MK2 Validation Analysis

Karthik Suresh

9/30/2020; updated: 11/18/2021

Contents

Data Import and cleaning

The data for these analyses came from cBioPortal. We downloaded the clinical data as tracks from the website along with mRNA data. The mRNA data was available in 2 forms: log-normalized to other genes, or non-scaled - i.e. log normalized RSEM expression that is not normalized to other genes. We chose to use the latter (although on comparison, the difference in actual values between the two forms of mRNA data was minimal, and did not significantly change our model results)

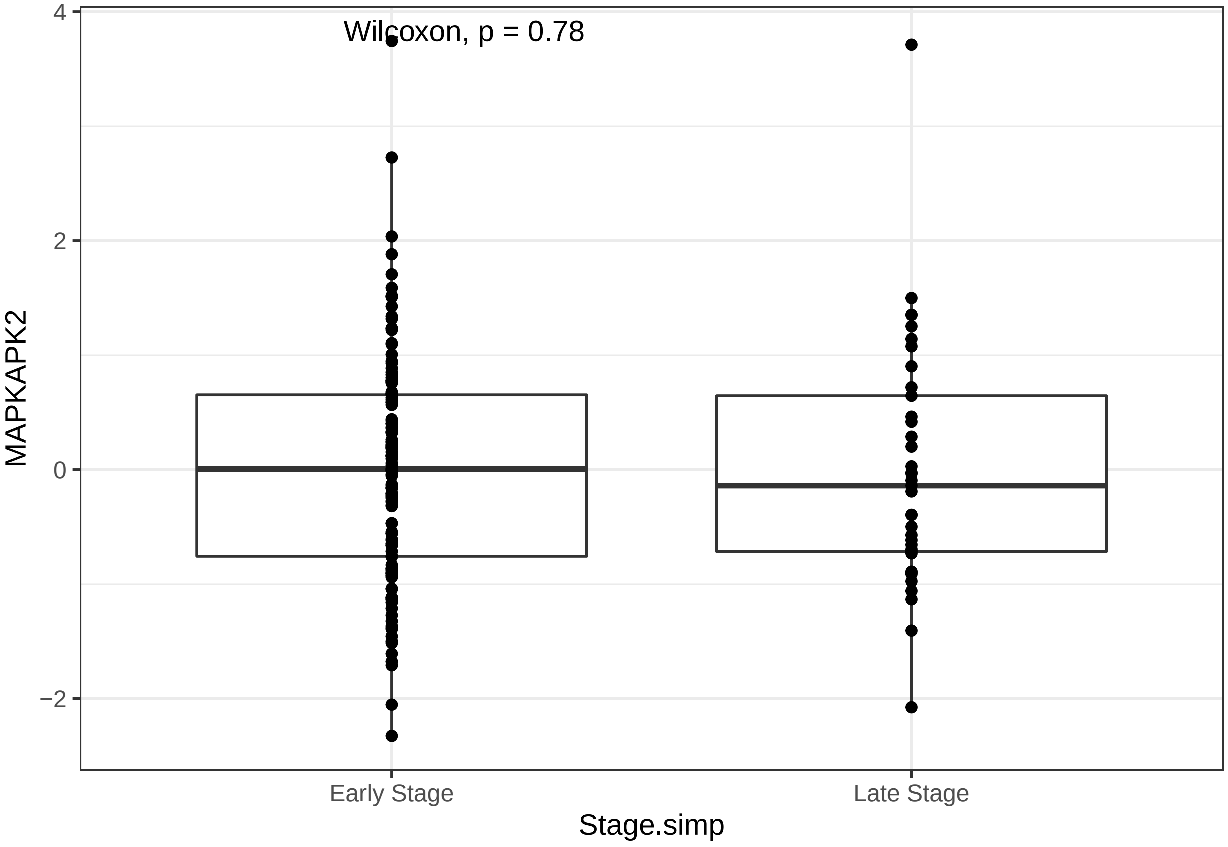
```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=70), tidy=TRUE)
```

```
knitr::opts_chunk$set(echo = TRUE)
library(data.table)
library(ggplot2)
library(ggpubr)
library(dplyr)
```

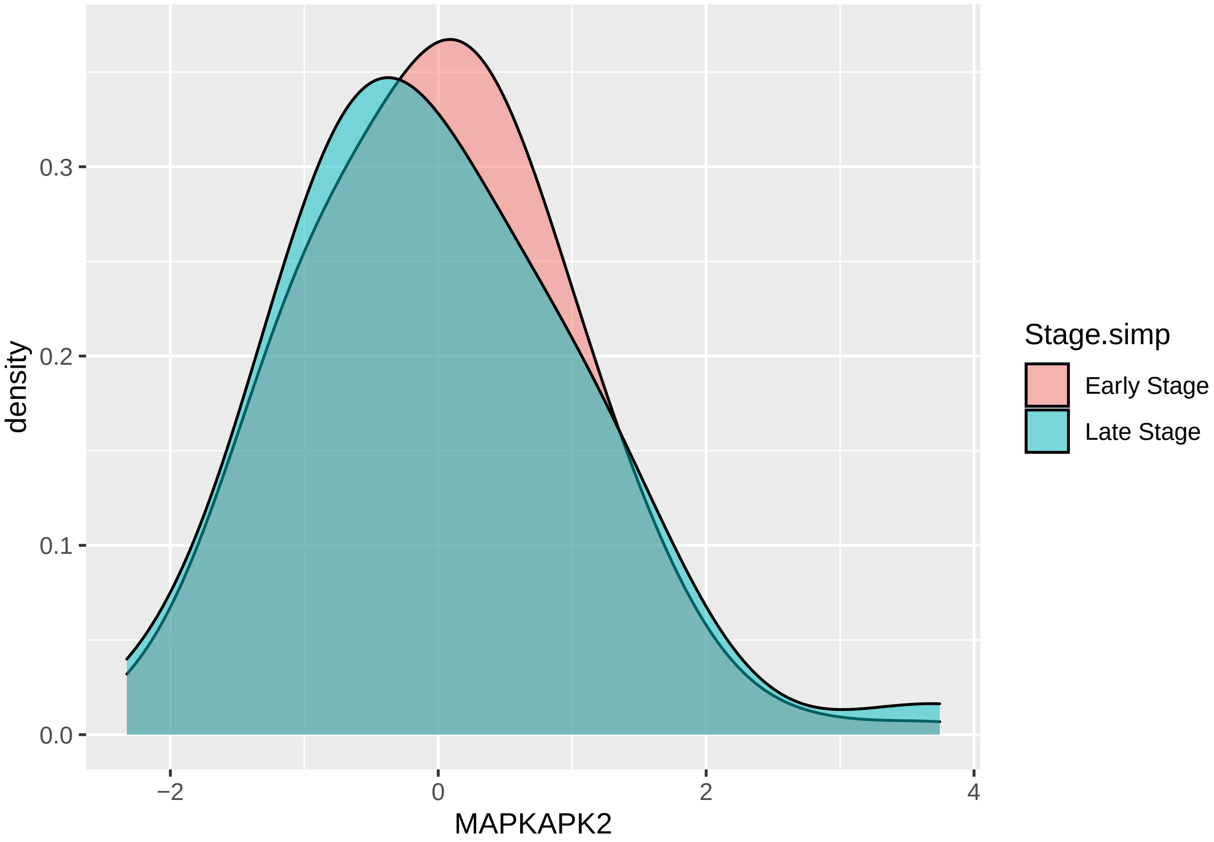
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

