Survival Analysis for HGG patients

This notebook does Kaplan Meier survival analysis on the following covariates: 1) reported_gender 2) CNS_region, 3) primary_site 4) tumor_descriptor 5) molecular_subtype 6) age (categorized)

Packages and functions Read in set up script.

```
library(survival)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggpubr)
```

```
root_dir <- rprojroot::find_root(rprojroot::has_dir(".git"))
analysis_dir <- file.path(root_dir, "analyses", "splicing_index")
data_dir <- file.path("~/OpenPedCan-analysis", "data")

input_dir <- file.path(analysis_dir, "results")

km_survival_plots_dir <- file.path(analysis_dir, "survival_plots")
if(!dir.exists(km_survival_plots_dir)){
  dir.create(km_survival_plots_dir)
}
```

Set up files and directories

```
histology_df <- readr::read_tsv(file.path(data_dir, "histologies.tsv"))
```

Read in files necessary for analyses

```
## Rows: 36150 Columns: 46

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (41): Kids_First_Biospecimen_ID, sample_id, aliquot_id, Kids_First_Parti...
## dbl  (5): OS_days, age_last_update_days, normal_fraction, tumor_fraction, tu...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sample_list <- readr::read_tsv(file.path(input_dir, "splicing_index.total.hgg_clusters.surv.txt"))
```

```
## Rows: 75 Columns: 2

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): Kids_First_Biospecimen_ID, Level

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
histology_df$PFS_days <- as.numeric(histology_df$PFS_days)
```

Calculate PFS status based on the PFS days and OS days

```
## Warning: NAs introduced by coercion
```

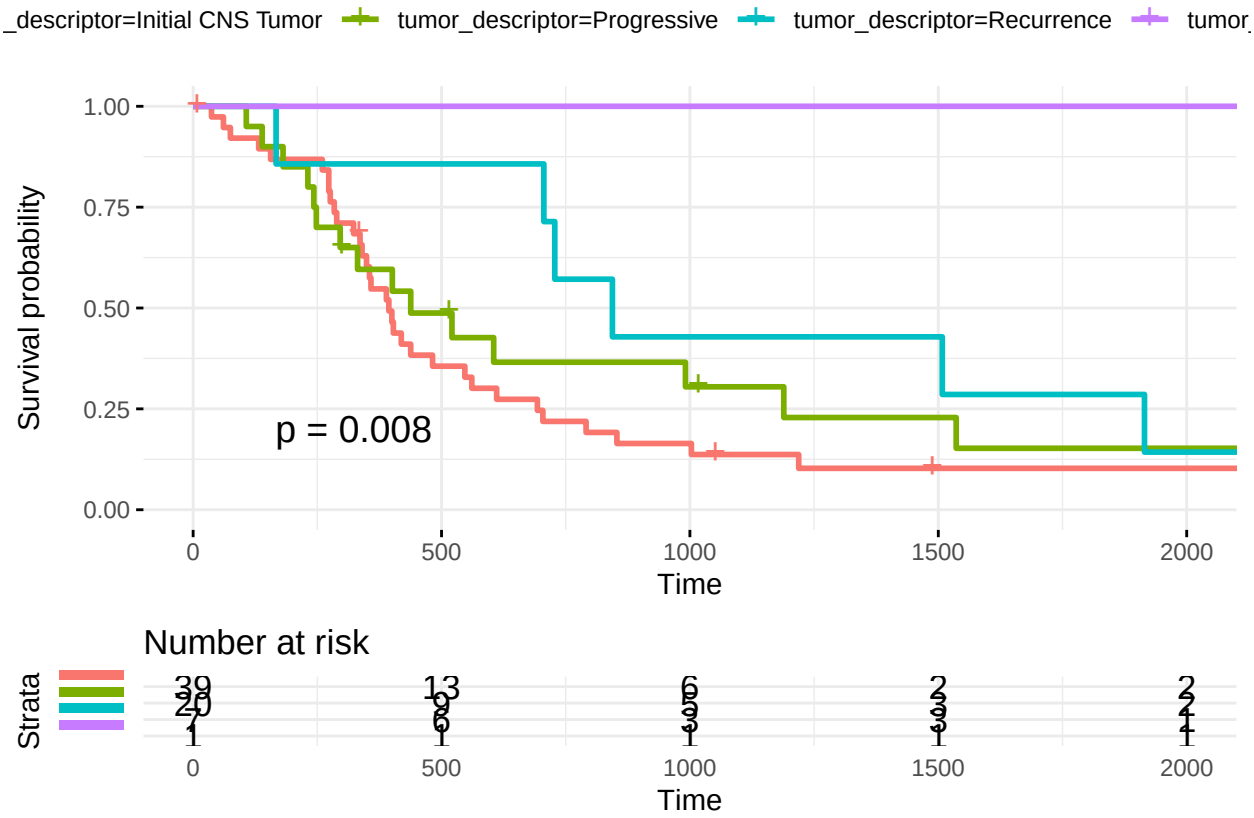
```
histology_df$age_at_diagnosis_days <- as.numeric(histology_df$age_at_diagnosis_days)
```

