SURVIVAL ANALYSIS - TCGA PRAD CANCER

Kelvin Ofori-Minta University of Texas at El Paso (UTEP)

June 20, 2022

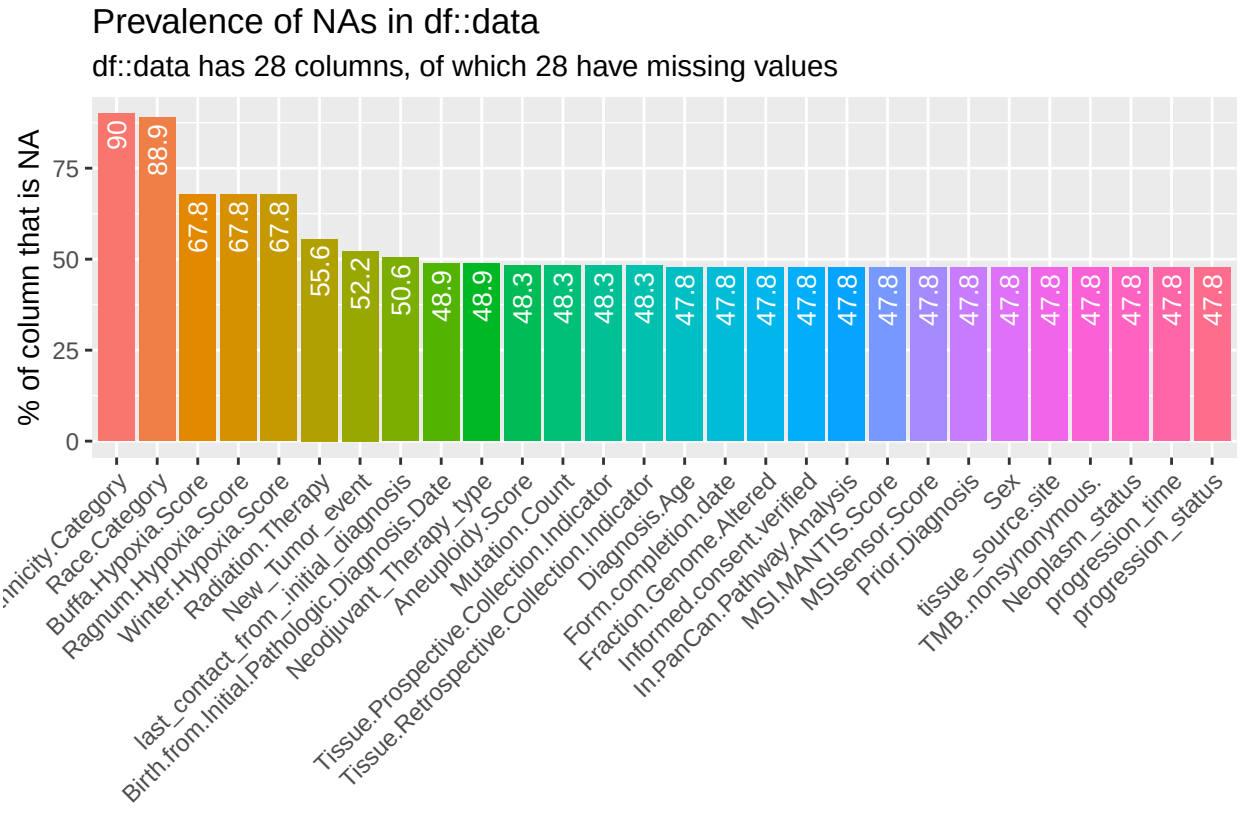
Contents

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))
```




```
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
Neodjuvant_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 Rename long variables

```
"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\nce
```

```
"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, bu
```