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

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

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

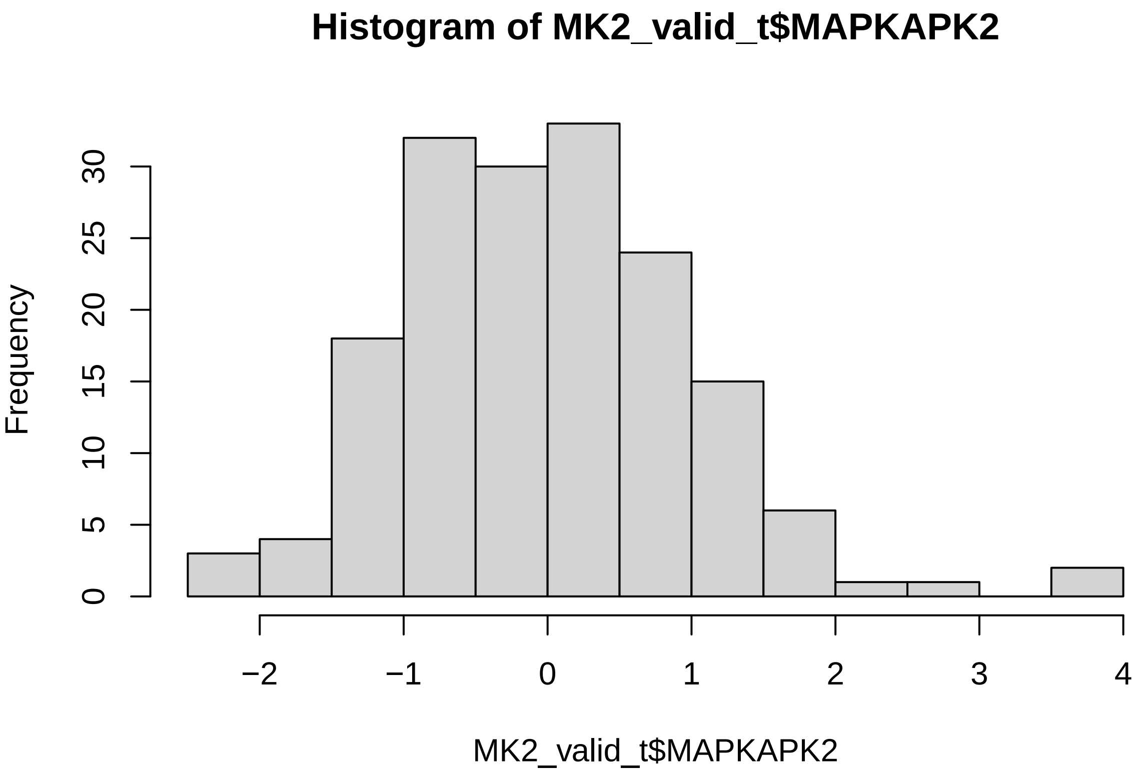
```r
# This is MAPKAPK2 mRNA and survival data obtained from cBioBortal -
# Study: OncoSG, Nat Genetics 2020

MK2_valid <- read.table("../rawdata/validation.dataset.MK2.Exp.tsv", sep = "\t")

# these are normalized across the genome. We want the non-scaled data
MK2_valid_nn <- read.csv("../rawdata/MAPKAPK2_valid_mRNA_nonnorm.txt",
    sep = "\t")
MK2_valid <- MK2_valid[c(1:8, 14), ]
MK2_valid_t <- transpose(MK2_valid)
colnames(MK2_valid_t) <- MK2_valid_t[1, ]
MK2_valid_t <- MK2_valid_t[-c(1:2), ]
MK2_valid_t <- left_join(MK2_valid_t, MK2_valid_nn, by = c(track_name = "Patient.ID"))

# switch the expression over to the non-scaled version
MK2_valid_t$MAPKAPK2 <- MK2_valid_t$MAPKAPK2..mRNA.Expression..log.RNA.Seq.V2.RSEM.

MK2_valid_t$Stage.simp <- lapply(MK2_valid_t$Stage, function(x) if (x ==
    "I" | x == "II") return("Early Stage") else return("Late Stage"))
MK2_valid_t$Stage.simp <- as.factor(unlist(MK2_valid_t$Stage.simp))
MK2_valid_t$MAPKAPK2 <- as.numeric(MK2_valid_t$MAPKAPK2)

# MK2 expression levels
ggplot(MK2_valid_t, aes(x = Stage.simp, y = MAPKAPK2)) + geom_boxplot(outlier.shape = NA) +
    geom_point() + theme_bw() + stat_compare_means()
```



