

# SURVIVAL ANALYSIS - TCGA PRAD CANCER

Kelvin Ofori-Minta

University of Texas at El Paso (UTEP)

June 11, 2022

## Contents

<b>1</b>	<b>Loading and Cleaning Data</b>	<b>2</b>
1.1	Inspecting dataframe for missing values . . . . .	2
1.1.1	Notes . . . . .	3
1.1.2	Re-coding variables . . . . .	4
<b>2</b>	<b>KM Curve- Default survival probability with Radiation Therapy</b>	<b>6</b>
<b>3</b>	<b>KM Curve- survival probability:Censored cases of Radiation therapy</b>	<b>8</b>
<b>4</b>	<b>KM Curve- survival probability:Censored cases of New tumor event</b>	<b>10</b>
<b>5</b>	<b>KM Curve- survival probability:Censored cases of Prior Diagnosis</b>	<b>12</b>

# 1 Loading and Cleaning Data

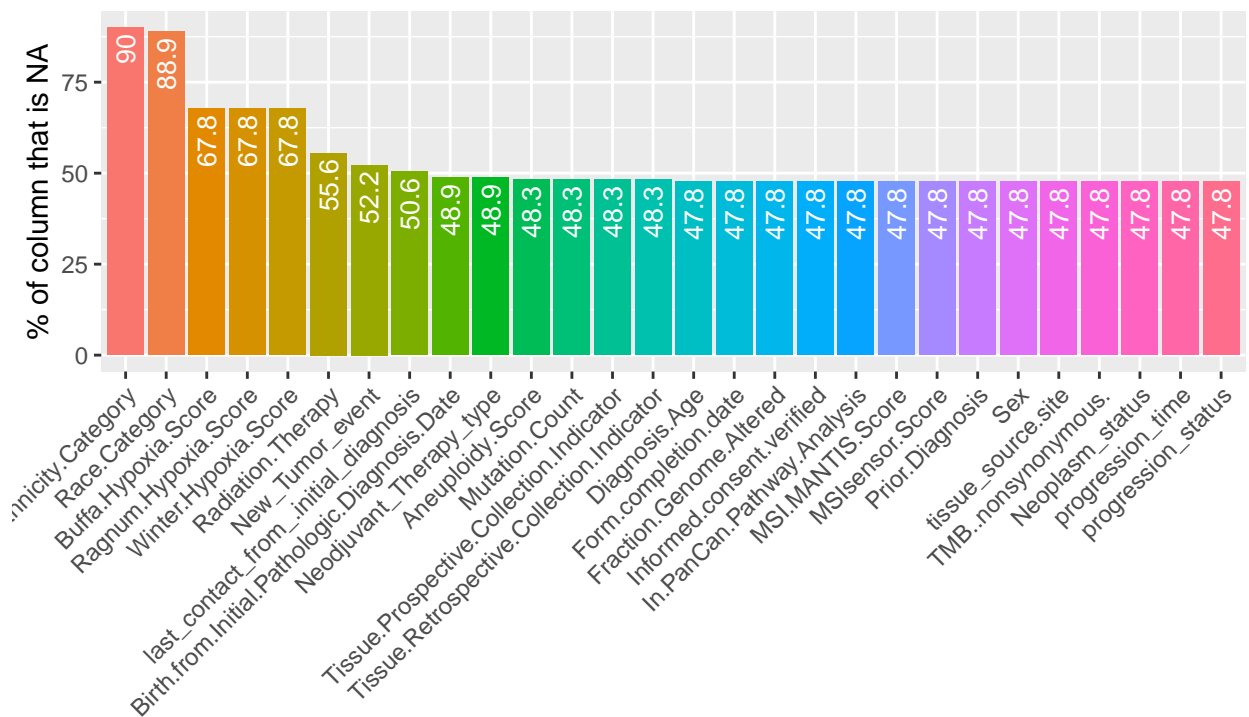
```
data <- read.csv("neoplasm.status_withtumor.csv", header = T, stringsAsFactors = F,
               na.strings = "NA")
#selecting columns of interest
#data<-s[, c(4,7,8,12,13,19,20,21,23,27:32,44:47,50,53:55,58,62,40:42)]
# write.csv(data,"C://Users//Kelvin//Desktop//Spring 2022//research with Dr. Leung//survival//
```