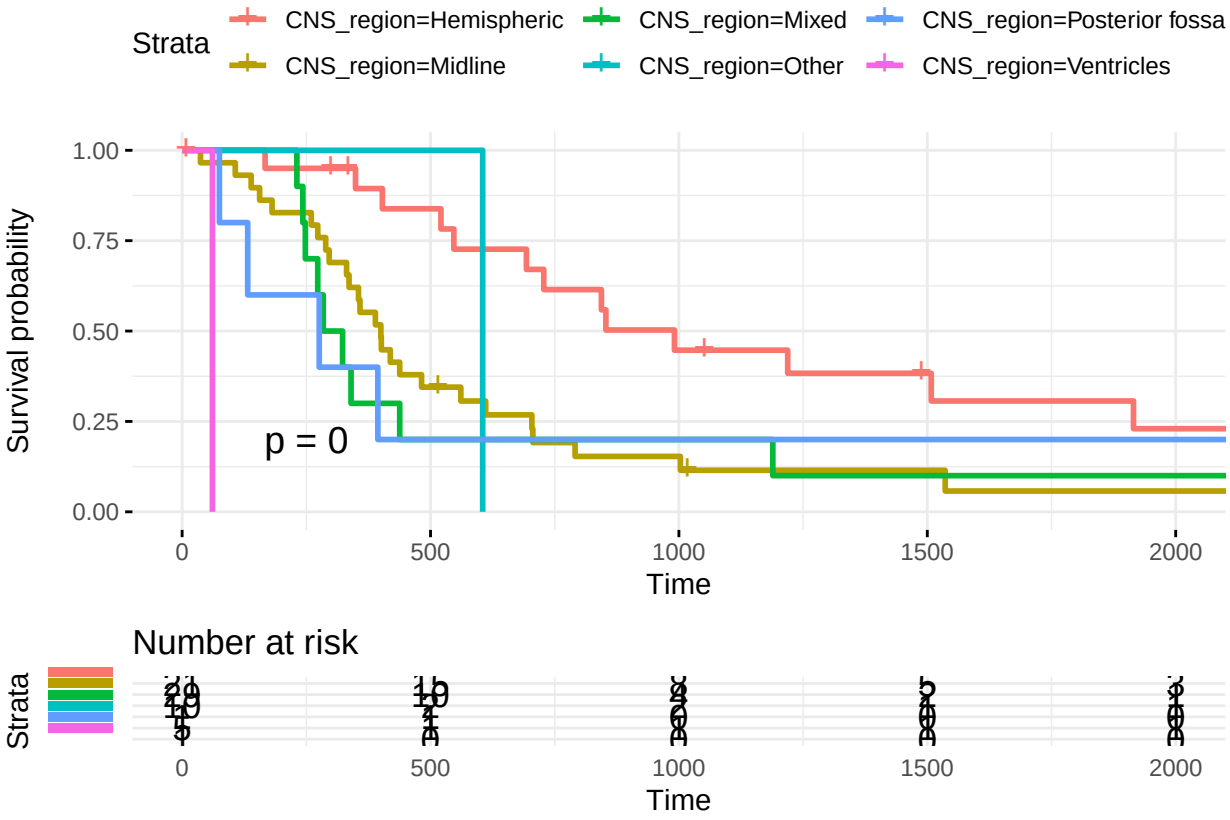
```
## Warning: NAs introduced by coercion
```

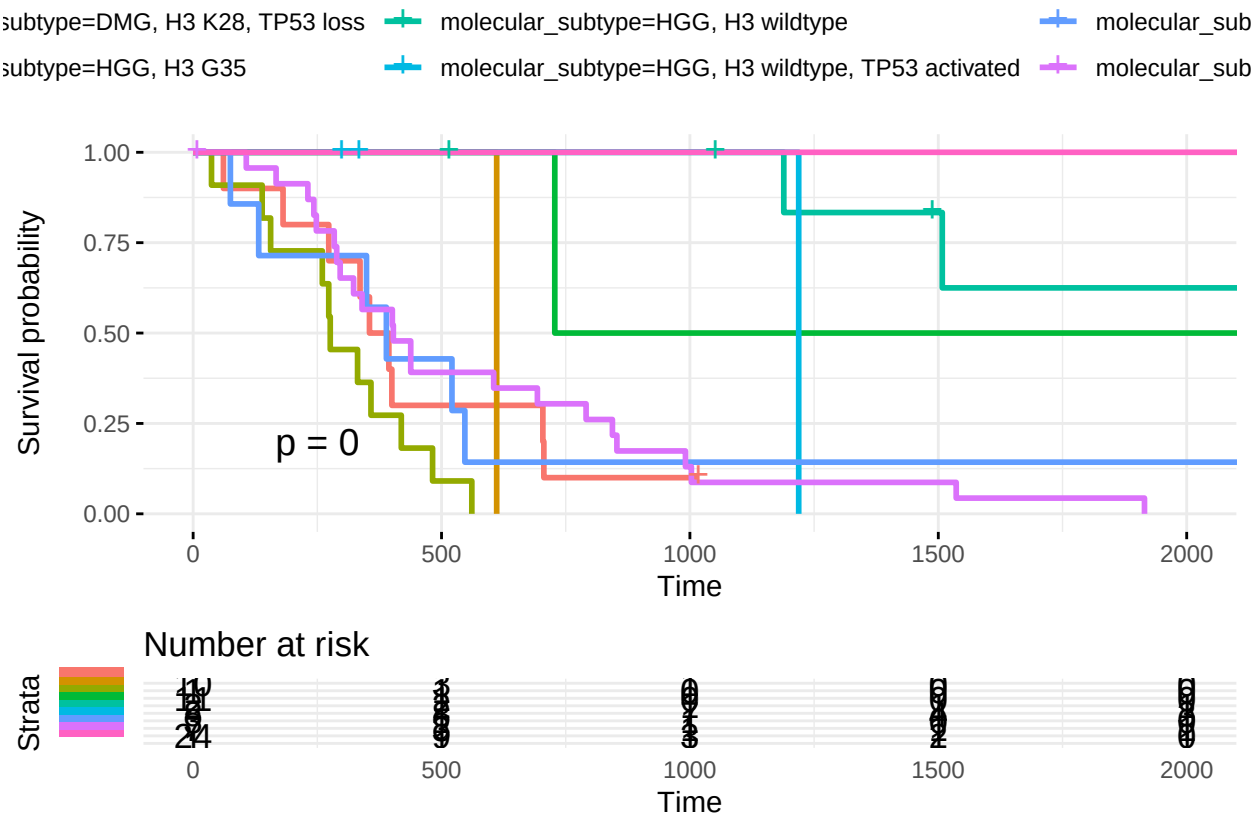
```
# filter to only samples of interest
histology_df <- histology_df %>%
  # keep only samples in the sample list
  dplyr::filter(Kids_First_Biospecimen_ID %in% sample_list$Kids_First_Biospecimen_ID) %>%
  # keep only unique Kids First Participant
  dplyr::distinct(Kids_First_Participant_ID, .keep_all = TRUE) %>%
  dplyr::filter(!is.na(OS_status)) %>%
  dplyr::mutate(os_status_level = case_when(
    OS_status == "LIVING" ~ 0,
    OS_status == "DECEASED" ~ 1)) %>%
  dplyr::mutate(PFS_status = if_else(PFS_days < OS_days, 1, 0)) %>%
  dplyr::mutate(age_group = case_when(
    age_at_diagnosis_days <= 5*365.25 ~ "1-5 years old",
    age_at_diagnosis_days <= 10*365.25 & age_at_diagnosis_days > 5*365.25 ~ "5-10 years old",
    age_at_diagnosis_days <= 15*365.25 & age_at_diagnosis_days > 10*365.25 ~ "10-15 years old",
    age_at_diagnosis_days > 15*365.25 ~ "over 15 years old"
  ))
```