1.1.2 Re-coding variables

```
# newdata$Neodjuvant_Therapy_type <- factor(newdata$Neodjuvant_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"),
                                    labels=c("No","Yes"))


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


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"))

#all white , no adjuvant therapy
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 -240
##  $ Ethnicity.Category                            : chr  NA "Not Hispanic Or Latino" "Not Hispa
##  $ 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 ...
##  $ Neodjuvant_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/ 2 levels "No","Yes": NA 1 1 1 NA
##  $ 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 Generate Random subsets from entire data

```
ndata<-data
set.seed(1)
library(dplyr)
# Generate different random rows as subsets to be used for analysis
 sample_data1<-sample_n(ndata, 100) #100
 sample_data2<-sample_n(ndata, 90) #90
 sample_data3<-sample_n(ndata, 90) #90
 sample_data4<-sample_n(ndata, 80) #80
```

3 KM Curve - Survival probability with Radiation Therapy of 100 sampled subjects

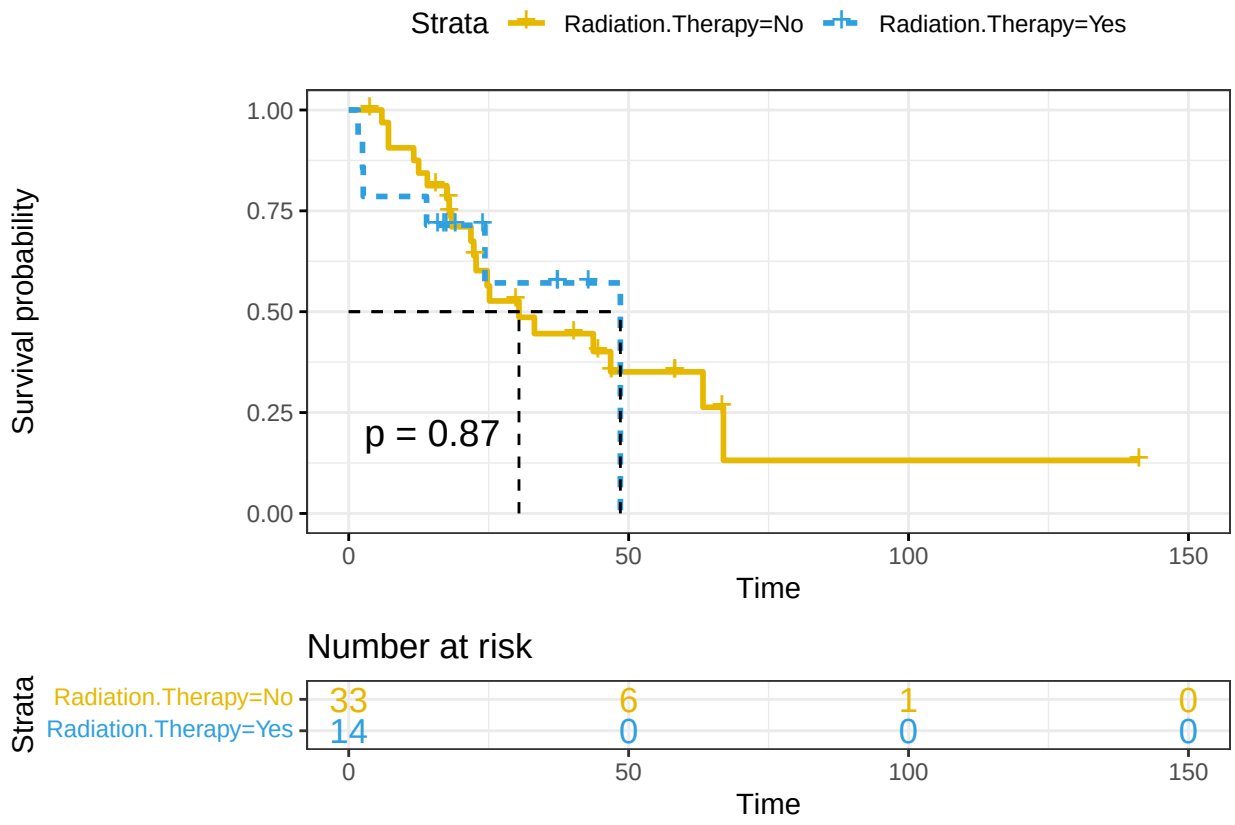
```
library("survival")
library("survminer")

fit2a<-survfit(Surv(sample_data1$progression_time, sample_data1$progression_status==2)~
                   sample_data1$Radiation.Therapy, data=sample_data1)
print(fit2a)
```