## 1.1 Inspecting dataframe for missing values

```
require(inspectdf)
show_plot(inspect_na(data))
```

Prevalence of NAs in df::data

df::data has 28 columns, of which 28 have missing values



```
missing = inspect_na(data)
missing[, 3] = round(missing[, 3], 2)
names(missing) = c("variable", "count", "proportion")
require(kableExtra)
# missing<-as.matrix.data.frame(missing)
kable(missing)
```

variable	count	proportion
Ethnicity.Category	162	90.00
Race.Category	160	88.89
Buffa.Hypoxia.Score	122	67.78
Ragnum.Hypoxia.Score	122	67.78
Winter.Hypoxia.Score	122	67.78
Radiation.Therapy	100	55.56
New_Tumor_event	94	52.22
last_contact_from_initial_diagnosis	91	50.56
Birth.from.Initial.Pathologic.Diagnosis.Date	88	48.89
Neoadjuvant_Therapy_type	88	48.89
Aneuploidy.Score	87	48.33
Mutation.Count	87	48.33
Tissue.Prospective.Collection.Indicator	87	48.33
Tissue.Retrospective.Collection.Indicator	87	48.33
Diagnosis.Age	86	47.78
Form.completion.date	86	47.78
Fraction.Genome.Altered	86	47.78
Informed.consent.verified	86	47.78
In.PanCan.Pathway.Analysis	86	47.78
MSI.MANTIS.Score	86	47.78
MSIsensor.Score	86	47.78
Prior.Diagnosis	86	47.78
Sex	86	47.78
tissue_source.site	86	47.78
TMB..nonsynonymous.	86	47.78
Neoplasm_status	86	47.78
progression_time	86	47.78
progression_status	86	47.78

```
# as.data.frame.matrix(missing)
# kable(as.da(missing))
```

### 1.1.1 Notes

"TMB-H means that the tumor has a high number of mutations. Doctors have found that certain immunotherapy drugs are more likely to work against TMB-H cancers. This is because the immune system may be able to find and attack cancer cells with high TMB more easily."

```
## [1] "TMB-H means that the tumor has a high number of mutations. Doctors have found that\nces
```

```
"Person neoplasm status..... You are correct, IMO: tumor free does not mean normal, but rath
```

```
## [1] "Person neoplasm status..... You are correct, IMO: tumor free does not mean normal, b
```

