Final Project 175

Branson Enani Mikaela Statner Leslie Liu

2022-11-18

Report on Competing Risks Model for Cardiovascular Disease

We aim to explore how BMI, Gender and Age affect Cardiovascular health, specifically the event of death from cardiovascular disease.

We will perform a competing risk analysis on data from 453 patients. The event that we are focusing on is death from CVD (cardiovascular disease). This event is of primary clinical interest and the other causes/events are "competing" with the primary event. The competing risk is death from other causes. We also want to take into account the effect of several covariates such as Sex, Age, and BMI. There are 453 observations from subjects age 20 to 114. Variables include gender, with 0 being coded as male and 1 being coded as female, Body Mass Index (BMI) in kilograms per m^2, follow up time in days, and the event type. For the event type, a 1 corresponds to death from a cardiovascular disease, a 2 corresponds to death from another cause, and 0 is censored.

The data for a competing risks model is different than other survival data because instead of each patient having two different outcomes (being censored or experiencing an inevitable event), the subject may also experience an alternative event, a competing risk. When we factor in these observations to the Kaplan-Meier model, it produces an estimate that is biased upward even though the events may be independent. Therefore, when doing regression on competing risks, we have two different options.

(1) We can model the effects of covariates on the cause-specific hazard of the outcome, which allows us to estimate the effect of the covariates on subjects who have not yet had an event.

(2) We can model the effects of covariates on the cumulative incidence function which allows us to estimate the effect of the covariates on the absolute risk of the outcome over time.

Below we are downloading the required packages for survival analysis and reading in the raw data.

```
comprisk_data <- read.csv("comprisk.dat", header = FALSE)
comprisk  <- read.table("comprisk.dat", quote = "\"", comment.char="")

names <- c('ID', 'Age', 'Gender', 'BMI', 'Time', 'Status')
colnames(comprisk) <- names
```

Here we convert the columns to numeric values under their respective covariates.

```
comprisk.age <-(comprisk$Age)
comprisk.bmi<-(comprisk$BMI)
comprisk.time <- (comprisk$Time)
comprisk.status <- (comprisk$Status)
comprisk.gender <- (comprisk$Gender)
```

The number of patients who died from a cardiovascular disease, died from another cause or were censored can be obtained using the table() function:

```
table(comprisk$Status)
```

```
##
##   0   1   2
## 116 167 170
```

There are 167 events due to death from cardiovascular disease, 170 events due to death from other causes, and 116 censored events.

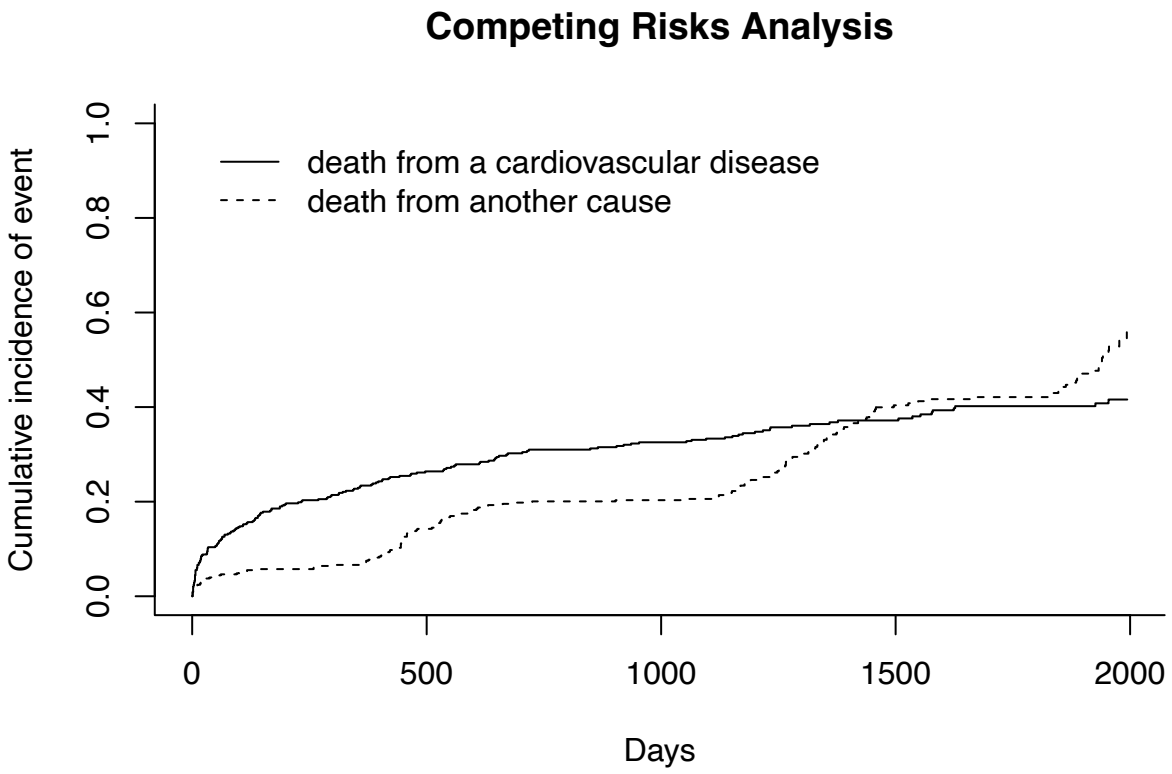
Cumulative Incidence Function

```
CI.overall <- cmprsk::cuminc(comprisk.time, fstatus = comprisk$Status)
CI.overall
```

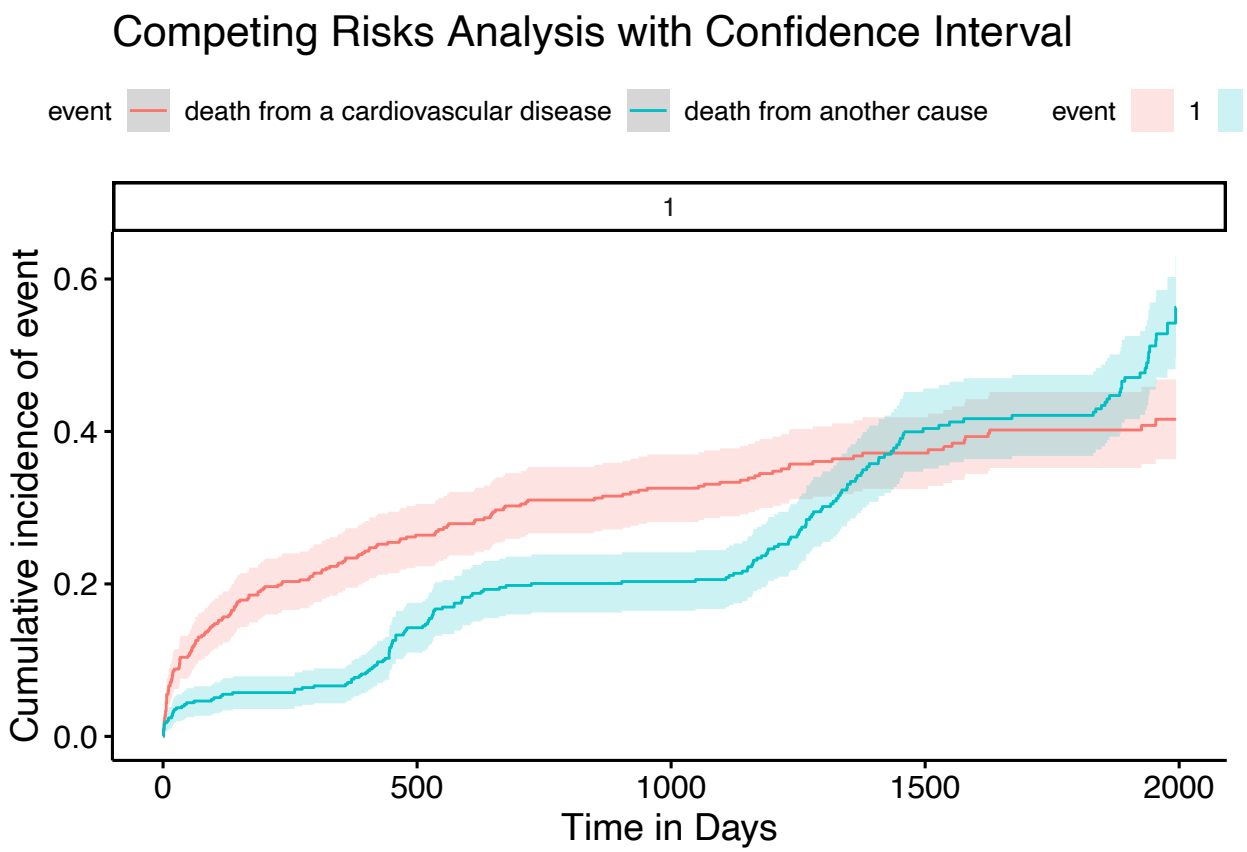
```
## Estimates and Variances:
## $est
##           500      1000      1500
## 1 1 0.2639860 0.3255391 0.3716929
## 1 2 0.1425074 0.2031471 0.4038211
##
## $var
##               500         1000         1500
## 1 1 0.0004328607 0.0005084718 0.0005758851
## 1 2 0.0002778898 0.0003822172 0.0007174177
```

The printed result shows the estimated marginal probability of each outcome (1=death from a cardiovascular disease, 2=death from another cause) at days 500, 1000, and 1500 along with the variance for each estimate. For example, the estimated marginal probability of death from a cardiovascular disease by day 1000 is 32.6%, and the estimated marginal probability of death from another cause by day 1000 is 20.3%. The information can be plotted to get an overall picture of the cumulative incidence of each event over the entire time course, given below.

```
plot(CI.overall, curvlab = c("death from a cardiovascular disease", "death from another cause"), xlab =
```




```
ggcompetingrisks(CI.overall,  conf.int = TRUE, xlab = "Time in Days",
                      ylab = "Cumulative incidence of event",
                      title = "Competing Risks Analysis with Confidence Interval") + scale_color_discre
```



Next, we can calculate the separate estimates of the cumulative incidence for death from CVD or death from other causes. This can be accomplished by using the group argument in the cmprsk() function

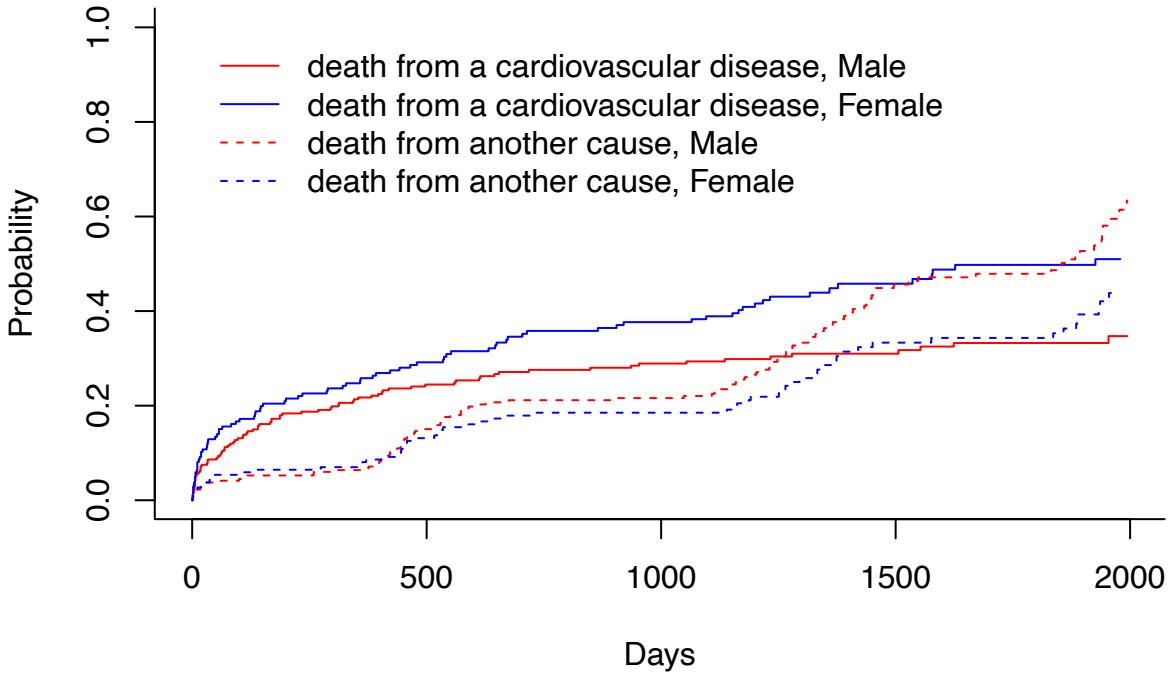
Cumulative Incidence analyzing GENDER

```
CI.gender <- cmprsk::cuminc(comprisk.time, fstatus = comprisk$Status,
group = comprisk$Gender)
CI.gender
```

```
## Tests:
##       stat          pv df
## 1 7.589692 0.005870298  1
## 2 5.540177 0.018584698  1
## Estimates and Variances:
## $est
##           500      1000      1500
## 0 1 0.2446275 0.2890803 0.3098829
## 1 1 0.2916255 0.3766371 0.4578942
## 0 2 0.1503320 0.2159594 0.4562472
## 1 2 0.1312371 0.1851320 0.3332720
##
## $var
##               500         1000         1500
## 0 1 0.0006996691 0.0008067514 0.0008671381
## 1 1 0.0011247376 0.0013296191 0.0015498318
## 0 2 0.0004964564 0.0006879503 0.0013063858
## 1 2 0.0006288858 0.0008581268 0.0015622761
```

