3: NHANES Data Analysis

15/06/2020

0: Data Preparation

Load relevant packages:

```
library(nhanesA)
library(skimr)
library(mstate)
library(cmprsk)
library(gridExtra)
library(finalfit)
library(tidyverse)
library(ggplot2)
```

In this data analysis, the NHANES 2005-2006 data was used [1], as well as the 2015 Linked Mortality File (LMF) [2] which is contained in the file `outcome.csv'.

Import NHANES tables

To import the NHANES 2005-2006 tables, a tutorial by C.Endres was followed [3].

Import age and gender from demographic table.

```
demo_d  <- nhanes('DEMO_D') %>% select('SEQN', 'RIDAGEYR', 'RIAGENDR') %>%
mutate(RIAGENDR = factor(RIAGENDR))

## Processing SAS dataset DEMO_D     ..
```

No missing data, 1= male, 2 = female, age between 0 and 85 (all above 84 given 85)

Load body measurement data

```
bmx_d  <- nhanes('BMX_D') %>% select('SEQN', 'BMXBMI')

## Processing SAS dataset BMX_D      ..

data0 <- merge(demo_d, bmx_d, by = 'SEQN')
```

About 10% BMI data is missing.

Load blood pressure measurements :

Systolic blood pressure mmHg 1st reading

```
systolic  <- nhanes('BPX_D') %>% select('SEQN', 'BPXSY1')

## Processing SAS dataset BPX_D      ..
```

Lab measurements for cholesterol levels:

```
ldl <- nhanes('TRIGLY_D') %>% select('SEQN', 'LBDLDL')

## Processing SAS dataset TRIGLY_D   ..

data1 <- merge(data0, ldl, by = 'SEQN')
```

Load questionnare diabetes data:

```
# Ever told by doctor you have diabetes?
diq_d <- nhanes('DIQ_D') %>% select('SEQN', 'DIQ010') %>% mutate(DIQ010 =
factor(DIQ010))

## Processing SAS dataset DIQ_D      ..

diabetes_d <- nhanesTranslate('DIQ_D', 'DIQ010', data=diq_d)

## Translated columns: DIQ010

data2 <- merge(data1, diabetes_d, by = 'SEQN')




# Cigarette use questionnaire data
# Smoked at least 100 cigarettes in life?

smq_d <- nhanes('SMQ_D') %>% select('SEQN', 'SMQ020') %>% mutate(SMQ020 =
factor(SMQ020))

## Processing SAS dataset SMQ_D      ..

smoke_d <- nhanesTranslate('SMQ_D', 'SMQ020', data=smq_d)

## Translated columns: SMQ020

data3 <- merge(data2, smoke_d)

data4 <- merge(data3, systolic)

covariates <- nhanesTranslate('DEMO_D', 'RIAGENDR', data=data4)

## Translated columns: RIAGENDR

summary(covariates)

##       SEQN          RIDAGEYR      RIAGENDR        BMXBMI           LBDLDL
##  Min.   :31130   Min.   :12.0   Male  :1658   Min.   : 13.43   Min.   :
19.0
##  1st Qu.:33730   1st Qu.:18.0   Female:1694   1st Qu.: 22.23   1st Qu.:
81.0
##  Median :36354   Median :33.0                 Median : 26.27   Median
```

```
:103.0
##  Mean   :36331   Mean   :37.8                Mean   : 27.37   Mean
:106.8
##  3rd Qu.:38900   3rd Qu.:54.0                3rd Qu.: 30.95   3rd
Qu.:130.0
##  Max.   :41474   Max.   :85.0                Max.   :130.21   Max.
:328.0
##                                              NA's   :50       NA's   :326
##         DIQ010            SMQ020         BPXSY1
##  Yes       : 245   Yes       :1114   Min.   : 74.0
##  No        :3063   No        :1172   1st Qu.:108.0
##  Borderline:  41   Refused   :   0   Median :116.0
##  Don't know:   3   Don't know:   3   Mean   :120.2
##                    NA's      :1063   3rd Qu.:128.0
##                                      Max.   :224.0
##                                      NA's   :448
```

```
skim(covariates)
```

```
## Warning: Couldn't find skimmers for class: labelled, integer, numeric; No
user-
## defined `sfl` provided. Falling back to `character`.

## Warning: Couldn't find skimmers for class: labelled, integer, numeric; No
user-
## defined `sfl` provided. Falling back to `character`.

## Warning: Couldn't find skimmers for class: labelled, integer, numeric; No
user-
## defined `sfl` provided. Falling back to `character`.

## Warning: Couldn't find skimmers for class: labelled, integer, numeric; No
user-
## defined `sfl` provided. Falling back to `character`.
```