## 1.1.2 Re-coding variables

```

# newdata$Neoadjuvant_Therapy_type <- factor(newdata$Neoadjuvant_Therapy_type,
#                                           levels=c("No", "Yes"),
#                                           labels=c("No", "Yes")) all were "no"

data$In.PanCan.Pathway.Analysis<-factor(data$In.PanCan.Pathway.Analysis,
                                         levels=c("No", "Yes"),
                                         labels=c("No", "Yes"))

data$Prior.Diagnosis<-factor(data$Prior.Diagnosis,
                              levels=c("No", "Yes", "Yes, History Of Synchronous And Or Bilateral",
                              labels=c("No", "Yes", "yes_history"))

data$tissue_source.site<-factor(data$tissue_source.site,
                                levels = c("university", "Biotech & Pharma", "Hospital", "Research",
                                labels=c("university", "biotech_pharma", "hospital", "research"))

data$New_Tumor_event <- factor(data$New_Tumor_event,
                                levels=c("No", "Yes"),
                                labels=c("No", "Yes"))

data$Radiation.Therapy <- factor(data$Radiation.Therapy,
                                levels=c("No", "Yes"),
                                labels=c("No", "Yes"))

str(data)

## 'data.frame':    180 obs. of  28 variables:
## $ Diagnosis.Age : int NA 64 65 48 NA 57 65 66 57 67 ...
## $ Aneuploidy.Score : int NA 0 3 0 NA 0 2 1 0 5 ...
## $ Buffa.Hypoxia.Score : int NA -31 -17 -13 NA -37 -29 -33 -31 -29
## $ last_contact_from_.initial_diagnosis : int NA 31 62 62 NA 91 1427 2118 1882 1115
## $ Birth.from.Initial.Pathologic.Diagnosis.Date: int NA -23649 -23803 -17807 NA -21002 -24
## $ Ethnicity.Category : chr NA "Not Hispanic Or Latino" "Not Hispanic" ...
## $ Form.completion.date : chr NA "3/21/2012" "3/21/2012" "3/16/2012" ...
## $ Fraction.Genome.Altered : num NA 0.0125 0.2071 0.0284 NA ...
## $ Neoadjuvant_Therapy_type : chr NA "No" "No" "No" ...
## $ Informed.consent.verified : chr NA "Yes" "Yes" "Yes" ...
## $ In.PanCan.Pathway.Analysis : Factor w/ 2 levels "No","Yes": NA 2 2 2 NA ...
## $ MSI.MANTIS.Score : num NA 0.266 0.272 0.34 NA ...
## $ MSIsensor.Score : num NA 0 0.01 0.2 NA 0 0.01 0 0 0.31 ...
## $ Mutation.Count : int NA 33 78 108 NA 34 40 31 37 59 ...
## $ New_Tumor_event : Factor w/ 2 levels "No","Yes": NA NA NA NA ...

```

```
## $ Prior.Diagnosis      : Factor w/ 3 levels "No","Yes","yes_history
## $ Race.Category        : chr  NA "white" "white" "white" ...
## $ Radiation.Therapy    : Factor w/ 2 levels "No","Yes": NA NA NA NA
## $ Ragnum.Hypoxia.Score : int  NA -20 -2 6 NA -20 -20 -20 -12 -8 ...
## $ Sex                  : chr  NA "Male" "Male" "Male" ...
## $ Tissue.Prospective.Collection.Indicator : chr  NA "Yes" "Yes" "Yes" ...
## $ Tissue.Retrospective.Collection.Indicator : chr  NA "No" "No" "No" ...
## $ tissue_source.site   : Factor w/ 4 levels "university","biotech_pl
## $ TMB..nonsynonymous.  : num  NA 1.1 2.6 3.57 NA ...
## $ Winter.Hypoxia.Score : int  NA -28 -16 -24 NA -40 -30 -26 -40 -32
## $ Neoplasm_status      : chr  NA "with_tumor" "with_tumor" "with_tu
## $ progression_time     : num  NA 1.02 2.04 2.04 NA ...
## $ progression_status   : int  NA 1 1 1 NA 1 1 2 2 2 ...
```