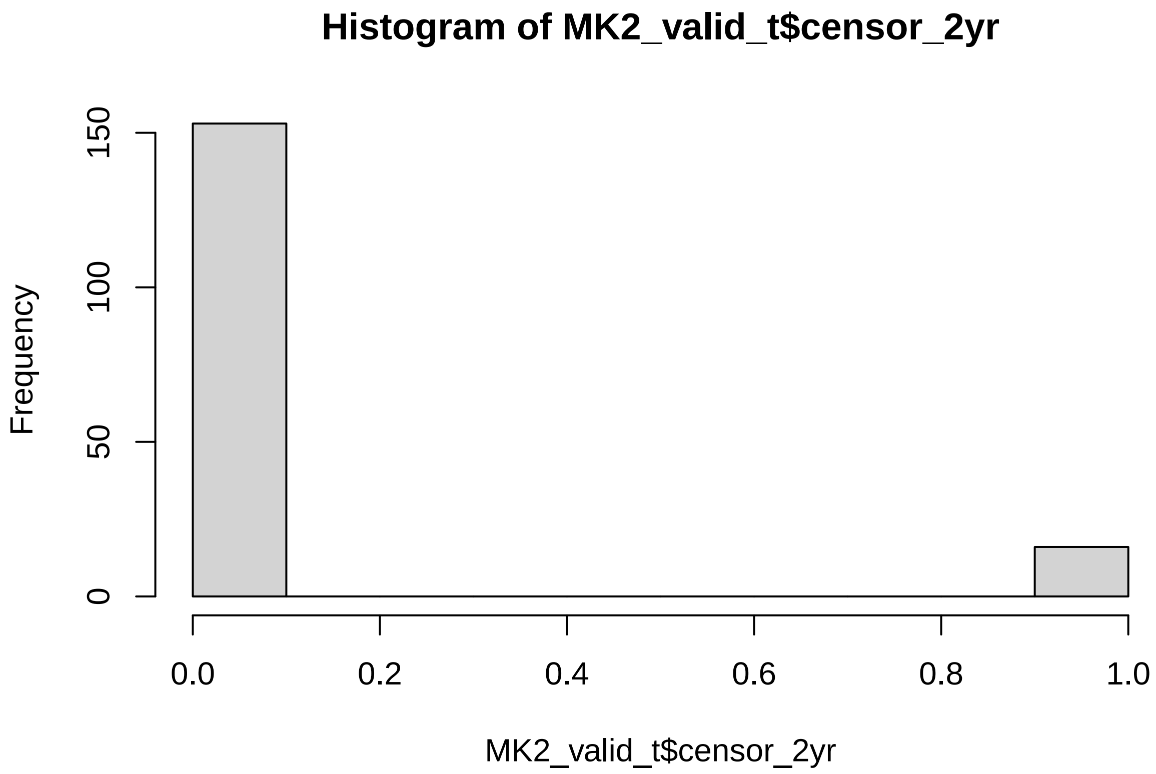
```r
ggplot(MK2_valid_t, aes(x = MAPKAPK2, fill = Stage.simp)) + geom_density(adjust = 1.5,
    alpha = 0.5)
```



```r
hist(MK2_valid_t$MAPKAPK2)
```



```r
# new variables for cox ph analysis
MK2_valid_t$censor <- lapply(MK2_valid_t$`Overall survival status`, function(x) if (x ==
    "0:LIVING") return(0) else return(1))
MK2_valid_t$time <- as.numeric(MK2_valid_t$`Overall survival months`)
MK2_valid_t$censor <- as.numeric(unlist(MK2_valid_t$censor))
MK2_valid_t$time <- as.numeric(MK2_valid_t$time)
MK2_valid_t$Sex <- as.factor(MK2_valid_t$Sex)
MK2_valid_t$Stage <- as.factor(MK2_valid_t$Stage)
MK2_valid_t$Smoking <- as.factor(MK2_valid_t$`Smoking status`)
MK2_valid_t$Age <- as.numeric(MK2_valid_t$Age)

# R censoring variable The problem is that there are very few events
# (deaths) at one year. So we can't R censor at 1 year like we did
# previously. So, we will try 2 years.
MK2_valid_t$censor_2yr <- mapply(function(dead, time) if (dead == 1 & time <
    25) return(1) else if (dead == 0 & time > 24) return(0) else if (dead ==
    0 & time < 25) return(0) else if (dead == 1 & time > 24) return(0),
    MK2_valid_t$censor, MK2_valid_t$time)
MK2_valid_t$censor_2yr <- as.numeric(unlist(MK2_valid_t$censor_2yr))


hist(MK2_valid_t$censor_2yr)
```



We will explore the data in 2 parts. In the first part - this is the part published in the paper - we really just look at the first two years (censoring at 2 years). This is in part to maintain symmetry with how we analyzed TCGA data.

Not shown here are our more extensive analyses sing the entire time period. The problem here is that MK2 expression (not unexpectedly) behaved in a time-varying fashion over this extended time period, and thus we used a variety of analyses to look at that relationship. If folks are interested in these analyses, they are available off the gitHub repo; you can look at the MK2analysis_validation.rmd file under ~/scripts/old.versions for these analyses.

Cox PH and Logistic Regression analysis with R censoring at 2 years

This will be done in two forms - using MK2 as a continuous variable, and again using top 1/3 vs. bottom 2/3 MK2 levels.