This cumulative incidence function that is accounting for difference in gender can be plotted:

```
plot(CI.gender, lty = c(1, 1, 2, 2), col = c("red", "blue", "red",
"blue"), curvlab = c("death from a cardiovascular disease, Male", "death from a cardiovascular disease,
"death from another cause, Male", "death from another cause, Female"), xlab = "Days")
```



The event of death from Cardiovascular Disease (CVD) is higher in female patients, and death from other causes is higher in male patients. Next we will perform a formal test for differences in follow-up time for death by CVD and death by other causes between gender.

Cumulative Incidence analyzing AGE

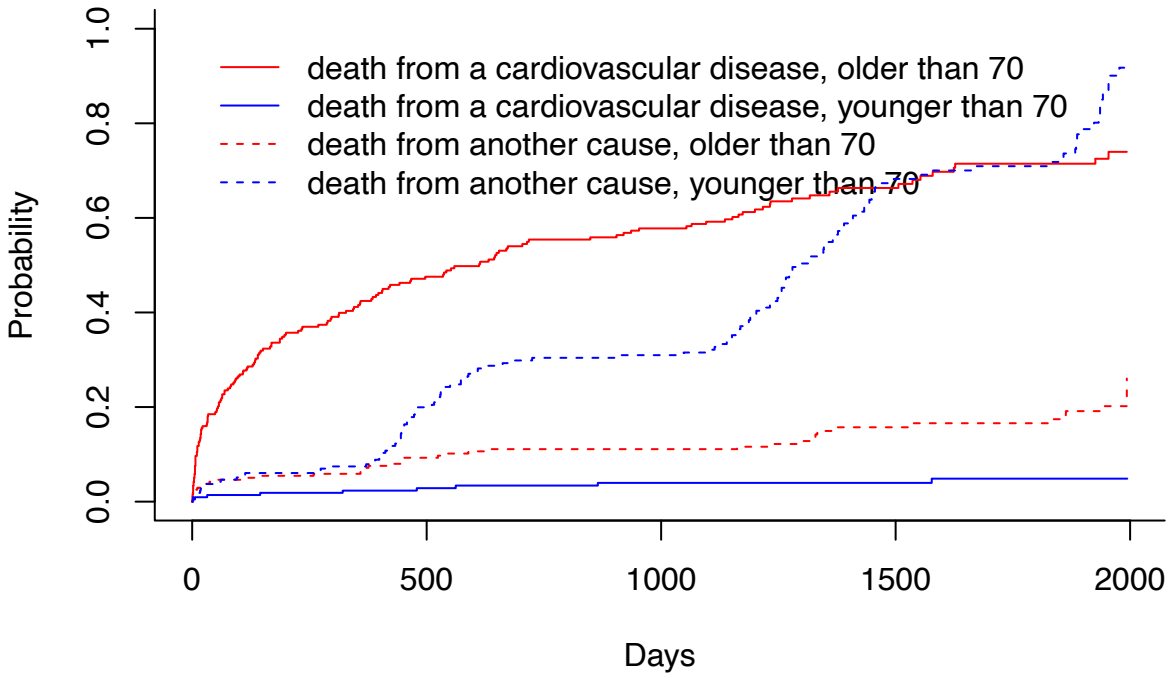
```
CI.age <- cmprsk::cuminc(comprisk.time, fstatus = comprisk$Status,
group = (comprisk.age < 70))
CI.age
```

```
## Tests:
##       stat pv df
## 1 186.2743  0  1
## 2 106.1219  0  1
## Estimates and Variances:
## $est
##                 500       1000       1500
## FALSE 1 0.47556716 0.57776084 0.66327429
## TRUE 1  0.02843758 0.03961262 0.03961262
## FALSE 2 0.09276620 0.11097626 0.15719531
## TRUE 2  0.19947909 0.30971326 0.68239237
##
## $var
##                   500          1000          1500
```

```
## FALSE 1 0.0010569578 0.0010706704 0.0010891989
## TRUE 1  0.0001319259 0.0001911007 0.0001911007
## FALSE 2 0.0003569888 0.0004247262 0.0006741142
## TRUE 2  0.0007867952 0.0011133416 0.0015681935
```

This cumulative incidence function that is accounting for if the subject is older or younger than 70 years can be plotted:

```
plot(CI.age, lty = c(1, 1, 2, 2), col = c("red", "blue", "red",
"blue"), curvlab = c("death from a cardiovascular disease, older than 70", "death from a cardiovascular
"death from another cause, older than 70", "death from another cause, younger than 70"), xlab = "Days")
```



The event of death from Cardiovascular Disease (CVD) is higher in patients older than 70, and death from other causes is higher in patients younger than 70. Death from Cardiovascular Disease (CVD) for subjects under 70 years old is almost completely horizontal at y=0. other causes is higher in male patients.

Cumulative Incidence analyzing BMI

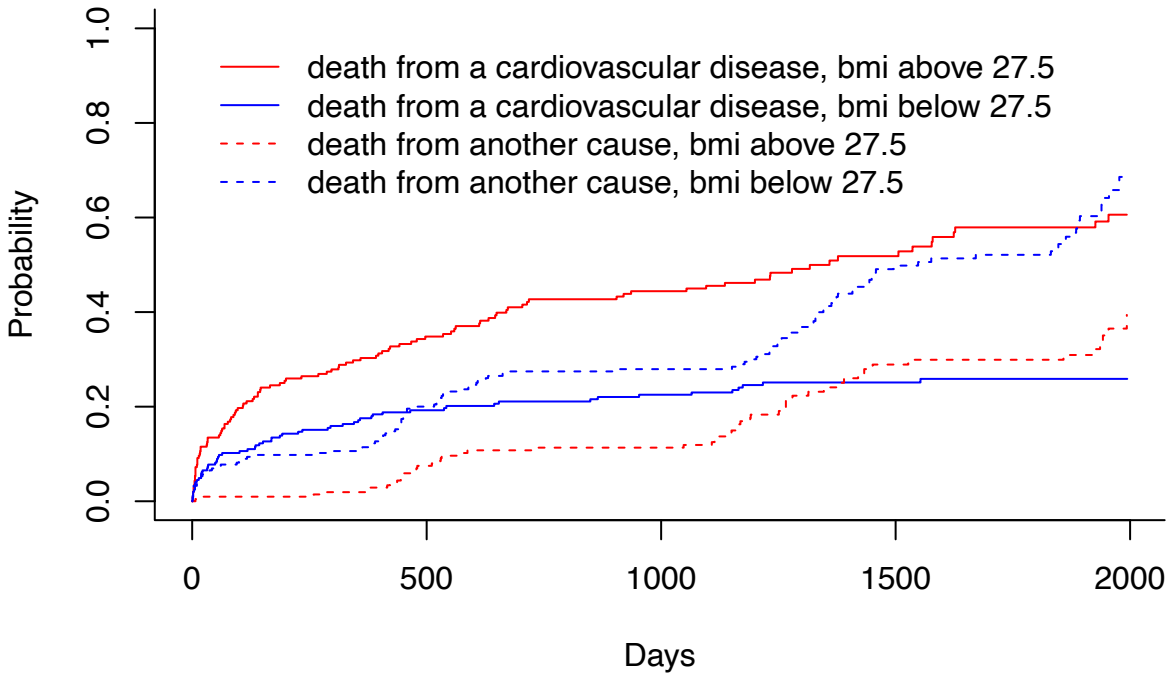
```
CI.bmi <- cmprsk::cuminc(comprisk.time, fstatus = comprisk$Status,
group = (comprisk.bmi < 27.5))
CI.bmi
```

```
## Tests:
##       stat           pv df
## 1 34.17397 5.039856e-09  1
## 2 27.80873 1.339204e-07  1
## Estimates and Variances:
## $est
##                500      1000      1500
## FALSE 1 0.34826167 0.4441827 0.5182604
```

```
## TRUE 1  0.19236961 0.2252399 0.2511420
## FALSE 2 0.07484931 0.1133376 0.2890252
## TRUE 2  0.20011330 0.2793253 0.4983299
##
## $var
##                   500         1000         1500
## FALSE 1 0.0011083689 0.0012717619 0.0014554480
## TRUE 1  0.0006391903 0.0007373499 0.0008195271
## FALSE 2 0.0003487553 0.0005243345 0.0014080397
## TRUE 2  0.0006730152 0.0008796853 0.0013654503
```

This cumulative incidence function that is accounting for if the subject has a bmi greater than 27.5 can be plotted:

```
plot(CI.bmi, lty = c(1, 1, 2, 2), col = c("red", "blue", "red",
"blue"), curvlab = c("death from a cardiovascular disease, bmi above 27.5", "death from a cardiovascular
"death from another cause, bmi above 27.5", "death from another cause, bmi below 27.5"), xlab = "Days")
```



The event of death from Cardiovascular Disease (CVD) is higher in patients with a bmi above 27.5, and death from other causes is higher in patients with a bmi below 27.5.

In doing the previous analysis, we show that, in the presence of competing risks, the basic descriptive statistic of event occurrence is not the survival function. Instead, it is the cumulative incidence function for each event type.

The reason that the competing risks plot is different from the normal survival curve is because the individual probabilities do not eventually flatten at probability = 1. If we only have 1 event involved, eventually the probability will be 1 because over a long enough time the event will occur. Given the fact that we have two events, each of them reach a probability that is less than 1 because they are competing

Since Status is non binary (0= censor, 1= CVD Death, 2=Non CVD Death) we need to split up the data into binary groups.

```r
status_cvd_event <- comprisk.status == 1
#Status 1 includes treats cvd events as the event of interest

status_noncvd_event <- comprisk.status == 2
#Status 2 includes treats non cvd events as the event of interest

status_every_outcome <- comprisk.status > 0
#Status 3 includes considers non-censored events as the events of interest
```

Now let's turn them into survival objects.