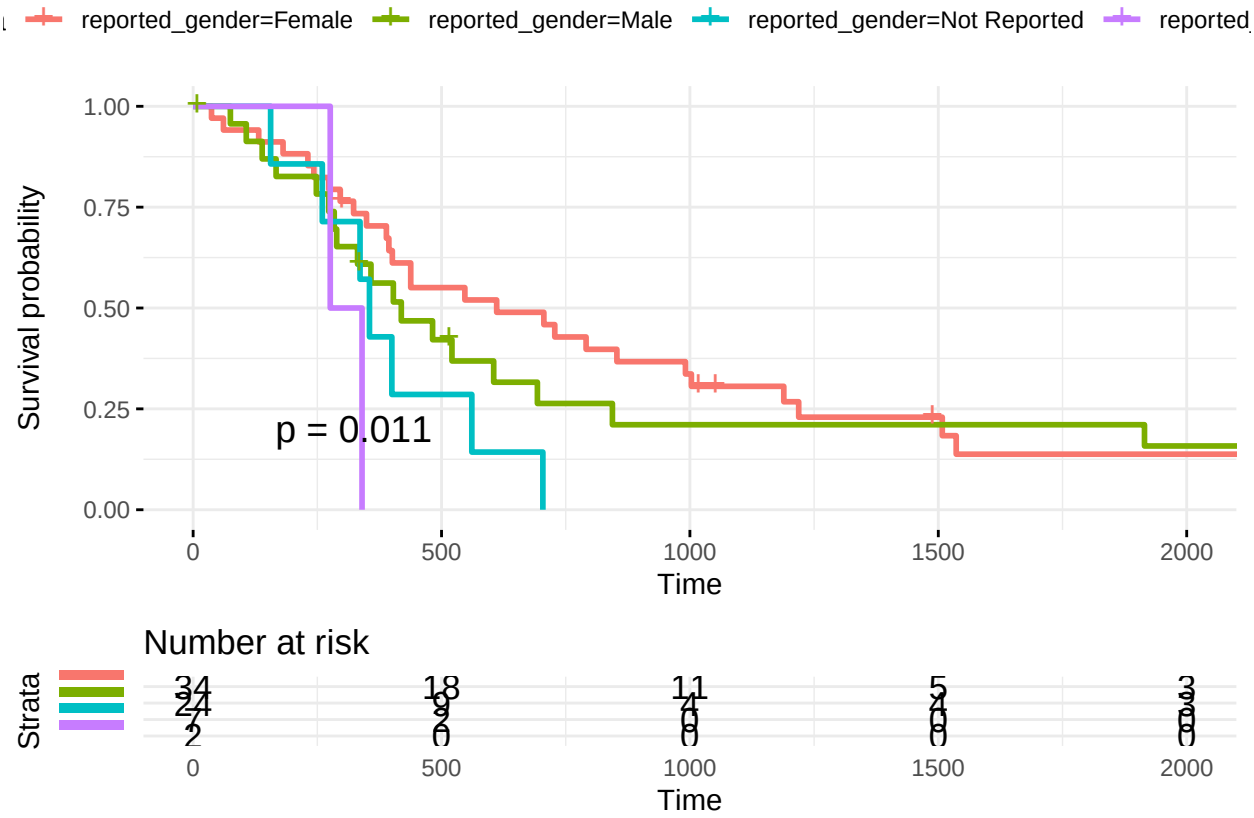
Survival analysis OS

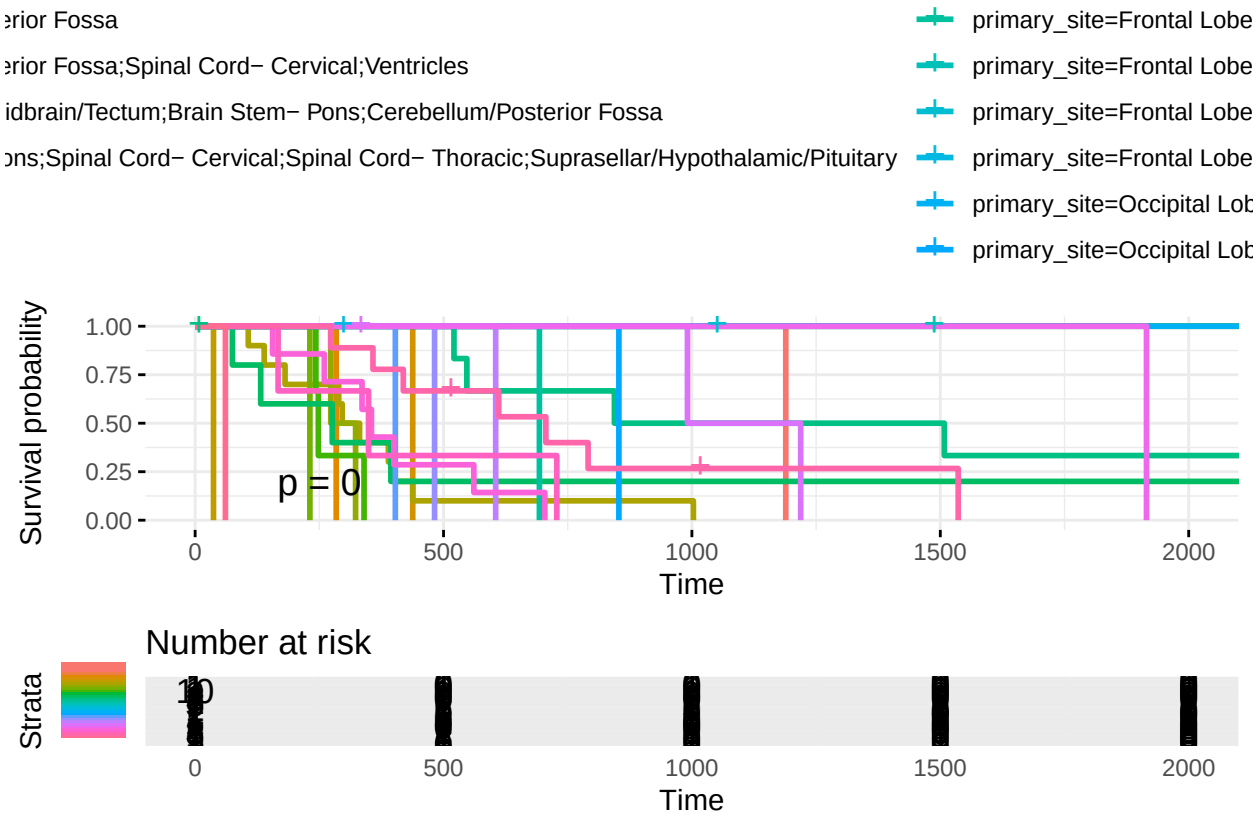
```
for(ind_var in c("tumor_descriptor", "CNS_region", "molecular_subtype",
                 "reported_gender", "primary_site", "age_group")){
  # generate the log-rank model
  fit_ind_var <- survival::survdiff(
    as.formula(paste0("survival::Surv(OS_days, os_status_level) ~ ", ind_var)),
    data = histology_df
  )

  # get the p.values and all statistics for the model
  fit_ind_var$p.value <- round(pchisq(fit_ind_var$chisq, df = 1, lower = FALSE), digits=3)
  fit_ind_var_df <- as.data.frame(fit_ind_var[c("n", "obs", "exp", "chisq", "p.value")])

  fit_ind_var_df <- fit_ind_var_df %>%
    mutate(covaraite = ind_var,
           model = "OS")

  # generate the kaplan-meier model
  kap_fit_ind_var <- survival::survfit(
    as.formula(paste0("survival::Surv(OS_days, os_status_level) ~ ", ind_var)),
    data = histology_df
  )