## 2 KM Curve- Default survival probability with Radiation Therapy

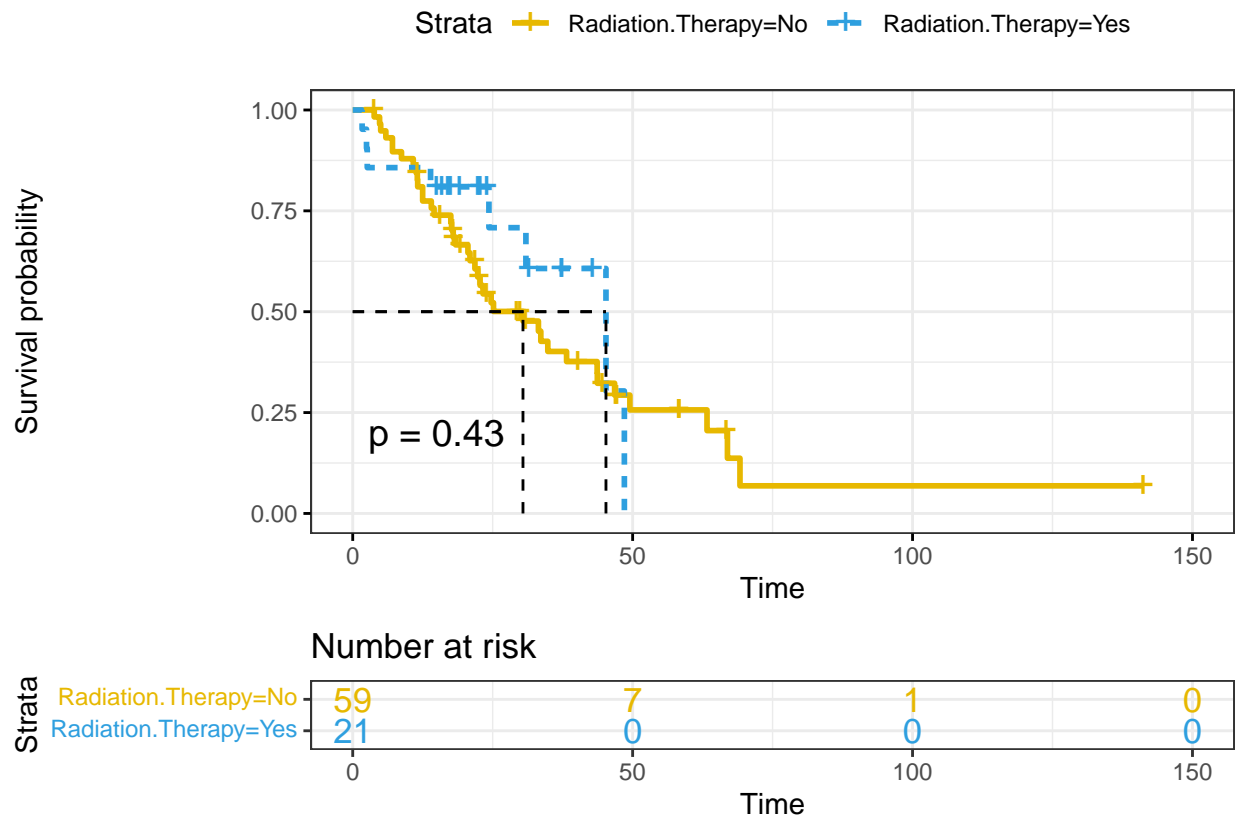
```
library("survival")
library("survminer")
ndata<-data
fit1a <- survfit(Surv(ndata$progression_time, ndata$progression_status) ~ ndata$Radiation.Therapy)
print(fit1a)
```

```
## Call: survfit(formula = Surv(ndata$progression_time, ndata$progression_status) ~
##      ndata$Radiation.Therapy, data = ndata)
##
##      100 observations deleted due to missingness
##              n events median 0.95LCL 0.95UCL
## ndata$Radiation.Therapy=No  59      39  30.4    21.8    43.7
## ndata$Radiation.Therapy=Yes 21       8  45.2    30.9     NA
```

```
summary(fit1a)$table
```

```
##              records n.max n.start events      *rmean *se(rmean)
## ndata$Radiation.Therapy=No      59      59      59      39 39.53371    6.151106
## ndata$Radiation.Therapy=Yes     21      21      21       8 35.03934    4.080074
##              median 0.95LCL 0.95UCL
## ndata$Radiation.Therapy=No 30.41063 21.82990 43.69267
## ndata$Radiation.Therapy=Yes 45.23786 30.93665      NA
```

```
ggsurvplot(fit1a,
            #legend.labs=c("tumor_free", "with_tumor"),
            pval = TRUE, conf.int = F,
            risk.table = TRUE, # Add risk table
            risk.table.col = "strata", # Change risk table color by groups
            linetype = "strata", # Change line type by groups
            surv.median.line = "hv", # Specify median survival
            ggtheme = theme_bw(), # Change ggplot2 theme
            palette = c("#E7B800", "#2E9FDF"))
```