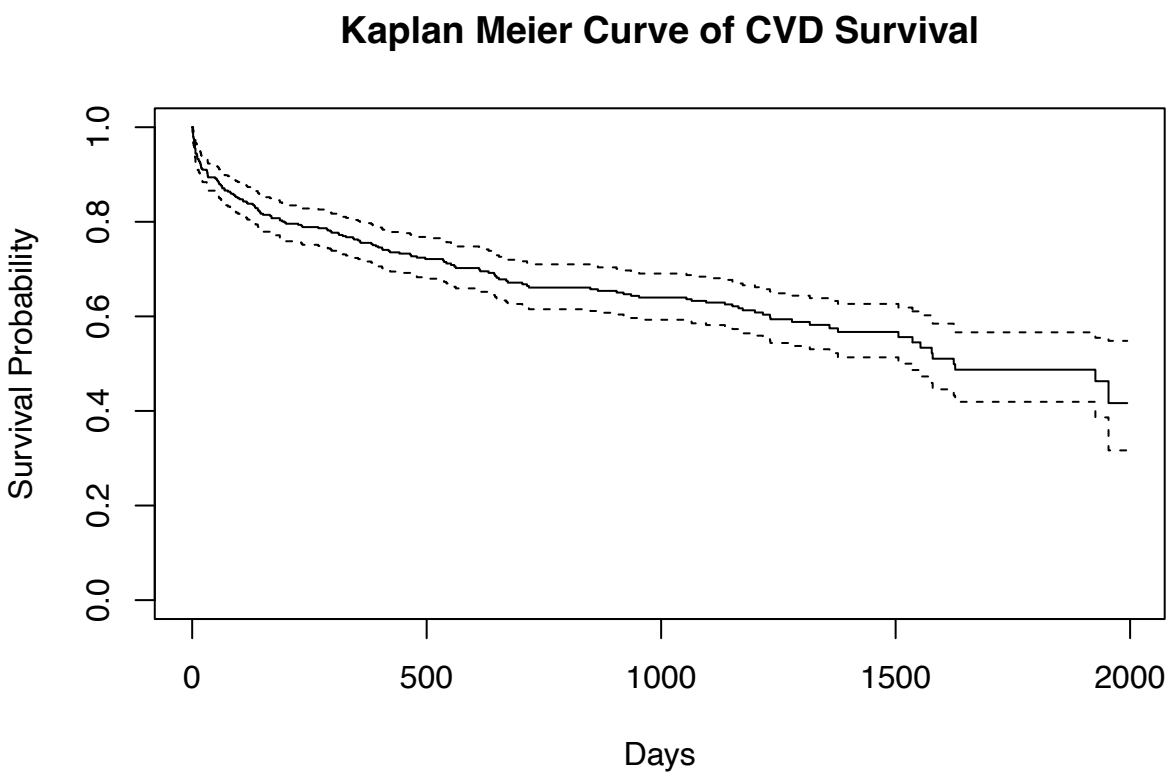
```r
vec1 <- Surv(comprisk.time, status_cvd_event)
vec2 <- Surv(comprisk.time, status_noncvd_event)
vec3 <- Surv(comprisk.time, status_every_outcome)
```

Kaplan-Meier Estimates

We are plotting a Kaplan-Meier estimate of the survival probability. The Kaplan–Meier estimator is used to estimate the survival function. The Kaplan–Meier estimator is a step-down function, with each step at each time an event of interest occurs. The height of a step at a given time is the proportion of subjects at risk just before the given time, who experience the event at that time"

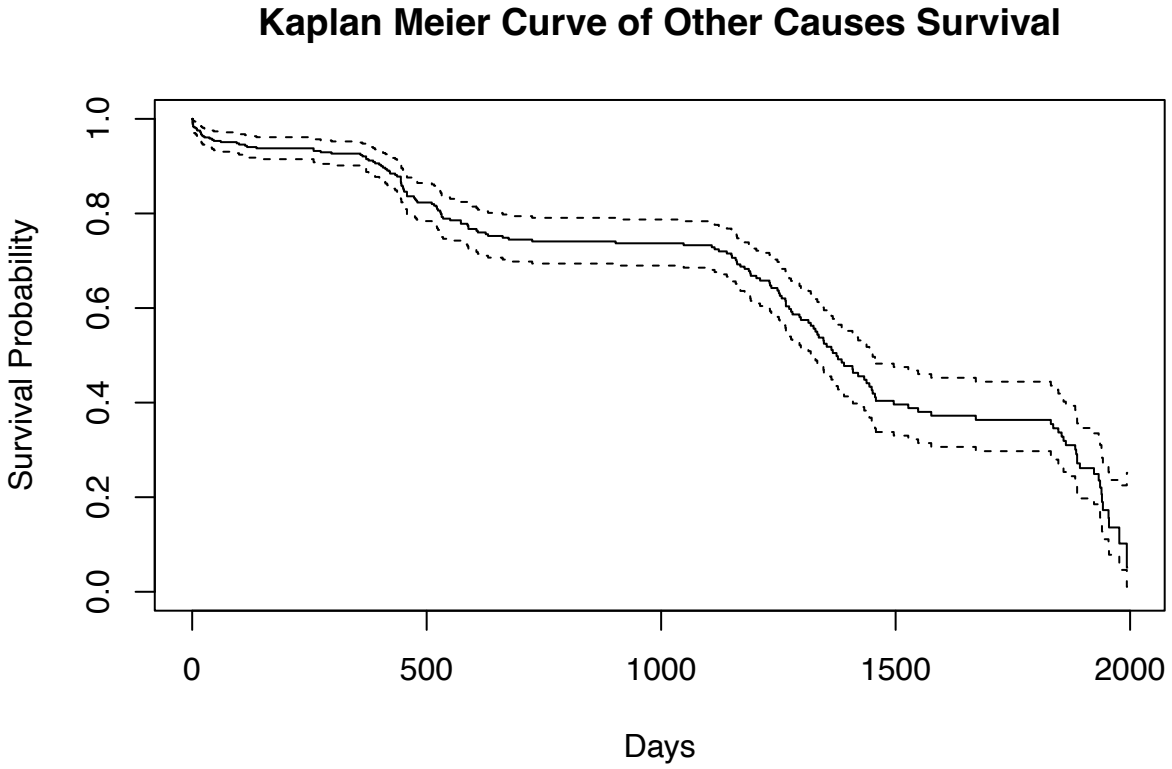
Here we want to get a Kaplan Meier Curve with our event of interest being CVD

```r
comprisk.survfit <- survfit(vec1 ~1, data = comprisk)
fit <- (comprisk.survfit)
plot(fit, main = "Kaplan Meier Curve of CVD Survival", ylab = 'Survival Probability', xlab = 'Days')
```



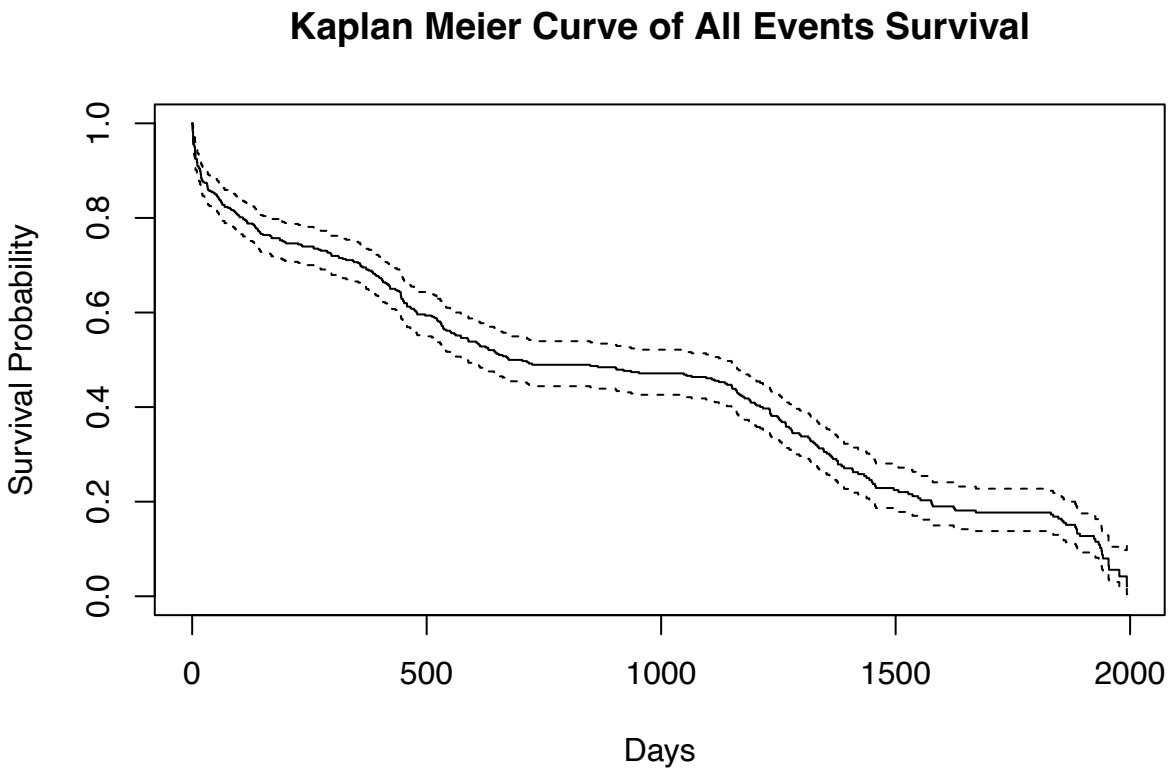
This curve shows the Kaplan Meier estimate of survival when a CVD event is treated as our event of interest.

```
comprisk.survfit2 <- survfit(vec2 ~1, data = comprisk)
fit2 <- (comprisk.survfit2)
plot(fit2, main = "Kaplan Meier Curve of Other Causes Survival", ylab = 'Survival Probability', xlab =
```



This curve shows the Kaplan Meier estimate of survival when "Other causes" is treated as our event of interest.

```
comprisk.survfit3 <- survfit(vec3 ~1, data = comprisk)
fit3 <- (comprisk.survfit3)
plot(fit3, main = "Kaplan Meier Curve of All Events Survival", ylab = 'Survival Probability', xlab = 'Da
```



This curve shows the Kaplan Meier estimate of survival when any outcome is treated as our event of interest.

Coxph Models

The Cox proportional hazards model estimates the hazard function. Proportional hazards models, or Cox proportional hazards models, allow us to investigate the association between a set of covariates and the event of interest.

Here are the various coxph models when we include gender as our covariate.

```
cox1 <- coxph(vec1~comprisk.gender, data = comprisk)
summary(cox1)
```

```
## Call:
## coxph(formula = vec1 ~ comprisk.gender, data = comprisk)
##
##   n= 453, number of events= 167
##
##                    coef exp(coef) se(coef)     z Pr(>|z|)
## comprisk.gender 0.4171    1.5176   0.1548 2.694  0.00705 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## comprisk.gender     1.518     0.6589      1.12     2.055
##
## Concordance= 0.539  (se = 0.02 )
## Likelihood ratio test= 7.21  on 1 df,   p=0.007
## Wald test            = 7.26  on 1 df,   p=0.007
## Score (logrank) test = 7.37  on 1 df,   p=0.007
```

We can interpret these results by noticing the coefficient = 0.4171. This number is positive, therefore it positively affects the hazard ration and hence negatively influences the survival function. Additionally, the exp(coef) = 1.1340 which means there is a 113% increase in the hazard rate for males. This means that being a female in the study positively affects your survival probability compared to being a male. p = 0.254 is greater than alpha = 0.05, so we fail to reject the null hypothesis and conclude that it is not statistically significant to include which sex in our model.

```
cox2 <- coxph(vec2~comprisk.gender, data = comprisk)
summary(cox2)
```

```
## Call:
## coxph(formula = vec2 ~ comprisk.gender, data = comprisk)
##
##   n= 453, number of events= 170
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## comprisk.gender -0.1749    0.8395   0.1609 -1.087    0.277
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## comprisk.gender    0.8395      1.191    0.6124     1.151
##
## Concordance= 0.514  (se = 0.023 )
## Likelihood ratio test= 1.2  on 1 df,   p=0.3
## Wald test            = 1.18  on 1 df,   p=0.3
## Score (logrank) test = 1.19  on 1 df,   p=0.3
```