  # generate and save the survival plot
  surv_plot <- survminer::ggsurvplot(kap_fit_ind_var,
                                     pval = round(fit_ind_var$p.value, digits=3), # use computed pval f
                                     data = histology_df,
                                     risk.table = TRUE,
                                     xlim = c(0, 2000),
                                     break.time.by = 500,
                                     ggtheme = theme_minimal(),
                                     risk.table.y.text.col = TRUE,
                                     risk.table.y.text = FALSE)
  print(surv_plot)
}
```

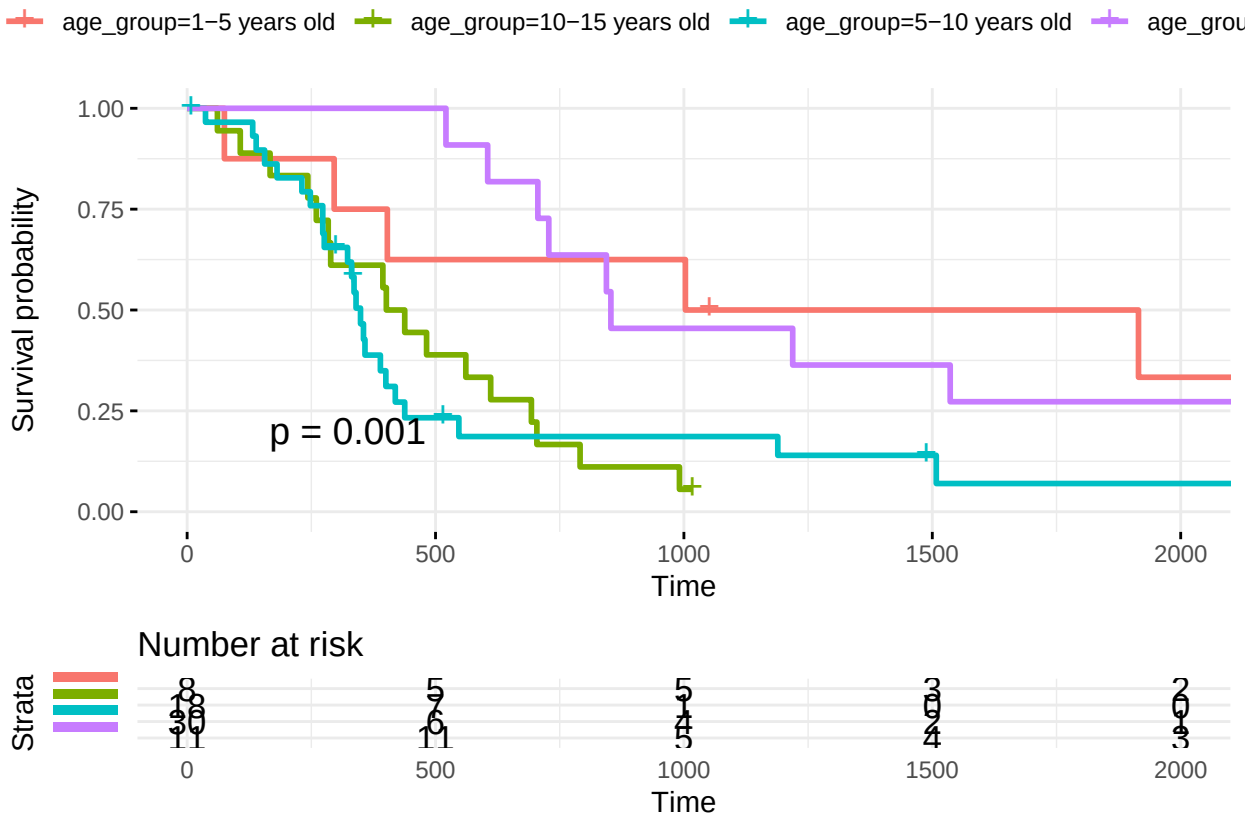

_descriptor=Initial CNS Tumor
tumor_descriptor=Progressive
tumor_descriptor=Recurrence
tumor_
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.008
0
500
1000
1500
2000
Time
Number at risk
Strata
39
13
6
2
2
28
9
5
3
2
20
6
3
3
1
7
1
1
1
1
1
0
500
1000
1500
2000
Time


