor_stage"
```

```r
# Recode the treatment prodecures: 0 - breast conserving, 1 - mastectomy
data_BIDC$treatment = 1*(data_BIDC$type_of_breast_surgery == "MASTECTOMY")
unique(data_BIDC$type_of_breast_surgery)
```

```
## [1] "MASTECTOMY"        "BREAST CONSERVING"
```

```r
data_BIDC$primary_tumor_laterality[data_BIDC$primary_tumor_laterality == ""] = NA
data_BIDC$inferred_menopausal_state[data_BIDC$inferred_menopausal_state == ""] = NA
data_BIDC$er_status_measured_by_ihc[data_BIDC$er_status_measured_by_ihc == ""] = NA


sum(is.na(data_BIDC$primary_tumor_laterality))
```

```
## [1] 83
```

```r
sum(is.na(data_BIDC$inferred_menopausal_state))
```

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

```r
sum(is.na(data_BIDC$er_status_measured_by_ihc))
```

```
## [1] 22
```

```r
sum(is.na(data_BIDC[,26:514]))
```

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

Data Imputation with MICE

```r
categorical_vars <- c("cellularity", "pam50_._claudin.low_subtype",
                      "neoplasm_histologic_grade", "tumor_other_histologic_subtype",
                      "integrative_cluster", "X3.gene_classifier_subtype")

other_categorical = c("inferred_menopausal_state",
                      "primary_tumor_laterality",
                      "pr_status",
                      "tumor_stage",
                      "her2_status",
                      "er_status_measured_by_ihc",
                      "radio_therapy", "hormone_therapy", "chemotherapy")

data_mice <- data_BIDC %>% mutate(across(all_of(unlist(c(categorical_vars, other_categorical))), factor

data_BIDC_2 = data_BIDC
data_BIDC_3 = data_BIDC
data_BIDC_4 = data_BIDC
data_BIDC_5 = data_BIDC

data_mice = data_mice[,1:24]
#sum(is.na(data_mice))
names(data_mice)
```

```
##  [1] "age_at_diagnosis"               "type_of_breast_surgery"
##  [3] "cellularity"                    "chemotherapy"
##  [5] "pam50_._claudin.low_subtype"    "er_status_measured_by_ihc"
##  [7] "er_status"                      "neoplasm_histologic_grade"
##  [9] "her2_status_measured_by_snp6"   "her2_status"
## [11] "tumor_other_histologic_subtype" "hormone_therapy"
## [13] "inferred_menopausal_state"      "integrative_cluster"
## [15] "primary_tumor_laterality"       "lymph_nodes_examined_positive"
## [17] "mutation_count"                 "nottingham_prognostic_index"
## [19] "overall_survival_months"        "pr_status"
## [21] "radio_therapy"                  "X3.gene_classifier_subtype"
## [23] "tumor_size"                     "tumor_stage"
```

```
set.seed(5)
mice_obj <- mice(data = data_mice, m = 5)
```

```
##
##  iter imp variable
##   1   1  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   1   2  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   1   3  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   1   4  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   1   5  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   2   1  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   2   2  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   2   3  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   2   4  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   2   5  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   3   1  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   3   2  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   3   3  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   3   4  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##   3   5  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_co
##     4   1  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     4   2  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     4   3  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     4   4  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     4   5  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     5   1  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     5   2  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     5   3  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     5   4  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
##     5   5  er_status_measured_by_ihc  neoplasm_histologic_grade  primary_tumor_laterality  mutation_cou
```

```
data_full <- complete(mice_obj,1)
data_full_2 <- complete(mice_obj,2)
data_full_3 <- complete(mice_obj,3)
data_full_4 <- complete(mice_obj,4)
data_full_5 <- complete(mice_obj,5)

data_BIDC[,1:24] = data_full
data_BIDC_2[,1:24] = data_full_2
data_BIDC_3[,1:24] = data_full_3
data_BIDC_4[,1:24] = data_full_4
data_BIDC_5[,1:24] = data_full_5

data_BIDC$tumor_stage = 0*(data_BIDC$tumor_stage == 0) + 1*(data_BIDC$tumor_stage == 1) + 2*(data_BIDC$


data_BIDC_2$tumor_stage = 0*(data_BIDC_2$tumor_stage == 0) + 1*(data_BIDC_2$tumor_stage == 1) + 2*(data_

data_BIDC_3$tumor_stage = 0*(data_BIDC_3$tumor_stage == 0) + 1*(data_BIDC_3$tumor_stage == 1) + 2*(data_

data_BIDC_4$tumor_stage = 0*(data_BIDC_4$tumor_stage == 0) + 1*(data_BIDC_4$tumor_stage == 1) + 2*(data_

data_BIDC_5$tumor_stage = 0*(data_BIDC_5$tumor_stage == 0) + 1*(data_BIDC_5$tumor_stage == 1) + 2*(data_
```

```
# 0 - pre, 1 - post menopause
data_BIDC$menopause = 1*(data_BIDC$inferred_menopausal_state == "Post")

# 0 - left, 1 - right
data_BIDC$tumor_laterality = 1*(data_BIDC$primary_tumor_laterality == "Right")

# Recode more variables
data_BIDC$progesterone_status = 1*(data_BIDC$pr_status == "Positive")
data_BIDC$HER2_status = 1*(data_BIDC$her2_status == "Positive")
data_BIDC$er_status_ihc = 1*(data_BIDC$er_status_measured_by_ihc == "Positive")

###################################################
# 0 - pre, 1 - post menopause
data_BIDC_2$menopause = 1*(data_BIDC_2$inferred_menopausal_state == "Post")

# 0 - left, 1 - right
data_BIDC_2$tumor_laterality = 1*(data_BIDC_2$primary_tumor_laterality == "Right")

# Recode more variables
data_BIDC_2$progesterone_status = 1*(data_BIDC_2$pr_status == "Positive")
data_BIDC_2$HER2_status = 1*(data_BIDC_2$her2_status == "Positive")
```

```r
data_BIDC_2$er_status_ihc = 1*(data_BIDC_2$er_status_measured_by_ihc == "Positive")
##################################################

###################################################
# 0 - pre, 1 - post menopause
data_BIDC_3$menopause = 1*(data_BIDC_3$inferred_menopausal_state == "Post")

# 0 - left, 1 - right
data_BIDC_3$tumor_laterality = 1*(data_BIDC_3$primary_tumor_laterality == "Right")

# Recode more variables
data_BIDC_3$progesterone_status = 1*(data_BIDC_3$pr_status == "Positive")
data_BIDC_3$HER2_status = 1*(data_BIDC_3$her2_status == "Positive")
data_BIDC_3$er_status_ihc = 1*(data_BIDC_3$er_status_measured_by_ihc == "Positive")
##################################################

###################################################
# 0 - pre, 1 - post menopause
data_BIDC_4$menopause = 1*(data_BIDC_4$inferred_menopausal_state == "Post")

# 0 - left, 1 - right
data_BIDC_4$tumor_laterality = 1*(data_BIDC_4$primary_tumor_laterality == "Right")

# Recode more variables
data_BIDC_4$progesterone_status = 1*(data_BIDC_4$pr_status == "Positive")
data_BIDC_4$HER2_status = 1*(data_BIDC_4$her2_status == "Positive")
data_BIDC_4$er_status_ihc = 1*(data_BIDC_4$er_status_measured_by_ihc == "Positive")
##################################################

###################################################
# 0 - pre, 1 - post menopause
data_BIDC_5$menopause = 1*(data_BIDC_5$inferred_menopausal_state == "Post")

# 0 - left, 1 - right
data_BIDC_5$tumor_laterality = 1*(data_BIDC_5$primary_tumor_laterality == "Right")

# Recode more variables
data_BIDC_5$progesterone_status = 1*(data_BIDC_5$pr_status == "Positive")
data_BIDC_5$HER2_status = 1*(data_BIDC_5$her2_status == "Positive")
data_BIDC_5$er_status_ihc = 1*(data_BIDC_5$er_status_measured_by_ihc == "Positive")
##################################################
```

```r
data_BIDC_excl = data_BIDC[data_BIDC$tumor_stage != 0 & data_BIDC$tumor_stage != 4,]

data_BIDC_excl_2 = data_BIDC[data_BIDC_2$tumor_stage != 0 & data_BIDC_2$tumor_stage != 4,]
data_BIDC_excl_3 = data_BIDC[data_BIDC_3$tumor_stage != 0 & data_BIDC_3$tumor_stage != 4,]
data_BIDC_excl_4 = data_BIDC[data_BIDC_4$tumor_stage != 0 & data_BIDC_4$tumor_stage != 4,]
data_BIDC_excl_5 = data_BIDC[data_BIDC_5$tumor_stage != 0 & data_BIDC_5$tumor_stage != 4,]
```

```r
# Mutation indicators
gene_expr_names = names(data_BIDC)[26:514]
mutation_names = names(data_BIDC)[515:687]

for (var in mutation_names) {
  data_BIDC[[var]] <- as.integer(data_BIDC[[var]] != 0)

  data_BIDC_2[[var]] <- as.integer(data_BIDC_2[[var]] != 0)
  data_BIDC_3[[var]] <- as.integer(data_BIDC_3[[var]] != 0)
  data_BIDC_4[[var]] <- as.integer(data_BIDC_4[[var]] != 0)
  data_BIDC_5[[var]] <- as.integer(data_BIDC_5[[var]] != 0)
}



# Exclude patients with rare mutations
p = ncol(data_BIDC)
n = nrow(data_BIDC)
include = rep(FALSE, n)

n_2 = nrow(data_BIDC_2)
n_3 = nrow(data_BIDC_3)
n_4 = nrow(data_BIDC_4)
n_5 = nrow(data_BIDC_5)

include_2 = rep(FALSE, nrow(data_BIDC_2))
include_3 = rep(FALSE, nrow(data_BIDC_3))
include_4 = rep(FALSE, nrow(data_BIDC_4))
include_5 = rep(FALSE, nrow(data_BIDC_5))

for (i in 1:n)
{
  if (sum(as.numeric(data_BIDC[i,c(608, 611:614, 616:687)])) == 0)
  {
    include[i] = TRUE
  }
}

sum(include)
```

```
## [1] 889
```

```
sum(as.numeric((data_BIDC[,608] != 0)))
```

```
## [1] 17
```

```
for (s in 611:614)
{
  print(sum(as.numeric((data_BIDC[,s] != 0))))
}
```

```
## [1] 21
## [1] 17
## [1] 19
## [1] 19
```

```
sum(include)
```

```
## [1] 889
```

```
############################ EARLY EDA ###########
data_check = data_BIDC[include,]
```

```
sum(data_check$treatment)
```

```
## [1] 522
```

```
nrow(data_check) - sum(data_check$treatment)
```

```
## [1] 367
```

```
colnames(data_check)[1:30]
```

```
##  [1] "age_at_diagnosis"               "type_of_breast_surgery"
##  [3] "cellularity"                    "chemotherapy"
##  [5] "pam50_._claudin.low_subtype"    "er_status_measured_by_ihc"
##  [7] "er_status"                      "neoplasm_histologic_grade"
##  [9] "her2_status_measured_by_snp6"   "her2_status"
## [11] "tumor_other_histologic_subtype" "hormone_therapy"
## [13] "inferred_menopausal_state"      "integrative_cluster"
## [15] "primary_tumor_laterality"       "lymph_nodes_examined_positive"
## [17] "mutation_count"                 "nottingham_prognostic_index"
## [19] "overall_survival_months"        "pr_status"
## [21] "radio_therapy"                  "X3.gene_classifier_subtype"
## [23] "tumor_size"                     "tumor_stage"
## [25] "death_from_cancer"              "brca1"
## [27] "brca2"                          "palb2"
## [29] "pten"                           "tp53"
```

```
sum(data_check$death_from_cancer == "Died of Disease" & data_check$treatment == 1) / sum(data_check$trea
```

```
## [1] 0.3984674
```

```
sum(data_check$death_from_cancer == "Died of Disease" & data_check$treatment == 0) / sum(data_check$trea
```

```
## [1] 0.2752044
```

```
sum(is.na(data_sel[include,8])) / 889
```

```
## [1] 0.03149606
```

```
sum(is.na(data_sel[include,17])) / 889
```

```
## [1] 0.03712036
```

```
sum(is.na(data_sel[include,23])) / 889
```

```
## [1] 0.005624297
```

```
sum(is.na(data_sel[include,24]))/ 889
```

```
## [1] 0.2542182
```

```
1 - (nrow(na.omit(data_sel[include,-17])) / nrow(data_sel[include,]))
```

```
## [1] 0.2744657
```

```
max(data_check$overall_survival_months)
```

```
## [1] 337.0333
```

```
min(data_check$overall_survival_months)
```

```
## [1] 0.7666667
```

```
data_check$death_from_cancer[data_check$overall_survival_months == max(data_check$overall_survival_mont
```

```
## [1] "Living"
```

```
data_check$death_from_cancer[data_check$overall_survival_months == min(data_check$overall_survival_mont
```

```
## [1] "Living"
```

```
################################################

data_BIDC_common = data_BIDC[data_BIDC$tumor_stage != 0 & data_BIDC$tumor_stage != 4 & include,]

data_BIDC_common_2 = data_BIDC_2[data_BIDC_2$tumor_stage != 0 & data_BIDC_2$tumor_stage != 4 & include,]
data_BIDC_common_3 = data_BIDC_3[data_BIDC_3$tumor_stage != 0 & data_BIDC_3$tumor_stage != 4 & include,]
data_BIDC_common_4 = data_BIDC_4[data_BIDC_4$tumor_stage != 0 & data_BIDC_4$tumor_stage != 4 & include,]
data_BIDC_common_5 = data_BIDC_5[data_BIDC_5$tumor_stage != 0 & data_BIDC_5$tumor_stage != 4 & include,]



nrow(data_BIDC_common)
```

```
## [1] 879
```

We have 19 variables describing patient characteristics, tumor characteristics and other treatments, 489 variables representing gene expression values and 173 indicators of common mutations for a subset of these genes.