Cox PH with R censoring at 2 years, MK2 as a continuous variable

```
# MK2 survival analysis using R censoring at 1 year
summary(coxph(Surv(time, censor_2yr) ~ MAPKAPK2, id = Sample.ID, data = MK2_valid_t))
```

```
## Call:
## coxph(formula = Surv(time, censor_2yr) ~ MAPKAPK2, data = MK2_valid_t,
##     id = Sample.ID)
##
##   n= 169, number of events= 16
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## MAPKAPK2 -0.6548    0.5196   0.2796 -2.341   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## MAPKAPK2    0.5196      1.925    0.3003    0.8988
##
## Concordance= 0.686  (se = 0.055 )
## Likelihood ratio test= 5.91  on 1 df,   p=0.02
## Wald test            = 5.48  on 1 df,   p=0.02
## Score (logrank) test = 5.38  on 1 df,   p=0.02
```

```
summary(coxph(Surv(time, censor_2yr) ~ MAPKAPK2 + Sex, id = Sample.ID,
    data = MK2_valid_t))
```

```
## Call:
## coxph(formula = Surv(time, censor_2yr) ~ MAPKAPK2 + Sex, data = MK2_valid_t,
##     id = Sample.ID)
##
##   n= 169, number of events= 16
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## MAPKAPK2 -0.7174    0.4880   0.2833 -2.532   0.0113 *
## SexMale   0.9137    2.4934   0.5204  1.756   0.0791 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## MAPKAPK2     0.488     2.0492    0.2801    0.8503
## SexMale      2.493     0.4011    0.8992    6.9145
##
## Concordance= 0.716  (se = 0.055 )
## Likelihood ratio test= 9.13  on 2 df,   p=0.01
## Wald test            = 8.55  on 2 df,   p=0.01
## Score (logrank) test = 8.26  on 2 df,   p=0.02
```

```
summary(coxph(Surv(time, censor_2yr) ~ MAPKAPK2 + Sex + Stage.simp, id = Sample.ID,
    data = MK2_valid_t))
```

```
## Call:
## coxph(formula = Surv(time, censor_2yr) ~ MAPKAPK2 + Sex + Stage.simp,
##     data = MK2_valid_t, id = Sample.ID)
##
##   n= 169, number of events= 16
##
##                         coef exp(coef) se(coef)      z Pr(>|z|)
## MAPKAPK2             -0.7861    0.4556   0.2984 -2.635  0.00842 **
## SexMale               0.9529    2.5933   0.5315  1.793  0.07299 .
## Stage.simpLate Stage  1.4174    4.1262   0.5032  2.816  0.00486 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## MAPKAPK2                0.4556     2.1948    0.2539    0.8176
## SexMale                 2.5933     0.3856    0.9150    7.3497
## Stage.simpLate Stage    4.1262     0.2424    1.5388   11.0638
```

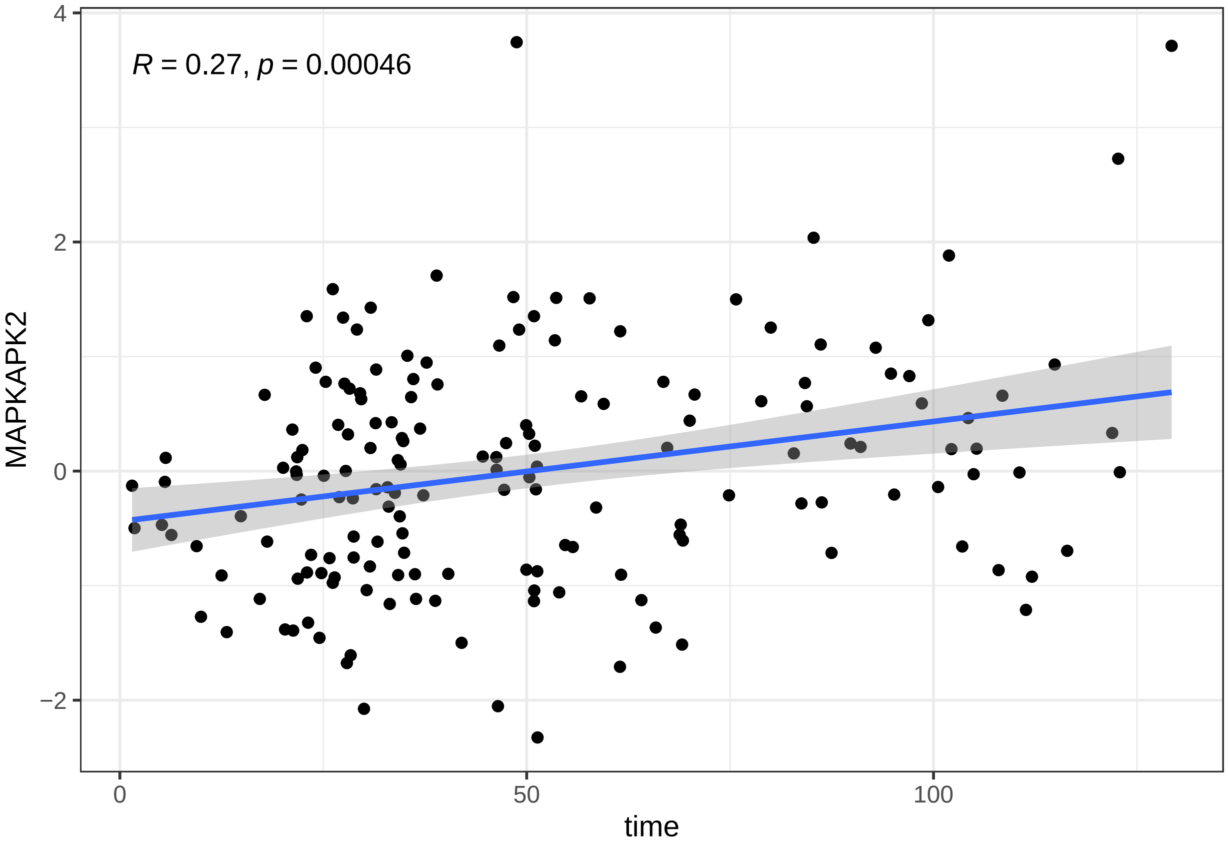
```
##
## Concordance= 0.795  (se = 0.044 )
## Likelihood ratio test= 16.5  on 3 df,   p=9e-04
## Wald test            = 14.81  on 3 df,   p=0.002
## Score (logrank) test = 16.91  on 3 df,   p=7e-04
```

```
summary(coxph(Surv(time, censor_2yr) ~ MAPKAPK2 + Age + Sex + Smoking +
    Stage.simp, id = Sample.ID, data = MK2_valid_t))
```

```
## Call:
## coxph(formula = Surv(time, censor_2yr) ~ MAPKAPK2 + Age + Sex +
##     Smoking + Stage.simp, data = MK2_valid_t, id = Sample.ID)
##
##   n= 169, number of events= 16
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## MAPKAPK2             -0.90164   0.40590  0.30940 -2.914  0.00357 **
## Age                   0.03292   1.03346  0.02668  1.234  0.21722
## SexMale               0.08766   1.09162  0.61359  0.143  0.88640
## SmokingYes            1.63958   5.15300  0.65482  2.504  0.01228 *
## Stage.simpLate Stage  1.49883   4.47645  0.51553  2.907  0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## MAPKAPK2                0.4059     2.4636    0.2213    0.7444
## Age                     1.0335     0.9676    0.9808    1.0889
## SexMale                 1.0916     0.9161    0.3279    3.6338
## SmokingYes              5.1530     0.1941    1.4278   18.5970
## Stage.simpLate Stage    4.4764     0.2234    1.6297   12.2958
##
## Concordance= 0.827  (se = 0.046 )
## Likelihood ratio test= 24.36  on 5 df,   p=2e-04
## Wald test            = 19.58  on 5 df,   p=0.001
## Score (logrank) test = 23.71  on 5 df,   p=2e-04
```

```
MK2_CoxPH_valid_RC <- coxph(Surv(time, censor_2yr) ~ MAPKAPK2 + Age + Sex +
    Smoking + Stage.simp, id = Sample.ID, data = MK2_valid_t)

ggplot(MK2_valid_t, aes(x = time, y = MAPKAPK2)) + geom_point() + geom_smooth(method = "lm") +
    theme_bw() + stat_cor()
```