Strata
CNS_region=Hemispheric
CNS_region=Midline
CNS_region=Mixed
CNS_region=Other
CNS_region=Posterior fossa
CNS_region=Ventricles
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0
0
500
1000
1500
2000
Time
Number at risk
Strata
0
500
1000
1500
2000
Time


subtype=DMG, H3 K28, TP53 loss
molecular_subtype=HGG, H3 wildtype
molecular_sub
subtype=HGG, H3 G35
molecular_subtype=HGG, H3 wildtype, TP53 activated
molecular_sub
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0
0
500
1000
1500
2000
Time
Number at risk
Strata
0
500
1000
1500
2000
Time


reported_gender=Female
reported_gender=Male
reported_gender=Not Reported
reported
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.011
0
500
1000
1500
2000
Time
Number at risk
Strata
34
18
11
5
3
24
9
4
2
1
7
2
0
0
0
2
0
0
0
0
0
500
1000
1500
2000
Time

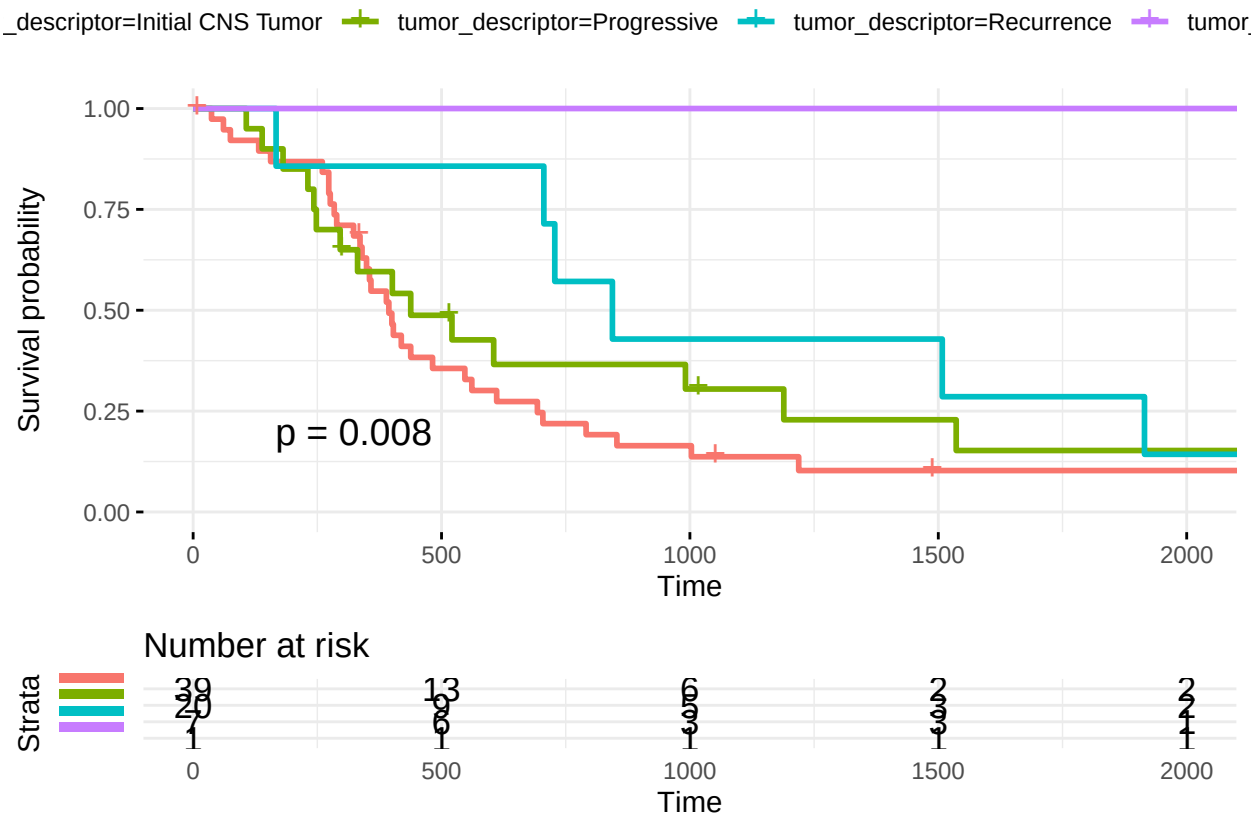

erior Fossa
erior Fossa;Spinal Cord− Cervical;Ventricles
idbrain/Tectum;Brain Stem− Pons;Cerebellum/Posterior Fossa
ons;Spinal Cord− Cervical;Spinal Cord− Thoracic;Suprasellar/Hypothalamic/Pituitary
primary_site=Frontal Lobe
primary_site=Frontal Lobe
primary_site=Frontal Lobe
primary_site=Frontal Lobe
primary_site=Occipital Lob
primary_site=Occipital Lob
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0
0
500
1000
1500
2000
Time
Number at risk
Strata
10
0
500
1000
1500
2000
Time