Data summary

Name	covariates
Number of rows	3352
Number of columns	8

Column type frequency:

character	4
factor	3
numeric	1

Group variables None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
SEQN	0	1.00	5	5	0	3352	0
RIDAGEYR	0	1.00	2	2	0	74	0
LBDLDL	326	0.90	2	3	0	196	0
BPXSY1	448	0.87	2	3	0	69	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
RIAGENDR	0	1.00	FALSE	2	Fem: 1694, Mal: 1658
DIQ010	0	1.00	FALSE	4	No: 3063, Yes: 245, Bor: 41, Don: 3
SMQ020	1063	0.68	FALSE	3	No: 1172, Yes: 1114, Don: 3, Ref: 0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
BMXBMI	50	0.99	27.37	7.28	13.43	22.25	26.265	30.94	130.21	▇▂▁

Load mortality data

```
outcome = read_csv('outcome.csv') %>% select(-'X1')

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   SEQN = col_double(),
##   eligstat = col_double(),
##   mortstat = col_double(),
##   ucod_leading = col_double(),
##   permth_int = col_double(),
##   permth_exm = col_double()
## )

# Convert data types
data <- merge(covariates, outcome, by = 'SEQN') %>% mutate(SEQN =
factor(SEQN)) %>%
  mutate(eligstat = factor(eligstat),
         mortstat = factor(mortstat),
```

```
         ucod_leading = factor(ucod_leading),
         permth_int = as.numeric(permth_int),
         permth_exm = as.numeric(permth_exm),
         RIDAGEYR = as.numeric(RIDAGEYR),
         RIAGENDR = factor(RIAGENDR),
         BMXBMI = as.numeric(BMXBMI),
         LBDLDL = as.numeric(LBDLDL),
         DIQ010 = factor(DIQ010),
         BPXSY1 = as.numeric(BPXSY1),
         SMQ020 = factor(SMQ020))

summary(data)
```

```
##       SEQN         RIDAGEYR      RIAGENDR        BMXBMI           LBDLDL
##  31130  :   1   Min.   :12.0   Male  :1658   Min.   : 13.43   Min.   :
19.0
##  31131  :   1   1st Qu.:18.0   Female:1694   1st Qu.: 22.23   1st Qu.:
81.0
##  31132  :   1   Median :33.0                 Median : 26.27   Median
:103.0
##  31133  :   1   Mean   :37.8                 Mean   : 27.37   Mean
:106.8
##  31134  :   1   3rd Qu.:54.0                 3rd Qu.: 30.95   3rd
Qu.:130.0
##  31139  :   1   Max.   :85.0                 Max.   :130.21   Max.
:328.0
##  (Other):3346                                NA's   :50       NA's   :326
##         DIQ010            SMQ020         BPXSY1      eligstat mortstat
##  Yes       : 245   Yes       :1114   Min.   : 74.0   1:2565   0   :2230
##  No        :3063   No        :1172   1st Qu.:108.0   2: 785   1   : 335
##  Borderline:  41   Don't know:   3   Median :116.0   3:   2   NA's: 787
##  Don't know:   3   NA's      :1063   Mean   :120.2
##                                      3rd Qu.:128.0
##                                      Max.   :224.0
##                                      NA's   :448
##   ucod_leading    permth_int      permth_exm
##  10     : 134   Min.   :  2.0   Min.   :  1.0
##  1      :  63   1st Qu.:112.0   1st Qu.:112.0
##  2      :  62   Median :119.0   Median :118.0
##  3      :  18   Mean   :113.2   Mean   :112.4
##  5      :  15   3rd Qu.:127.0   3rd Qu.:126.0
##  (Other):  43   Max.   :132.0   Max.   :132.0
##  NA's   :3017   NA's   :787     NA's   :787
```

```
skim(data)
```

Data summary

Name	data
Number of rows	3352

Number of columns	13

Column type frequency:

factor	7
numeric	6

Group variables	None

Variable type: factor

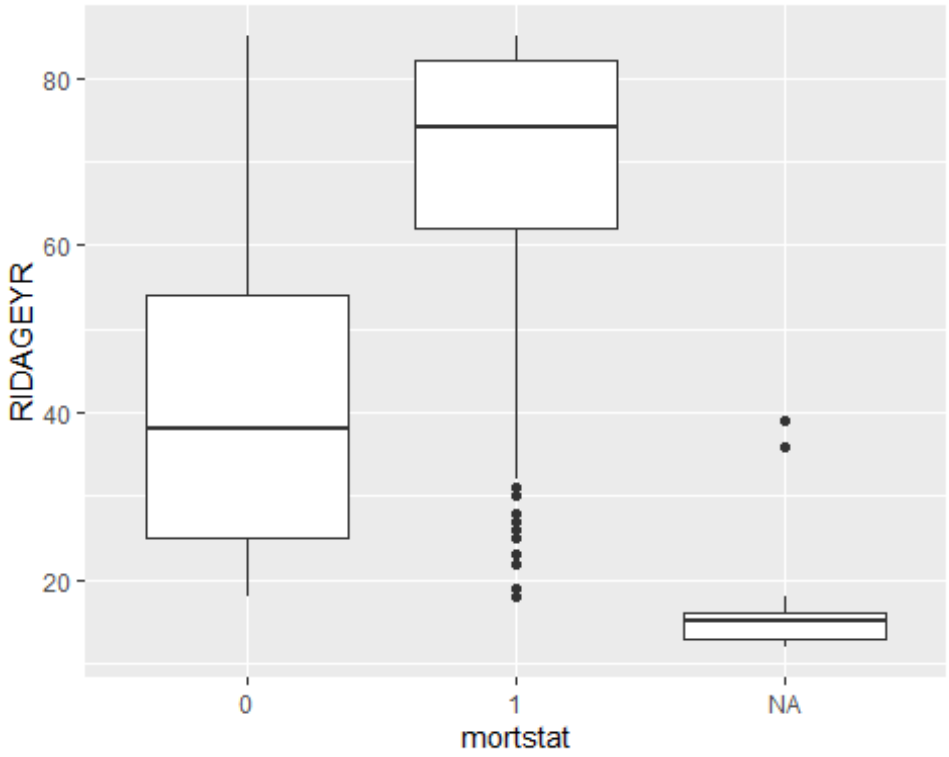
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
SEQN	0	1.00	FALSE	3352	311: 1, 311: 1, 311: 1, 311: 1
RIAGENDR	0	1.00	FALSE	2	Fem: 1694, Mal: 1658
DIQ010	0	1.00	FALSE	4	No: 3063, Yes: 245, Bor: 41, Don: 3
SMQ020	1063	0.68	FALSE	3	No: 1172, Yes: 1114, Don: 3
eligstat	0	1.00	FALSE	3	1: 2565, 2: 785, 3: 2
mortstat	787	0.77	FALSE	2	0: 2230, 1: 335
ucod_leading	3017	0.10	FALSE	10	10: 134, 1: 63, 2: 62, 3: 18

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
RIDAGEYR	0	1.00	37.80	21.84	12.00	18.00	33.00	54.00	85.00	▇▅▃▂▂
BMXBMI	50	0.99	27.37	7.28	13.43	22.22	26.27	30.95	130.21	▇▂▁▁▁
LBDLDL	326	0.90	106.78	35.83	19.00	81.00	103.00	130.00	328.00	▃▇▅▂▁
BPXSY1	448	0.87	120.22	18.69	74.00	108.00	116.00	128.00	224.00	▂▇▅▂▁
permth_int	787	0.77	113.19	24.33	2.00	112.00	119.00	127.00	132.00	▁▁▁▁▇
permth_exm	787	0.77	112.38	24.36	1.00	112.00	118.00	126.00	132.00	▁▁▁▁▇

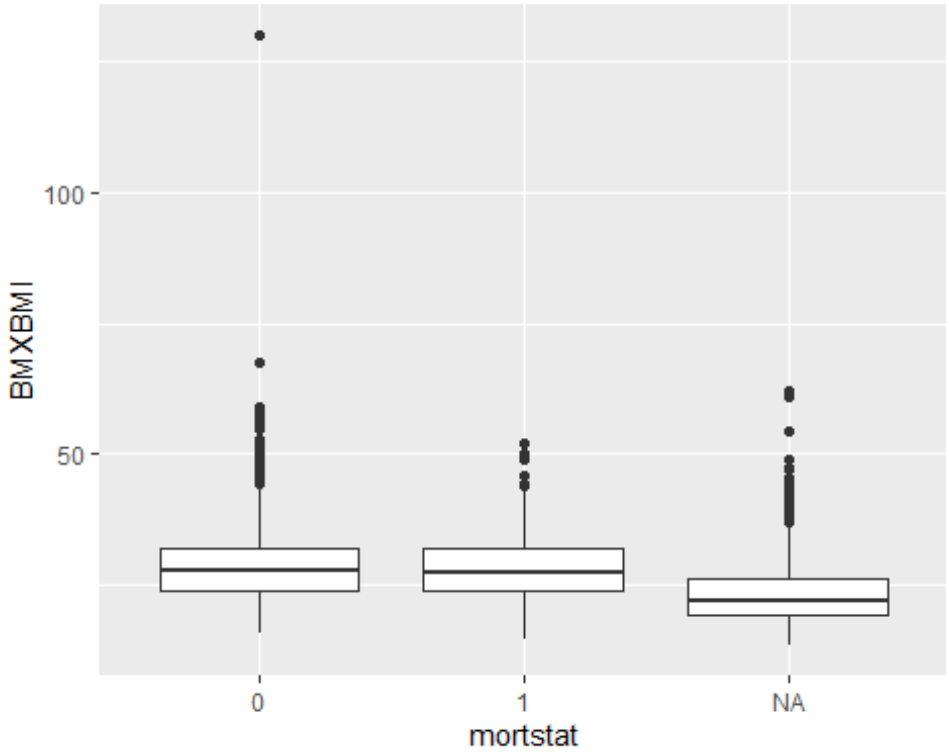
Data checks and exploration

```
ggplot(data, aes(x = mortstat, y = RIDAGEYR)) + geom_boxplot()
```