We can interpret these results by noticing the coefficient = -0.1529. This number is negative, therefore it negatively affects the hazard ration and hence positively influences the survival function. Additionally, the exp(coef) = 0.8582 which means there is a 86% increase in the hazard rate for males. This means that being a female in the study positively affects your survival probability compared to being a male. p = 0.217 is greater than alpha = 0.05, so we fail to reject the null hypothesis and conclude that it is not statistically significant to include which sex in our model.

```
cox3 <- coxph(vec3~comprisk.gender, data = comprisk)
summary(cox3)
```

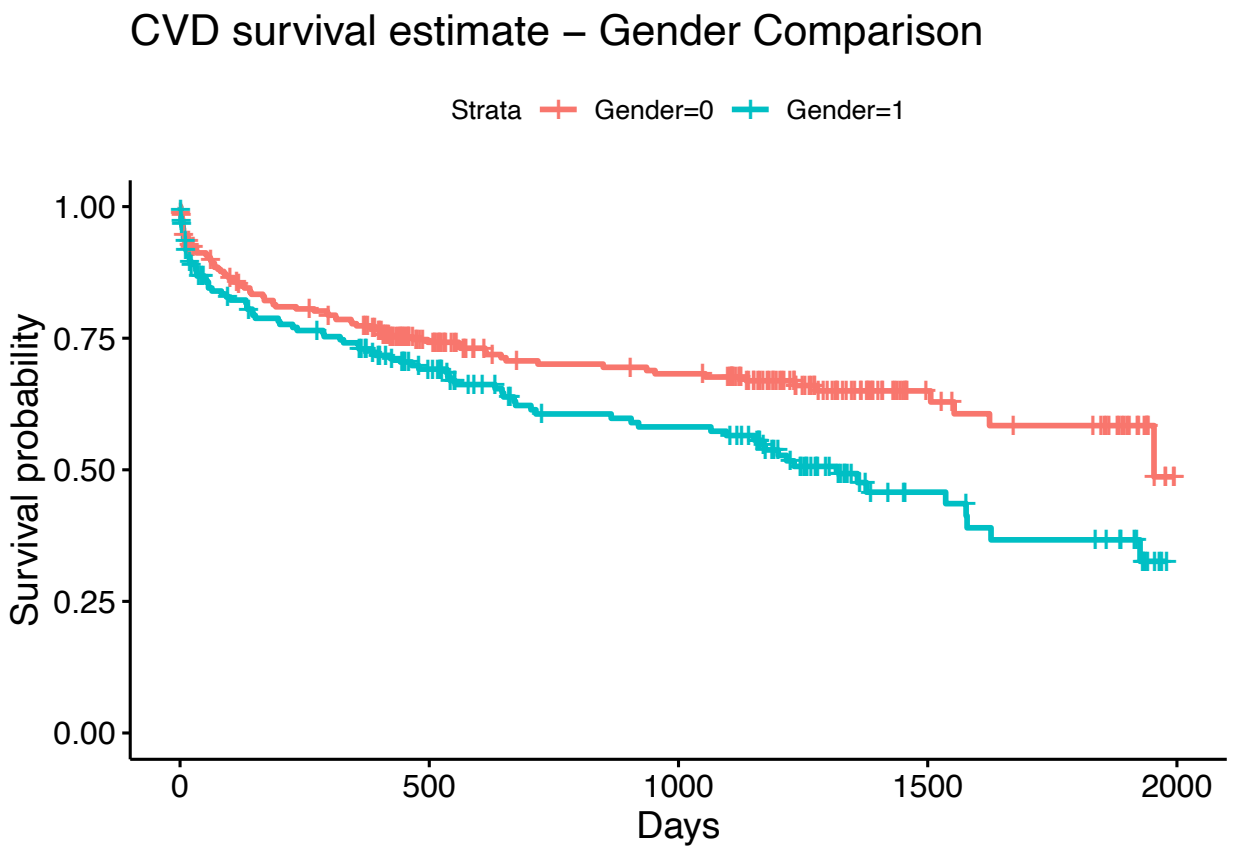
```
## Call:
## coxph(formula = vec3 ~ comprisk.gender, data = comprisk)
##
##   n= 453, number of events= 337
##
##                     coef exp(coef) se(coef)    z Pr(>|z|)
## comprisk.gender 0.1258    1.1340   0.1103 1.14    0.254
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## comprisk.gender     1.134     0.8818    0.9136     1.408
##
## Concordance= 0.519  (se = 0.015 )
## Likelihood ratio test= 1.29  on 1 df,   p=0.3
## Wald test            = 1.3  on 1 df,   p=0.3
## Score (logrank) test = 1.3  on 1 df,   p=0.3
```

We can interpret these results by noticing the coefficient = 0.1258. This number is positive, therefore it positively affects the hazard ration and hence negatively influences the survival function. Additionally, the

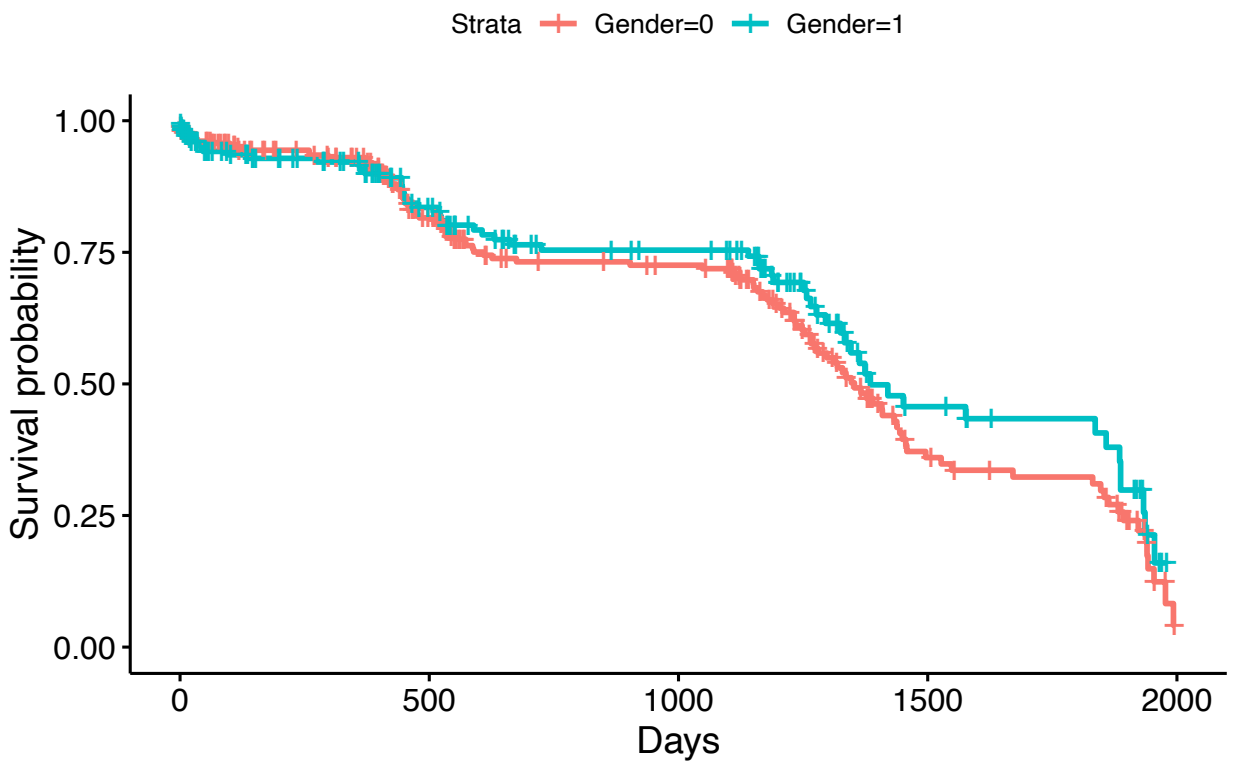
exp(coef) = 1.226 which means there is a 123% increase in the hazard rate for males. This means that being a female in the study positively affects your survival probability compared to being a male. p = 0.0892 is greater than alpha = 0.05, so we fail to reject the null hypothesis and conclude that it is not statistically significant to include which sex in our model.

We want to compare how Gender has an effect on Survival from a CVD event versus a Non Cvd Event

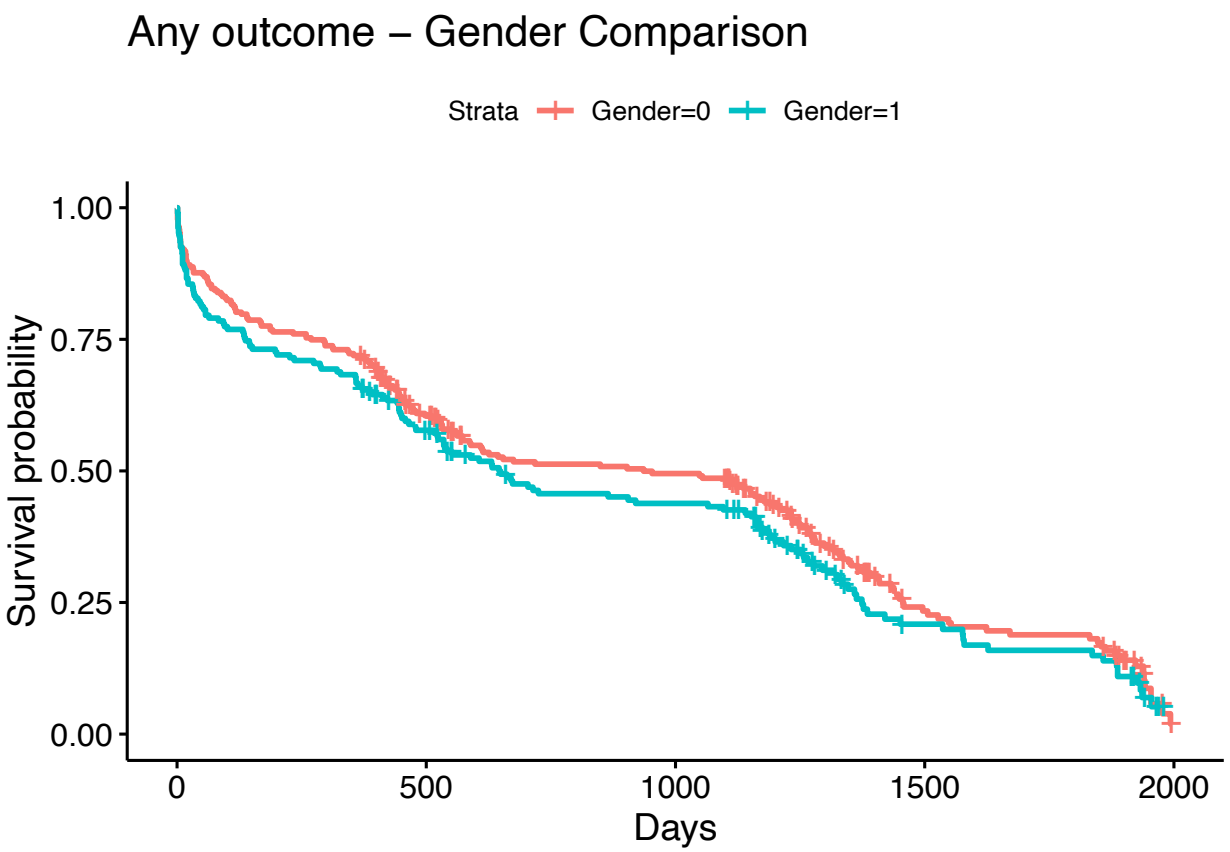
```
fit1_gender <- survfit(vec1~Gender, data = comprisk)
fit2_gender <- survfit(vec2~Gender, data = comprisk)
fit3_gender <- survfit(vec3~Gender, data = comprisk)

ggsurvplot(fit1_gender, comprisk, title = "CVD survival estimate - Gender Comparison", xlab = "Days"  )
```




```
ggsurvplot(fit2_gender, comprisk, title = "Other Causes survival estimate - Gender Comparison", xlab =
```




```
ggsurvplot(fit3_gender, comprisk , title ="Any outcome - Gender Comparison", xlab = "Days")
```



Each of these Curves compare Male (0) and Female (1), for each of our different survival vectors. The first plot compares gender when considering CVD as our event, the second plot compares gender when considering other causes as our event, and the third plot compares gender for any outcome.