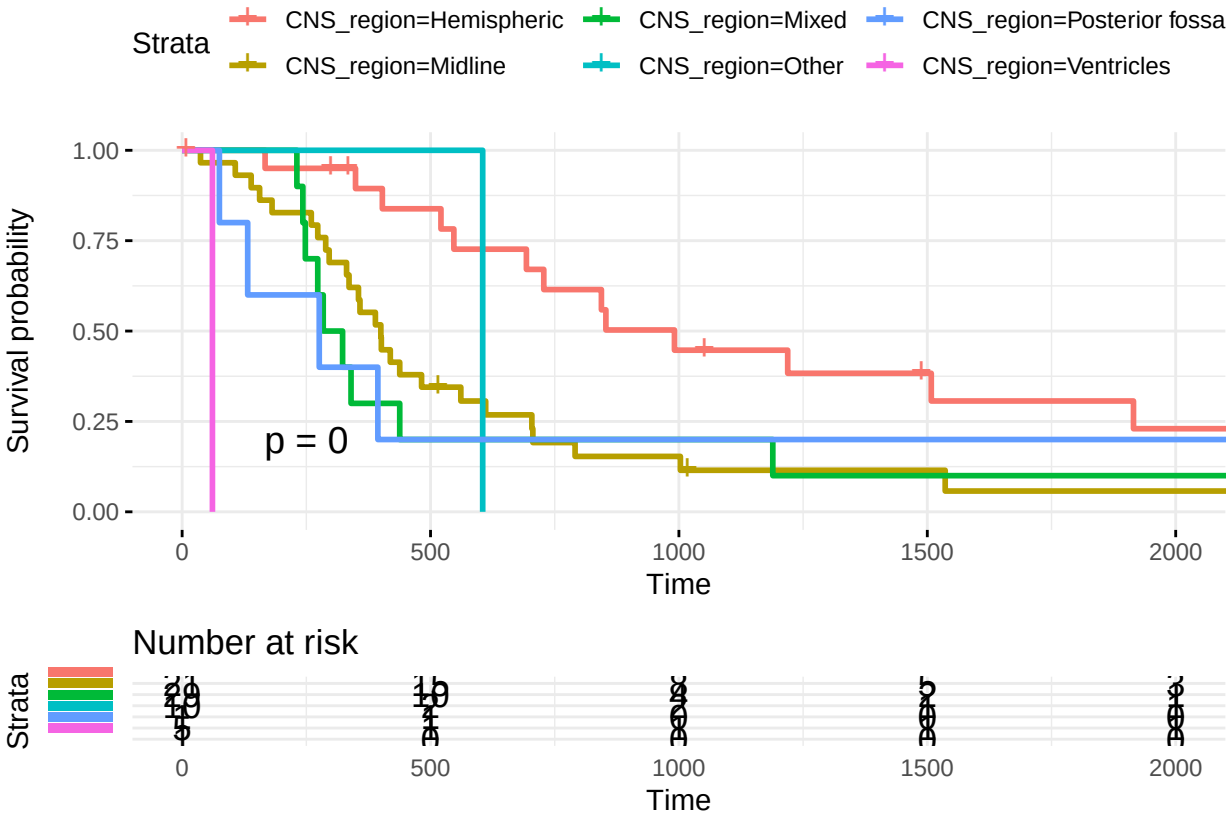


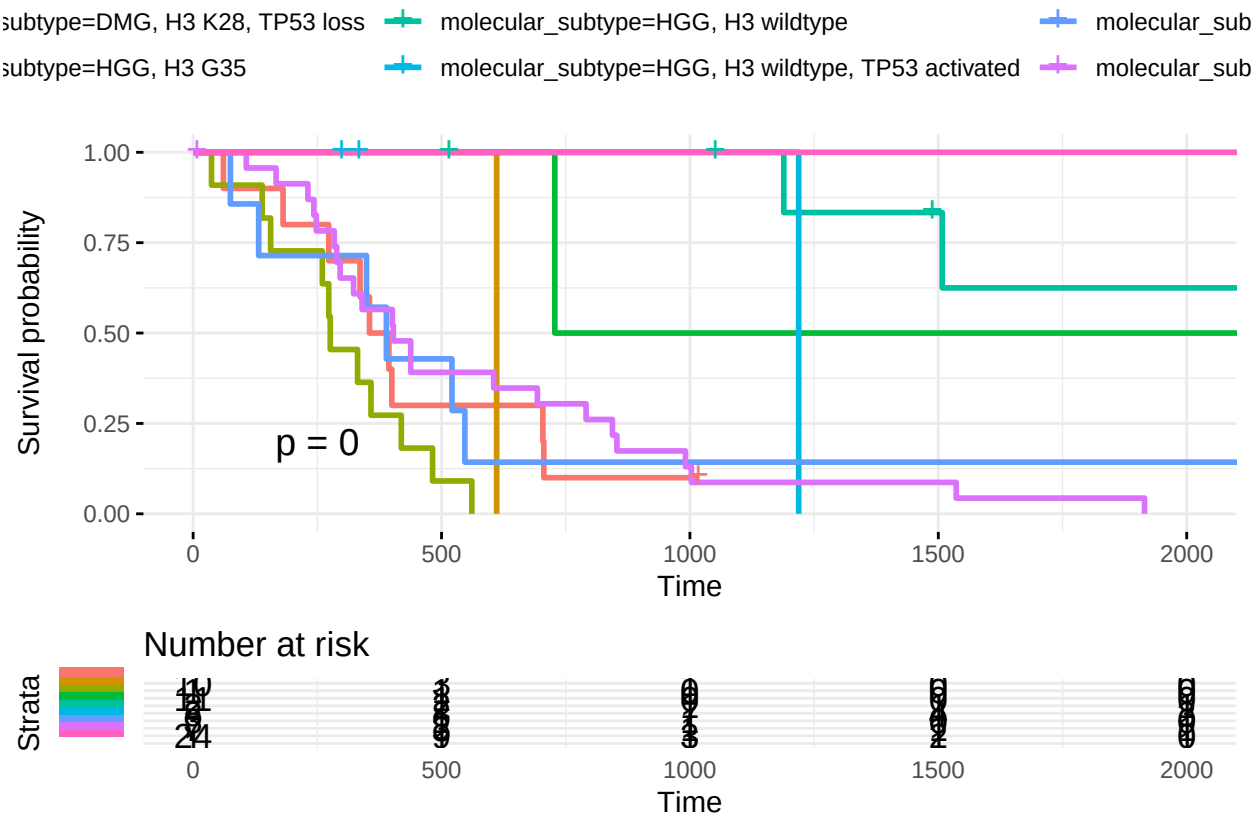
Survival analysis PFS

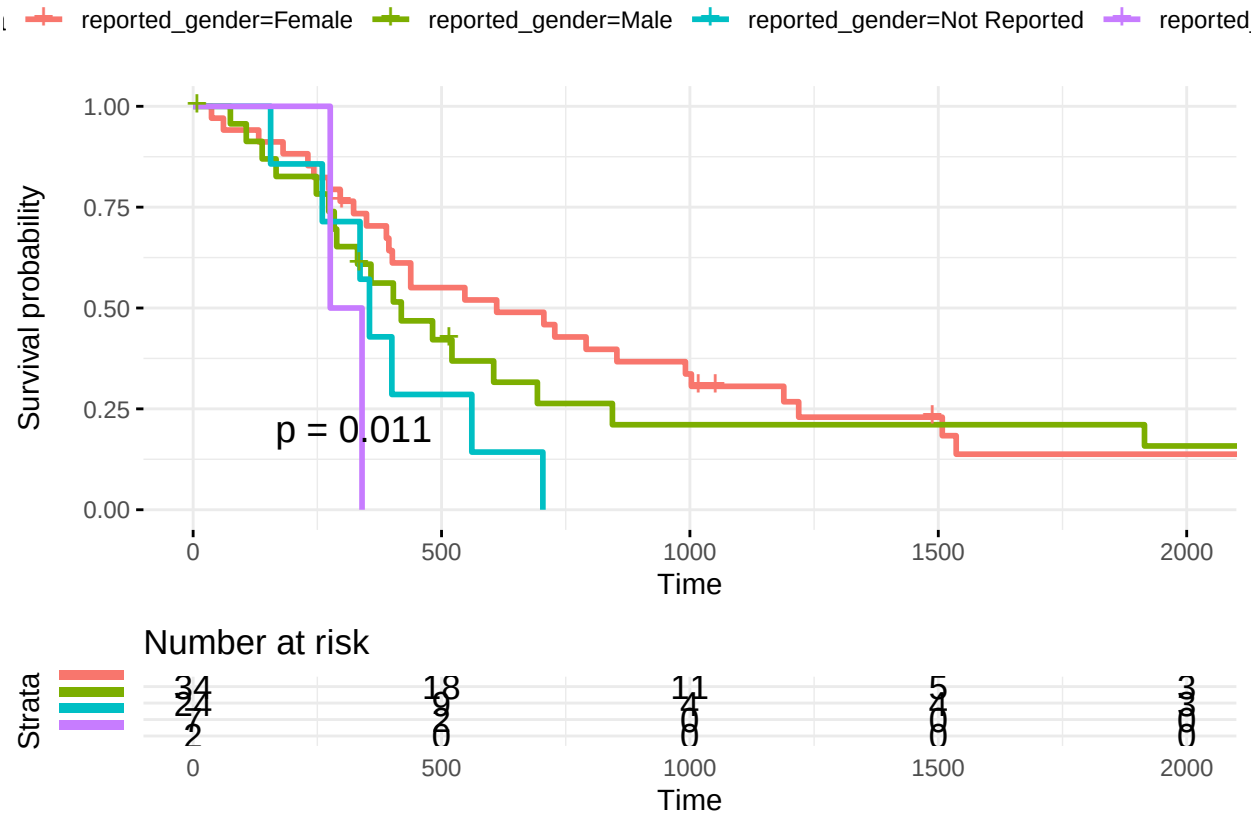
```
for(ind_var in c("tumor_descriptor", "CNS_region", "molecular_subtype",
                 "reported_gender", "primary_site", "age_group")){
    # generate the log-rank model
  fit_ind_var <- survival::survdiff(
    as.formula(paste0("survival::Surv(OS_days, os_status_level) ~ ", ind_var)),
    data = histology_df
  )

  # get the p.values and all statistics for the model
  fit_ind_var$p.value <- round(pchisq(fit_ind_var$chisq, df = 1, lower = FALSE), digits=3)
  fit_ind_var_df <- as.data.frame(fit_ind_var[c("n", "obs", "exp", "chisq", "p.value")])

  fit_ind_var_df <- fit_ind_var_df %>%
    mutate(covaraite = ind_var,
           model = "OS")

  # generate the kaplan-meier model
  kap_fit_ind_var <- survival::survfit(