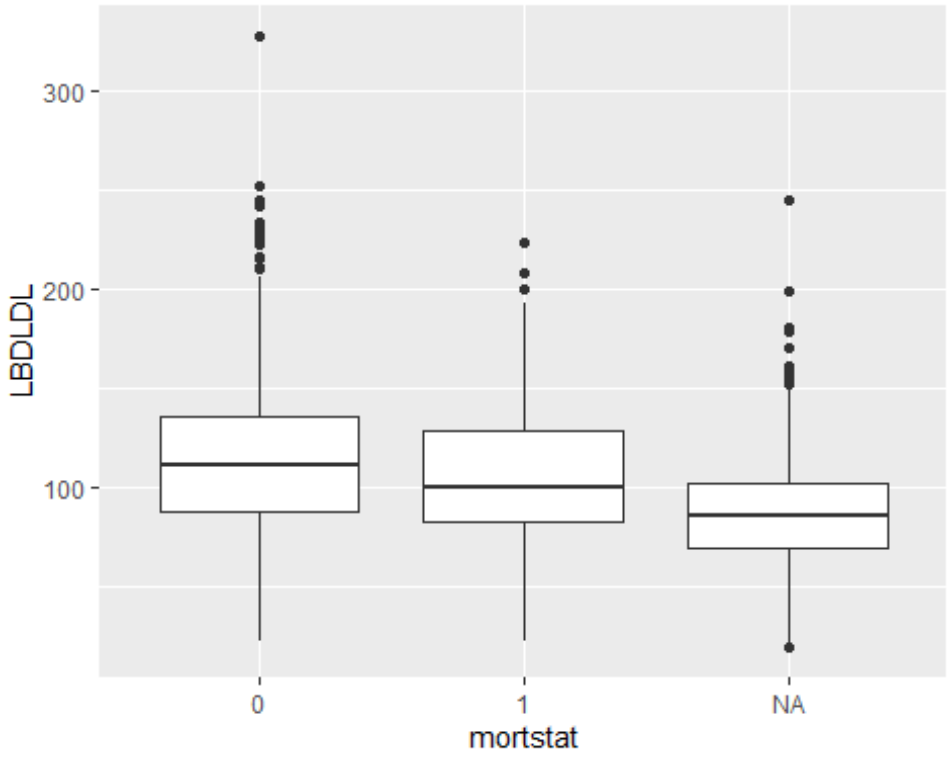

```
ggplot(data, aes(x = mortstat, y = BMXBMI)) + geom_boxplot()
```

```
## Warning: Removed 50 rows containing non-finite values (stat_boxplot).
```