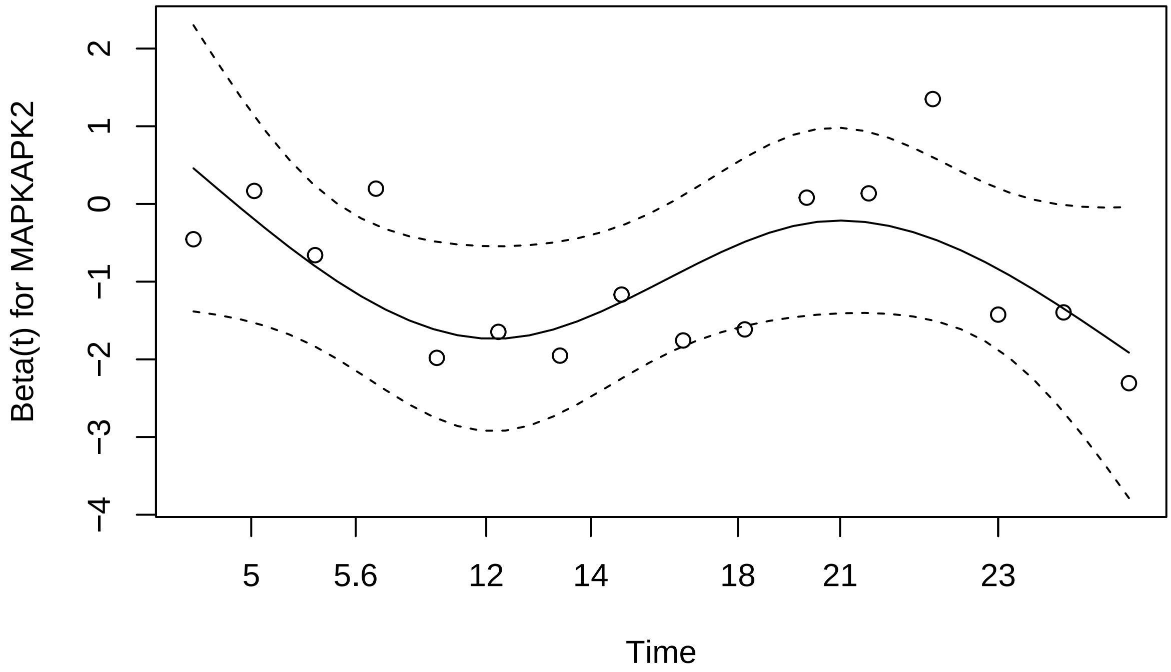
```
## `geom_smooth()` using formula 'y ~ x'
```



```
cox.zph(MK2_CoxPH_valid_RC)
```

```
##             chisq df    p
## MAPKAPK2    0.117  1 0.73
## Age         0.624  1 0.43
## Sex         1.305  1 0.25
## Smoking     0.910  1 0.34
## Stage.simp  0.681  1 0.41
## GLOBAL      2.783  5 0.73
```

```
plot(cox.zph(MK2_CoxPH_valid_RC)[1])
```