```
# delete mutations which did not occur
mutation_names = setdiff(mutation_names, c("hras_mut", "siah1_mut", "smarcb1_mut", "stmn2_mut", "foxo1_

# delete rare mutations
mutation_names = setdiff(mutation_names, names(data_BIDC)[c(608, 611:614, 616:685)])
```

Creating model matrix

```
# Convert the mutation variables, categorical variables, and treatment to factors

categorical_vars <- c("cellularity", "pam50_._claudin.low_subtype",
                      "neoplasm_histologic_grade", "tumor_other_histologic_subtype",
                      "integrative_cluster", "X3.gene_classifier_subtype", 'tumor_stage')

categorical_recoded <- c("menopause", "HER2_status", "progesterone_status", "tumor_laterality", "radio_


data_BIDC_common <- data_BIDC_common %>% mutate(across(all_of(unlist(c(categorical_vars))), as.factor))

########################
data_BIDC_common_2 <- data_BIDC_common_2 %>% mutate(across(all_of(unlist(c(categorical_vars))), as.facto
data_BIDC_common_3 <- data_BIDC_common_3 %>% mutate(across(all_of(unlist(c(categorical_vars))), as.facto
data_BIDC_common_4 <- data_BIDC_common_4 %>% mutate(across(all_of(unlist(c(categorical_vars))), as.facto
data_BIDC_common_5 <- data_BIDC_common_5 %>% mutate(across(all_of(unlist(c(categorical_vars))), as.facto

#########################
# Create the interaction terms

interaction_terms <- lapply(1:length(gene_expr_names), function(i) {
  if (paste0(gene_expr_names[i], "_mut") %in% mutation_names) {
    return(paste0(gene_expr_names[i], ":", gene_expr_names[i], "_mut"))
  }
})
interaction_terms <- unlist(interaction_terms)

# Define the main effects variables
main_effects <- c("age_at_diagnosis", "lymph_nodes_examined_positive", "tumor_size")

# Create the formula for the model matrix

vars1 = paste(main_effects, collapse = " + ")
vars2 = paste(categorical_vars, collapse = " + ")
vars3 = paste(categorical_recoded, collapse = " + ")
vars3 = paste(gene_expr_names, collapse = " + ")
vars4 = paste(interaction_terms, collapse = " + ")


vars_all = paste(vars1, vars2, vars3, vars4, sep = " + ")
form = paste(" ~  ", "treatment + ", vars_all, " + treatment:(", vars_all, ")")

# STANDARDIZE THE VARIABLES FOLLOWING GELMAN
data_BIDC_common_std = data_BIDC_common

data_BIDC_common_std_2 = data_BIDC_common_2
data_BIDC_common_std_3 = data_BIDC_common_3
data_BIDC_common_std_4 = data_BIDC_common_4
data_BIDC_common_std_5 = data_BIDC_common_5

cont.names = c(main_effects, gene_expr_names)

for (i in 1:ncol(data_BIDC_common))
{
  if (names(data_BIDC_common)[i] %in% cont.names)
  {
    data_BIDC_common_std[,i] = (data_BIDC_common[,i] - mean(data_BIDC_common[,i]))/ (2*sd(data_BIDC_comm

    data_BIDC_common_std_2[,i] = (data_BIDC_common_2[,i] - mean(data_BIDC_common_2[,i]))/ (2*sd(data_BID

    data_BIDC_common_std_3[,i] = (data_BIDC_common_3[,i] - mean(data_BIDC_common_3[,i]))/ (2*sd(data_BID

        data_BIDC_common_std_4[,i] = (data_BIDC_common_4[,i] - mean(data_BIDC_common_4[,i]))/ (2*sd(data

            data_BIDC_common_std_5[,i] = (data_BIDC_common_5[,i] - mean(data_BIDC_common_5[,i]))/ (2*sd

  }
}



model_matrix <- model.matrix(as.formula(form), data = data_BIDC_common_std)

model_matrix_2 <- model.matrix(as.formula(form), data = data_BIDC_common_std_2)
```

```
model_matrix_3 <- model.matrix(as.formula(form), data = data_BIDC_common_std_3)
model_matrix_4 <- model.matrix(as.formula(form), data = data_BIDC_common_std_4)
model_matrix_5 <- model.matrix(as.formula(form), data = data_BIDC_common_std_5)


delta_2 = 1*(data_BIDC_common_2$death_from_cancer == "Died of Disease")
delta_3 = 1*(data_BIDC_common_3$death_from_cancer == "Died of Disease")
delta_4 = 1*(data_BIDC_common_4$death_from_cancer == "Died of Disease")
delta_5 = 1*(data_BIDC_common_5$death_from_cancer == "Died of Disease")

y_2 = Surv(data_BIDC_common_2$overall_survival_months, delta_2)
y_3 = Surv(data_BIDC_common_3$overall_survival_months, delta_3)
y_4 = Surv(data_BIDC_common_4$overall_survival_months, delta_4)
y_5 = Surv(data_BIDC_common_5$overall_survival_months, delta_5)

treatment_2 = data_BIDC_common_2$treatment
treatment_3 = data_BIDC_common_3$treatment
treatment_4 = data_BIDC_common_4$treatment
treatment_5 = data_BIDC_common_5$treatment
```

```
dataX = model_matrix[,2:ncol(model_matrix)]
logY = log(data_BIDC_common$overall_survival_months)
delta = 1*(data_BIDC_common$death_from_cancer == "Died of Disease")

x = model_matrix
y = Surv(data_BIDC_common$overall_survival_months, delta)

p = ncol(dataX)

#weight.set <- list("w" = c(0, rep(1, p-1)))
#set.seed(3)
#fit.en.cv <- penAFT.cv(dataX, logY, delta, alpha = 0.5, nlambda = 30, nfolds = 5)
```

Out-of-sample performance - concordance

```
#########################################
####### OUT OF SAMPLE PERFORMANCE ######
#########################################

get.concordance = function(pred_test, truth_test, death)
{
  nvalid = length(pred_test)
  agree.count = 0
  pair.count = 0
  for (i in 2:nvalid)
  {
    for (j in 1:(i-1))
    {
      pair.count = pair.count + death[j]
      agree.count = agree.count + death[j]*((pred_test[i] >= pred_test[j]) == (truth_test[i] >= truth_te
    }
  }
  concord = agree.count / pair.count
  return(concord)
}

ndata = nrow(data_BIDC_common)
nvalid = floor(0.3*nrow(data_BIDC_common))

#ndata = nrow(dataX)
#nvalid = floor(0.3*nrow(dataX))

#set.seed(4)

set.seed(4)
index = sample(1:ndata, size = nvalid, replace = FALSE)
test_data = data_BIDC_common[index,]
train_data = data_BIDC_common[-index,]

#en.gehan = penAFT(dataX[-index,], logY[-index], delta[-index], alpha = 0.6, lambda = c(lambda.gehan))

set.seed(8)
en.gehan.cv = penAFT.cv(dataX[-index,], logY[-index], delta[-index], alpha = 0.5, nlambda = 30, nfold =
```

```
## CV through:  ###                 20 %
## CV through:  ### ###             40 %
## CV through:  ### ### ###         60 %
## CV through:  ### ### ### ###     80 %
## CV through:  ### ### ### ### ###  100 %
```

```
############################################################
###### Change after fixing package #########################

lambda = en.gehan.cv$full.fit$lambda
cv.err.linPred = en.gehan.cv$cv.err.linPred
lambda[which(cv.err.linPred == min(cv.err.linPred))]
```

```
## [1] 0.02274304
```

```
best.ind = which.min(cv.err.linPred)
lambda.min = lambda[which.min(cv.err.linPred)]
lambda.gehan = lambda.min

#beta.gehan = fit.en.cv$full.fit$beta[,best.ind]
############################################################

#saveRDS(en.gehan.cv, file = 'gehan_train.RDS')
#en.gehan = penAFT(dataX[-index,], logY[-index], delta[-index], alpha = 0.6, nlambda = 10)

x = model_matrix

set.seed(6)
p = ncol(x)
penalty = c(0, rep(1, p-1))
en.cox.cv =  cv.glmnet(x[-index,], y[-index,], family = "cox", alpha = 0.5, nlambda = 30, nfold = 5)
en.cox.cv$index
```

```
##     Lambda
## min      6
## 1se      3
```

```
beta.cox = en.cox.cv$glmnet.fit$beta[,8]
lambda.cox = en.cox.cv$glmnet.fit$lambda[8]
preds.cox <- x[index,]%*%beta.cox

truth_test = logY[index]
death = delta[index]


preds.gehan <- penAFT.predict(en.gehan.cv, Xnew = dataX[index,], lambda = lambda.gehan)

get.concordance(-preds.cox, truth_test, death)
```

```
## [1] 0.677399
```

```
get.concordance(preds.gehan, truth_test, death)
```

```
## [1] 0.6635101
```

Causal Inference

Propensity Scores

```
x = model_matrix

delta = 1*(data_BIDC_common$death_from_cancer == "Died of Disease")
y = Surv(data_BIDC_common$overall_survival_months, delta)

# PROPENSITY SCORES

vars1 = paste(main_effects, collapse = " + ")
vars2 = paste(categorical_vars, collapse = " + ")
vars3 = paste(categorical_recoded, collapse = " + ")
vars3 = paste(gene_expr_names, collapse = " + ")
vars4 = paste(interaction_terms, collapse = " + ")

x = model_matrix

vars_all = paste(vars1, vars2, vars3, vars4, sep = " + ")
form_2 = paste("treatment ~  ", vars_all)

x_prop <- model.matrix(as.formula(form_2), data = data_BIDC_common_std)
treatment = data_BIDC_common$treatment

# ridge regression
set.seed(3)
prop.cv  = cv.glmnet(x_prop, treatment, family = "binomial", alpha = 0, nlambda = 30, nfold = 5)
e.vec = predict(prop.cv, newx = x_prop, s = prop.cv$lambda.min, type = 'response')
omega = 1
W.vec = omega / (treatment*e.vec + (1-treatment)*e.vec)
```

```
W.vec = c(W.vec)
```

EDA

```
data_EDA = data_BIDC_common
data_EDA$prop_score = e.vec

data_EDA$death = ifelse(delta, "Yes", "No")


print(nrow(data_EDA))
```

```
## [1] 879
```

```
print(sum(delta))
```

```
## [1] 303
```

```
print(sum(data_EDA$treatment))
```

```
## [1] 513
```

```
print(nrow(data_EDA) - sum(data_EDA$treatment))
```

```
## [1] 366
```

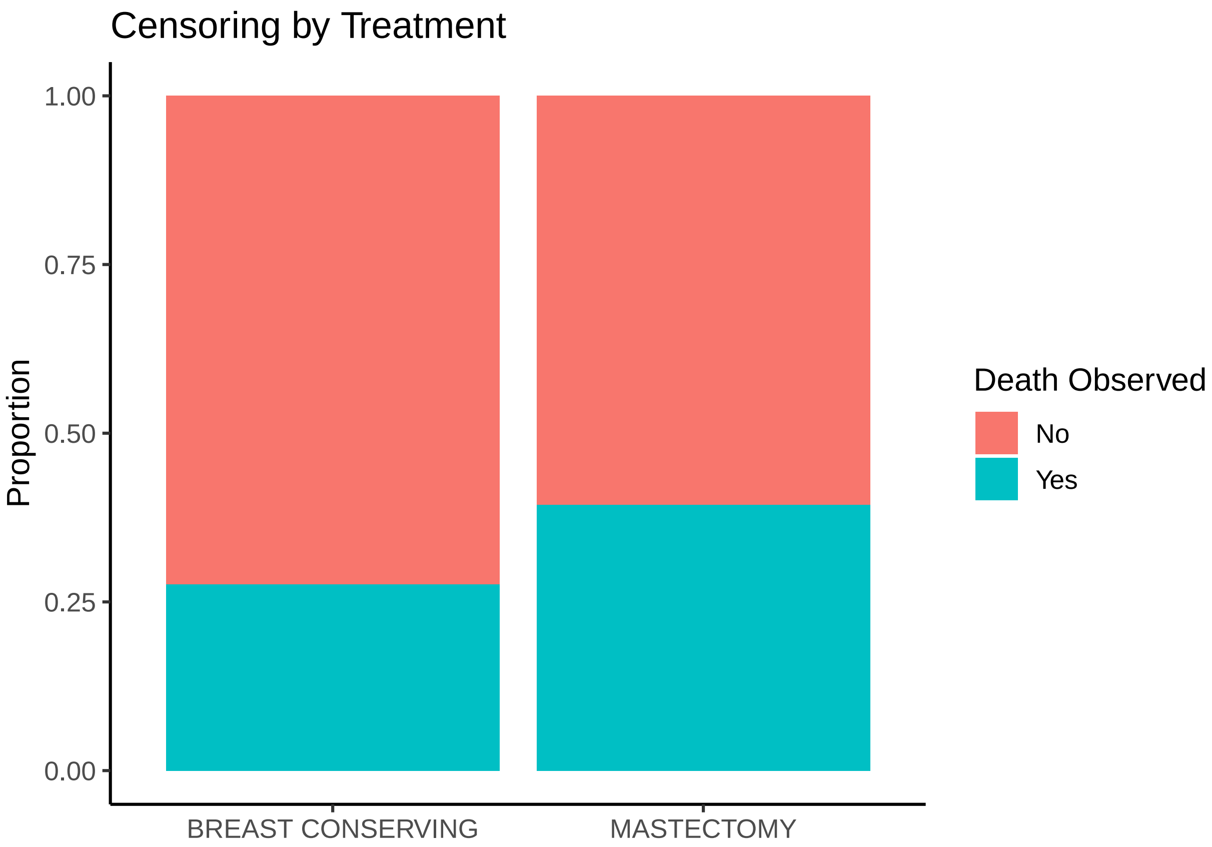
```
print(sum(data_EDA$treatment == 1 & delta))
```

```
## [1] 202
```

```
print(sum(data_EDA$treatment == 0 & delta))
```

```
## [1] 101
```

```
data_EDA$treatment_name = ifelse(data_EDA$type_of_breast_surgery == 'MASTECTOMY', 'Mastectomy', 'BCS')


g1 <- ggplot(data_EDA, aes(x = type_of_breast_surgery, fill = death)) +
  geom_bar(position = "fill") +
  ggtitle("Censoring by Treatment") +
  theme(axis.title.x = element_blank()) +
  labs(fill = "Death Observed", y = "Proportion")


# pdf("plots/censoring_by_trt.pdf", height = 3.5, width = 6)
# g1
# dev.off()

g1
```

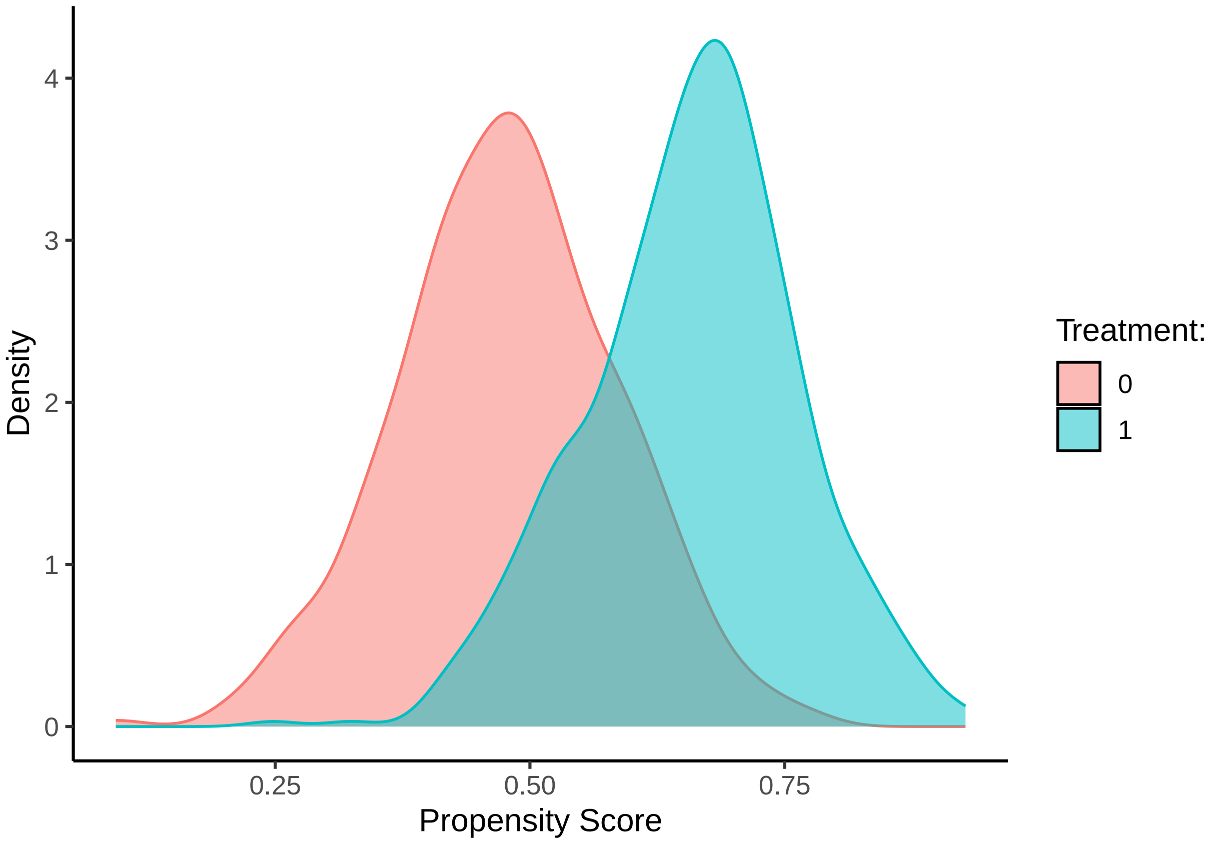


```
#######################################

# w.out1 <- WeightIt::weightit(
#     treatment ~ age + income + dnr1 + cat2 + gastr +
#           aps1 + sod1 + ninsclas,
#     data = rhc_processed, estimand = "ATE", method = "ps")
#
# w.out1 <- WeightIt::weightit(as.formula(form_2), data = data_BIDC_common_std, estimand = "ATE", metho
#
# love.plot(w.out1)
```