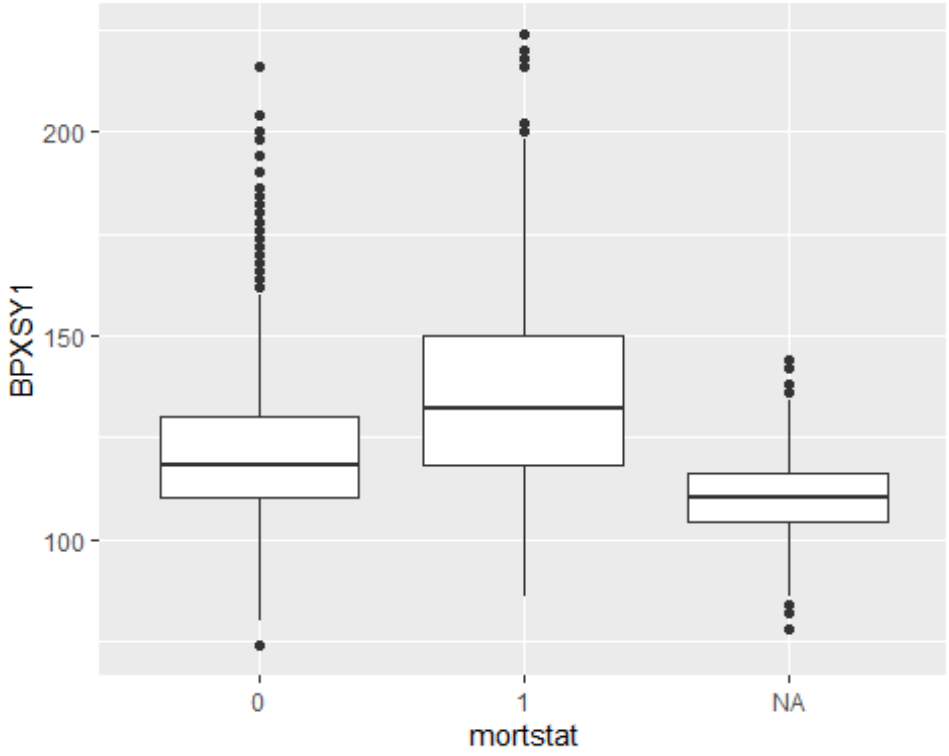

```
ggplot(data, aes(x = mortstat, y = LBDLDL)) + geom_boxplot()
```

```
## Warning: Removed 326 rows containing non-finite values (stat_boxplot).
```



```r
ggplot(data, aes(x = mortstat, y = BPXSY1)) + geom_boxplot()
```

```
## Warning: Removed 448 rows containing non-finite values (stat_boxplot).
```



Change data format

Change format of ucod_leading so that :

- 0 = Ineligible, under age 18, assumed alive, or no cause of death data
- 1 = death from disease of the heart
- 2 = maligant neoplasms (cancer)
- 3 = all other causes e.g. accidents, respiratory, diabetes etc.

```r
# First convert 0s to NA's
summary(data$ucod_leading)
```

```
##    1    2    3    4    5    6    7    8    9   10 NA's
##   63   62   18   10   15    8    9    6   10  134 3017
```

```r
levels(data$ucod_leading)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```

```r
levels(data$ucod_leading) <- c(levels(data$ucod_leading),"0")

data$ucod_leading[is.na(data$ucod_leading)] <- 0
summary(data$ucod_leading)
```

```
##    1    2    3    4    5    6    7    8    9   10    0
##   63   62   18   10   15    8    9    6   10  134 3017

# Convert 3-10 to 3
cause <- data.frame(ucod_leading=c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10), cause
=c(0, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3))
df2 <- merge(data,cause, by = 'ucod_leading') %>%
  mutate(SEQN = factor(SEQN),
         eligstat = factor(eligstat),
         mortstat = factor(mortstat),
         ucod_leading = factor(ucod_leading),
         permth_int = as.numeric(permth_int),
         permth_exm = as.numeric(permth_exm),
         RIDAGEYR = as.numeric(RIDAGEYR)/10,
         RIAGENDR = factor(RIAGENDR),
         BMXBMI = as.numeric(BMXBMI),
         LBDLDL = as.numeric(LBDLDL)/10,
         DIQ010 = factor(DIQ010),
         BPXSY1 = as.numeric(BPXSY1)/10,
         SMQ020 = factor(SMQ020),
         cause = factor(cause))


summary(df2)

##   ucod_leading       SEQN         RIDAGEYR      RIAGENDR        BMXBMI
##  0      :3017   31130  :   1   Min.   :1.20   Male  :1658   Min.   : 13.43
##  10     : 134   31131  :   1   1st Qu.:1.80   Female:1694   1st Qu.: 22.23
##  1      :  63   31132  :   1   Median :3.30                 Median : 26.27
##  2      :  62   31133  :   1   Mean   :3.78                 Mean   : 27.37
##  3      :  18   31134  :   1   3rd Qu.:5.40                 3rd Qu.: 30.95
##  5      :  15   31139  :   1   Max.   :8.50                 Max.   :130.21
##  (Other):  43   (Other):3346                                NA's   :50
##      LBDLDL             DIQ010            SMQ020         BPXSY1
eligstat
##  Min.   : 1.90   Yes       : 245   Yes       :1114   Min.   : 7.40
1:2565
##  1st Qu.: 8.10   No        :3063   No        :1172   1st Qu.:10.80   2:
785
##  Median :10.30   Borderline:  41   Don't know:   3   Median :11.60   3:
2
##  Mean   :10.68   Don't know:   3   NA's      :1063   Mean   :12.02
##  3rd Qu.:13.00                                       3rd Qu.:12.80
##  Max.   :32.80                                       Max.   :22.40
##  NA's   :326                                         NA's   :448
##  mortstat      permth_int      permth_exm    cause
##  0   :2230   Min.   :  2.0   Min.   :  1.0   0:3017
##  1   : 335   1st Qu.:112.0   1st Qu.:112.0   1:  63
##  NA's: 787   Median :119.0   Median :118.0   2:  62
##              Mean   :113.2   Mean   :112.4   3: 210
```