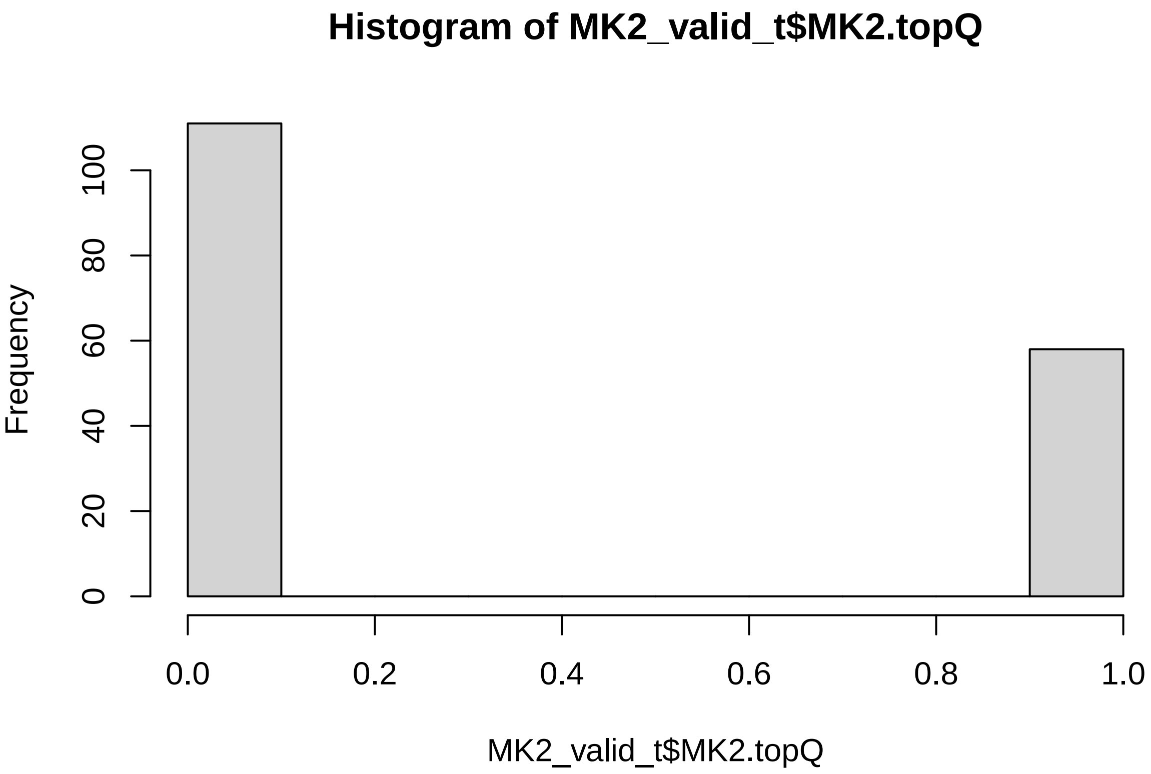


```
summary(MK2_CoxPH_valid_RC)
```

```
## Call:
## coxph(formula = Surv(time, censor_2yr) ~ MAPKAPK2 + Age + Sex +
##     Smoking + Stage.simp, data = MK2_valid_t, id = Sample.ID)
##
##   n= 169, number of events= 16
##
##                         coef exp(coef) se(coef)      z Pr(>|z|)
## MAPKAPK2            -0.90164   0.40590  0.30940 -2.914  0.00357 **
## Age                  0.03292   1.03346  0.02668  1.234  0.21722
## SexMale              0.08766   1.09162  0.61359  0.143  0.88640
## SmokingYes           1.63958   5.15300  0.65482  2.504  0.01228 *
## Stage.simpLate Stage  1.49883   4.47645  0.51553  2.907  0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## MAPKAPK2               0.4059     2.4636    0.2213    0.7444
## Age                    1.0335     0.9676    0.9808    1.0889
## SexMale                1.0916     0.9161    0.3279    3.6338
## SmokingYes             5.1530     0.1941    1.4278   18.5970
## Stage.simpLate Stage    4.4764     0.2234    1.6297   12.2958
##
## Concordance= 0.827  (se = 0.046 )
## Likelihood ratio test= 24.36  on 5 df,   p=2e-04
## Wald test            = 19.58  on 5 df,   p=0.001
## Score (logrank) test = 23.71  on 5 df,   p=2e-04
```

Cox PH with R censoring at 2 years, MK2 as a “hi” vs. “low” (dichotomous) variable

```
MK2_valid_t$MK2.topQ <- lapply(MK2_valid_t$MAPKAPK2, function(x) if (x >
    quantile(MK2_valid_t$MAPKAPK2, 0.66)) return(1) else return(0))
MK2_valid_t$MK2.topQ <- as.numeric(unlist(MK2_valid_t$MK2.topQ))
hist(MK2_valid_t$MK2.topQ)
```



```
MK2_CoxPH_valid_RC_topQ <- coxph(Surv(time, censor_2yr) ~ MK2.topQ + Age +
    Sex + Smoking + Stage.simp, id = Sample.ID, data = MK2_valid_t)
cox.zph(MK2_CoxPH_valid_RC_topQ)
```

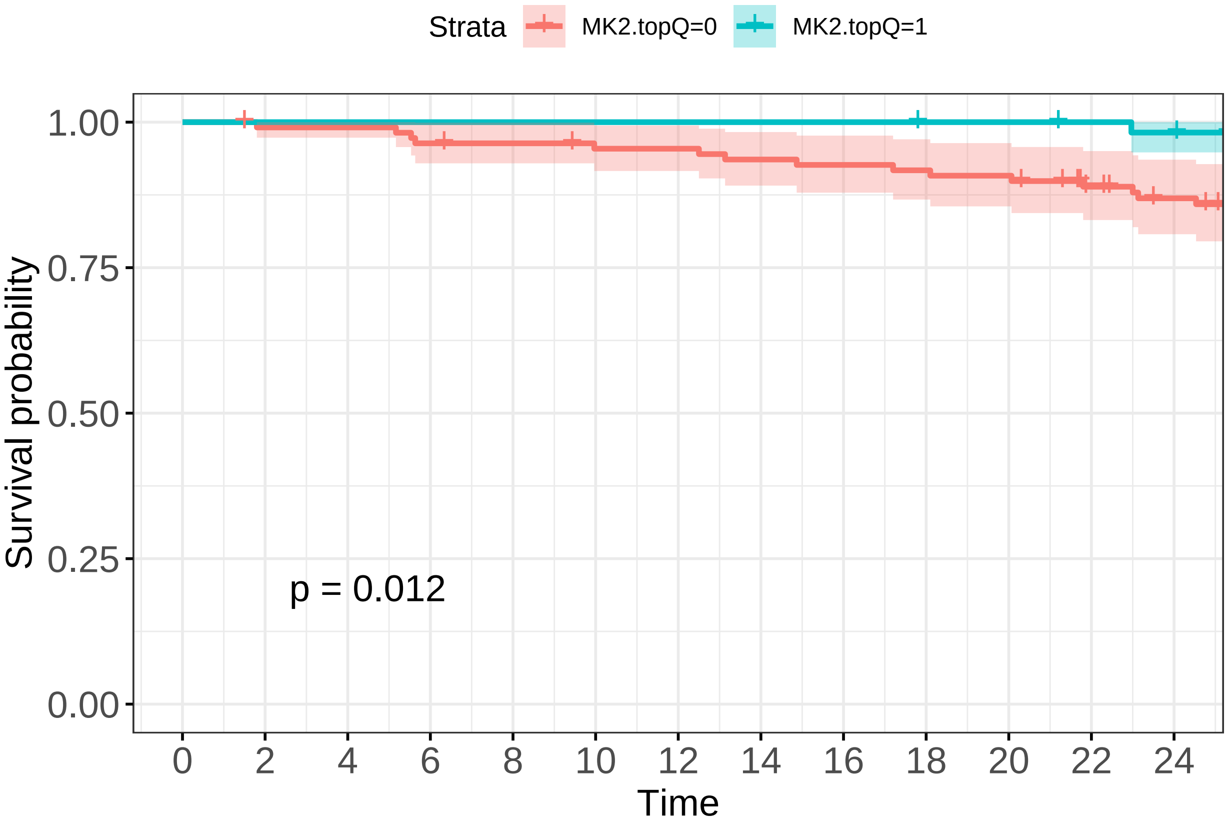
```
##            chisq df    p
## MK2.topQ   0.849  1 0.36
## Age        0.493  1 0.48
## Sex        1.393  1 0.24
## Smoking    0.769  1 0.38
## Stage.simp 0.596  1 0.44
## GLOBAL     3.335  5 0.65
```

```
summary(MK2_CoxPH_valid_RC_topQ)
```