```
sum(data_EDA$treatment)
```

```
## [1] 513
```

```
data_EDA = data_BIDC_common
data_EDA$prop_score = e.vec
data_EDA$Mastectomy = as.factor(data_EDA$treatment)

p_prop_score <- ggplot(data_EDA, aes(x = prop_score, fill = as.factor(treatment),
                                 color = as.factor(treatment))) +
  geom_density(alpha = 0.5) +
  labs(x = "Propensity Score",
       y = "Density",
       fill = "Treatment:") +
  scale_color_discrete(guide = "none")
p_prop_score
```

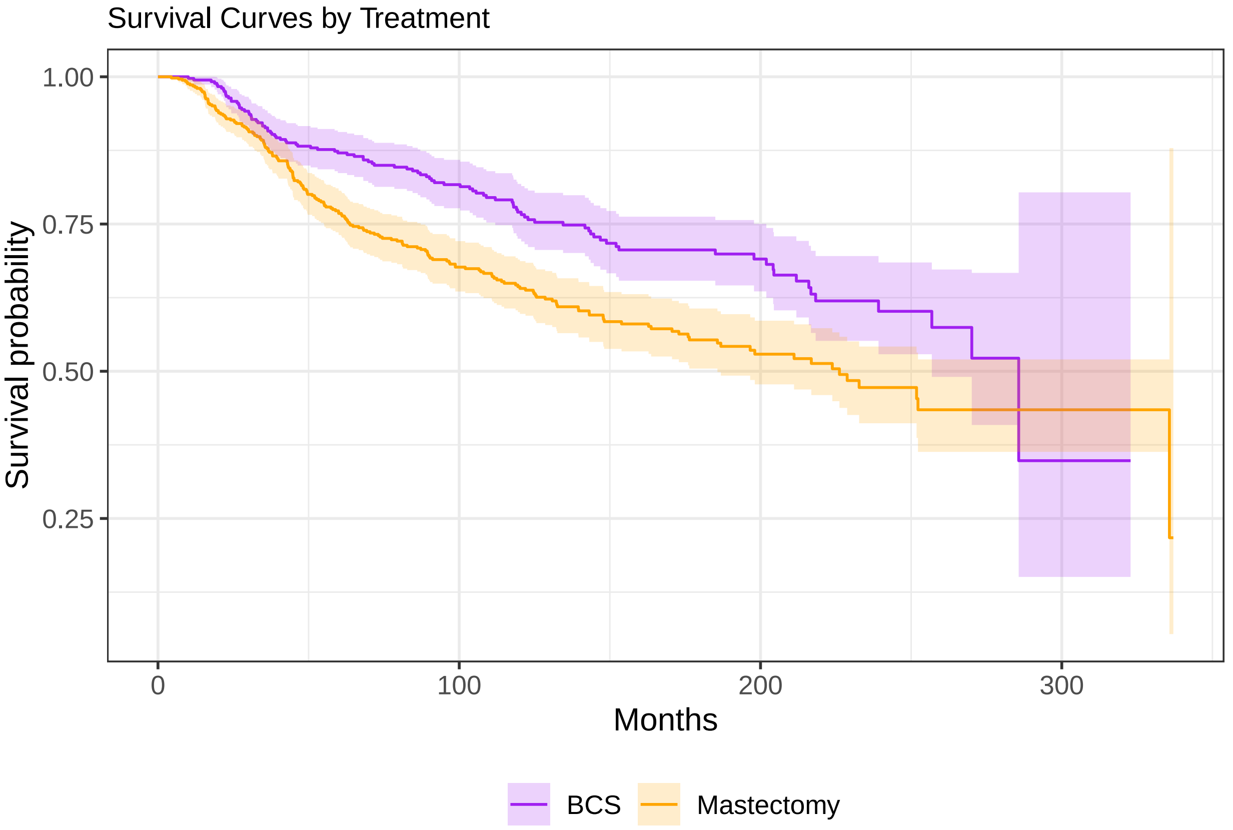


```
#ggsave("plots/prop_scores.pdf", p_prop_score, width = 8, height = 4)
```

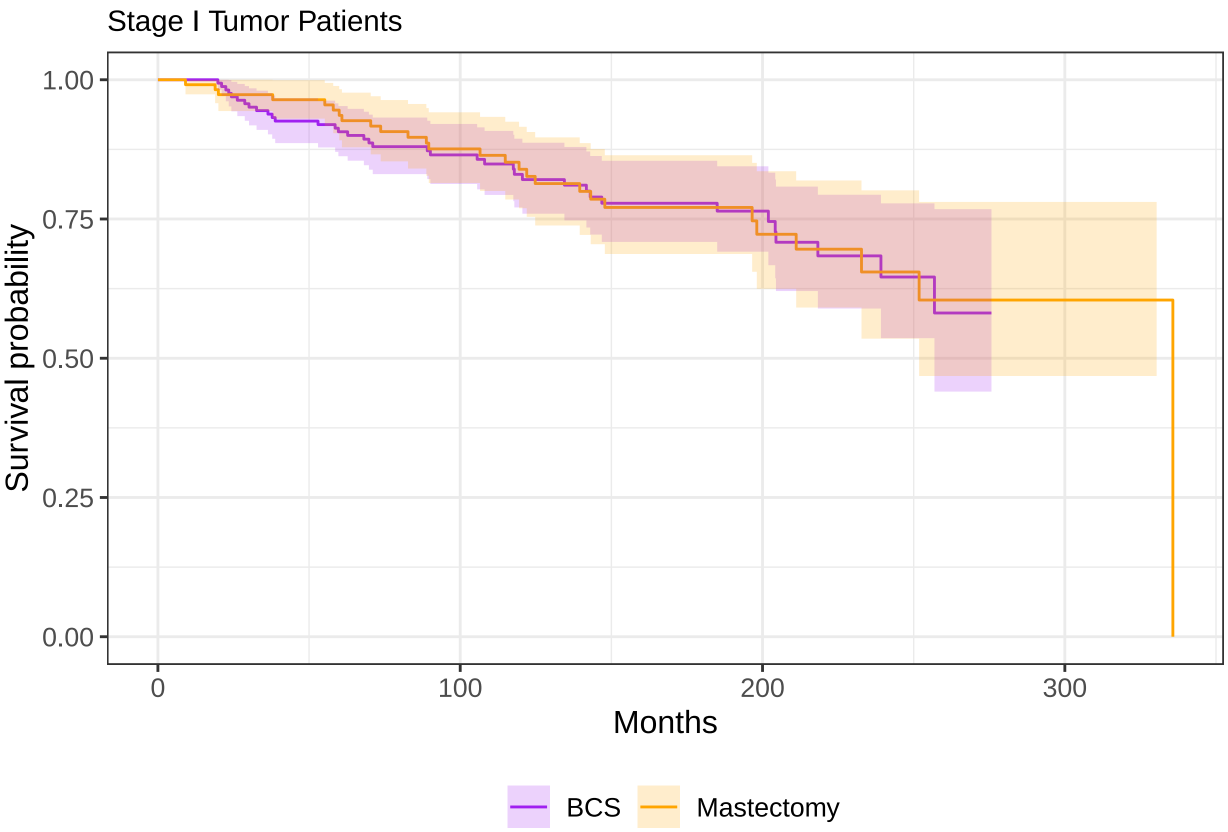
```
PlotKMCurve <- function(group_var, var_name = NULL, data, y) {
  if (is.null(var_name)) var_name <- group_var
  survfit2(y ~ get(group_var), data = data) %>%
    ggsurvfit() +
    add_confidence_interval() +
    labs(
      x = "Months",
      y = "Survival probability",
      fill = element_blank(), color = element_blank()
    )
}

fig1 <- PlotKMCurve("Mastectomy", var_name = NULL, data_EDA, y) +
  labs(subtitle = "Survival Curves by Treatment") +
  scale_color_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy")) +
  scale_fill_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy"))

pdf("plots/surv_curve.pdf", height = 3.5, width = 6)
fig1
dev.off()
```

```
## pdf
##   2
```

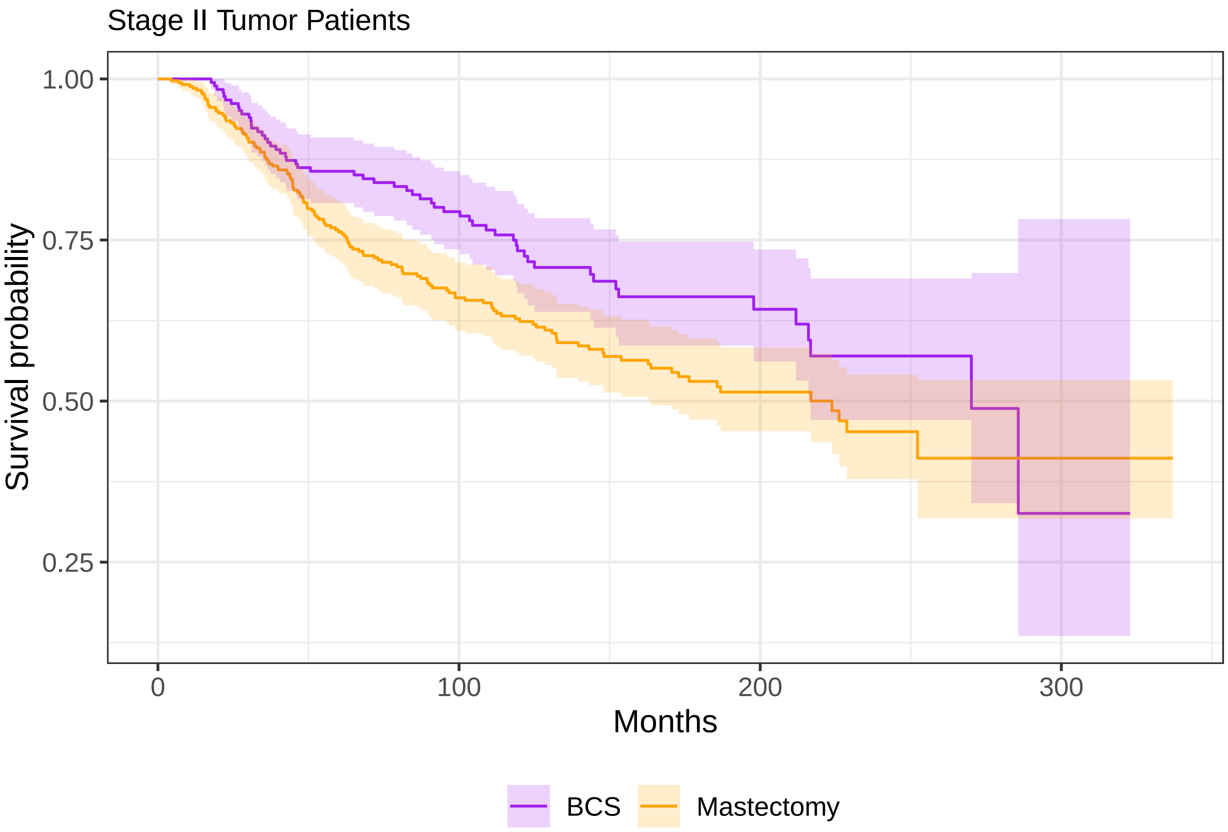
```
fig1
```



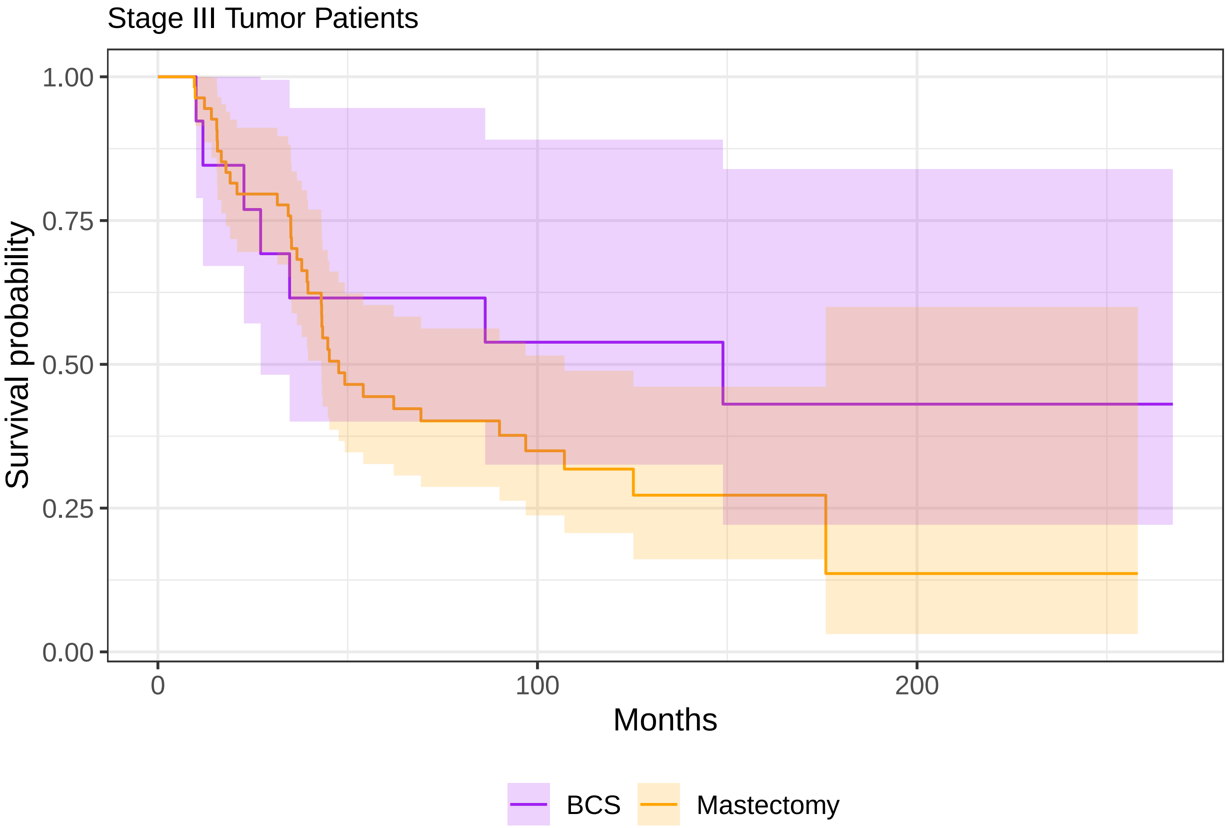
```
fig_stage1 <- PlotKMCurve("Mastectomy", var_name = NULL, data_EDA[data_EDA$tumor_stage == 1,], y[data_EI
  labs(subtitle = "Stage I Tumor Patients") +
  scale_color_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy")) +
  scale_fill_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy"))