After analysis gender is not statistically significant from these p-values.

Proportional hazards models, or Cox proportional hazards models, allow us to investigate the association between a set of covariates and the event of interest. The Cox proportional hazards model estimates the hazard function.

Now we are creating coxph models with age as our covariate of interest.

```
coxage1 <- coxph(vec1~comprisk.age, data = comprisk)
summary(coxage1)
```

```
## Call:
## coxph(formula = vec1 ~ comprisk.age, data = comprisk)
##
##   n= 453, number of events= 167
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## comprisk.age 0.084115  1.087754 0.005687 14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## comprisk.age     1.088     0.9193     1.076       1.1
##
```

```
## Concordance= 0.848  (se = 0.013 )
## Likelihood ratio test= 288.4  on 1 df,   p=<2e-16
## Wald test            = 218.8  on 1 df,   p=<2e-16
## Score (logrank) test = 259.9  on 1 df,   p=<2e-16
```

We can interpret these results by noticing the coefficient = 0.084115. This number is positive, therefore it positively affects the hazard ration and hence negatively influences the survival function. p = 1.06e-13 is less than alpha = 0.05, so we reject the null hypothesis and conclude that it is statistically significant to include age in our model.

```
coxage2 <- coxph(vec2~comprisk.age, data = comprisk)
summary(coxage2)
```

```
## Call:
## coxph(formula = vec2 ~ comprisk.age, data = comprisk)
##
##   n= 453, number of events= 170
##
##                   coef exp(coef)  se(coef)      z Pr(>|z|)
## comprisk.age -0.027887  0.972498  0.004581 -6.087 1.15e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## comprisk.age    0.9725      1.028    0.9638    0.9813
##
## Concordance= 0.617  (se = 0.023 )
## Likelihood ratio test= 38.02  on 1 df,   p=7e-10
## Wald test            = 37.06  on 1 df,   p=1e-09
## Score (logrank) test = 37.61  on 1 df,   p=9e-10
```

We can interpret these results by noticing the coefficient = -0.027887. This number is negative, therefore it negatively affects the hazard ration and hence positively influences the survival function. p = 8.77e-06 is less than alpha = 0.05, so we reject the null hypothesis and conclude that it is statistically significant to include age in our model.

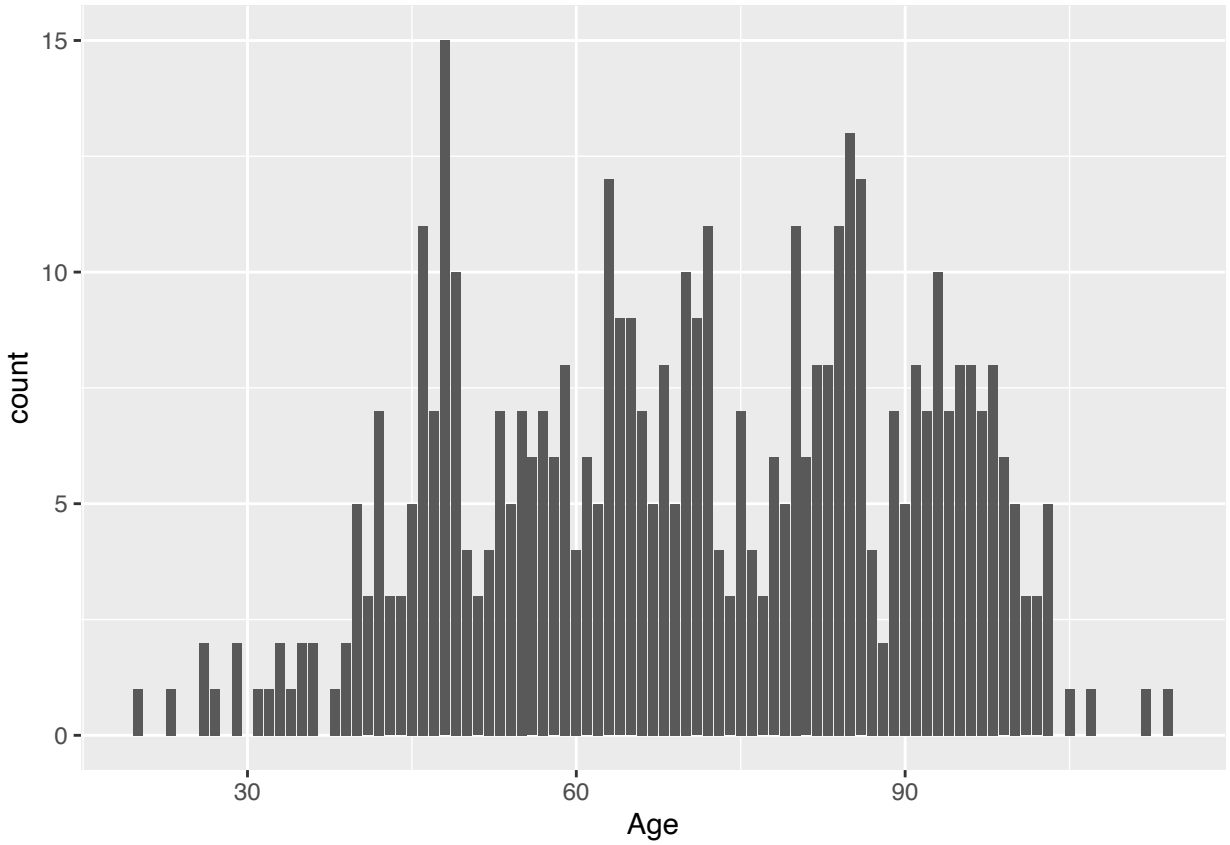
```
coxage3 <- coxph(vec3~comprisk.age, data = comprisk)
summary(coxage3)
```

```
## Call:
## coxph(formula = vec3 ~ comprisk.age, data = comprisk)
##
##   n= 453, number of events= 337
##
##                  coef exp(coef) se(coef)     z Pr(>|z|)
## comprisk.age 0.02334   1.02361  0.00314 7.433 1.06e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## comprisk.age     1.024     0.9769     1.017      1.03
##
```

```
## Concordance= 0.666  (se = 0.014 )
## Likelihood ratio test= 56.67  on 1 df,   p=5e-14
## Wald test            = 55.24  on 1 df,   p=1e-13
## Score (logrank) test = 56.24  on 1 df,   p=6e-14
```

We can interpret these results by noticing the coefficient = 0.02334. This number is positive, therefore it positively affects the hazard ration and hence negatively influences the survival function. p <2e-16 is less than alpha = 0.05, so we reject the null hypothesis and conclude that it is statistically significant to include age in our model.

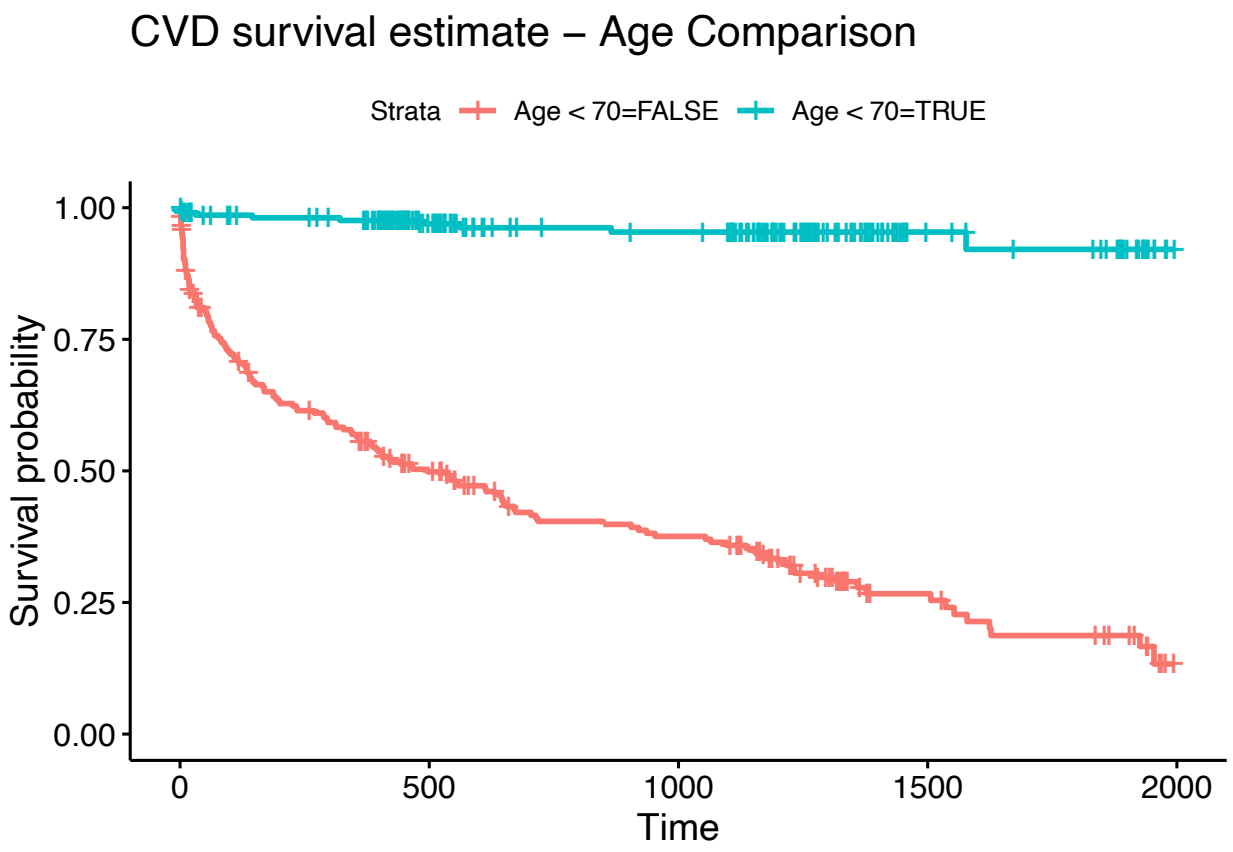
```
ggplot(comprisk)+
  geom_bar(aes(x =Age))
```