```
##              3rd Qu.:127.0   3rd Qu.:126.0
##              Max.   :132.0   Max.   :132.0
##              NA's   :787     NA's   :787
```

Data in long format for mstate

Convert data into long format with K rows per individual

```
head(df2)
```

```
##   ucod_leading  SEQN RIDAGEYR RIAGENDR BMXBMI LBDLDL DIQ010 SMQ020 BPXSY1
## 1            0 31133      1.6   Female  16.79    8.1     No   <NA>   12.0
## 2            0 31134      7.3     Male  30.63    9.8     No     No   13.0
## 3            0 31155      3.8     Male  25.61    9.1     No     No   12.6
## 4            0 31158      7.1     Male  21.65   11.0     No    Yes   15.4
## 5            0 31139      1.8   Female  29.45     NA     No   <NA>   11.0
## 6            0 31132      7.0     Male  24.74    7.5    Yes     No   13.8
##   eligstat mortstat permth_int permth_exm cause
## 1        2     <NA>         NA         NA     0
## 2        1        0        124        123     0
## 3        1        0        124        123     0
## 4        1        0        120        119     0
## 5        1        0        122        122     0
## 6        1        0        129        129     0
```

```
data_short <- df2 %>%
  select(-c(permth_int, mortstat, eligstat, ucod_leading)) %>%
  rename(time = permth_exm, status = cause)


causes <- data.frame(status=c(0, 1, 2, 3), cause = c("event-free", "heart",
"cancer", "other"))

data_short <- merge(data_short, causes, by = "status")
head(data_short)
```

```
##   status  SEQN RIDAGEYR RIAGENDR BMXBMI LBDLDL DIQ010 SMQ020 BPXSY1 time
## 1      0 31133      1.6   Female  16.79    8.1     No   <NA>   12.0   NA
## 2      0 31134      7.3     Male  30.63    9.8     No     No   13.0  123
## 3      0 31155      3.8     Male  25.61    9.1     No     No   12.6  123
## 4      0 31158      7.1     Male  21.65   11.0     No    Yes   15.4  119
## 5      0 31139      1.8   Female  29.45     NA     No   <NA>   11.0  122
## 6      0 31132      7.0     Male  24.74    7.5    Yes     No   13.8  129
##        cause
## 1 event-free
## 2 event-free
## 3 event-free
## 4 event-free
## 5 event-free
## 6 event-free
```

```r
# Individuals with missing time are inelidgible or under 18
summary(data_short$time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0   112.0   118.0   112.4   126.0   132.0     787
```

```r
# Remove ineligible individuals
data_short <- data_short[!is.na(data_short$time), ]

# Remove individuals with missing covariates
data_short <- na.omit(data_short)

# Remove individuals in 'Don't know' categories
data_short <- data_short %>%
  filter(!(DIQ010 == "Don't know" | SMQ020 == "Don't know"))


data_short$DIQ010 <- droplevels(data_short$DIQ010, exclude = "Don't know")
data_short$SMQ020 <- droplevels(data_short$SMQ020, exclude = "Don't know")



# Reorder
data_short <- data_short %>% select(SEQN, time, status, cause, RIAGENDR,
RIDAGEYR, BMXBMI, LBDLDL, DIQ010, BPXSY1, SMQ020) %>% mutate(SEQN =
as.integer(SEQN))

#Order factor levels to make output easier to read
data_short$DIQ010 = factor(data_short$DIQ010 ,
                    levels=c("No", "Borderline", "Yes"))

data_short$SMQ020 = factor(data_short$SMQ020 ,
                    levels=c("No", "Yes"))



summary(data_short)
```

```
##       SEQN           time       status          cause        RIAGENDR
##  Min.   :   2   Min.   :  1.0   0:1568   cancer    :  46   Male  :884
##  1st Qu.: 886   1st Qu.:112.0   1:  49   event-free:1568   Female:925
##  Median :1697   Median :119.0   2:  46   heart     :  49
##  Mean   :1697   Mean   :112.8   3: 146   other     : 146
##  3rd Qu.:2528   3rd Qu.:126.0
##  Max.   :3351   Max.   :132.0
##     RIDAGEYR         BMXBMI           LBDLDL             DIQ010
##  Min.   :2.000   Min.   : 15.68   Min.   : 2.20   No        :1611
##  1st Qu.:3.200   1st Qu.: 24.20   1st Qu.: 8.80   Borderline:  30
##  Median :4.600   Median : 27.64   Median :11.30   Yes       : 168
##  Mean   :4.795   Mean   : 28.76   Mean   :11.49
##  3rd Qu.:6.300   3rd Qu.: 31.89   3rd Qu.:13.80
##  Max.   :8.500   Max.   :130.21   Max.   :32.80
```