fig_stage2 <- PlotKMCurve("Mastectomy", var_name = NULL, data_EDA[data_EDA$tumor_stage == 2,], y[data_EI
  labs(subtitle = "Stage II Tumor Patients") +
  scale_color_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy")) +
  scale_fill_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy"))


fig_stage3 <- PlotKMCurve("Mastectomy", var_name = NULL, data_EDA[data_EDA$tumor_stage == 3,], y[data_EI
  labs(subtitle = "Stage III Tumor Patients") +
  scale_color_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy")) +
  scale_fill_manual(values = c("purple", "orange"), labels = c("BCS", "Mastectomy"))


fig_stage1
```


Stage I Tumor Patients
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Months
BCS
Mastectomy

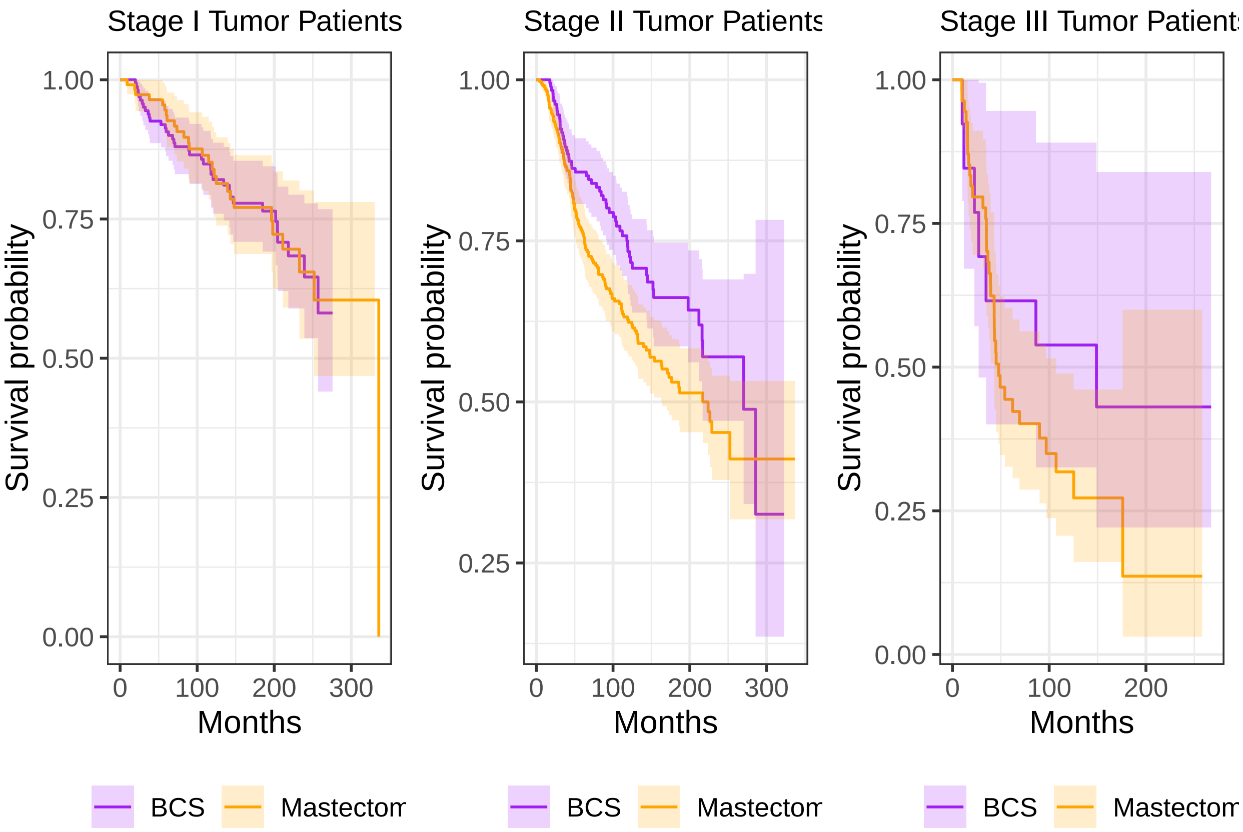
```
fig_stage2
```


Stage II Tumor Patients
Survival probability
1.00
0.75
0.50
0.25
0
100
200
300
Months
BCS
Mastectomy

```
fig_stage3
```



```
# Arrange the plots side by side
combined_plots <- grid.arrange(fig_stage1, fig_stage2, fig_stage3, ncol = 3)
```



```
# Save the arranged plots as a PDF
ggsave("plots/stages_surv.pdf", combined_plots, width = 10, height = 4)


pdf("plots/stage1_diff.pdf", height = 3.5, width = 6)
fig_stage1
dev.off()
```

```
## pdf
##   2
```

```
pdf("plots/stage2_diff.pdf", height = 3.5, width = 6)
fig_stage2
dev.off()
```

```
## pdf
##   2
```

```
pdf("plots/stage3_diff.pdf", height = 3.5, width = 6)
fig_stage3
dev.off()
```

```
## pdf
##   2
```

```
# Exclude treatment from the penalized predictors
p = ncol(x)
penalty = c(0, 0, rep(1, p-2))
```

```
set.seed(4)
cox.cv = cv.glmnet(x, y, weights = W.vec, penalty.factor = penalty, family = "cox", alpha = 0.5, nlambd
lambda.cox = cox.cv$lambda.min

cox.fit =  glmnet(x, y, weights = W.vec, family = "cox", alpha = 0.5, penalty.factor = penalty, lambda =
```

```
cox.fit$beta[,1]['treatment']
```

```
## treatment
## 0.3477823
```

```
cox.en.beta = cox.fit$beta[,1]
cox.en.beta.nonzero = cox.fit$beta[,1][abs(cox.fit$beta[,1]) > 1e-10][order(cox.fit$beta[,1][abs(cox.fi

length(cox.en.beta.nonzero)
```

```
## [1] 51
```

```
cox.en.beta.nonzero
```

```
##   lymph_nodes_examined_positive                 kdm3a:kdm3a_mut
##                     0.429991335                     0.369957710
##                       treatment            integrative_cluster5
##                     0.347782309                     0.261403318
##                           gsk3b                  treatment:e2f8
##                     0.239318961                     0.219624668
##                           prkg1            treatment:tumor_size
##                     0.199972830                     0.189831732
##                        eif4ebp1                treatment:zfyve9
##                     0.129527071                     0.084331358
##                            e2f7                            dll3
##                     0.060377905                     0.054955479
##                  treatment:e2f7                      tumor_size
##                     0.040005288                     0.039285186
##                           smad7                 treatment:nrarp
##                     0.031579997                     0.031220094
##                          prkacg                           pdgfb
##                     0.026250727                     0.015773806
##                 treatment:ush2a                            akt3
##                     0.013131015                     0.007516268
##                            men1                           smad6
##                     0.006066667                     0.003035938
##                treatment:srd5a3               treatment:hsd17b7
##                    -0.002806890                    -0.003366814
##                  treatment:nrg3                          map3k1
##                    -0.008037592                    -0.011575291
##                           casp7                           mmp25
##                    -0.012710693                    -0.014298532
##                treatment:acvr1c                            igf1
##                    -0.014745488                    -0.015017205
##      treatment:erbb2:erbb2_mut                  treatment:rpgr
##                    -0.030886788                    -0.031102534
##            integrative_cluster3               treatment:ugt2b17
##                    -0.036916691                    -0.041345522
##                           brca2                         ugt2b17
##                    -0.042959447                    -0.043295059
##                  treatment:inha                            mapt
##                    -0.052234841                    -0.055817428
##                            e2f1         treatment:chd1:chd1_mut
##                    -0.057084879                    -0.059168382
##                          diras3                   tbx3:tbx3_mut
##                    -0.078511004                    -0.079733374
##                          acvr1c                            cul1
##                    -0.105162501                    -0.106889044
##                          stat5a                   tp53:tp53_mut
##                    -0.111515329                    -0.121472412
##                 gata3:gata3_mut                  treatment:cul1
##                    -0.138823153                    -0.145590316
## pam50_._claudin.low_subtypeLumA               treatment:cyp11a1
##                    -0.152032481                    -0.154061815
##                  treatment:mapt
##                    -0.317144317
```

Sensitivity Anaysis - different MICE imputations

```
x_2 = model_matrix_2
x_3 = model_matrix_3
x_4 = model_matrix_4
x_5 = model_matrix_5

x_prop_2 <- model.matrix(as.formula(form_2), data = data_BIDC_common_std_2)
x_prop_3 <- model.matrix(as.formula(form_2), data = data_BIDC_common_std_3)
x_prop_4 <- model.matrix(as.formula(form_2), data = data_BIDC_common_std_4)
x_prop_5 <- model.matrix(as.formula(form_2), data = data_BIDC_common_std_5)

# ridge regression
set.seed(3)
prop.cv_2  = cv.glmnet(x_prop_2, treatment_2, family = "binomial", alpha = 0, nlambda = 30, nfold = 5)
e.vec_2 = predict(prop.cv_2, newx = x_prop_2, s = prop.cv_2$lambda.min, type = 'response')
omega  = 1
W.vec_2 = omega / (treatment_2*e.vec_2 + (1-treatment_2)*e.vec_2)

set.seed(3)
prop.cv_3  = cv.glmnet(x_prop_3, treatment_3, family = "binomial", alpha = 0, nlambda = 30, nfold = 5)
e.vec_3 = predict(prop.cv_3, newx = x_prop_3, s = prop.cv_3$lambda.min, type = 'response')
omega  = 1
W.vec_3 = omega / (treatment_3*e.vec_3 + (1-treatment_3)*e.vec_3)

set.seed(3)
prop.cv_4  = cv.glmnet(x_prop_4, treatment_4, family = "binomial", alpha = 0, nlambda = 30, nfold = 5)
e.vec_4 = predict(prop.cv_4, newx = x_prop_4, s = prop.cv_4$lambda.min, type = 'response')
omega  = 1
W.vec_4 = omega / (treatment_4*e.vec_4 + (1-treatment_4)*e.vec_4)

set.seed(3)
prop.cv_5  = cv.glmnet(x_prop_5, treatment_5, family = "binomial", alpha = 0, nlambda = 30, nfold = 5)
e.vec_5 = predict(prop.cv_5, newx = x_prop_5, s = prop.cv_5$lambda.min, type = 'response')
omega  = 1
W.vec_5 = omega / (treatment_5*e.vec_5 + (1-treatment_5)*e.vec_5)
```

```
set.seed(4)
cox.cv_2 = cv.glmnet(x_2, y_2, weights = W.vec_2, penalty.factor = penalty, family = "cox", alpha = 0.5
lambda.cox_2 = cox.cv_2$lambda.min

cox.fit_2 =  glmnet(x_2, y_2, weights = W.vec_2, family = "cox", alpha = 0.5, penalty.factor = penalty,

set.seed(4)
cox.cv_3 = cv.glmnet(x_3, y_3, weights = W.vec_3, penalty.factor = penalty, family = "cox", alpha = 0.5
lambda.cox_3 = cox.cv_3$lambda.min

cox.fit_3 =  glmnet(x_3, y_3, weights = W.vec_3, family = "cox", alpha = 0.5, penalty.factor = penalty,

set.seed(4)
cox.cv_4 = cv.glmnet(x_4, y_4, weights = W.vec_4, penalty.factor = penalty, family = "cox", alpha = 0.5
lambda.cox_4 = cox.cv_4$lambda.min

cox.fit_4 =  glmnet(x_4, y_4, weights = W.vec_4, family = "cox", alpha = 0.5, penalty.factor = penalty,

set.seed(4)
cox.cv_5 = cv.glmnet(x_5, y_5, weights = W.vec_5, penalty.factor = penalty, family = "cox", alpha = 0.5
lambda.cox_5 = cox.cv_5$lambda.min

cox.fit_5 =  glmnet(x_5, y_5, weights = W.vec_5, family = "cox", alpha = 0.5, penalty.factor = penalty,
```

```
cox.en.beta_2 = cox.fit_2$beta[,1]
cox.en.beta.nonzero_2 = cox.fit_2$beta[,1][abs(cox.fit_2$beta[,1]) > 1e-10][order(cox.fit_2$beta[,1][abs

cox.en.beta_3 = cox.fit_3$beta[,1]
cox.en.beta.nonzero_3 = cox.fit_3$beta[,1][abs(cox.fit_3$beta[,1]) > 1e-10][order(cox.fit_3$beta[,1][abs

cox.en.beta_2 = cox.fit_2$beta[,1]
cox.en.beta.nonzero_2 = cox.fit_2$beta[,1][abs(cox.fit_2$beta[,1]) > 1e-10][order(cox.fit_2$beta[,1][abs

cox.en.beta.nonzero_2
```

```
##   lymph_nodes_examined_positive                       treatment
##                    0.3962247677                    0.3643216028
##                 kdm3a:kdm3a_mut            integrative_cluster5
##                    0.3415824700                    0.2874316842
##                           gsk3b            treatment:tumor_size
##                    0.2434052539                    0.2022740288
##                  treatment:e2f8                           prkg1
##                    0.1786587844                    0.1395393081
##                        eif4ebp1                    tumor_stage3
##                    0.1299748576                    0.1213385278
##                treatment:zfyve9                 treatment:nrarp
##                    0.1081327068                    0.0703986734
##                 treatment:prkcz                    tumor_stage2
##                    0.0565429854                    0.0500720413
##                      tumor_size                  treatment:e2f7
##                    0.0357496013                    0.0321913606
##                            e2f7                            dll3
##                     0.0280666054                    0.0266294111
##                  treatment:ush2a                 treatment:cxcl8
##                     0.0240795448                    0.0103307838
##                            pdgfb              treatment:eif4ebp1
##                     0.0094421874                    0.0080216058
##                             men1                  treatment:terc
##                     0.0077136094                    0.0076179627
##                            smad7               shank2:shank2_mut
##                     0.0025447624                   -0.0001583666
##                  treatment:pbrm1                           nr3c1
##                    -0.0011144275                   -0.0034206502
##                  birc6:birc6_mut                treatment:srd5a3
##                    -0.0042340371                   -0.0080621912
##                            casp7                treatment:acvr1c
##                    -0.0114110918                   -0.0166340409
##                             igf1            integrative_cluster3
##                    -0.0196002170                   -0.0197492308
##                            brca2                            e2f1
##                    -0.0318839150                   -0.0329041489
##                           map3k1                            mapt
##                    -0.0350371329                   -0.0382757821
##                   treatment:nrg3                 gata3:gata3_mut
##                    -0.0468981186                   -0.0539752577
##                   treatment:inha                         ugt2b17
##                    -0.0581064544                   -0.0653846921
##                treatment:hsd17b7         treatment:chd1:chd1_mut
##                    -0.0665996709                   -0.0768081300
##                           diras3                   tp53:tp53_mut
##                    -0.0804763607                   -0.0813106049
##                treatment:cyp11a1                     atr:atr_mut
##                    -0.0860935572                   -0.0878894843
##                             cul1                          acvr1c
##                    -0.0944239601                   -0.0997078098
##                    tbx3:tbx3_mut                  treatment:cul1
##                    -0.1021060807                   -0.1254065359
##                           stat5a pam50_._claudin.low_subtypeLumA
##                    -0.1442501922                   -0.1679222503
##                   treatment:mapt
##                    -0.3331440694
```