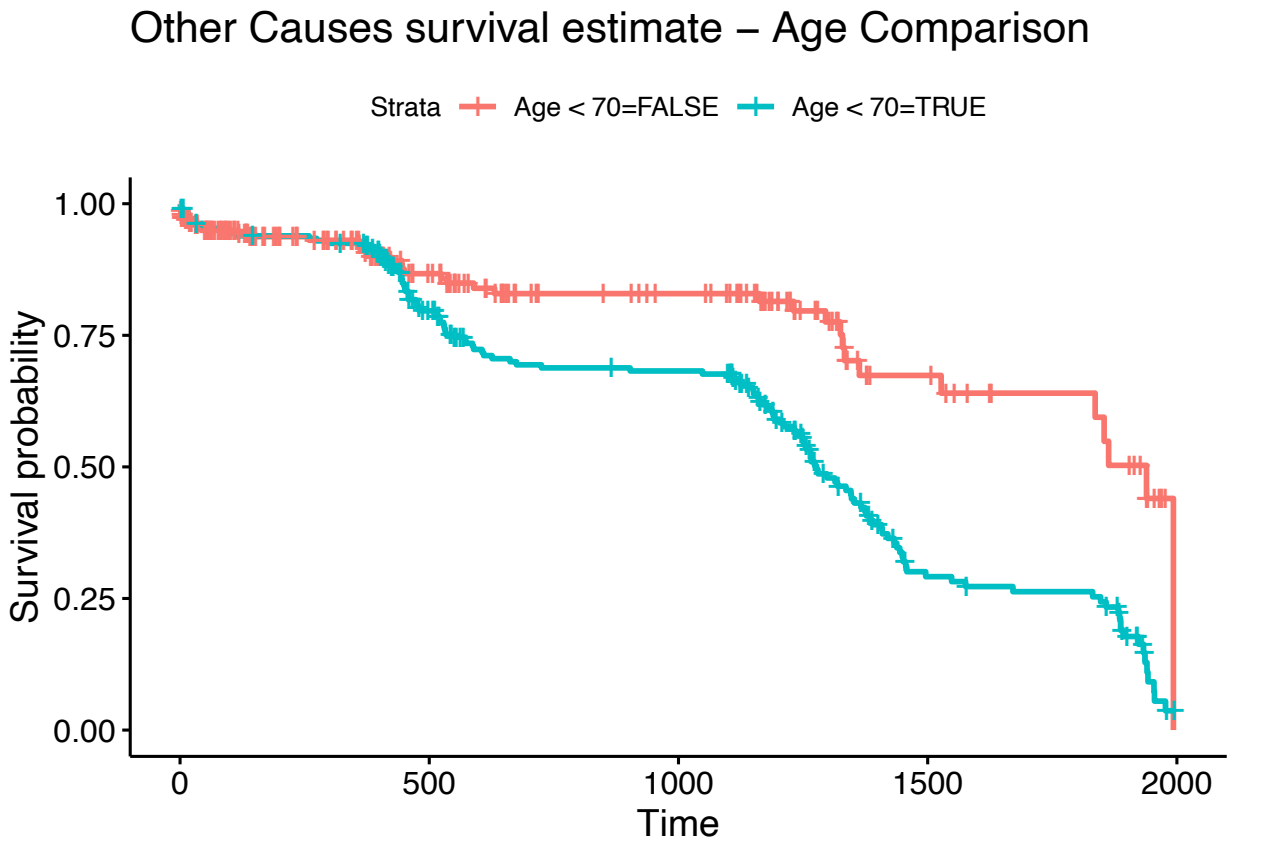

```
mean(comprisk.age)
```

```
## [1] 70.48344
```

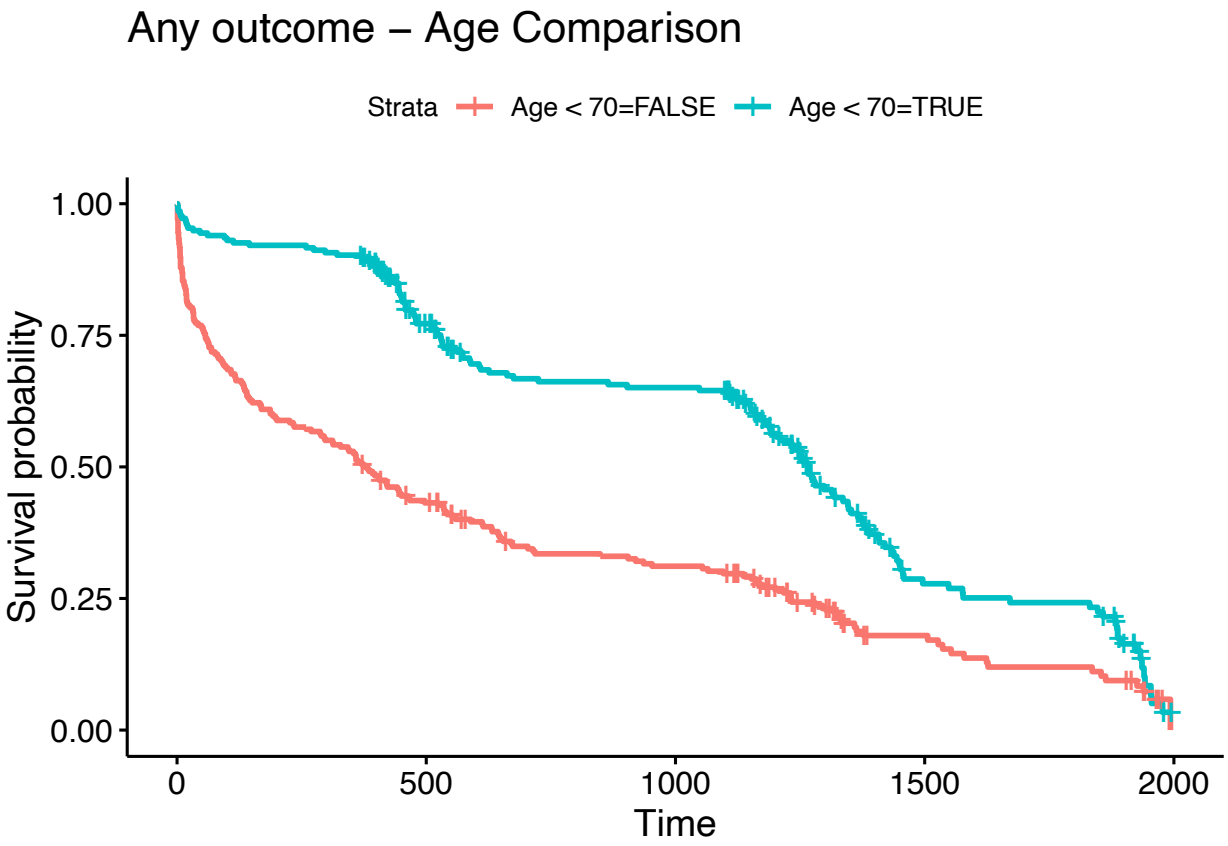
```
fit1_age <- survfit(vec1~Age<70, data = comprisk)
fit2_age <- survfit(vec2~Age<70, data = comprisk)
fit3_age <- survfit(vec3~Age<70, data = comprisk)

ggsurvplot(fit1_age, comprisk, title = "CVD survival estimate - Age Comparison"  )
```




```
ggsurvplot(fit2_age, comprisk, title = "Other Causes survival estimate - Age Comparison" )
```




```
ggsurvplot(fit3_age, comprisk , title ="Any outcome - Age Comparison")
```



In these plots we firstly plotted our Age distribution to see how many patients of each age there were. Then using the mean of 70 years old, we plotted survival curves. Each of the plots compare subjects under 70 years of age, and those 70 years or older. The first plot compares this age split when considering CVD as our event, the second plot compares the age split when considering other causes as our event, and the third plot compares the age split for any outcome.

Age has a statistically significant effect on our survival time for all of our coxph models.

We are creating coxph models with bmi as the covariate of interest.

```
coxbmi1 <- coxph(vec1~comprisk.bmi, data = comprisk)
summary(coxbmi1)
```

```
## Call:
## coxph(formula = vec1 ~ comprisk.bmi, data = comprisk)
##
##   n= 453, number of events= 167
##
##                 coef exp(coef) se(coef)     z Pr(>|z|)
## comprisk.bmi 0.06247   1.06446  0.01279 4.885 1.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## comprisk.bmi     1.064     0.9394     1.038     1.091
##
## Concordance= 0.587  (se = 0.023 )
## Likelihood ratio test= 22.22  on 1 df,   p=2e-06
```

```
## Wald test            = 23.86  on 1 df,   p=1e-06
## Score (logrank) test = 23.85  on 1 df,   p=1e-06
```

We can interpret these results by noticing the coefficient = 0.003229. This number is positive, therefore it positively affects the hazard ration and hence negatively influences the survival function. p = 0.747 is greater than alpha = 0.05, so we fail to reject the null hypothesis and conclude that it is statistically significant to include bmi in our model.

```
coxbmi2 <- coxph(vec2~comprisk.bmi, data = comprisk)
summary(coxbmi2)
```

```
## Call:
## coxph(formula = vec2 ~ comprisk.bmi, data = comprisk)
##
##   n= 453, number of events= 170
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## comprisk.bmi -0.07002   0.93237  0.01577 -4.439 9.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## comprisk.bmi    0.9324      1.073     0.904    0.9616
##
## Concordance= 0.634  (se = 0.025 )
## Likelihood ratio test= 21.09  on 1 df,   p=4e-06
## Wald test            = 19.7  on 1 df,   p=9e-06
## Score (logrank) test = 19.69  on 1 df,   p=9e-06
```

We can interpret these results by noticing the coefficient = -0.03118. This number is negative, therefore it negatively affects the hazard ration and hence positively influences the survival function. p = 0.00695 is less than alpha = 0.05, so we reject the null hypothesis and conclude that it is statistically significant to include bmi in our model.

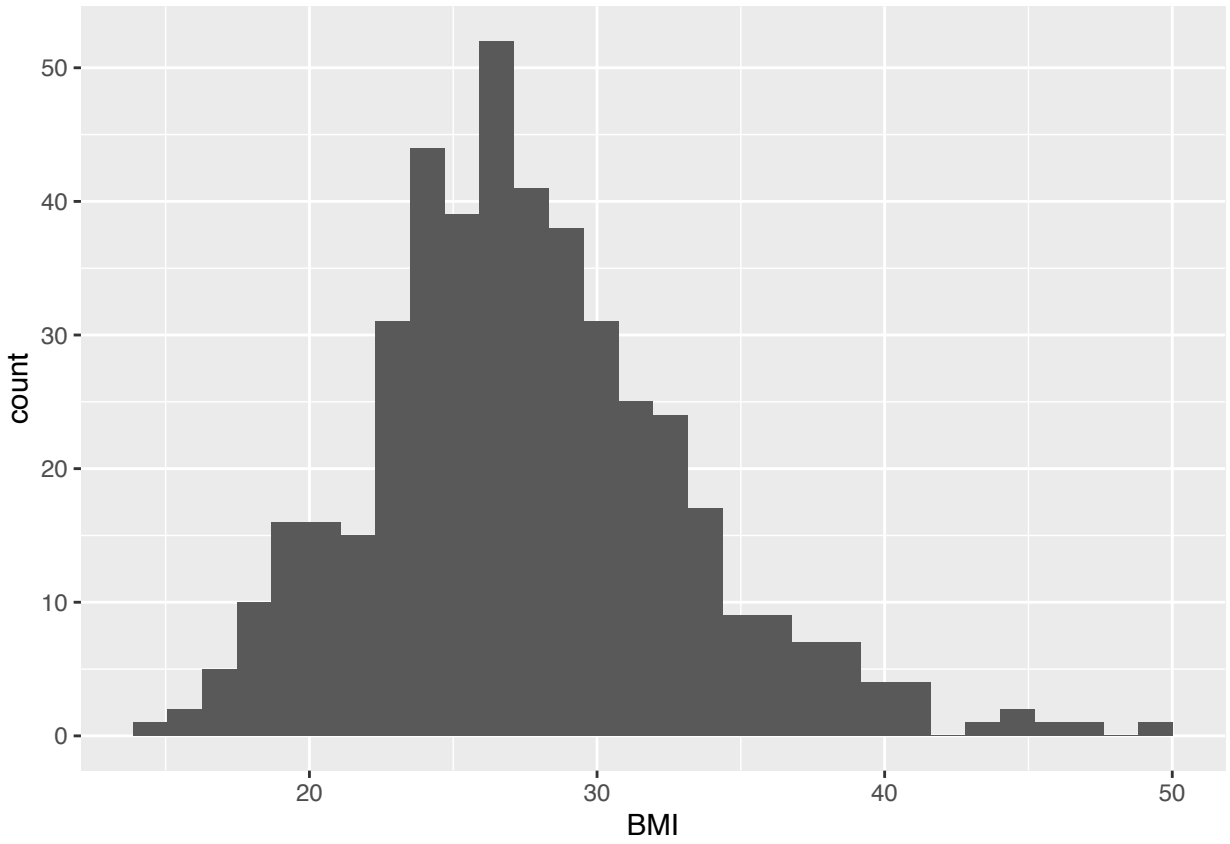
```
coxbmi3 <- coxph(vec3~comprisk.bmi, data = comprisk)
summary(coxbmi3)
```

```
## Call:
## coxph(formula = vec3 ~ comprisk.bmi, data = comprisk)
##
##   n= 453, number of events= 337
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## comprisk.bmi 0.003229  1.003234 0.010004 0.323    0.747
##
##              exp(coef) exp(-coef) lower .95 upper .95
## comprisk.bmi     1.003     0.9968    0.9838     1.023
##
## Concordance= 0.5  (se = 0.017 )
## Likelihood ratio test= 0.1  on 1 df,   p=0.7
## Wald test            = 0.1  on 1 df,   p=0.7
## Score (logrank) test = 0.1  on 1 df,   p=0.7
```