```
## Call: survfit(formula = Surv(sample_data1$progression_time, sample_data1$progression_status
##     2) ~ sample_data1$Radiation.Therapy, data = sample_data1)
##
##    53 observations deleted due to missingness
##                                      n events median 0.95LCL 0.95UCL
## sample_data1$Radiation.Therapy=No  33     20   30.4    22.3      NA
## sample_data1$Radiation.Therapy=Yes 14      6   48.5    24.3      NA
```

```
summary(fit2a)$table
```

```
##                                     records n.max n.start events    rmean
## sample_data1$Radiation.Therapy=No        33    33      33     20 47.76400
## sample_data1$Radiation.Therapy=Yes       14    14      14      6 32.67909
##                                     se(rmean)   median  0.95LCL 0.95UCL
## sample_data1$Radiation.Therapy=No   10.165516 30.41063 22.32304      NA
## sample_data1$Radiation.Therapy=Yes   5.631744 48.52550 24.32850      NA
```

```
ggsurvplot(fit2a,
          #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 with Radiation Therapy of 90 sampled subjects

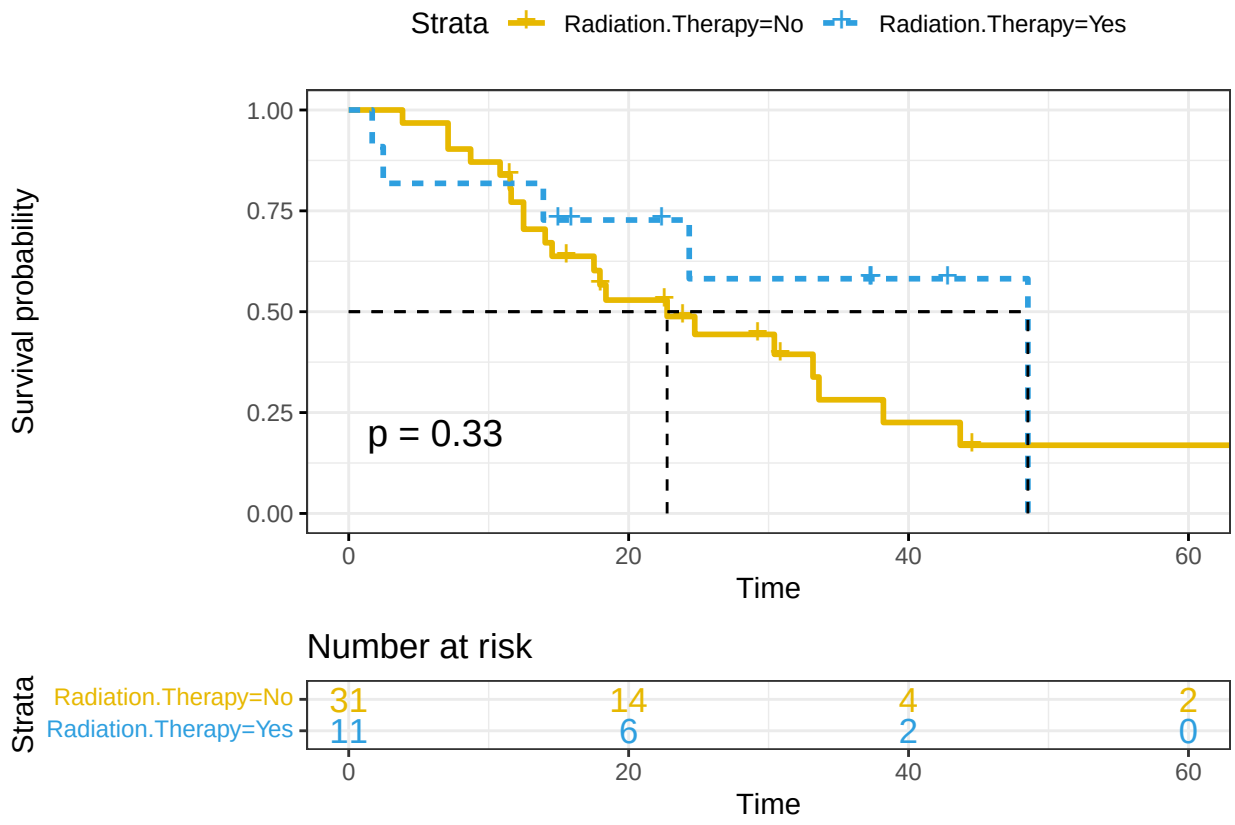
```
fit2b <- survfit(Surv(sample_data2$progression_time,sample_data2$progression_status==2) ~
                   sample_data2$Radiation.Therapy, data=sample_data2)
print(fit2b)
```