```
cox.en.beta.nonzero_3
```

```
##    lymph_nodes_examined_positive                       treatment
##                     0.3841860734                    0.3476927998
##             integrative_cluster5                           gsk3b
##                     0.2384409930                    0.2284688367
##                     tumor_stage3                  treatment:e2f8
##                     0.2141502347                    0.1785825332
##             treatment:tumor_size                        eif4ebp1
##                     0.1514550933                    0.1186634378
##                            prkg1                            e2f7
##                     0.1026237250                    0.0375293907
##                            aurka                            dll3
##                     0.0241813119                    0.0221239338
##                 treatment:zfyve9                  treatment:e2f7
##                    0.0100678051                    0.0090238419
##                           smad7                           pdgfb
##                    0.0069072699                    0.0009589283
##                           brca2                  treatment:rpgr
##                   -0.0016476231                   -0.0058930115
##                            e2f1                            mapt
##                   -0.0065817850                   -0.0207472337
##                         ugt2b17                   tp53:tp53_mut
##                   -0.0228313861                   -0.0299565115
##                 gata3:gata3_mut                          map3k1
##                   -0.0303628989                   -0.0335650384
##                            cul1               treatment:hsd17b7
##                   -0.0352237413                   -0.0394587768
##                          diras3                          acvr1c
##                   -0.0605019989                   -0.0774130175
##                          stat5a               treatment:cyp11a1
##                   -0.1088171940                   -0.1148574350
##                  treatment:cul1 pam50_._claudin.low_subtypeLumA
##                   -0.1158318310                   -0.1289206891
##                  treatment:mapt
##                   -0.2602185733
```

```
nonzero_2 = as.numeric(which(abs(cox.fit_2$beta[,1]) > 1e-10))
nonzero_3 = as.numeric(which(abs(cox.fit_3$beta[,1]) > 1e-10))
nonzero_4 = as.numeric(which(abs(cox.fit_4$beta[,1]) > 1e-10))
nonzero_5 = as.numeric(which(abs(cox.fit_5$beta[,1]) > 1e-10))
```

```
common = intersect(nonzero_2, intersect(nonzero_3, intersect(nonzero_4, nonzero_5)))

cox.fit_2$beta[common,1]
```

```
##                       treatment   lymph_nodes_examined_positive
##                     0.364321603                     0.396224768
## pam50_._claudin.low_subtypeLumA            integrative_cluster5
##                    -0.167922250                     0.287431684
##                            e2f7                          stat5a
##                     0.028066605                    -0.144250192
##                            cul1                          acvr1c
##                    -0.094423960                    -0.099707810
##                          diras3                        eif4ebp1
##                    -0.080476361                     0.129974858
##                           gsk3b                          map3k1
##                     0.243405254                    -0.035037133
##                           pdgfb                            mapt
##                     0.009442187                    -0.038275782
##                           prkg1                         ugt2b17
##                     0.139539308                    -0.065384692
##                   tp53:tp53_mut                 gata3:gata3_mut
##                    -0.081310605                    -0.053975258
##            treatment:tumor_size                  treatment:e2f8
##                     0.202274029                     0.178658784
##                  treatment:cul1                treatment:zfyve9
##                    -0.125406536                     0.108132707
##                  treatment:mapt               treatment:cyp11a1
##                    -0.333144069                    -0.086093557
##                treatment:hsd17b7
##                     -0.066599671
```

```
cox.fit_2$beta[common]
```

```
##  [1]  0.364321603  0.396224768 -0.167922250  0.287431684  0.028066605
##  [6] -0.144250192 -0.094423960 -0.099707810 -0.080476361  0.129974858
## [11]  0.243405254 -0.035037133  0.009442187 -0.038275782  0.139539308
## [16] -0.065384692 -0.081310605 -0.053975258  0.202274029  0.178658784
## [21] -0.125406536  0.108132707 -0.333144069 -0.086093557 -0.066599671
```

```
cox.fit_3$beta[common]
```

```
##  [1]  0.3476927998  0.3841860734 -0.1289206891  0.2384409930  0.0375293907
##  [6] -0.1088171940 -0.0352237413 -0.0774130175 -0.0605019989  0.1186634378
## [11]  0.2284688367 -0.0335650384  0.0009589283 -0.0207472337  0.1026237250
## [16] -0.0228313861 -0.0299565115 -0.0303628989  0.1514550933  0.1785825332
## [21] -0.1158318310  0.0100678051 -0.2602185733 -0.1148574350 -0.0394587768
```

```
cox.fit_4$beta[common]
```

```
##  [1]  0.34996087  0.37844573 -0.13994924  0.31536490  0.05016256 -0.10684454
##  [7] -0.05271470 -0.07773049 -0.07644247  0.10434735  0.20454000 -0.03935772
## [13]  0.01722024 -0.03311242  0.12420754 -0.02813309 -0.07297772 -0.03921558
## [19]  0.23477622  0.17115736 -0.13667896  0.08778544 -0.25136692 -0.11628356
## [25] -0.02933571
```

```
cox.fit_5$beta[common]
```

```
##  [1]  0.331994022  0.396280534 -0.144294892  0.240822534  0.029789071
##  [6] -0.132510561 -0.100009043 -0.102728346 -0.078702380  0.122959331
## [11]  0.243855479 -0.022577776  0.001276889 -0.033054844  0.146136591
## [16] -0.070581607 -0.061837565 -0.093933651  0.122027642  0.196330185
## [21] -0.148044856  0.100526881 -0.309994398 -0.165852232 -0.058253673
```

```
sum((cox.fit$beta[common] > 0 & cox.fit_2$beta[common] <= 0) | (cox.fit$beta[common] <= 0 & cox.fit_2$be
```

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

```
sum((cox.fit_2$beta[common] > 0 & cox.fit_3$beta[common] <= 0) | (cox.fit_2$beta[common] <= 0 & cox.fit_
```

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

```
sum((cox.fit_2$beta[common] > 0 & cox.fit_4$beta[common] <= 0) | (cox.fit_2$beta[common] <= 0 & cox.fit_
```

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

```
sum((cox.fit_2$beta[common] > 0 & cox.fit_5$beta[common] <= 0) | (cox.fit_2$beta[common] <= 0 & cox.fit_
```

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

```
get.race.unpenalized = function(x, model, weights)
{
  n = nrow(x)
  ind = which(names(x) == 'treatment')
  x0 = x
  x1 = x
  x0[,ind] = 0
  x1[,ind] = 1

  scurve0 = survfit(model, newdata = x0, weights = W.vec)
  scurve1 = survfit(model, newdata = x1, weights = W.vec)

  AUC0 = sum(rowMeans(scurve0$surv) * c(scurve0$time[1], diff(scurve0$time)))
  AUC1 = sum(rowMeans(scurve1$surv) * c(scurve1$time[1], diff(scurve1$time)))

  RACE = AUC1 - AUC0
  return(RACE)

}
```

Fitting the Cox model with selected variables

```
data_matrix = data.frame(model_matrix)
```

```
coxmodel <- coxph(y ~ kdm3a.kdm3a_mut  + lymph_nodes_examined_positive +
                      treatment      +     integrative_cluster5 +
                    integrative_cluster3 +
                          gsk3b      +      treatment.tumor_size +
                          prkg1     +             treatment.e2f8 +
                       eif4ebp1      +          treatment.zfyve9 +
                treatment.ush2a       +                     dll3 +
                treatment.nrarp        +              tumor_size +
                 treatment.e2f7         +                   e2f7 +
                          smad7          +                  men1 +
                         prkacg           +     treatment.notch1 +
                          pdgfb            +        tumor_stage2 +
                          mmp25             +   treatment.acvr1c +
               treatment.srd5a3             +  treatment.hsd17b7 +
                           mlh1             +              casp7 +
                          brca2             +     treatment.rpgr +
              treatment.ugt2b17             +               igf1 +
                 treatment.nrg3             +             map3k1 +
                           mapt             +               e2f1 +
                        ugt2b17 +
                 treatment.inha             +             diras3 +
                  tp53.tp53_mut             +        atr.atr_mut +
                gata3.gata3_mut             +             acvr1c +
                           cul1             +             stat5a +
                 treatment.cul1 + pam50_._claudin.low_subtypeLumA +
                birc6.birc6_mut           +    treatment.cyp11a1 +
                  tbx3.tbx3_mut           +       treatment.mapt, weights = c(W.vec), data = data_matri
```

```
summary(coxmodel)
```

```
## Call:
## coxph(formula = y ~ kdm3a.kdm3a_mut + lymph_nodes_examined_positive +
##     treatment + integrative_cluster5 + integrative_cluster3 +
##     gsk3b + treatment.tumor_size + prkg1 + treatment.e2f8 + eif4ebp1 +
##     treatment.zfyve9 + treatment.ush2a + dll3 + treatment.nrarp +
##     tumor_size + treatment.e2f7 + e2f7 + smad7 + men1 + prkacg +
##     treatment.notch1 + pdgfb + tumor_stage2 + mmp25 + treatment.acvr1c +
##     treatment.srd5a3 + treatment.hsd17b7 + mlh1 + casp7 + brca2 +
##     treatment.rpgr + treatment.ugt2b17 + igf1 + treatment.nrg3 +
##     map3k1 + mapt + e2f1 + ugt2b17 + treatment.inha + diras3 +
##     tp53.tp53_mut + atr.atr_mut + gata3.gata3_mut + acvr1c +
##     cul1 + stat5a + treatment.cul1 + pam50_._claudin.low_subtypeLumA +
##     birc6.birc6_mut + treatment.cyp11a1 + tbx3.tbx3_mut + treatment.mapt,
##     data = data_matrix, weights = c(W.vec))
##
##   n= 879, number of events= 303
##
##                                    coef exp(coef) se(coef) robust se      z
## kdm3a.kdm3a_mut                 3.43128  30.91621  0.87534   0.87844  3.906
## lymph_nodes_examined_positive   0.72412   2.06292  0.07472   0.09717  7.452
## treatment                       0.26090   1.29810  0.11607   0.15719  1.660
## integrative_cluster5            0.40535   1.49983  0.13883   0.18966  2.137
## integrative_cluster3           -0.32163   0.72497  0.21011   0.25933 -1.240
## gsk3b                           0.15196   1.16411  0.11760   0.16454  0.923
## treatment.tumor_size            0.71672   2.04770  0.26737   0.39037  1.836
## prkg1                           0.56714   1.76322  0.09176   0.12583  4.507
## treatment.e2f8                  0.34472   1.41159  0.13453   0.19751  1.745
## eif4ebp1                        0.16990   1.18518  0.10304   0.14410  1.179
## treatment.zfyve9                0.16526   1.17970  0.12904   0.17868  0.925
## treatment.ush2a                 0.46345   1.58955  0.15592   0.21685  2.137
## dll3                            0.14217   1.15278  0.09442   0.13962  1.018
## treatment.nrarp                 0.32735   1.38729  0.13669   0.17981  1.820
## tumor_size                     -0.20912   0.81129  0.24517   0.36376 -0.575
## treatment.e2f7                 -0.05427   0.94718  0.18771   0.25431 -0.213
## e2f7                            0.09505   1.09971  0.14381   0.19582  0.485
## smad7                           0.18508   1.20331  0.10507   0.16864  1.097
## men1                            0.14010   1.15039  0.10113   0.14157  0.990
## prkacg                          0.28760   1.33322  0.09446   0.12491  2.302
## treatment.notch1                0.22936   1.25780  0.16225   0.23105  0.993
## pdgfb                           0.07733   1.08040  0.10269   0.15567  0.497
## tumor_stage2                    0.22849   1.25670  0.10079   0.13628  1.677
## mmp25                          -0.26227   0.76930  0.10686   0.14638 -1.792
## treatment.acvr1c               -0.09333   0.91089  0.20565   0.28847 -0.324
## treatment.srd5a3               -0.28892   0.74907  0.12379   0.16516 -1.749
## treatment.hsd17b7               0.07064   1.07319  0.15018   0.19819  0.356
## mlh1                           -0.16297   0.84962  0.10250   0.15217 -1.071
## casp7                          -0.21199   0.80898  0.10512   0.13960 -1.519
## brca2                          -0.27453   0.75993  0.09850   0.13500 -2.033
## treatment.rpgr                 -0.06215   0.93974  0.15395   0.20772 -0.299
## treatment.ugt2b17               0.10298   1.10847  0.28363   0.41469  0.248
## igf1                           -0.11583   0.89063  0.13754   0.18137 -0.639
## treatment.nrg3                 -0.52297   0.59276  0.18325   0.28169 -1.857
## map3k1                          0.07408   1.07690  0.11572   0.15638  0.474
## mapt                           -0.10249   0.90258  0.16762   0.23227 -0.441
## e2f1                           -0.27502   0.75956  0.09713   0.13588 -2.024
## ugt2b17                        -0.73711   0.47849  0.21794   0.32051 -2.300
## treatment.inha                 -0.41856   0.65799  0.13396   0.15524 -2.696
## diras3                         -0.31422   0.73036  0.13913   0.18277 -1.719
## tp53.tp53_mut                  -0.44054   0.64369  0.10476   0.16012 -2.751
## atr.atr_mut                    -0.80931   0.44517  0.51444   0.46068 -1.757
## gata3.gata3_mut                -0.93305   0.39335  0.37258   0.42745 -2.183
## acvr1c                         -0.36373   0.69508  0.14311   0.21455 -1.695
## cul1                           -0.51979   0.59465  0.15983   0.25120 -2.069
## stat5a                         -0.20024   0.81854  0.10493   0.16514 -1.213
## treatment.cul1                 -0.07283   0.92976  0.19944   0.28862 -0.252
## pam50_._claudin.low_subtypeLumA -0.21972   0.80274  0.13744   0.18119 -1.213
## birc6.birc6_mut                 -0.75198   0.47143  0.49710   0.76160 -0.987
## treatment.cyp11a1               -0.46299   0.62940  0.15905   0.19968 -2.319
## tbx3.tbx3_mut                   -1.80113   0.16511  0.49935   0.80544 -2.236
## treatment.mapt                  -0.64863   0.52276  0.23365   0.32369 -2.004
##                                 Pr(>|z|)
## kdm3a.kdm3a_mut                 9.38e-05 ***
## lymph_nodes_examined_positive   9.21e-14 ***
## treatment                        0.09695 .
## integrative_cluster5             0.03257 *
## integrative_cluster3             0.21488
## gsk3b                            0.35575
## treatment.tumor_size             0.06636 .
## prkg1                           6.56e-06 ***
## treatment.e2f8                   0.08093 .
## eif4ebp1                         0.23839
## treatment.zfyve9                 0.35503
## treatment.ush2a                  0.03258 *
## dll3                             0.30854
## treatment.nrarp                  0.06868 .
## tumor_size                       0.56537
## treatment.e2f7                   0.83103
## e2f7                             0.62741
## smad7                            0.27244
## men1                             0.32236
## prkacg                           0.02131 *
## treatment.notch1                 0.32085
## pdgfb                            0.61934
## tumor_stage2                     0.09363 .
## mmp25                            0.07317 .
## treatment.acvr1c                 0.74628
## treatment.srd5a3                 0.08023 .
## treatment.hsd17b7                0.72154
## mlh1                             0.28417
## casp7                            0.12887
## brca2                            0.04200 *
## treatment.rpgr                   0.76477
## treatment.ugt2b17                0.80387
## igf1                             0.52307
## treatment.nrg3                   0.06338 .
## map3k1                           0.63569
## mapt                             0.65902
## e2f1                             0.04298 *
## ugt2b17                          0.02146 *
## treatment.inha                   0.00701 **
## diras3                           0.08558 .
## tp53.tp53_mut                    0.00594 **
## atr.atr_mut                      0.07896 .
## gata3.gata3_mut                  0.02905 *
## acvr1c                           0.09001 .
## cul1                             0.03853 *
## stat5a                           0.22530
## treatment.cul1                   0.80078
## pam50_._claudin.low_subtypeLumA  0.22525
## birc6.birc6_mut                  0.32346
## treatment.cyp11a1                0.02042 *
## tbx3.tbx3_mut                    0.02534 *
## treatment.mapt                   0.04508 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                 exp(coef) exp(-coef) lower .95 upper .95
## kdm3a.kdm3a_mut                   30.9162    0.03235   5.52654  172.9494
## lymph_nodes_examined_positive      2.0629    0.48475   1.70517    2.4957
## treatment                          1.2981    0.77036   0.95392    1.7665
## integrative_cluster5               1.4998    0.66674   1.03420    2.1751
## integrative_cluster3               0.7250    1.37938   0.43609    1.2052
## gsk3b                              1.1641    0.85903   0.84321    1.6071
## treatment.tumor_size               2.0477    0.48835   0.95275    4.4010
## prkg1                              1.7632    0.56714   1.37786    2.2564
## treatment.e2f8                     1.4116    0.70842   0.95849    2.0789
## eif4ebp1                           1.1852    0.84375   0.89357    1.5720
## treatment.zfyve9                   1.1797    0.84767   0.83115    1.6744
## treatment.ush2a                    1.5895    0.62911   1.03918    2.4314
## dll3                               1.1528    0.86747   0.87680    1.5156
## treatment.nrarp                    1.3873    0.72083   0.97523    1.9734
## tumor_size                         0.8113    1.23260   0.39769    1.6551
## treatment.e2f7                     0.9472    1.05576   0.57539    1.5592
## e2f7                               1.0997    0.90933   0.74919    1.6142
## smad7                              1.2033    0.83104   0.86463    1.6747
## men1                               1.1504    0.86927   0.87164    1.5183
## prkacg                             1.3332    0.75006   1.04370    1.7030
## treatment.notch1                   1.2578    0.79504   0.79973    1.9782
## pdgfb                              1.0804    0.92558   0.79631    1.4658
## tumor_stage2                       1.2567    0.79574   0.96211    1.6415
## mmp25                              0.7693    1.29988   0.57743    1.0249
## treatment.acvr1c                   0.9109    1.09783   0.51752    1.6033
## treatment.srd5a3                   0.7491    1.33499   0.54192    1.0354
## treatment.hsd17b7                  1.0732    0.93180   0.72774    1.5826
## mlh1                               0.8496    1.17700   0.63052    1.1448
## casp7                              0.8090    1.23613   0.61534    1.0636
## brca2                              0.7599    1.31591   0.58326    0.9901
## treatment.rpgr                     0.9397    1.06413   0.62545    1.4119
## treatment.ugt2b17                  1.1085    0.90214   0.49174    2.4987
## igf1                               0.8906    1.12280   0.62419    1.2708
## treatment.nrg3                     0.5928    1.68703   0.34127    1.0296
## map3k1                             1.0769    0.92859   0.79261    1.4631
## mapt                               0.9026    1.10793   0.57250    1.4230
## e2f1                               0.7596    1.31655   0.58197    0.9913
## ugt2b17                            0.4785    2.08989   0.25531    0.8968
## treatment.inha                     0.6580    1.51977   0.48538    0.8920
## diras3                             0.7304    1.36919   0.51046    1.0450
## tp53.tp53_mut                      0.6437    1.55355   0.47031    0.8810
## atr.atr_mut                        0.4452    2.24636   0.18046    1.0981
## gata3.gata3_mut                    0.3934    2.54225   0.17019    0.9091
## acvr1c                             0.6951    1.43869   0.45647    1.0584
## cul1                               0.5946    1.68167   0.36344    0.9729
## stat5a                             0.8185    1.22169   0.59221    1.1314
## treatment.cul1                     0.9298    1.07555   0.52808    1.6370
## pam50_._claudin.low_subtypeLumA    0.8027    1.24573   0.56279    1.1450
## birc6.birc6_mut                    0.4714    2.12120   0.10596    2.0975
## treatment.cyp11a1                  0.6294    1.58882   0.42556    0.9309
## tbx3.tbx3_mut                      0.1651    6.05646   0.03406    0.8005
## treatment.mapt                     0.5228    1.91293   0.27719    0.9859
##
## Concordance= 0.808  (se = 0.013 )
## Likelihood ratio test= 694.6  on 52 df,   p=<2e-16
## Wald test            = 502.7  on 52 df,   p=<2e-16
## Score (logrank) test = 834.7  on 52 df,   p=<2e-16,   Robust = 214.2  p=<2e-16
##
##   (Note: the likelihood ratio and score tests assume independence of
##      observations within a cluster, the Wald and robust score tests do not).
```