#1 - censored & 2- progression  
#1 - tumor\_free & 2 with tumor .....neoplasm status  
#1 - NO & 2-YES .....TREATMENT CODE

### 3 KM Curve- survival probability:Censored cases of Radiation therapy

```
library("survival")
library("survminer")

fit1b <- survfit(Surv(ndata$progression_time, ndata$progression_status==1) ~ ndata$Radiation.Therapy,
                 data = ndata)
print(fit1b)
```

```
## Call: survfit(formula = Surv(ndata$progression_time, ndata$progression_status ==
##      1) ~ ndata$Radiation.Therapy, data = ndata)
##
##      100 observations deleted due to missingness
##
##              n events median 0.95LCL 0.95UCL
## ndata$Radiation.Therapy=No  59      20   47.1    40.2     NA
## ndata$Radiation.Therapy=Yes 21      13   23.9    22.4     NA
```

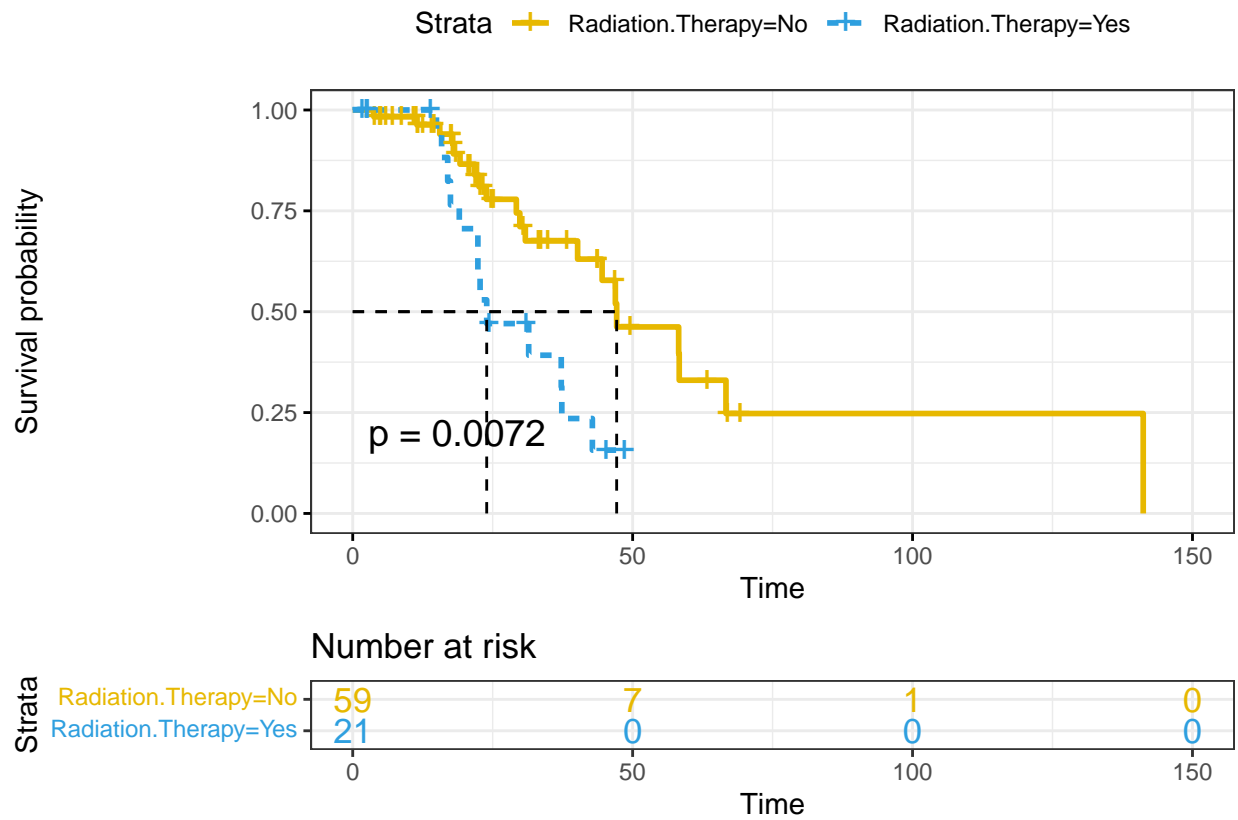
```
summary(fit1b)$table
```

	records	n.max	n.start	events	*rmean	*se(rmean)
ndata\$Radiation.Therapy=No	59	59	59	20	64.84633	10.35374
ndata\$Radiation.Therapy=Yes	21	21	21	13	44.15745	11.38353

```
##
##              median 0.95LCL 0.95UCL
## ndata$Radiation.Therapy=No 47.14469 40.17490    NA
## ndata$Radiation.Therapy=Yes 23.93398 22.35592    NA
```

```
ggsurvplot(fit1b,
            #legend.labs=c("tumor_free", "with_tumor"),
            pval = TRUE, conf.int = F,
            risk.table = TRUE, # Add risk table
            risk.table.col = "strata", # Change risk table color by groups
            linetype = "strata", # Change line type by groups
            surv.median.line = "hv", # Specify median survival
            ggtheme = theme_bw(), # Change ggplot2 theme
            palette = c("#E7B800", "#2E9FDF"))
```





#1 - censored & 2- progression  
 #1 - tumor\_free & 2 with tumor .....neoplasm status  
 #1 - NO & 2-YES .....TREATMENT CODE