```
##      BPXSY1      SMQ020
##  Min.   : 8.00   No :923
##  1st Qu.:11.20   Yes:886
##  Median :12.00
##  Mean   :12.41
##  3rd Qu.:13.40
##  Max.   :22.40
```

```
skim(data_short)
```

Data summary

Name	data_short
Number of rows	1809
Number of columns	11

Column type frequency:

factor	5
numeric	6

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
status	0	1	FALSE	4	0: 1568, 3: 146, 1: 49, 2: 46
cause	0	1	FALSE	4	eve: 1568, oth: 146, hea: 49, can: 46
RIAGENDR	0	1	FALSE	2	Fem: 925, Mal: 884
DIQ010	0	1	FALSE	3	No: 1611, Yes: 168, Bor: 30
SMQ020	0	1	FALSE	2	No: 923, Yes: 886

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
SEQN	0	1	1697.48	961.02	2.00	886.0	1697.00	2528.00	3351.00	▇▇▇▇▇
time	0	1	112.80	24.05	1.00	112.0	119.00	126.00	132.00	▁▁▁▁▇

RIDAGEYR	0	1	4.79	1.88	2.00	3.2	4.60	6.30	8.50	▆▇▇▅▂
BMXBMI	0	1	28.76	7.02	15.68	24.2	27.64	31.89	130.21	▇▁▁▁▁
LBDLDL	0	1	11.49	3.68	2.20	8.8	11.30	13.80	32.80	▂▇▅▁▁
BPXSY1	0	1	12.41	1.98	8.00	11.2	12.00	13.40	22.40	▂▇▅▁▁

```
# Check correlation between covariates:

round(cor(data_short %>% select(-c(time, SEQN, DIQ010, SMQ020, RIAGENDR,
cause, status))), 2)

##          RIDAGEYR BMXBMI LBDLDL BPXSY1
## RIDAGEYR     1.00   0.03   0.04   0.48
## BMXBMI       0.03   1.00   0.02   0.15
## LBDLDL       0.04   0.02   1.00   0.03
## BPXSY1       0.48   0.15   0.03   1.00
```

Systolic blood pressure appears to be positively correlated with BMI, and age.

The code below estimating the Cox model regression coefficients and the cumulative incidence for a given covariate using the `mstate` package has been adapted from a tutorial by H.Putter [4].

```
# Number of each event:
table(data_short$status)

##
##    0    1    2    3
## 1568   49   46  146

# Competing risk transition matrix
tmat <- trans.comprisk(3, names = c("event-free", "heart", "cancer",
"other"))
tmat

##             to
## from         event-free heart cancer other
##   event-free         NA     1      2     3
##   heart              NA    NA     NA    NA
##   cancer             NA    NA     NA    NA
##   other              NA    NA     NA    NA

# Can only go from event free to one of the K=3 causes

# Indicator columns for each of the 3 causes of deaths
data_short$stat1 <- as.numeric(data_short$status == 1)
data_short$stat2 <- as.numeric(data_short$status == 2)
```

```r
data_short$stat3 <- as.numeric(data_short$status == 3)

head(data_short)
```

```
##   SEQN time status      cause RIAGENDR RIDAGEYR BMXBMI LBDLDL DIQ010
BPXSY1
## 1    5  123      0 event-free     Male      7.3  30.63    9.8     No
13.0
## 2   12  123      0 event-free     Male      3.8  25.61    9.1     No
12.6
## 3   13  119      0 event-free     Male      7.1  21.65   11.0     No
15.4
## 4    3  129      0 event-free     Male      7.0  24.74    7.5    Yes
13.8
## 5   18  126      0 event-free   Female      3.3  25.70    8.6     No
11.4
## 6   20  129      0 event-free   Female      2.2  25.26   11.2     No
10.8
##   SMQ020 stat1 stat2 stat3
## 1     No     0     0     0
## 2     No     0     0     0
## 3    Yes     0     0     0
## 4     No     0     0     0
## 5     No     0     0     0
## 6     No     0     0     0
```