```
mod = coxmodel
sum.cox = summary(coxmodel)

sum.cox$coefficients
```

```
##                                        coef  exp(coef)   se(coef)  robust se
## kdm3a.kdm3a_mut                  3.43128063 30.9162096 0.87533658 0.87844389
## lymph_nodes_examined_positive    0.72412208  2.0629192 0.07471657 0.09717444
## treatment                        0.26090224  1.2981008 0.11606538 0.15718601
## integrative_cluster5             0.40535206  1.4998304 0.13883499 0.18965755
## integrative_cluster3            -0.32163112  0.7249656 0.21011256 0.25932535
## gsk3b                            0.15195506  1.1641079 0.11760331 0.16454298
## treatment.tumor_size             0.71671626  2.0476980 0.26737407 0.39037447
## prkg1                            0.56714275  1.7632219 0.09176190 0.12582532
## treatment.e2f8                   0.34471991  1.4115945 0.13452821 0.19751076
## eif4ebp1                         0.16989771  1.1851836 0.10304041 0.14409919
## treatment.zfyve9                 0.16525917  1.1796988 0.12904053 0.17868104
## treatment.ush2a                  0.46344797  1.5895453 0.15591579 0.21684668
## dll3                             0.14217304  1.1527761 0.09442464 0.13961913
## treatment.nrarp                  0.32735148  1.3872890 0.13669412 0.17981464
## tumor_size                      -0.20912356  0.8112950 0.24516521 0.36376281
## treatment.e2f7                  -0.05426543  0.9471807 0.18770652 0.25431345
## e2f7                             0.09504816  1.0997118 0.14380892 0.19582417
## smad7                            0.18507795  1.2033122 0.10506580 0.16864305
## men1                             0.14010413  1.1503936 0.10112524 0.14157371
## prkacg                           0.28759578  1.3332183 0.09446375 0.12491001
## treatment.notch1                 0.22936329  1.2577989 0.16225380 0.23104541
## pdgfb                            0.07733314  1.0804019 0.10268740 0.15566633
## tumor_stage2                     0.22848613  1.2566961 0.10079390 0.13628255
## mmp25                           -0.26226930  0.7693038 0.10686085 0.14637605
## treatment.acvr1c                -0.09333213  0.9108909 0.20564650 0.28846578
## treatment.srd5a3                -0.28892172  0.7490708 0.12379355 0.16516027
## treatment.hsd17b7                0.07063676  1.0731913 0.15017965 0.19819419
## mlh1                            -0.16297151  0.8496154 0.10250335 0.15216652
## casp7                           -0.21198567  0.8089763 0.10511610 0.13959532
## brca2                           -0.27452524  0.7599328 0.09850426 0.13500276
## treatment.rpgr                  -0.06215372  0.9397384 0.15394690 0.20772003
## treatment.ugt2b17                0.10298370  1.1084733 0.28363251 0.41469452
## igf1                            -0.11582710  0.8906292 0.13753803 0.18137107
## treatment.nrg3                  -0.52297085  0.5927569 0.18324556 0.28169400
## map3k1                           0.07408287  1.0768960 0.11571953 0.15638240
## mapt                            -0.10249324  0.9025842 0.16761950 0.23227235
## e2f1                            -0.27501543  0.7595604 0.09712506 0.13588075
## ugt2b17                         -0.73710986  0.4784948 0.21794217 0.32050869
## treatment.inha                  -0.41856185  0.6579924 0.13396076 0.15523830
## diras3                          -0.31421748  0.7303602 0.13912730 0.18277159
## tp53.tp53_mut                   -0.44053949  0.6436891 0.10475601 0.16012105
## atr.atr_mut                     -0.80931011  0.4451651 0.51443996 0.46068497
## gata3.gata3_mut                 -0.93305056  0.3933519 0.37257845 0.42744665
## acvr1c                          -0.36373322  0.6950766 0.14311082 0.21454785
## cul1                            -0.51978521  0.5946483 0.15982677 0.25120008
## stat5a                          -0.20023811  0.8185358 0.10493288 0.16513718
## treatment.cul1                  -0.07282955  0.9297593 0.19943786 0.28861813
## pam50_._claudin.low_subtypeLumA -0.21972172  0.8027422 0.13744390 0.18118632
## birc6.birc6_mut                 -0.75198266  0.4714309 0.49710438 0.76160103
## treatment.cyp11a1               -0.46298904  0.6293995 0.15905476 0.19968264
## tbx3.tbx3_mut                   -1.80112564  0.1651129 0.49934880 0.80544247
## treatment.mapt                  -0.64863482  0.5227589 0.23364665 0.32369130
##                                          z     Pr(>|z|)
## kdm3a.kdm3a_mut                  3.9060897 9.380171e-05
## lymph_nodes_examined_positive    7.4517750 9.209268e-14
## treatment                        1.6598312 9.694842e-02
## integrative_cluster5             2.1372841 3.257489e-02
## integrative_cluster3            -1.2402610 2.148789e-01
## gsk3b                            0.9234977 3.557479e-01
## treatment.tumor_size             1.8359711 6.636193e-02
## prkg1                            4.5073818 6.563246e-06
## treatment.e2f8                   1.7453222 8.092880e-02
## eif4ebp1                         1.1790330 2.383850e-01
## treatment.zfyve9                 0.9248836 3.550264e-01
## treatment.ush2a                  2.1372150 3.258051e-02
## dll3                             1.0182920 3.085392e-01
## treatment.nrarp                  1.8204940 6.868381e-02
## tumor_size                      -0.5748899 5.653658e-01
## treatment.e2f7                  -0.2133801 8.310305e-01
## e2f7                             0.4853750 6.274103e-01
## smad7                            1.0974538 2.724431e-01
## men1                             0.9896197 3.223601e-01
## prkacg                           2.3024237 2.131129e-02
## treatment.notch1                 0.9927195 3.208466e-01
## pdgfb                            0.4967878 6.193387e-01
## tumor_stage2                     1.6765619 9.362819e-02
## mmp25                           -1.7917501 7.317300e-02
## treatment.acvr1c                -0.3235466 7.462813e-01
## treatment.srd5a3                -1.7493415 8.023200e-02
## treatment.hsd17b7                0.3564018 7.215397e-01
## mlh1                            -1.0710076 2.841660e-01
## casp7                           -1.5185730 1.288700e-01
## brca2                           -2.0334786 4.200419e-02
## treatment.rpgr                  -0.2992187 7.647732e-01
## treatment.ugt2b17                0.2483363 8.038742e-01
## igf1                            -0.6386195 5.230705e-01
## treatment.nrg3                  -1.8565211 6.337933e-02
## map3k1                           0.4737290 6.356932e-01
## mapt                            -0.4412632 6.590225e-01
## e2f1                            -2.0239470 4.297560e-02
## ugt2b17                         -2.2998124 2.145885e-02
## treatment.inha                  -2.6962538 7.012422e-03
## diras3                          -1.7191812 8.558139e-02
## tp53.tp53_mut                   -2.7512903 5.936102e-03
## atr.atr_mut                     -1.7567539 7.895976e-02
## gata3.gata3_mut                 -2.1828468 2.904709e-02
## acvr1c                          -1.6953478 9.000947e-02
## cul1                            -2.0692080 3.852658e-02
## stat5a                          -1.2125562 2.252995e-01
## treatment.cul1                  -0.2523388 8.007792e-01
## pam50_._claudin.low_subtypeLumA -1.2126838 2.252507e-01
## birc6.birc6_mut                 -0.9873709 3.234609e-01
## treatment.cyp11a1               -2.3186244 2.041541e-02
## tbx3.tbx3_mut                   -2.2361940 2.533906e-02
## treatment.mapt                  -2.0038686 4.508414e-02
```

```
sum.cox$conf.int
```

```
##                                  exp(coef) exp(-coef) lower .95   upper .95
## kdm3a.kdm3a_mut                 30.9162096 0.03234549 5.5265416 172.9493927
## lymph_nodes_examined_positive    2.0629192 0.48474996 1.7051681   2.4957280
## treatment                        1.2981008 0.77035622 0.9539188   1.7664664
## integrative_cluster5             1.4998304 0.66674204 1.0342020   2.1750987
## integrative_cluster3             0.7249656 1.37937586 0.4360931   1.2051900
## gsk3b                            1.1641079 0.85902688 0.8432066   1.6071355
## treatment.tumor_size             2.0476980 0.48835325 0.9527491   4.4010193
## prkg1                            1.7632219 0.56714360 1.3778574   2.2563666
## treatment.e2f8                   1.4115945 0.70841875 0.9584920   2.0788895
## eif4ebp1                         1.1851836 0.84375112 0.8935692   1.5719658
## treatment.zfyve9                 1.1796988 0.84767399 0.8311465   1.6744212
## treatment.ush2a                  1.5895453 0.62911074 1.0391845   2.4313816
## dll3                             1.1527761 0.86747113 0.8768008   1.5156153
## treatment.nrarp                  1.3872890 0.72083034 0.9752331   1.9734469
## tumor_size                       0.8112950 1.23259728 0.3976888   1.6550620
## treatment.e2f7                   0.9471807 1.05576480 0.5753880   1.5592108
## e2f7                             1.0997118 0.90932914 0.7491919   1.6142274
## smad7                            1.2033122 0.83103950 0.8646276   1.6746635
## men1                             1.1503936 0.86926772 0.8716431   1.5182882
## prkacg                           1.3332183 0.75006472 1.0437048   1.7030400
## treatment.notch1                 1.2577989 0.79503965 0.7997330   1.9782329
## pdgfb                            1.0804019 0.92558145 0.7963096   1.4658475
## tumor_stage2                     1.2566961 0.79573733 0.9621136   1.6414747
## mmp25                            0.7693038 1.29987656 0.5774340   1.0249281
## treatment.acvr1c                 0.9108909 1.09782629 0.5175161   1.6032782
## treatment.srd5a3                 0.7490708 1.33498722 0.5419238   1.0353986
## treatment.hsd17b7                1.0731913 0.93180029 0.7277362   1.5826335
## mlh1                             0.8496154 1.17700315 0.6305187   1.1448452
## casp7                            0.8089763 1.23613018 0.6153356   1.0635540
## brca2                            0.7599328 1.31590578 0.5832579   0.9901244
## treatment.rpgr                   0.9397384 1.06412591 0.6254540   1.4119477
## treatment.ugt2b17                1.1084733 0.90214168 0.4917412   2.4986986
## igf1                             0.8906292 1.12280173 0.6241854   1.2708090
## treatment.nrg3                   0.5927569 1.68703213 0.3412701   1.0295680
## map3k1                          1.0768960 0.92859474 0.7926124   1.4631428
## mapt                            0.9025842 1.10792982 0.5725022   1.4229785
## e2f1                            0.7595604 1.31655099 0.5819698   0.9913436
## ugt2b17                         0.4784948 2.08988670 0.2553053   0.8967982
## treatment.inha                  0.6579924 1.51977432 0.4853799   0.8919900
## diras3                          0.7303602 1.36918747 0.5104599   1.0449910
## tp53.tp53_mut                   0.6436891 1.55354512 0.4703063   0.8809909
## atr.atr_mut                     0.4451651 2.24635772 0.1804618   1.0981378
## gata3.gata3_mut                 0.3933519 2.54225266 0.1701916   0.9091270
## acvr1c                          0.6950766 1.43869036 0.4564668   1.0584154
## cul1                            0.5946483 1.68166640 0.3634446   0.9729310
## stat5a                          0.8185358 1.22169363 0.5922058   1.1313649
## treatment.cul1                  0.9297593 1.07554720 0.5280783   1.6369775
## pam50_._claudin.low_subtypeLumA 0.8027422 1.24573001 0.5627947   1.1449912
## birc6.birc6_mut                 0.4714309 2.12120147 0.1059600   2.0974618
## treatment.cyp11a1               0.6293995 1.58881593 0.4255555   0.9308862
## tbx3.tbx3_mut                   0.1651129 6.05646100 0.0340555   0.8005251
## treatment.mapt                  0.5227589 1.91292756 0.2771883   0.9858890
```