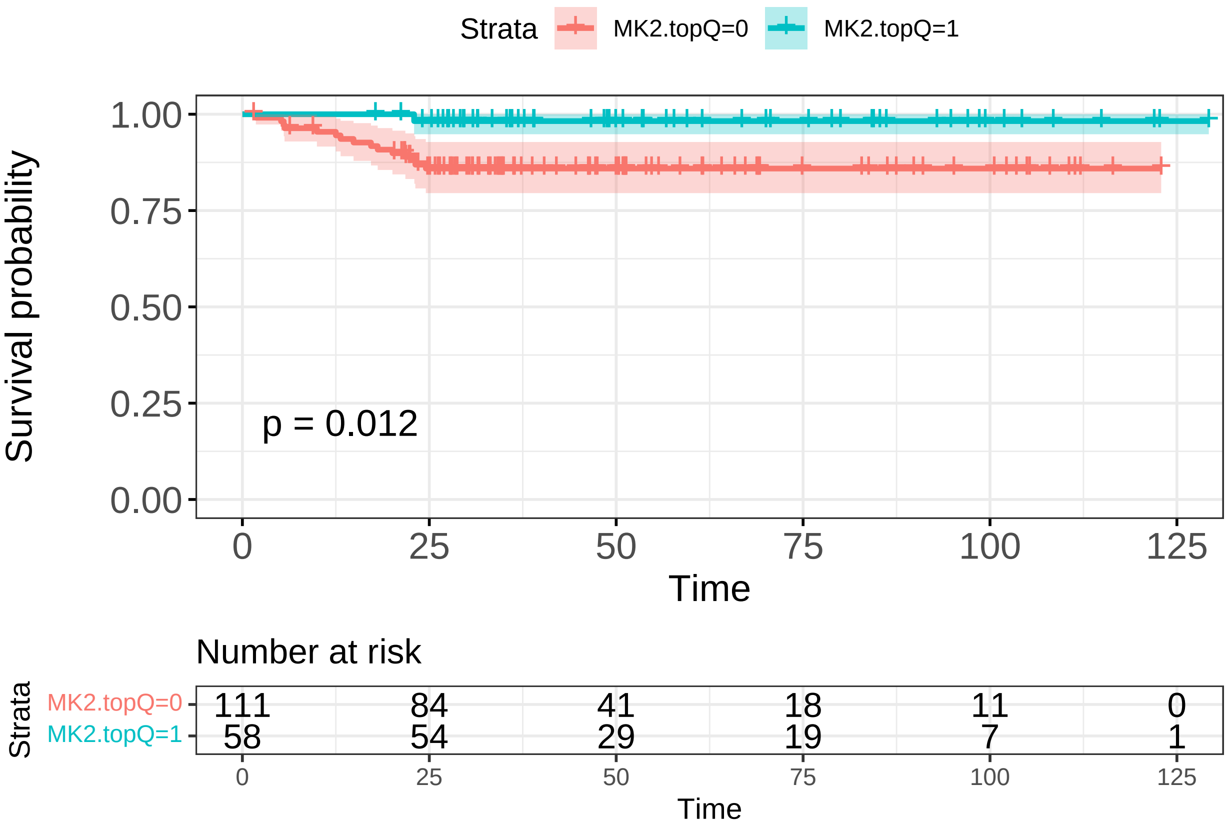
```
## Call:
## coxph(formula = Surv(time, censor_2yr) ~ MK2.topQ + Age + Sex +
##     Smoking + Stage.simp, data = MK2_valid_t, id = Sample.ID)
##
##   n= 169, number of events= 16
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## MK2.topQ             -2.24096   0.10636  1.03755 -2.160  0.03078 *
## Age                   0.02850   1.02891  0.02482  1.148  0.25087
## SexMale               0.26098   1.29820  0.59959  0.435  0.66337
## SmokingYes            1.23779   3.44798  0.62527  1.980  0.04775 *
## Stage.simpLate Stage  1.33972   3.81799  0.50235  2.667  0.00766 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## MK2.topQ                0.1064     9.4024   0.01392    0.8127
## Age                     1.0289     0.9719   0.98005    1.0802
## SexMale                 1.2982     0.7703   0.40084    4.2045
## SmokingYes              3.4480     0.2900   1.01235   11.7435
## Stage.simpLate Stage    3.8180     0.2619   1.42637   10.2197
##
## Concordance= 0.836  (se = 0.043 )
## Likelihood ratio test= 23.95  on 5 df,   p=2e-04
## Wald test            = 19.84  on 5 df,   p=0.001
## Score (logrank) test = 24.04  on 5 df,   p=2e-04
```

Model graphics for the Cox PH model above - KM curves

```
km_valid <- survfit(Surv(time, censor_2yr) ~ MK2.topQ, data = MK2_valid_t,
    type = "kaplan-meier")
ggsurvplot(km_valid, conf.int = TRUE, xlim = c(0, 24), break.x.by = 2,
    pval = TRUE, font.y = 14, font.x = 14, font.tickslab = 14, ggtheme = theme_bw())
```



```
ggsurvplot(km_valid, conf.int = TRUE, risk.table = TRUE, pval = TRUE, font.y = 14,
    font.x = 14, font.tickslab = 14, ggtheme = theme_bw())
```