We can interpret these results by noticing the coefficient = 0.04637. This number is positive, therefore it positively affects the hazard ration and hence negatively influences the survival function. p = 5.93e-06 is less than alpha = 0.05, so we reject the null hypothesis and conclude that it is statistically significant to include bmi in our model.

```
ggplot(comprisk)+
  geom_histogram(aes(x =BMI))
```

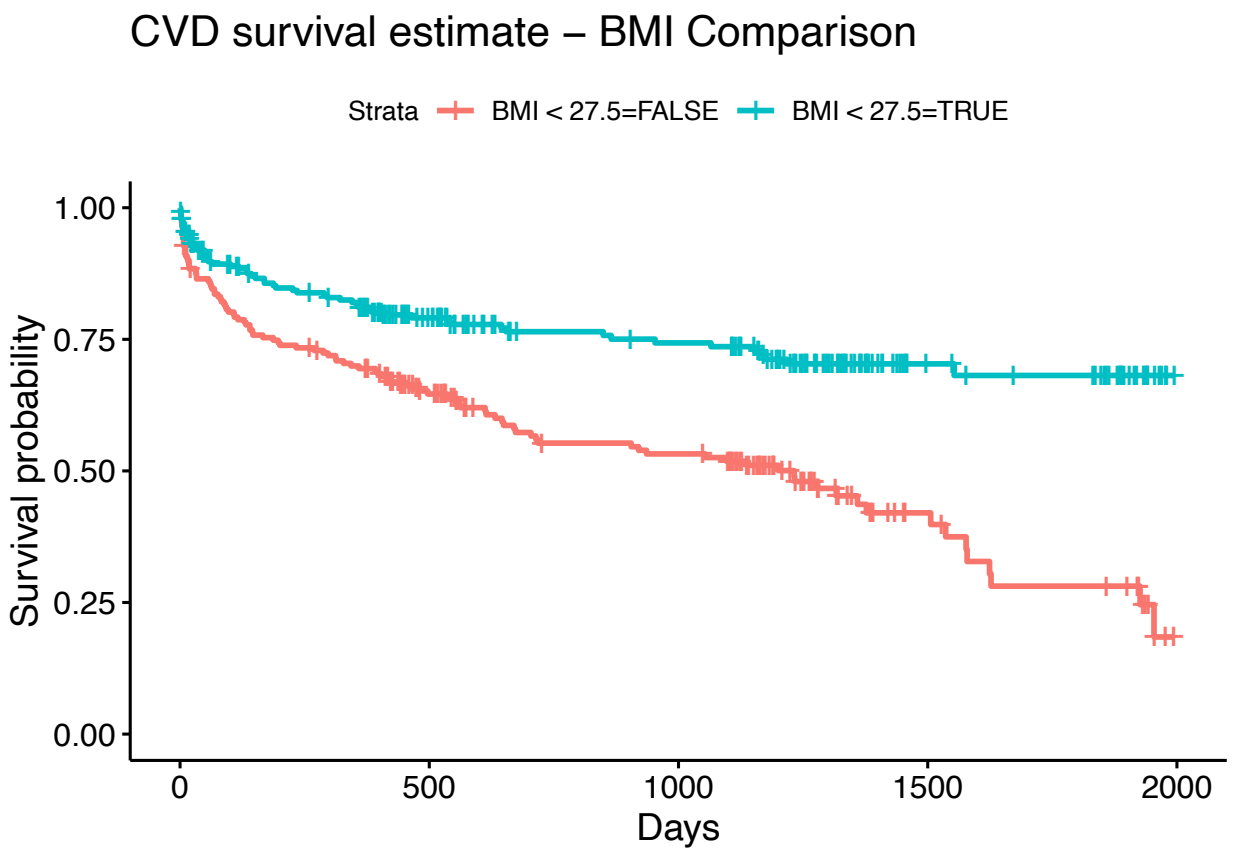
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



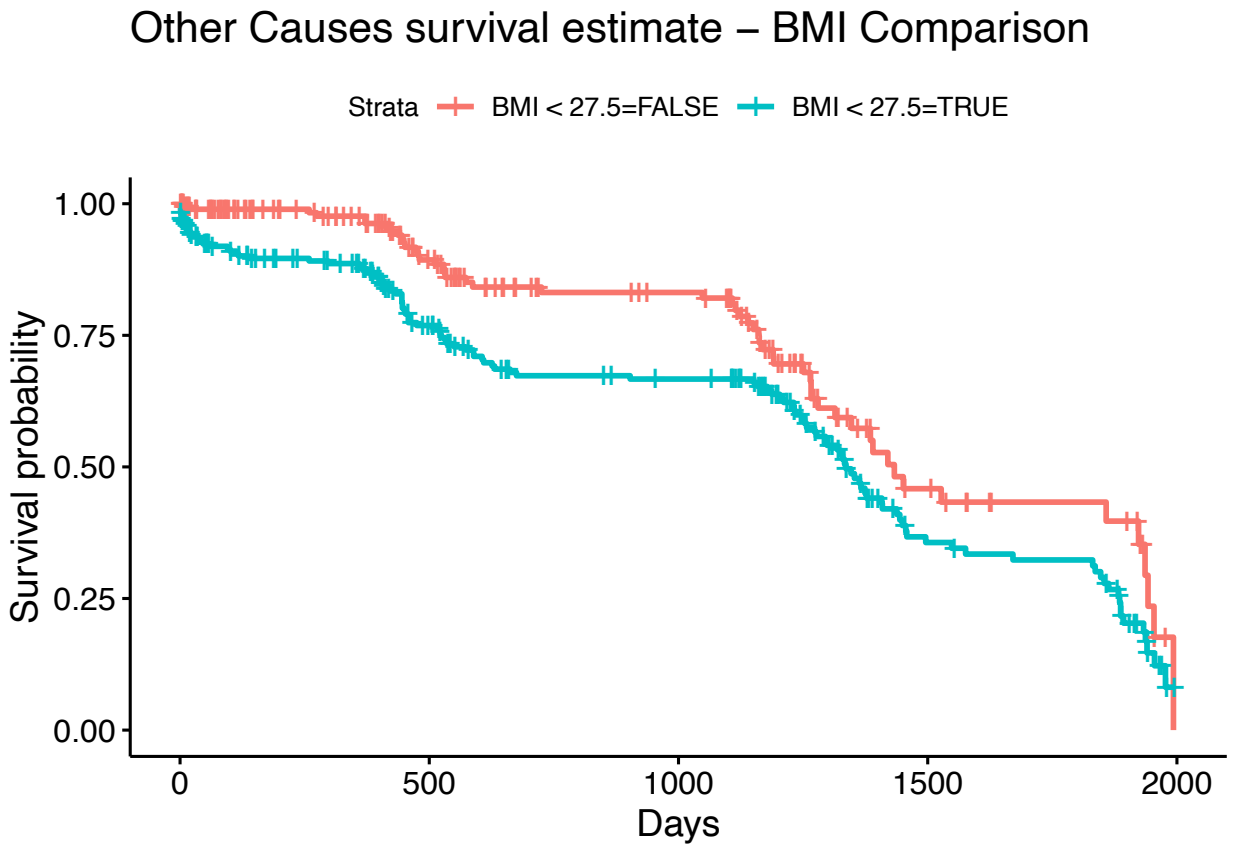

```
mean(comprisk.bmi)
```

```
## [1] 27.59502
```

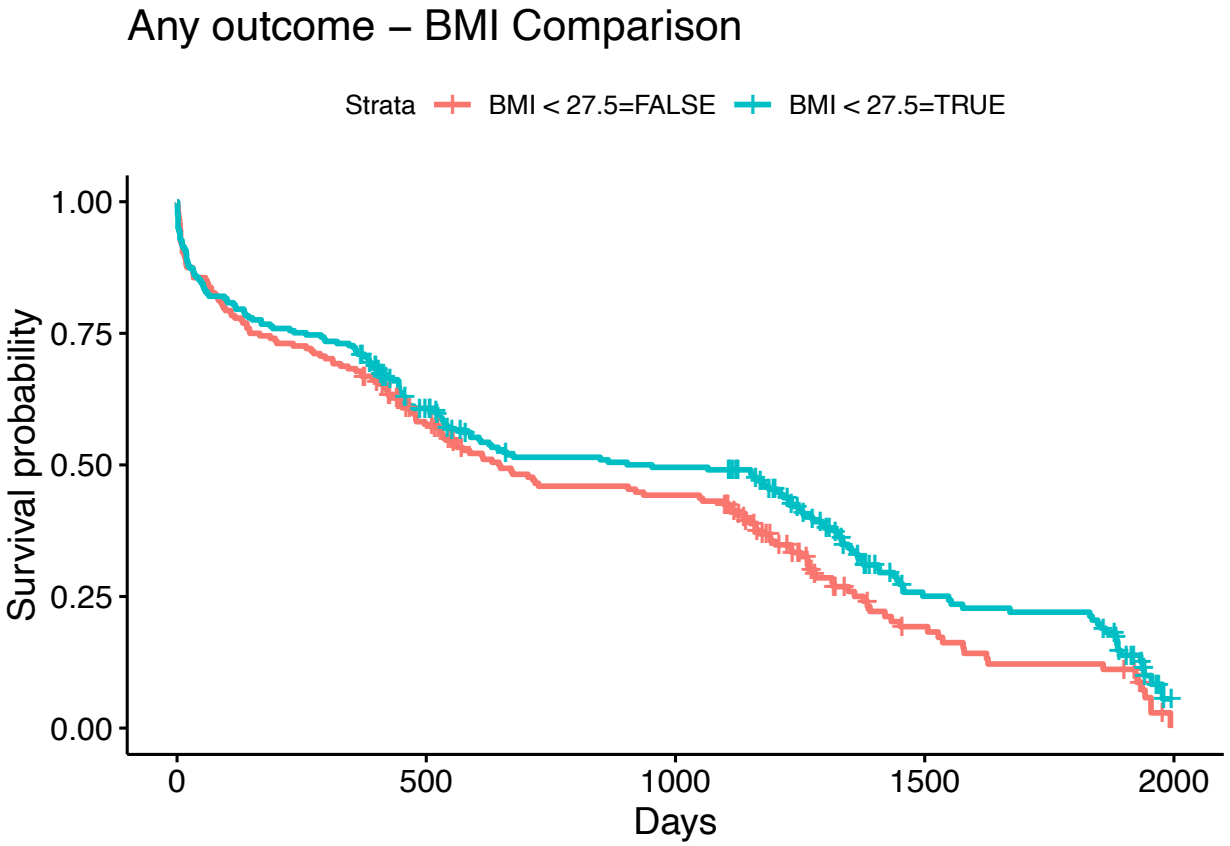
```
fit1_age <- survfit(vec1~BMI<27.5, data = comprisk)
fit2_age <- survfit(vec2~BMI<27.5, data = comprisk)
fit3_age <- survfit(vec3~BMI<27.5, data = comprisk)

ggsurvplot(fit1_age, comprisk, title = "CVD survival estimate - BMI Comparison", xlab = "Days"  )
```




```
ggsurvplot(fit2_age, comprisk, title = "Other Causes survival estimate - BMI Comparison" , xlab = "Days
```




```
ggsurvplot(fit3_age, comprisk , title ="Any outcome - BMI Comparison", xlab = "Days" )
```



In plotting our BMI comparisons we took a similar approach as we did with age. We analyzed the distribution of BMI among our subject, then calculated the mean which was equal to about 27.5. Each of the plots compare subjects under 27.5 BMI, and those over 27.5 BMI. The first plot compares this BMI split when considering CVD as our event, the second plot compares the BMI split when considering other causes as our event, and the third plot compares the BMI split for any outcome.

BMI DOES seem to have a statistically significant effect on our survival time model.