```
coefs = tibble(coef = rownames(sum.cox$coefficients), estimate = unname(sum.cox$coefficients[,1]), exp_
               robust_se = unname(sum.cox$coefficients[,4]), pval = unname(sum.cox$coefficients[,6]))

xtable(coefs, caption = "...", digits = 3, display = c('g','g', 'g', 'g', 'g', 'g'))
```

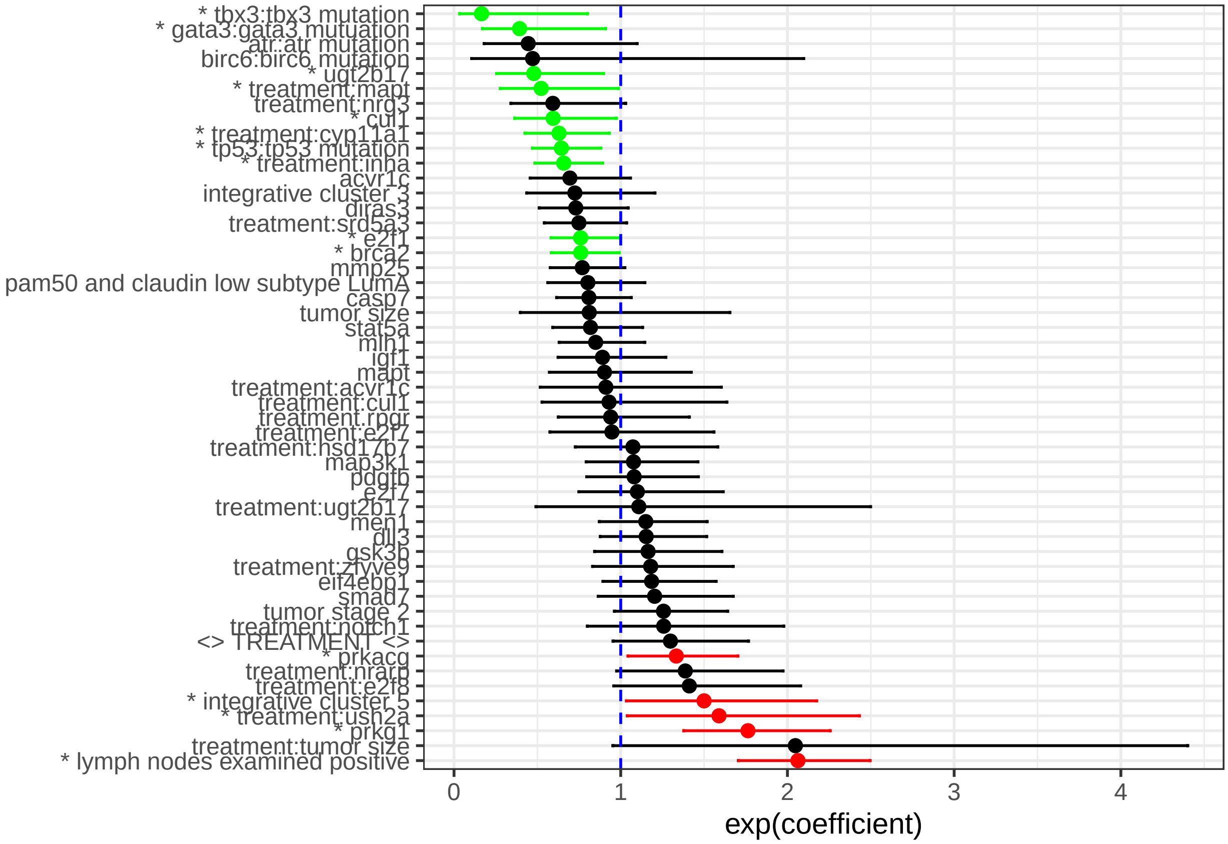
```
## % latex table generated in R 4.1.3 by xtable 1.8-4 package
## % Sat Apr 29 12:52:20 2023
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlrrrr}
##   \hline
##  & coef & estimate & exp\_estimate & robust\_se & pval \\
##   \hline
## 1 & kdm3a.kdm3a\_mut & 3.43 & 30.9 & 0.878 & 9.38e-05 \\
##   2 & lymph\_nodes\_examined\_positive & 0.724 & 2.06 & 0.0972 & 9.21e-14 \\
##   3 & treatment & 0.261 &  1.3 & 0.157 & 0.0969 \\
##   4 & integrative\_cluster5 & 0.405 &  1.5 & 0.19 & 0.0326 \\
##   5 & integrative\_cluster3 & -0.322 & 0.725 & 0.259 & 0.215 \\
##   6 & gsk3b & 0.152 & 1.16 & 0.165 & 0.356 \\
##   7 & treatment.tumor\_size & 0.717 & 2.05 & 0.39 & 0.0664 \\
##   8 & prkg1 & 0.567 & 1.76 & 0.126 & 6.56e-06 \\
##   9 & treatment.e2f8 & 0.345 & 1.41 & 0.198 & 0.0809 \\
##   10 & eif4ebp1 & 0.17 & 1.19 & 0.144 & 0.238 \\
##   11 & treatment.zfyve9 & 0.165 & 1.18 & 0.179 & 0.355 \\
##   12 & treatment.ush2a & 0.463 & 1.59 & 0.217 & 0.0326 \\
##   13 & dll3 & 0.142 & 1.15 & 0.14 & 0.309 \\
##   14 & treatment.nrarp & 0.327 & 1.39 & 0.18 & 0.0687 \\
##   15 & tumor\_size & -0.209 & 0.811 & 0.364 & 0.565 \\
##   16 & treatment.e2f7 & -0.0543 & 0.947 & 0.254 & 0.831 \\
##   17 & e2f7 & 0.095 &  1.1 & 0.196 & 0.627 \\
##   18 & smad7 & 0.185 &  1.2 & 0.169 & 0.272 \\
##   19 & men1 & 0.14 & 1.15 & 0.142 & 0.322 \\
##   20 & prkacg & 0.288 & 1.33 & 0.125 & 0.0213 \\
##   21 & treatment.notch1 & 0.229 & 1.26 & 0.231 & 0.321 \\
##   22 & pdgfb & 0.0773 & 1.08 & 0.156 & 0.619 \\
##   23 & tumor\_stage2 & 0.228 & 1.26 & 0.136 & 0.0936 \\
##    24 & mmp25 & -0.262 & 0.769 & 0.146 & 0.0732 \\
##    25 & treatment.acvr1c & -0.0933 & 0.911 & 0.288 & 0.746 \\
##    26 & treatment.srd5a3 & -0.289 & 0.749 & 0.165 & 0.0802 \\
##    27 & treatment.hsd17b7 & 0.0706 & 1.07 & 0.198 & 0.722 \\
##    28 & mlh1 & -0.163 & 0.85 & 0.152 & 0.284 \\
##    29 & casp7 & -0.212 & 0.809 & 0.14 & 0.129 \\
##    30 & brca2 & -0.275 & 0.76 & 0.135 & 0.042 \\
##    31 & treatment.rpgr & -0.0622 & 0.94 & 0.208 & 0.765 \\
##    32 & treatment.ugt2b17 & 0.103 & 1.11 & 0.415 & 0.804 \\
##    33 & igf1 & -0.116 & 0.891 & 0.181 & 0.523 \\
##    34 & treatment.nrg3 & -0.523 & 0.593 & 0.282 & 0.0634 \\
##    35 & map3k1 & 0.0741 & 1.08 & 0.156 & 0.636 \\
##    36 & mapt & -0.102 & 0.903 & 0.232 & 0.659 \\
##    37 & e2f1 & -0.275 & 0.76 & 0.136 & 0.043 \\
##    38 & ugt2b17 & -0.737 & 0.478 & 0.321 & 0.0215 \\
##    39 & treatment.inha & -0.419 & 0.658 & 0.155 & 0.00701 \\
##    40 & diras3 & -0.314 & 0.73 & 0.183 & 0.0856 \\
##    41 & tp53.tp53\_mut & -0.441 & 0.644 & 0.16 & 0.00594 \\
##    42 & atr.atr\_mut & -0.809 & 0.445 & 0.461 & 0.079 \\
##    43 & gata3.gata3\_mut & -0.933 & 0.393 & 0.427 & 0.029 \\
##    44 & acvr1c & -0.364 & 0.695 & 0.215 & 0.09 \\
##    45 & cul1 & -0.52 & 0.595 & 0.251 & 0.0385 \\
##    46 & stat5a & -0.2 & 0.819 & 0.165 & 0.225 \\
##    47 & treatment.cul1 & -0.0728 & 0.93 & 0.289 & 0.801 \\
##    48 & pam50\_.\_claudin.low\_subtypeLumA & -0.22 & 0.803 & 0.181 & 0.225 \\
##    49 & birc6.birc6\_mut & -0.752 & 0.471 & 0.762 & 0.323 \\
##    50 & treatment.cyp11a1 & -0.463 & 0.629 &  0.2 & 0.0204 \\
##    51 & tbx3.tbx3\_mut & -1.8 & 0.165 & 0.805 & 0.0253 \\
##    52 & treatment.mapt & -0.649 & 0.523 & 0.324 & 0.0451 \\
##     \hline
## \end{tabular}
## \caption{...}
## \end{table}
```

```r
CI_tabl = sum.cox$conf.int[2:52,c(1,3,4)]
colnames(CI_tabl) = c('Estimate', 'Lower CI', 'Upper CI')

#rownames(CI_tabl)
#coefs
```

```r
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment'] = '<> TREATMENT <>'
rownames(CI_tabl)[rownames(CI_tabl) == 'tbx3.tbx3_mut'] = '* tbx3:tbx3 mutation'
rownames(CI_tabl)[rownames(CI_tabl) == 'gata3.gata3_mut'] = '* gata3:gata3 mutuation'
rownames(CI_tabl)[rownames(CI_tabl) == 'ugt2b17'] = '* ugt2b17'
rownames(CI_tabl)[rownames(CI_tabl) == 'cul1'] = '* cul1'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.cyp11a1'] = '* treatment:cyp11a1'
rownames(CI_tabl)[rownames(CI_tabl) == 'tp53.tp53_mut'] = '* tp53:tp53 mutation'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.inha'] = '* treatment:inha'
rownames(CI_tabl)[rownames(CI_tabl) == 'e2f1'] = '* e2f1'
rownames(CI_tabl)[rownames(CI_tabl) == 'brca2'] = '* brca2'
rownames(CI_tabl)[rownames(CI_tabl) == 'prkacg'] = '* prkacg'
rownames(CI_tabl)[rownames(CI_tabl) == 'integrative_cluster5'] = '* integrative cluster 5'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.ush2a'] = '* treatment:ush2a'
rownames(CI_tabl)[rownames(CI_tabl) == 'prkg1'] = '* prkg1'
rownames(CI_tabl)[rownames(CI_tabl) == 'lymph_nodes_examined_positive'] = '* lymph nodes examined posit
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.mapt'] = '* treatment:mapt'

########################
rownames(CI_tabl)[rownames(CI_tabl) == 'integrative_cluster3'] = 'integrative cluster 3'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.tumor_size'] = 'treatment:tumor size'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.e2f8'] = 'treatment:e2f8'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.zfyve9'] = 'treatment:zfyve9'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.nrarp'] = 'treatment:nrarp'
rownames(CI_tabl)[rownames(CI_tabl) == 'tumor_size'] = 'tumor size'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.e2f7'] = 'treatment:e2f7'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.notch1'] = 'treatment:notch1'
rownames(CI_tabl)[rownames(CI_tabl) == 'tumor_stage2'] = 'tumor stage 2'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.acvr1c'] = 'treatment:acvr1c'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.srd5a3'] = 'treatment:srd5a3'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.hsd17b7'] = 'treatment:hsd17b7'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.ugt2b17'] = 'treatment:ugt2b17'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.nrg3'] = 'treatment:nrg3'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.inha'] = 'treatment:inha'
rownames(CI_tabl)[rownames(CI_tabl) == 'atr.atr_mut'] = 'atr:atr mutation'
rownames(CI_tabl)[rownames(CI_tabl) == 'treatment.cul1'] = 'treatment:cul1'

rownames(CI_tabl)[rownames(CI_tabl) == 'pam50_._claudin.low_subtypeLumA'] = 'pam50 and claudin low subty
rownames(CI_tabl)[rownames(CI_tabl) == 'birc6.birc6_mut'] = 'birc6:birc6 mutation'

CI_tabl_df <- as.data.frame(CI_tabl)
CI_tabl_df$Covariate <- rownames(CI_tabl)

# ggplot(data = CI_tabl_df, aes(x = Covariate, y = Estimate)) +
#   geom_point(size = 3) +
#   geom_errorbar(aes(ymin = `Lower CI`, ymax = `Upper CI`), width = 0.2) +
#   labs(x = "Covariate", y = "Estimate") +
#   theme_bw() +
#   theme(axis.text.x = element_text(angle = 45, hjust = 1))
#
#
# ggplot(data = CI_tabl_df, aes(x = Covariate, y = Estimate)) +
#   geom_point(size = 3) +
#   geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2) +
#   labs(x = "Covariate", y = "Estimate") +
#   coord_flip() +
#   theme_bw() +
#   theme(axis.text.y = element_text(hjust = 1))
#
#
# ggplot(data = CI_tabl_df, aes(y = Covariate, x = Estimate)) +
#   geom_point(size = 1) +
#   geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2) +
#   labs(y = "Covariate", x = "Estimate") +
#   theme_bw() +
#   theme(axis.text.y = element_text(hjust = 1))
#

#######################
```

```r
# Custom function to determine bar and text color
color_CI <- function(lower, upper, threshold = 1) {
  if (lower > threshold) {
    return("red")
  } else if (upper < threshold) {
    return("green")
  } else {
    return("black")
  }
}

# Apply custom function to the data frame
CI_tabl_df$Color <- mapply(color_CI, CI_tabl_df$`Lower CI`, CI_tabl_df$`Upper CI`)