    as.formula(paste0("survival::Surv(OS_days, os_status_level) ~ ", ind_var)),
    data = histology_df
  )
```

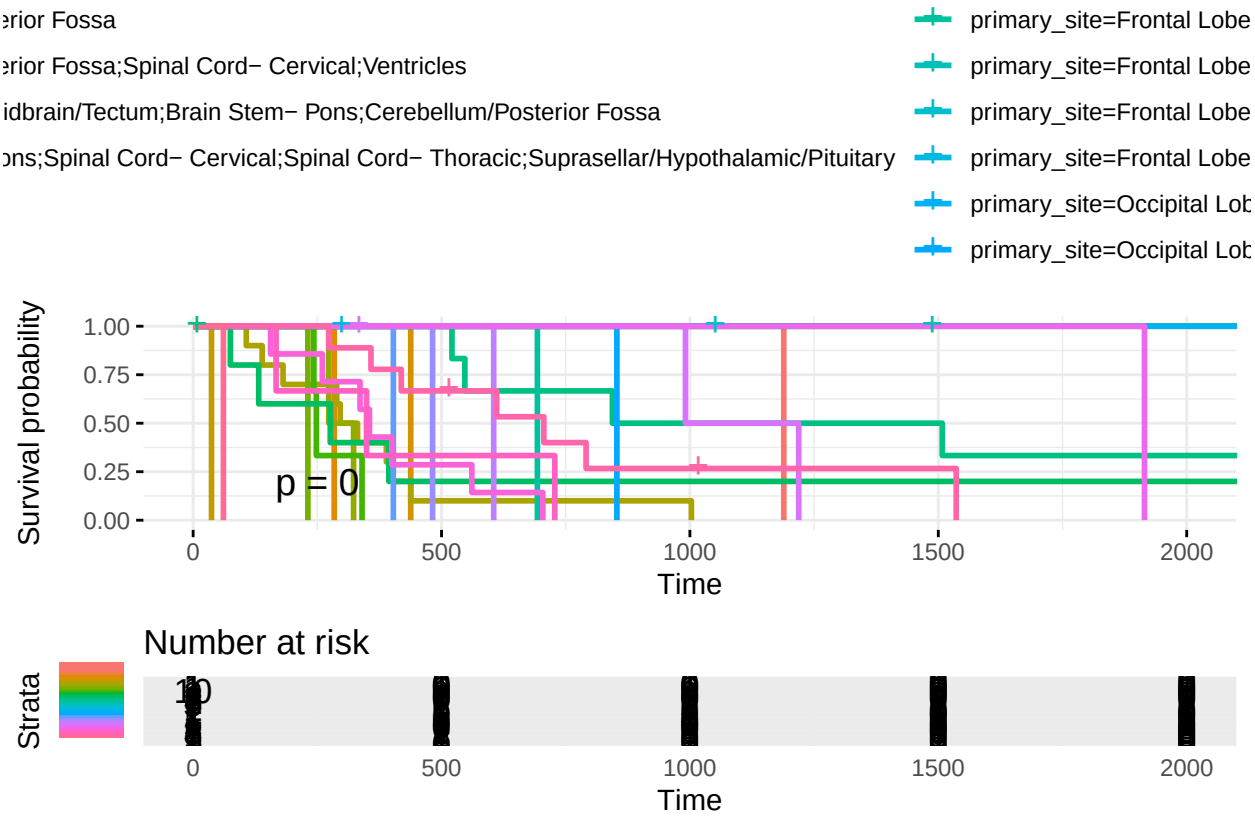
```r
  # generate and save the survival plot
  surv_plot <- survminer::ggsurvplot(kap_fit_ind_var,
                                     pval = round(fit_ind_var$p.value, digits=3), # use computed pval f
                                     data = histology_df,
                                     risk.table = TRUE,
                                     xlim = c(0, 2000),
                                     break.time.by = 500,
                                     ggtheme = theme_minimal(),
                                     risk.table.y.text.col = TRUE,
                                     risk.table.y.text = FALSE)

  print(surv_plot)
}
```