## 4 KM Curve- survival probability:Censored cases of New tumor event

```
library("survival")
library("survminer")

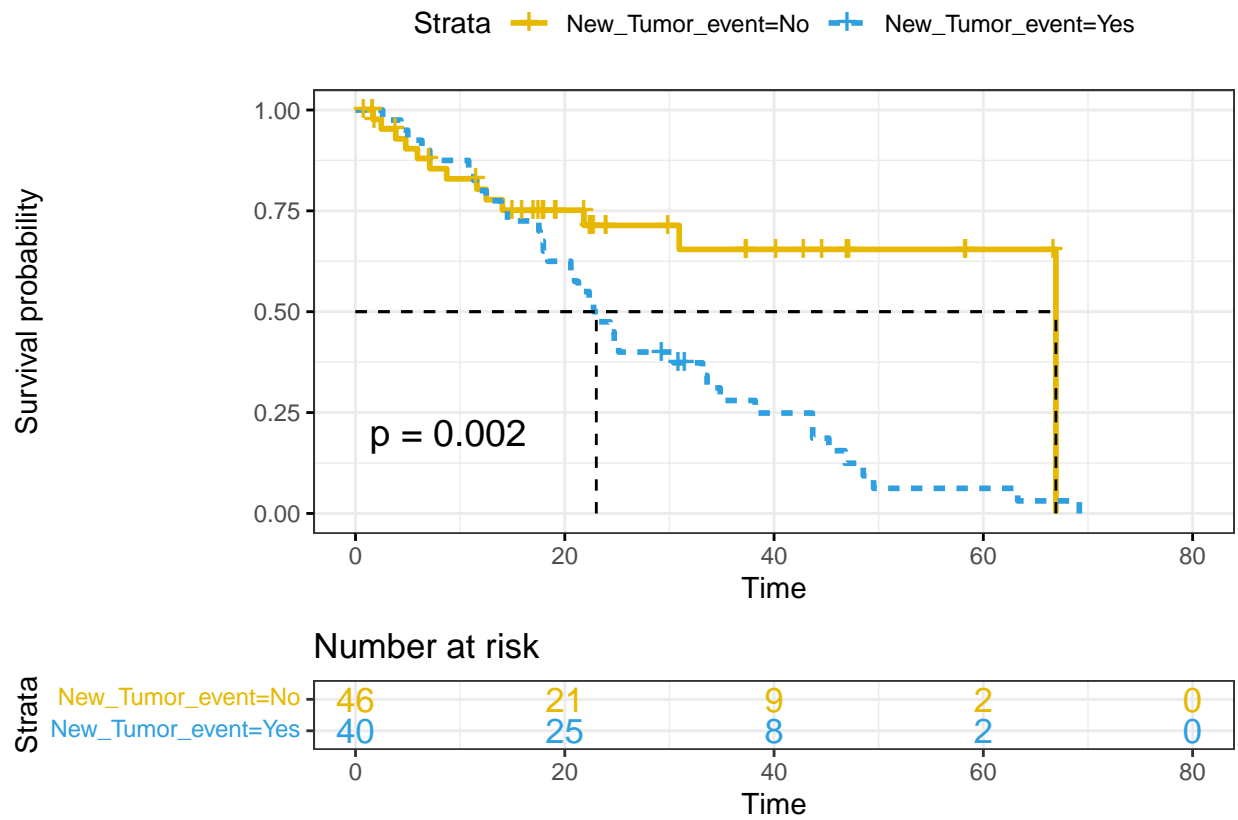
fit1c <- survfit(Surv(ndata$progression_time, ndata$progression_status==2) ~ ndata$New_Tumor_event,
                 data = ndata)
print(fit1c)

## Call: survfit(formula = Surv(ndata$progression_time, ndata$progression_status ==
##      2) ~ ndata$New_Tumor_event, data = ndata)
##
##      94 observations deleted due to missingness
##
##              n events median 0.95LCL 0.95UCL
## ndata$New_Tumor_event=No  46      13   66.9      NA      NA
## ndata$New_Tumor_event=Yes 40      37   23.0    18.4    34.8

summary(fit1c)$table

##              records n.max n.start events   *rmean *se(rmean)
## ndata$New_Tumor_event=No      46      46      46      13 48.31179   4.520889
## ndata$New_Tumor_event=Yes     40      40      40      37 27.33404   2.697264
##
##              median 0.95LCL 0.95UCL
## ndata$New_Tumor_event=No 66.93625      NA      NA
## ndata$New_Tumor_event=Yes 23.01345 18.37788 34.84893

ggsurvplot(fit1c,
            #legend.labs=c("tumor_free", "with_tumor"),
            pval = TRUE, conf.int = F,
            risk.table = TRUE, # Add risk table
            risk.table.col = "strata", # Change risk table color by groups
            linetype = "strata", # Change line type by groups
            surv.median.line = "hv", # Specify median survival
            ggtheme = theme_bw(), # Change ggplot2 theme
            palette = c("#E7B800", "#2E9FDF"))
```



#1 - censored & 2- progression  
 #1 - tumor\_free & 2 with tumor .....neoplasm status  
 #1 - NO & 2-YES .....TREATMENT CODE

## 5 KM Curve- survival probability:Censored cases of Prior Diagnosis

```
library("survival")
library("survminer")

fit1d <- survfit(Surv(ndata$progression_time, ndata$progression_status==2) ~ ndata$Prior.Diagnosis,
                 data = ndata)
print(fit1d)
```