Logistic regression for death at 2 year

```
glm_valid_RC_topQ <- glm(censor_2yr ~ MK2.topQ + Age + Sex + Smoking +
    Stage.simp, data = MK2_valid_t, family = binomial(link = "logit"))
ResourceSelection::hoslem.test(glm_valid_RC_topQ$y, fitted(glm_valid_RC_topQ))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  glm_valid_RC_topQ$y, fitted(glm_valid_RC_topQ)
## X-squared = 5.1862, df = 8, p-value = 0.7375
```

```
summary(glm_valid_RC_topQ)
```

```
##
## Call:
## glm(formula = censor_2yr ~ MK2.topQ + Age + Sex + Smoking + Stage.simp,
##     family = binomial(link = "logit"), data = MK2_valid_t)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.39541  -0.44702  -0.25153  -0.09478   2.67335
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.58292    2.03162  -2.748  0.00600 **
## MK2.topQ             -2.43052    1.08706  -2.236  0.02536 *
## Age                   0.03575    0.02906   1.230  0.21852
## SexMale               0.27854    0.69235   0.402  0.68746
## SmokingYes            1.47837    0.70376   2.101  0.03567 *
## Stage.simpLate Stage  1.60790    0.60011   2.679  0.00738 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.869  on 168  degrees of freedom
## Residual deviance:  81.567  on 163  degrees of freedom
## AIC: 93.567
##
## Number of Fisher Scoring iterations: 7
```

```
exp(coefficients(glm_valid_RC_topQ))
```

```
##          (Intercept)             MK2.topQ                  Age
##          0.003761567          0.087990908          1.036401141
##              SexMale           SmokingYes Stage.simpLate Stage
##          1.321193700          4.385791796          4.992317722
```

```
exp(confint(glm_valid_RC_topQ))
```

```
## Waiting for profiling to be done...
```

```
##                             2.5 %     97.5 %
## (Intercept)          4.839724e-05  0.1541378
## MK2.topQ             4.595678e-03  0.4995502
## Age                  9.810628e-01  1.1011159
## SexMale              3.394220e-01  5.2842893
## SmokingYes           1.153863e+00 18.8912207
## Stage.simpLate Stage 1.545490e+00 16.8396107
```

```
listVars <- c("Age", "MAPKAPK2", "Chemotherapy", "Stage.simp", "Sex", "Smoking")
catVars <- c("Chemotherapy", "Stage.simp", "Sex", "Smoking")
table1_validationcohort <- tableone::CreateTableOne(listVars, MK2_valid_t,
    catVars, strata = c("MK2.topQ"), addOverall = TRUE)
print(table1_validationcohort)
```

```
##                               Stratified by MK2.topQ
##                                Overall       0             1             p
##   n                              169           111           58
##   Age (mean (SD))              64.00 (9.47)  63.96 (10.23) 64.07 (7.91)   0.946
##   MAPKAPK2 (mean (SD))          0.00 (1.00)  -0.56 (0.60)   1.07 (0.69)  <0.001
##   Chemotherapy = Yes (%)          53 (31.4)     29 (26.1)     24 (41.4)   0.064
##   Stage.simp = Late Stage (%)     37 (21.9)     25 (22.5)     12 (20.7)   0.938
##   Sex = Male (%)                  75 (44.4)     44 (39.6)     31 (53.4)   0.121
##   Smoking = Yes (%)               61 (36.1)     41 (36.9)     20 (34.5)   0.883
##                               Stratified by MK2.topQ
##                                test
##   n
##   Age (mean (SD))
##   MAPKAPK2 (mean (SD))
##   Chemotherapy = Yes (%)
##   Stage.simp = Late Stage (%)
##   Sex = Male (%)
##   Smoking = Yes (%)
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## 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] survminer_0.4.9   survival_3.2-11   dplyr_1.0.7       ggpubr_0.4.0
## [5] ggplot2_3.3.5     data.table_1.14.0 knitr_1.33
##
## loaded via a namespace (and not attached):
##  [1] tidyr_1.1.3             splines_4.1.0           carData_3.0-4
##  [4] assertthat_0.2.1        highr_0.9               cellranger_1.1.0
##  [7] yaml_2.2.1              pillar_1.6.2            backports_1.2.1
## [10] lattice_0.20-44         glue_1.4.2              digest_0.6.27
## [13] ggsignif_0.6.2          gridtext_0.1.4          ResourceSelection_0.3-5
## [16] colorspace_2.0-2        htmltools_0.5.2         Matrix_1.3-4
## [19] survey_4.0              pkgconfig_2.0.3         labelled_2.8.0
## [22] broom_0.7.8             haven_2.4.1             purrr_0.3.4
## [25] xtable_1.8-4            scales_1.1.1            km.ci_0.5-2
## [28] openxlsx_4.2.4          rio_0.5.27              KMsurv_0.1-5
## [31] proxy_0.4-26            tibble_3.1.4            mgcv_1.8-36
## [34] generics_0.1.0          farver_2.1.0            car_3.0-11
## [37] ellipsis_0.3.2          withr_2.4.2             magrittr_2.0.1
## [40] crayon_1.4.1            readxl_1.3.1            ggtext_0.1.1
## [43] evaluate_0.14           fansi_0.5.0             nlme_3.1-152
## [46] MASS_7.3-54             class_7.3-19            rstatix_0.7.0
## [49] forcats_0.5.1           xml2_1.3.2              foreign_0.8-81
## [52] tableone_0.13.0         tools_4.1.0             hms_1.1.0
## [55] mitools_2.4             formatR_1.11            lifecycle_1.0.0
## [58] stringr_1.4.0           munsell_0.5.0           zip_2.2.0
## [61] e1071_1.7-7             compiler_4.1.0          rlang_0.4.11
## [64] grid_4.1.0              labeling_0.4.2          rmarkdown_2.11
## [67] gtable_0.3.0            abind_1.4-5             DBI_1.1.1
## [70] curl_4.3.2              markdown_1.1            R6_2.5.1
## [73] gridExtra_2.3           zoo_1.8-9               fastmap_1.1.0
## [76] survMisc_0.5.5          utf8_1.2.2              stringi_1.7.4
## [79] Rcpp_1.0.7              vctrs_0.3.8             tidyselect_1.1.1
## [82] xfun_0.25
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.