```
## Call: survfit(formula = Surv(sample_data2$progression_time, sample_data2$progression_status
##     2) ~ sample_data2$Radiation.Therapy, data = sample_data2)
##
##    48 observations deleted due to missingness
##                                      n events median 0.95LCL 0.95UCL
## sample_data2$Radiation.Therapy=No  31     22   22.8    14.5    43.7
## sample_data2$Radiation.Therapy=Yes 11      5   48.5    24.3      NA
```

```
summary(fit2b)$table
```

```
##                                     records n.max n.start events    rmean
## sample_data2$Radiation.Therapy=No        31    31      31     22 28.95121
## sample_data2$Radiation.Therapy=Yes       11    11      11      5 33.41254
##                                     se(rmean)   median  0.95LCL  0.95UCL
## sample_data2$Radiation.Therapy=No    4.046563 22.75044 14.53135 43.69267
## sample_data2$Radiation.Therapy=Yes   6.022552 48.52550 24.32850       NA
```

```
ggsurvplot(fit2b,
          #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 with Radiation Therapy of 90 sampled subjects

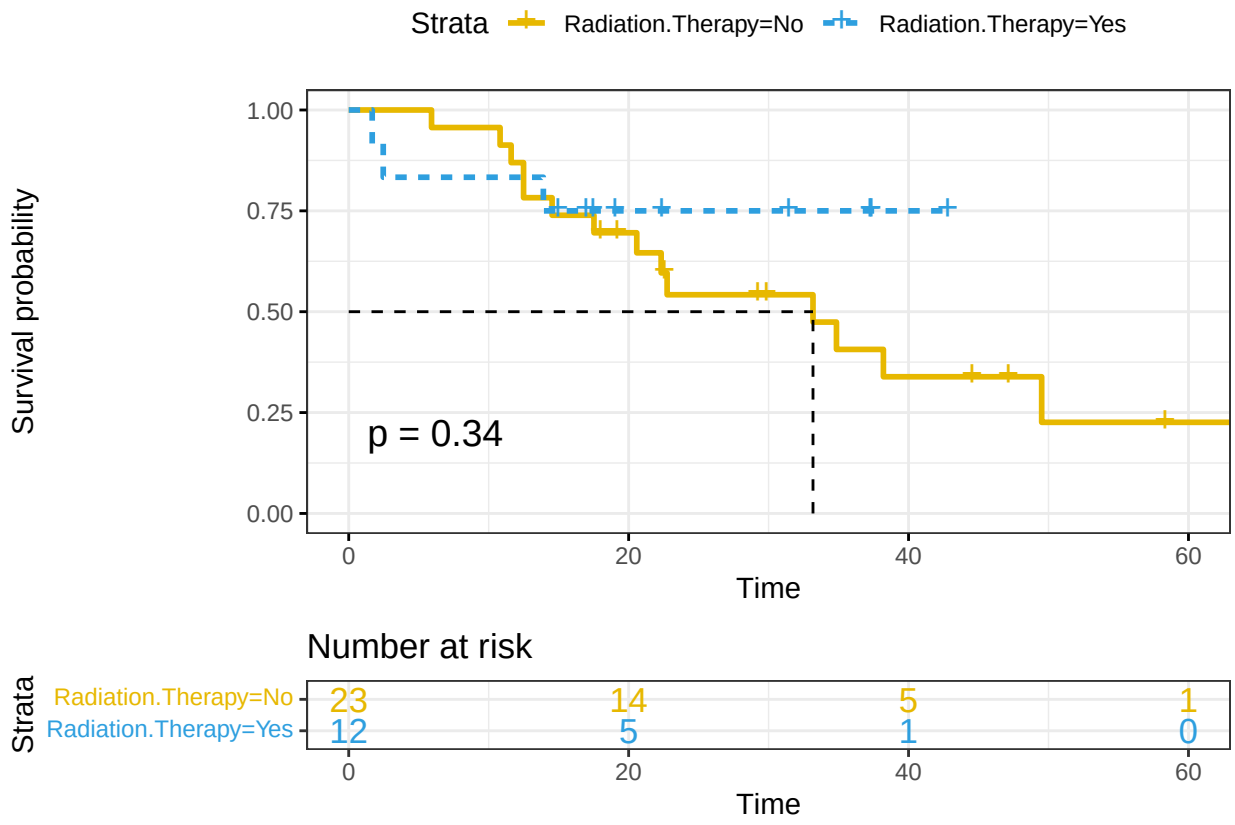
```
fit2c <- survfit(Surv(sample_data3$progression_time,sample_data3$progression_status==2) ~
                   sample_data3$Radiation.Therapy, data=sample_data3)
print(fit2c)
```