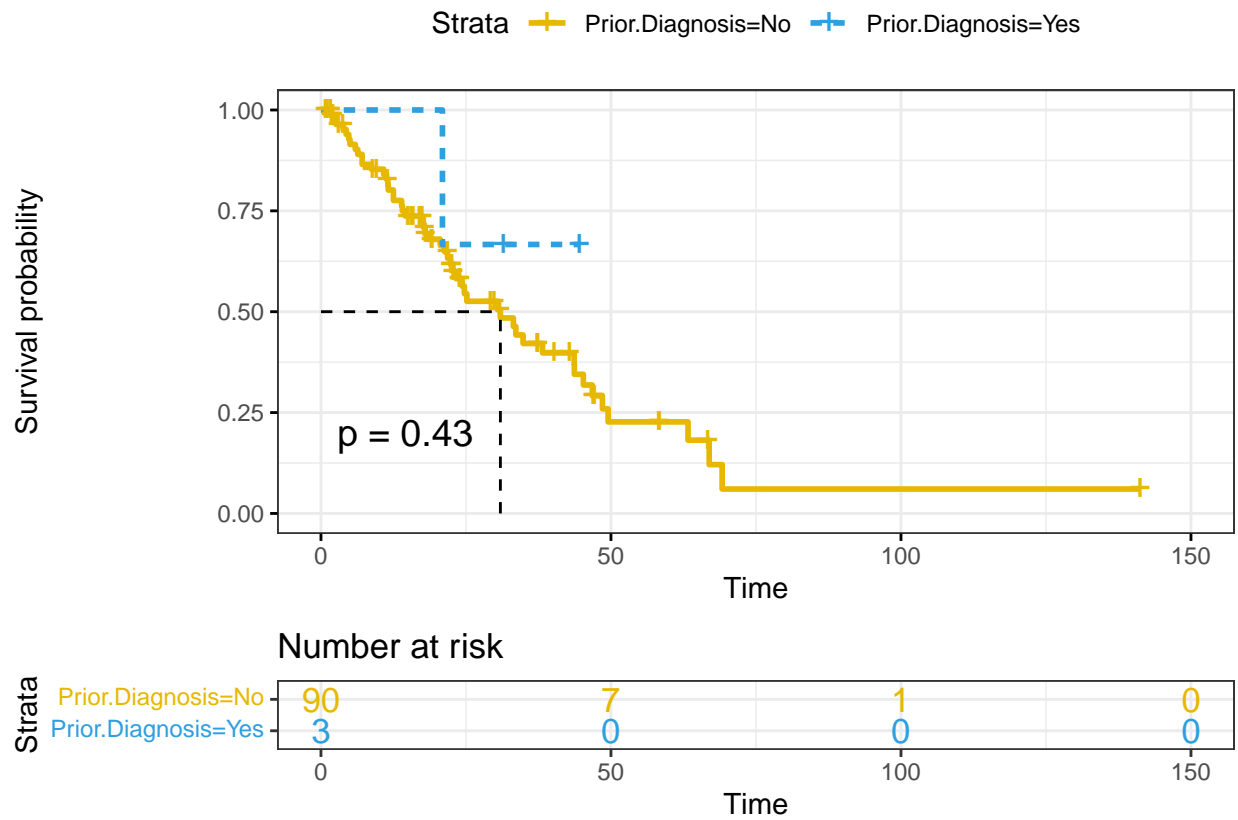
```
## Call: survfit(formula = Surv(ndata$progression_time, ndata$progression_status ==
##      2) ~ ndata$Prior.Diagnosis, data = ndata)
##
##      87 observations deleted due to missingness
##              n events median 0.95LCL 0.95UCL
## ndata$Prior.Diagnosis=No  90      49   30.9    22.8    45.2
## ndata$Prior.Diagnosis=Yes  3       1    NA    20.9     NA
```

```
summary(fit1d)$table
```

	records	n.max	n.start	events	*rmean	*se(rmean)
ndata\$Prior.Diagnosis=No	90	90	90	49	38.65491	5.451505
ndata\$Prior.Diagnosis=Yes	3	3	3	1	101.11670	32.731088

```
##              median 0.95LCL 0.95UCL
## ndata$Prior.Diagnosis=No 30.93665 22.75044 45.23786
## ndata$Prior.Diagnosis=Yes  NA 20.94224    NA
```

```
ggsurvplot(fit1d,
            #legend.labs=c("tumor_free", "with_tumor"),
            pval = TRUE, conf.int = F,
            risk.table = TRUE, # Add risk table
            risk.table.col = "strata", # Change risk table color by groups
            linetype = "strata", # Change line type by groups
            surv.median.line = "hv", # Specify median survival
            ggtheme = theme_bw(), # Change ggplot2 theme
            palette = c("#E7B800", "#2E9FDF"))
```



#1 - censored & 2- progression