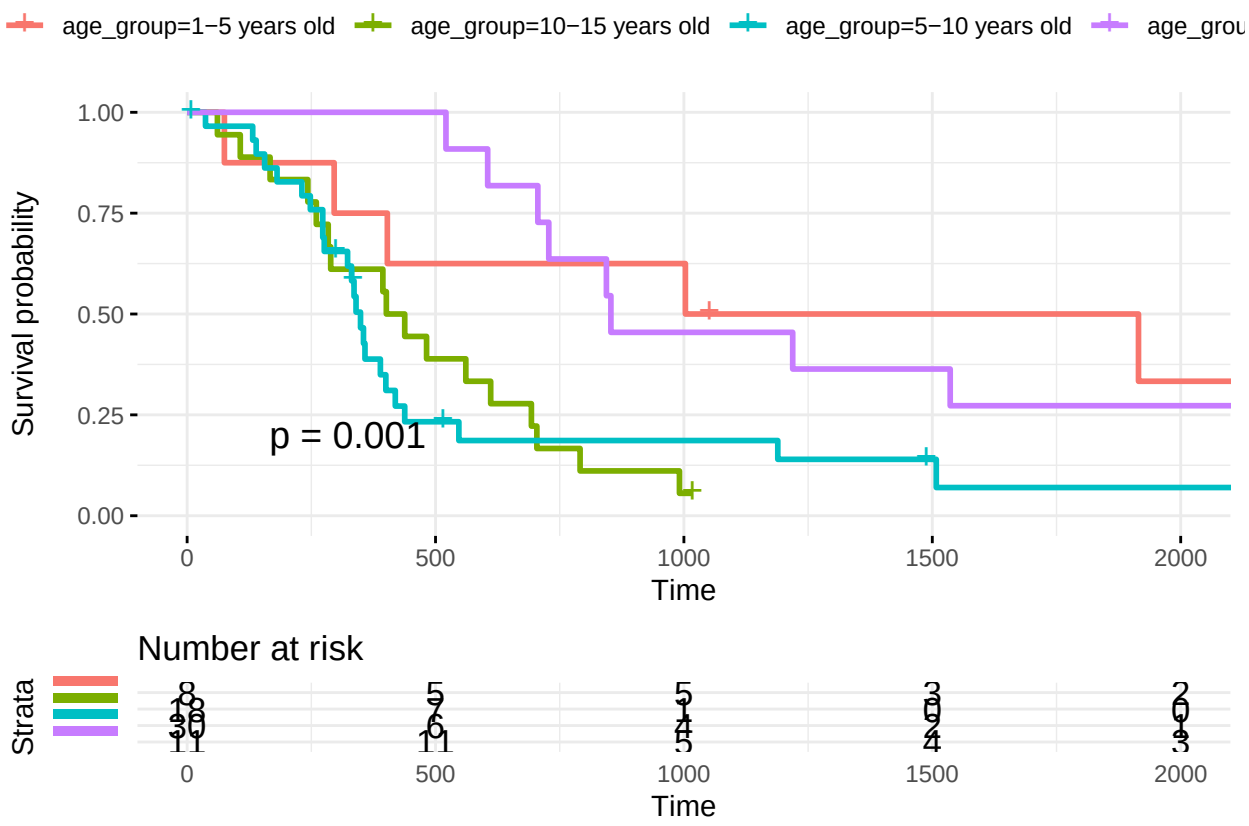



Strata
CNS_region=Hemispheric
CNS_region=Midline
CNS_region=Mixed
CNS_region=Other
CNS_region=Posterior fossa
CNS_region=Ventricles
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0
0
500
1000
1500
2000
Time
Number at risk
Strata
0
500
1000
1500
2000
Time


subtype=DMG, H3 K28, TP53 loss
molecular_subtype=HGG, H3 wildtype
molecular_sub
subtype=HGG, H3 G35
molecular_subtype=HGG, H3 wildtype, TP53 activated
molecular_sub
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0
0
500
1000
1500
2000
Time
Number at risk
Strata
0
500
1000
1500
2000
Time


reported_gender=Female
reported_gender=Male
reported_gender=Not Reported
reported
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.011
0
500
1000
1500
2000
Time
Number at risk
Strata
34
18
11
5
3
24
9
4
2
1
7
2
0
0
0
2
0
0
0
0
0
500
1000
1500
2000
Time


erior Fossa
erior Fossa;Spinal Cord− Cervical;Ventricles
idbrain/Tectum;Brain Stem− Pons;Cerebellum/Posterior Fossa
ons;Spinal Cord− Cervical;Spinal Cord− Thoracic;Suprasellar/Hypothalamic/Pituitary
primary_site=Frontal Lobe
primary_site=Frontal Lobe
primary_site=Frontal Lobe
primary_site=Frontal Lobe
primary_site=Occipital Lob
primary_site=Occipital Lob
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0
0
500
1000
1500
2000
Time
Number at risk
Strata
10
0
500
1000
1500
2000
Time


age_group=1−5 years old
age_group=10−15 years old
age_group=5−10 years old
age_grou
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.001
0
500
1000
1500
2000
Time
Number at risk
Strata
8
5
5
3
2
18
7
1
0
0
30
6
4
2
1
11
11
5
4
3
0
500
1000
1500
2000
Time