```
## Call: survfit(formula = Surv(sample_data3$progression_time, sample_data3$progression_status
##     2) ~ sample_data3$Radiation.Therapy, data = sample_data3)
##
##    55 observations deleted due to missingness
##                                      n events median 0.95LCL 0.95UCL
## sample_data3$Radiation.Therapy=No   23     14   33.2    20.6      NA
## sample_data3$Radiation.Therapy=Yes  12      3     NA      NA      NA
```

```
summary(fit2c)$table
```

```
##                                     records n.max n.start events    rmean
## sample_data3$Radiation.Therapy=No        23    23      23     14 34.92506
## sample_data3$Radiation.Therapy=Yes       12    12      12      3 51.50902
##                                     se(rmean)   median 0.95LCL 0.95UCL
## sample_data3$Radiation.Therapy=No    4.847872 33.17224 20.5806      NA
## sample_data3$Radiation.Therapy=Yes   7.624896       NA      NA      NA
```

```
ggsurvplot(fit2c,
          #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
```

6 KM Curve - Survival probability with Radiation Therapy of 80 sampled subjects

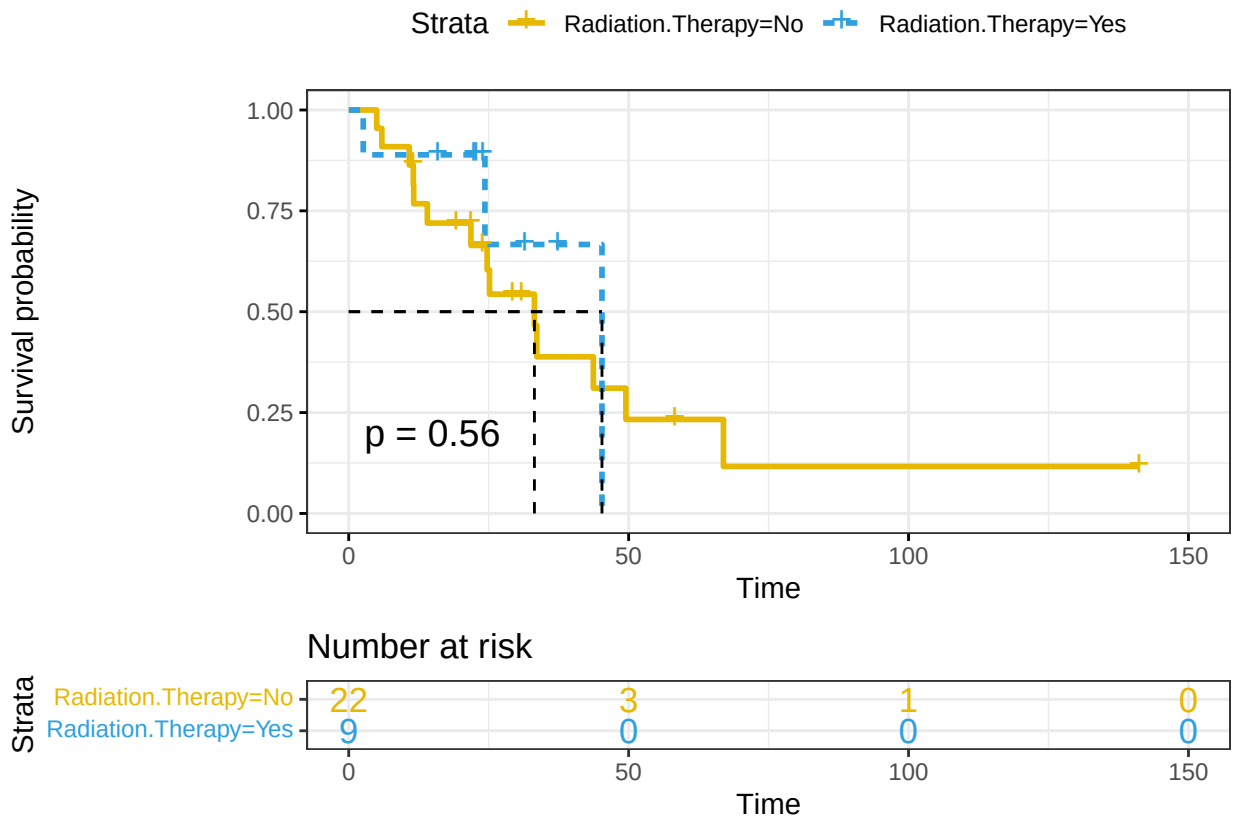
```
fit2d <- survfit(Surv(sample_data4$progression_time,sample_data4$progression_status==2) ~
                   sample_data4$Radiation.Therapy, data=sample_data4)
print(fit2d)
```

```
## Call: survfit(formula = Surv(sample_data4$progression_time, sample_data4$progression_status
##     2) ~ sample_data4$Radiation.Therapy, data = sample_data4)
##
##    49 observations deleted due to missingness
##                                      n events median 0.95LCL 0.95UCL
## sample_data4$Radiation.Therapy=No   22     14   33.2    21.8      NA
## sample_data4$Radiation.Therapy=Yes  9      3    45.2    24.3      NA
```

```
summary(fit2d)$table
```

```
##                                     records n.max n.start events    rmean
## sample_data4$Radiation.Therapy=No        22    22      22     14 43.65753
## sample_data4$Radiation.Therapy=Yes        9     9       9      3 35.85349
##                                     se(rmean)   median 0.95LCL 0.95UCL
## sample_data4$Radiation.Therapy=No   10.806475 33.17224 21.8299      NA
## sample_data4$Radiation.Therapy=Yes   5.617246 45.23786 24.3285      NA
```

```
ggsurvplot(fit2d,
          #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
```