# # Create the plot
# ggplot(data = CI_tabl_df, aes(y = Covariate, x = Estimate, color = Color)) +
#   geom_point(size = 1) +
#   geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2, color = CI_tabl_df$Color) +
#   geom_segment(aes(x = `Lower CI`, xend = `Lower CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_segment(aes(x = `Upper CI`, xend = `Upper CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_vline(xintercept = 1, linetype = "dashed", color = "blue") +
#   labs(y = "Covariate", x = "Estimate") +
#   scale_color_identity() +
#   theme_bw() +
#   theme(axis.text.y = element_text(hjust = 1))
#
# #####################################
#
# # Create a data frame for axis labels
# axis_labels <- data.frame(Covariate = CI_tabl_df$Covariate, Color = CI_tabl_df$Color)
# rownames(axis_labels) <- axis_labels$Covariate
#
# # Create the plot
# ggplot(data = CI_tabl_df, aes(y = Covariate, x = Estimate, color = Color)) +
#   geom_point(size = 1) +
#   geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2, color = CI_tabl_df$Color) +
#   geom_segment(aes(x = `Lower CI`, xend = `Lower CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_segment(aes(x = `Upper CI`, xend = `Upper CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_vline(xintercept = 1, linetype = "dashed", color = "blue") +
#   labs(y = NULL, x = "Estimate") +
#   scale_y_discrete(labels = axis_labels$Covariate) +
#   scale_color_identity() +
#   theme_bw() +
#   theme(axis.text.y = element_markdown(color = axis_labels$Color))


###########################
#
# # Sort the data frame by increasing point estimates
# CI_tabl_df <- CI_tabl_df %>% arrange(Estimate)
#
# # Update the axis labels data frame
# axis_labels <- data.frame(Covariate = CI_tabl_df$Covariate, Color = CI_tabl_df$Color)
# rownames(axis_labels) <- axis_labels$Covariate
#
# # Create the plot
# ggplot(data = CI_tabl_df, aes(y = Covariate, x = Estimate, color = Color)) +
#   geom_point(size = 3) +
#   geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2, color = CI_tabl_df$Color) +
#   geom_segment(aes(x = `Lower CI`, xend = `Lower CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_segment(aes(x = `Upper CI`, xend = `Upper CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_vline(xintercept = 1, linetype = "dashed", color = "blue") +
#   labs(y = NULL, x = "Estimate") +
#   scale_y_discrete(labels = axis_labels$Covariate) +
#   scale_color_identity() +
#   theme_bw() +
#   theme(axis.text.y = # Sort the data frame by increasing point estimates
# CI_tabl_df <- CI_tabl_df %>% arrange(Estimate)

# Convert the 'Covariate' column to a factor and specify the levels in the desired order
# CI_tabl_df$Covariate <- factor(CI_tabl_df$Covariate, levels = CI_tabl_df$Covariate)
#
# # Update the axis labels data frame
# axis_labels <- data.frame(Covariate = CI_tabl_df$Covariate, Color = CI_tabl_df$Color)
# rownames(axis_labels) <- axis_labels$Covariate
#
# # Create the plot
# ggplot(data = CI_tabl_df, aes(y = Covariate, x = Estimate, color = Color)) +
#   geom_point(size = 3) +
#   geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2, color = CI_tabl_df$Color) +
#   geom_segment(aes(x = `Lower CI`, xend = `Lower CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_segment(aes(x = `Upper CI`, xend = `Upper CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_vline(xintercept = 1, linetype = "dashed", color = "blue") +
#   labs(y = NULL, x = "Estimate") +
#   scale_color_identity() +
#   theme_bw() +
#   theme(axis.text.y = element_markdown(color = axis_labels$Color))
#  element_text(hjust = 1))

##############################################

# # Sort the data frame by increasing point estimates
# CI_tabl_df <- CI_tabl_df %>% arrange(Estimate)
#
# # Convert the 'Covariate' column to a factor and specify the levels in the desired order
# CI_tabl_df$Covariate <- factor(CI_tabl_df$Covariate, levels = CI_tabl_df$Covariate)
#
# # Update the axis labels data frame
# axis_labels <- data.frame(Covariate = CI_tabl_df$Covariate, Color = CI_tabl_df$Color)
# rownames(axis_labels) <- axis_labels$Covariate
#
# # Create the plot
# ggplot(data = CI_tabl_df, aes(y = Covariate, x = Estimate, color = Color)) +
#   geom_point(size = 3) +
#   geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2, color = CI_tabl_df$Color) +
#   geom_segment(aes(x = `Lower CI`, xend = `Lower CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
```

```r
#   geom_segment(aes(x = `Upper CI`, xend = `Upper CI`, y = as.numeric(Covariate) - 0.1, yend = as.nume
#   geom_vline(xintercept = 1, linetype = "dashed", color = "blue") +
#   labs(y = NULL, x = "Estimate") +
#   scale_color_identity() +
#   theme_bw() +
#   theme(axis.text.y = # Sort the data frame by increasing point estimates

CI_tabl_df <- CI_tabl_df %>% arrange(-Estimate)

##############################################

# Convert the 'Covariate' column to a factor and specify the levels in the desired order
CI_tabl_df$Covariate <- factor(CI_tabl_df$Covariate, levels = CI_tabl_df$Covariate)

# CI_tabl_df$Covariate[CI_tabl_df$Covariate == 'treatment'] = 'TREATMENT'
# CI_tabl_df$Covariate[CI_tabl_df$Covariate == 'tbx.tbx3_mut'] = '* tbx.tbx3_mut'
# CI_tabl_df$Covariate[CI_tabl_df$Covariate == 'gata.gata3_mut'] = '* gata.gata3_mut'
# CI_tabl_df$Covariate[CI_tabl_df$Covariate == 'ugt2b17'] = '* ugt2b17'


# Update the axis labels data frame
#axis_labels <- data.frame(Covariate = CI_tabl_df$Covariate, Color = CI_tabl_df$Color)
#rownames(axis_labels) <- axis_labels$Covariate

# Create the plot
res_plot = ggplot(data = CI_tabl_df, aes(y = Covariate, x = Estimate, color = Color)) +
  geom_point(size = 2) +
  geom_errorbarh(aes(xmin = `Lower CI`, xmax = `Upper CI`), height = 0.2, color = CI_tabl_df$Color) +
  geom_segment(aes(x = `Lower CI`, xend = `Lower CI`, y = as.numeric(Covariate) - 0.1, yend = as.numeric
  geom_segment(aes(x = `Upper CI`, xend = `Upper CI`, y = as.numeric(Covariate) - 0.1, yend = as.numeric
  geom_vline(xintercept = 1, linetype = "dashed", color = "blue") +
  labs(y = NULL, x = "exp(coefficient)") +
  scale_color_identity() +
  theme_bw() +
  theme(axis.text.y =  element_text(hjust = 1))

res_plot
```



```
ggsave("plots/res_plot.pdf", res_plot, width = 8, height = 6)
```

```
#survfit(formula = coxmodel, newdata = data_matrix)

get.race.unpenalized(data_matrix, coxmodel, W.vec)
```

```
## [1] -15.73215
```

Sensitivity Analysis - Propensity Scores

```
coxmodel_overlap <- coxph(y[e.vec > 0.40 & e.vec < 0.75] ~ kdm3a.kdm3a_mut  + lymph_nodes_examined_posi
                      treatment      +     as.factor(integrative_cluster5) +
                    as.factor(integrative_cluster3) +
                          gsk3b      +      treatment.tumor_size +
                          prkg1     +             treatment.e2f8 +
                       eif4ebp1      +          treatment.zfyve9 +
                treatment.ush2a       +                     dll3 +
                treatment.nrarp        +              tumor_size +
                 treatment.e2f7         +                   e2f7 +
                          smad7          +                  men1 +
                         prkacg           +     treatment.notch1 +
                          pdgfb            +        as.factor(tumor_stage2) +
                          mmp25             +   treatment.acvr1c +
               treatment.srd5a3             +  treatment.hsd17b7 +
                           mlh1             +              casp7 +
                          brca2             +     treatment.rpgr +
              treatment.ugt2b17             +               igf1 +
                 treatment.nrg3             +             map3k1 +
                           mapt             +               e2f1 +
                        ugt2b17 +
                 treatment.inha             +             diras3 +
                  tp53.tp53_mut             +        atr.atr_mut +
                gata3.gata3_mut             +             acvr1c +
                           cul1             +             stat5a +
                 treatment.cul1 + as.factor(pam50_._claudin.low_subtypeLumA) +
                birc6.birc6_mut           +    treatment.cyp11a1 +
                  tbx3.tbx3_mut           +       treatment.mapt, weights = W.vec[e.vec > 0.40 & e.vec
```

```
get.race.unpenalized(data_matrix[e.vec > 0.40 & e.vec < 0.75,], coxmodel_overlap, W.vec[e.vec > 0.40 &
```

```
## [1] -21.59136
```

In interest of saving computational time we run bootstrap beforehand and load the results here. The reader can uncomment the sections of code below if they wish to run the bootstrap.

```
# BOOTSTARP
n = nrow(data_matrix)
S = 1000
boot.RACE.unpenalized = vector(length = S)
ind_matrix = array(rep(NA, n*S), c(S,n))

# set.seed(4)
# for (i in 1:S)
# {
#   cat('Progress: ', i/S, '\n')
#   boot_ind = sample(1:n, size = n, replace = TRUE)
#   data_matrix_boot = x[boot_ind,]
#   y_boot = y[boot_ind,]
#   W_boot = W.vec[boot_ind]
#
#   ind_matrix[i,] = boot_ind
#
#   coxmodel.boot <- coxph(y_boot ~ kdm3a.kdm3a_mut  + lymph_nodes_examined_positive +
#                       treatment      +     as.factor(integrative_cluster5) +
#                     as.factor(integrative_cluster3) +
#                           gsk3b      +      treatment.tumor_size +
#                           prkg1     +             treatment.e2f8 +
#                        eif4ebp1      +          treatment.zfyve9 +
#                 treatment.ush2a       +                     dll3 +
#                 treatment.nrarp        +              tumor_size +
#                  treatment.e2f7         +                   e2f7 +
#                           smad7          +                  men1 +
#                          prkacg           +     treatment.notch1 +
#                           pdgfb            +        as.factor(tumor_stage2) +
#                           mmp25             +   treatment.acvr1c +
#                treatment.srd5a3             +  treatment.hsd17b7 +
#                            mlh1             +              casp7 +
#                           brca2             +     treatment.rpgr +
#               treatment.ugt2b17             +               igf1 +
#                  treatment.nrg3             +             map3k1 +
#                            mapt             +               e2f1 +
#                         ugt2b17 +
#                  treatment.inha            +            diras3 +
#                   tp53.tp53_mut            +        atr.atr_mut +
#                 gata3.gata3_mut            +            acvr1c +
#                            cul1            +            stat5a +
#                  treatment.cul1 + as.factor(pam50_._claudin.low_subtypeLumA) +
#                 birc6.birc6_mut          +    treatment.cyp11a1 +
#                   tbx3.tbx3_mut          +       treatment.mapt, weights = W_boot, data = data_matri
#
#
#   boot.RACE.unpenalized[i] = get.race.unpenalized(data_matrix_boot, coxmodel.boot, W_boot)
# }


#saveRDS(boot.RACE.unpenalized, file = 'boot_RACE_unpenalized.RDS')

boot.RACE.unpenalized = readRDS(file = 'boot_RACE_unpenalized.RDS')
```

```
#boot.RACE.unpenalized
quantile(boot.RACE.unpenalized, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -35.055441   3.424606
```

Diagnostics

Cox-Snell Residuals

```
source("http://myweb.uiowa.edu/pbreheny/7210/f18/notes/fun.R")
sfit <- survfit(coxmodel)
H0 <- -log(sfit$surv)
H <- approxfun(c(0, sfit$time), c(0, H0), method='constant')
e1 <- H(coxmodel$y[,1])*exp(coxmodel$linear.predictors)
e2 <- coxmodel$y[,2]-residuals(coxmodel)
head(e1)
```

```
## [1] 2.342085e-01 4.409510e-02 8.648566e-02 2.494864e-02 8.578785e-05
## [6] 2.084147e-01
```

```
head(e2)
```

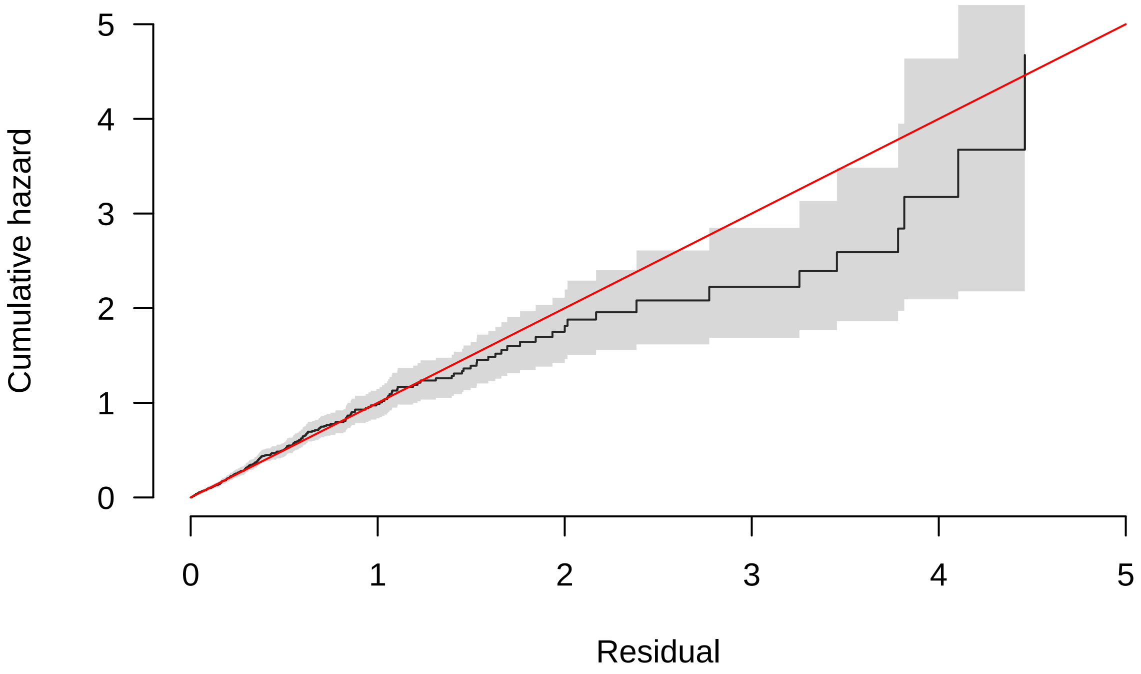
```
##            1            2            3            7           12           13
## 2.342085e-01 4.409510e-02 8.648566e-02 2.494864e-02 8.578785e-05 2.084147e-01
```

```
efit <- survfit(Surv(e1, coxmodel$y[,2])~1)
lim <- c(0,5)
pdf("plots/cox_snell_resid.pdf", height = 3.5, width = 6)
plot(efit, fun='cumhaz', mark.time=FALSE, bty='n', conf.int=FALSE, lwd=1, las=1,
     xlab='Residual', ylab='Cumulative hazard', xlim=lim, ylim=lim)
ciband(efit, fun=function(x) -log(x))
lines(lim, lim, col='red', lwd=1)
dev.off()
```

```
## pdf
##   2
```

```
plot(efit, fun='cumhaz', mark.time=FALSE, bty='n', conf.int=FALSE, lwd=1, las=1,
     xlab='Residual', ylab='Cumulative hazard', xlim=lim, ylim=lim)
```

```
ciband(efit, fun=function(x) -log(x))
lines(lim, lim, col='red', lwd=1)
```


Cumulative hazard
Residual