AIC and Stepwise Function

```
comprisk.surv <- Surv(comprisk.time,status_cvd_event)
fit1 <- coxph(comprisk.surv ~ comprisk.gender+ comprisk.age+ comprisk.bmi, data = comprisk)
fit2 <- coxph(comprisk.surv ~ 1, comprisk)
stepAIC(fit2, direction = "forward",
        scope = list(upper=fit1, lower=fit2))
```

```
## Start:  AIC=1871.32
## comprisk.surv ~ 1
##
##                    Df    AIC
## + comprisk.age      1 1584.9
## + comprisk.bmi      1 1851.1
## + comprisk.gender   1 1866.1
## <none>                1871.3
##
## Step:  AIC=1584.89
## comprisk.surv ~ comprisk.age
##
```

```
##                   Df    AIC
## + comprisk.bmi     1 1570.0
## <none>               1584.9
## + comprisk.gender  1 1586.1
##
## Step:  AIC=1569.97
## comprisk.surv ~ comprisk.age + comprisk.bmi
##
##                   Df    AIC
## <none>               1570.0
## + comprisk.gender  1 1571.1
```

```
## Call:
## coxph(formula = comprisk.surv ~ comprisk.age + comprisk.bmi,
##     data = comprisk)
##
##                   coef exp(coef) se(coef)      z        p
## comprisk.age 0.084645  1.088331 0.005884 14.386  < 2e-16
## comprisk.bmi 0.061571  1.063506 0.014548  4.232 2.31e-05
##
## Likelihood ratio test=305.3  on 2 df, p=< 2.2e-16
## n= 453, number of events= 167
```

```
coxph(formula = comprisk.surv ~ comprisk.age + comprisk.bmi,
    data = comprisk)
```

```
## Call:
## coxph(formula = comprisk.surv ~ comprisk.age + comprisk.bmi,
##     data = comprisk)
##
##                   coef exp(coef) se(coef)      z        p
## comprisk.age 0.084645  1.088331 0.005884 14.386  < 2e-16
## comprisk.bmi 0.061571  1.063506 0.014548  4.232 2.31e-05
##
## Likelihood ratio test=305.3  on 2 df, p=< 2.2e-16
## n= 453, number of events= 167
```

We have discovered the best model for competing risk model from the previous line of code, and we conclude it is the best because it has the lowest AIC (1569.72). The model includes the covariates age and bmi, while excluding gender from the model. This conclusion is in agreement with our conclusions drawn from our coxph analysis.

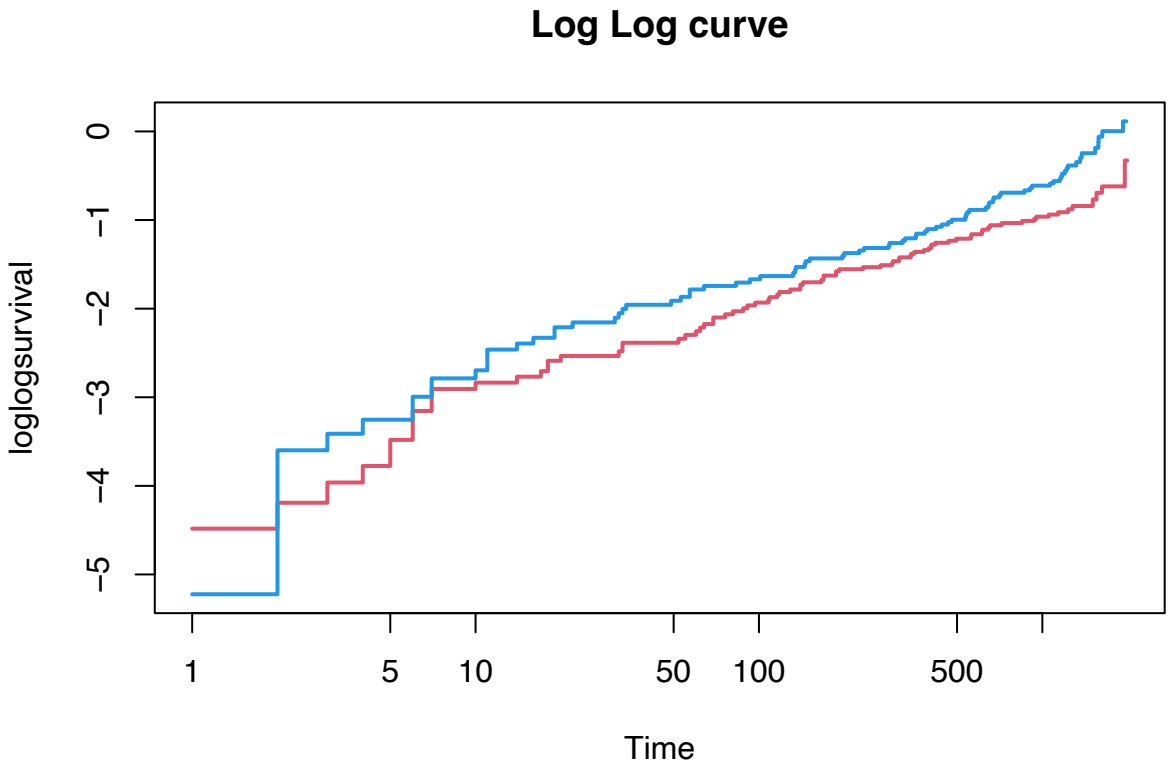
Testing Proportional Hazards Assumptions

The assumptions necessary for inference include:

Analyzing a log-log plot, and analyzing the coxzph.

This is the log-log Plot for Gender.

```
plot(survfit(vec1~Gender,data = comprisk), fun = 'cloglog', xlab = 'Time', col = c(2,4),lwd = 2, ylab =
```



Since gender is our only discrete variable that we are looking at, we created a log-log plot to check our proportional hazards assumption. From the plot, we can see that the lines are mostly parallel and intersect only due to the effect being not particularly significant.

```
coxph(formula = comprisk.surv ~ comprisk.age + comprisk.bmi,
    data = comprisk)
```

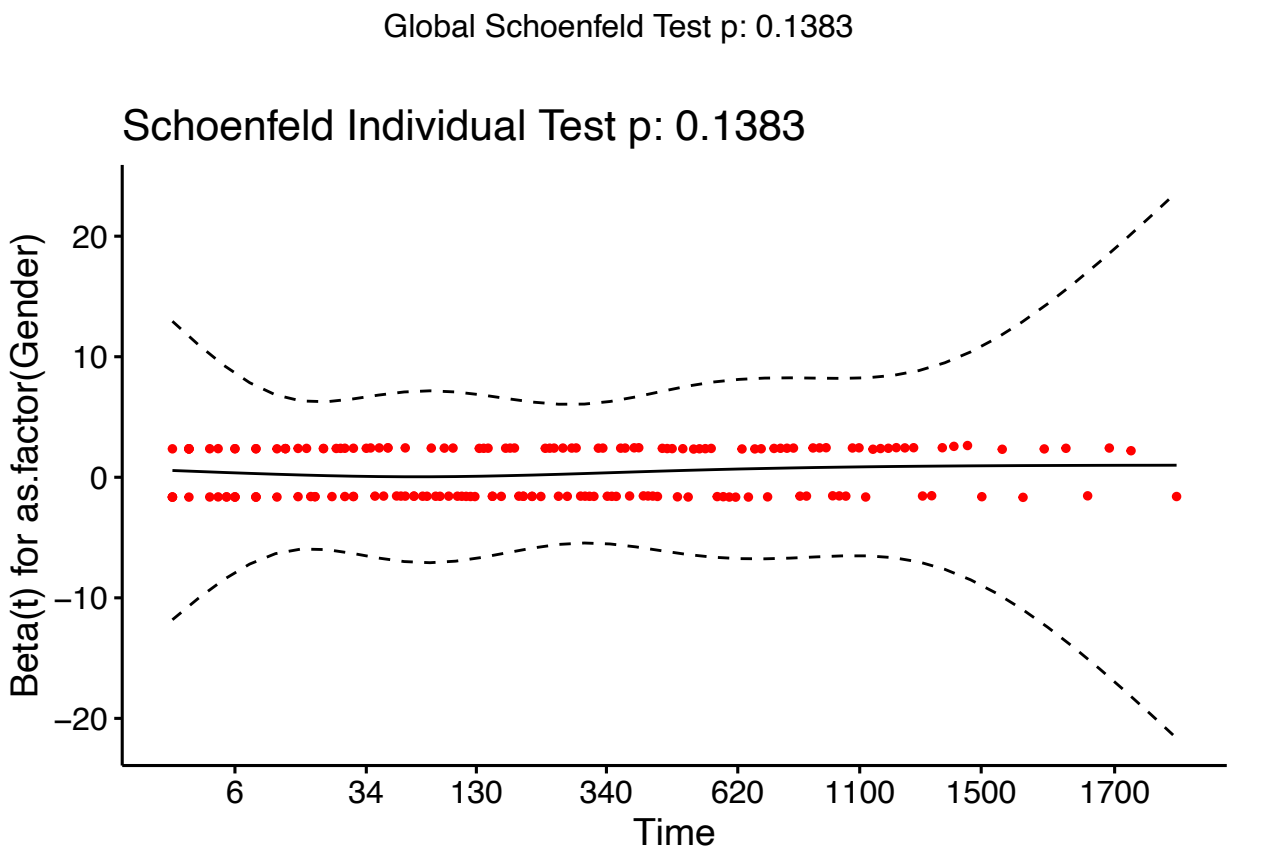
```
## Call:
## coxph(formula = comprisk.surv ~ comprisk.age + comprisk.bmi,
##     data = comprisk)
##
##                   coef exp(coef) se(coef)      z        p
## comprisk.age 0.084645  1.088331 0.005884 14.386  < 2e-16
## comprisk.bmi 0.061571  1.063506 0.014548  4.232 2.31e-05
##
## Likelihood ratio test=305.3  on 2 df, p=< 2.2e-16
## n= 453, number of events= 167
```

```
comprisk.surv.zph <- coxph(Surv(comprisk$Time,status_cvd_event)  ~ as.factor(Gender), data = comprisk)
test.ph <- cox.zph(comprisk.surv.zph)
print(test.ph)
```

```
##                     chisq df    p
## as.factor(Gender)     2.2  1 0.14
## GLOBAL                2.2  1 0.14
```

With a p-value = 0.7421, which is greater than 0.05, we fail to reject the null hypothesis and conclude that the proportional hazards assumption is met.

```
ggcoxzph(test.ph)
```



The residuals (red dots) are essentially parallel to the line y=0 which shows that there is not a significant pattern to the deviation.

To summarize this report, we've performed Cox Proportional Hazard models, indicating that age and bmi were the only statistically significant covariates in our data. We also performed AIC Step-wise selection, indicating the same results from our COXPH, that the most efficient model would be: coxph(formula = comprisk.surv ~ comprisk.age + comprisk.bmi, data = comprisk). Furthermore our cumulative incidence analysis found that the event of death from Cardiovascular Disease (CVD) is higher in female patients, and death from other causes is higher in male patients. The event of death from Cardiovascular Disease (CVD) is higher in patients older than 70, and death from other causes is higher in patients younger than 70.The event of death from Cardiovascular Disease (CVD) is higher in patients with a bmi above , and death from other causes is higher in patients with a bmi below 27.5.

References:

Hosmer, D.W. and Lemeshow, S. and May, S.
(2008) Applied
Survival Analysis: Regression Modeling of Time to Event Data: Second
Edition,
John Wiley and Sons Inc., New York, NY

Statistical Methods for Cohort Studies of CKD: Survival Analysis in the Setting of Competing Risks Jesse Yenchih Hsu,*† Jason A. Roy,*† Dawei Xie,*† Wei Yang,*† Haochang Shou,*† Amanda Hyre Anderson,*† J. Richard Landis,*† Christopher Jepson,*† Myles Wolf,‡ Tamara Isakova,§| Mahboob Rahman,¶ **†† and Harold I. Feldman,*† and on behalf of the Chronic Renal Insufficiency Cohort (CRIC) Study Investigators