```r
# Convert data into long format using msprep:
data_long <- msprep(time = c(NA, "time", "time", "time"), status = c(NA,
"stat1", "stat2", "stat3"), data = data_short, keep = c("RIDAGEYR",
"RIAGENDR", "BMXBMI", "LBDLDL", "DIQ010", "BPXSY1", "SMQ020"), trans = tmat)

tail(data_long)
```

```
## An object of class 'msdata'
##
## Data:
##        id from to trans Tstart Tstop time status RIDAGEYR RIAGENDR BMXBMI
## 5422 1808    1  2     1      0    58   58      0      7.8   Female  33.42
## 5423 1808    1  3     2      0    58   58      0      7.8   Female  33.42
## 5424 1808    1  4     3      0    58   58      1      7.8   Female  33.42
## 5425 1809    1  2     1      0    93   93      0      7.7   Female  33.26
## 5426 1809    1  3     2      0    93   93      0      7.7   Female  33.26
## 5427 1809    1  4     3      0    93   93      1      7.7   Female  33.26
##      LBDLDL DIQ010 BPXSY1 SMQ020
## 5422   11.4    Yes   11.2    Yes
## 5423   11.4    Yes   11.2    Yes
## 5424   11.4    Yes   11.2    Yes
## 5425    4.0    Yes   15.8     No
## 5426    4.0    Yes   15.8     No
## 5427    4.0    Yes   15.8     No
```

```r
# Check number of events same as before:
events(data_long)
```

```
## $Frequencies
##             to
## from         event-free heart cancer other no event total entering
##   event-free          0    49     46   146     1568           1809
##   heart               0     0      0     0       49             49
##   cancer              0     0      0     0       46             46
##   other               0     0      0     0      146            146
##
## $Proportions
##             to
## from         event-free      heart     cancer      other   no event
##   event-free 0.00000000 0.02708679 0.02542841 0.08070757 0.86677722
##   heart      0.00000000 0.00000000 0.00000000 0.00000000 1.00000000
##   cancer     0.00000000 0.00000000 0.00000000 0.00000000 1.00000000
##   other      0.00000000 0.00000000 0.00000000 0.00000000 1.00000000
```

```r
# Add cause-specific covariates for regression:
data_long <- expand.covs(data_long, covs =
c("RIDAGEYR","RIAGENDR","BMXBMI","LBDLDL", "DIQ010", "BPXSY1", "SMQ020"))

head(data_long)
```

```
## An object of class 'msdata'
##
## Data:
##   id from to trans Tstart Tstop time status RIDAGEYR RIAGENDR BMXBMI
LBDLDL
## 1  1    1  2     1      0   123  123      0      7.3     Male  30.63
9.8
## 2  1    1  3     2      0   123  123      0      7.3     Male  30.63
9.8
## 3  1    1  4     3      0   123  123      0      7.3     Male  30.63
9.8
## 4  2    1  2     1      0   123  123      0      3.8     Male  25.61
9.1
## 5  2    1  3     2      0   123  123      0      3.8     Male  25.61
9.1
## 6  2    1  4     3      0   123  123      0      3.8     Male  25.61
9.1
##   DIQ010 BPXSY1 SMQ020 RIDAGEYR.1 RIDAGEYR.2 RIDAGEYR.3 RIAGENDRFemale.1
## 1     No   13.0     No        7.3        0.0        0.0                0
## 2     No   13.0     No        0.0        7.3        0.0                0
## 3     No   13.0     No        0.0        0.0        7.3                0
## 4     No   12.6     No        3.8        0.0        0.0                0
## 5     No   12.6     No        0.0        3.8        0.0                0
## 6     No   12.6     No        0.0        0.0        3.8                0
##   RIAGENDRFemale.2 RIAGENDRFemale.3 BMXBMI.1 BMXBMI.2 BMXBMI.3 LBDLDL.1
```

```
## 1                0                0    30.63     0.00     0.00      9.8
## 2                0                0     0.00    30.63     0.00      0.0
## 3                0                0     0.00     0.00    30.63      0.0
## 4                0                0    25.61     0.00     0.00      9.1
## 5                0                0     0.00    25.61     0.00      0.0
## 6                0                0     0.00     0.00    25.61      0.0
##   LBDLDL.2 LBDLDL.3 DIQ010Borderline.1 DIQ010Borderline.2
DIQ010Borderline.3
## 1      0.0      0.0                  0                  0
0
## 2      9.8      0.0                  0                  0
0
## 3      0.0      9.8                  0                  0
0
## 4      0.0      0.0                  0                  0
0
## 5      9.1      0.0                  0                  0
0
## 6      0.0      9.1                  0                  0
0
##   DIQ010Yes.1 DIQ010Yes.2 DIQ010Yes.3 BPXSY1.1 BPXSY1.2 BPXSY1.3
SMQ020Yes.1
## 1           0           0           0     13.0      0.0      0.0
0
## 2           0           0           0      0.0     13.0      0.0
0
## 3           0           0           0      0.0      0.0     13.0
0
## 4           0           0           0     12.6      0.0      0.0
0
## 5           0           0           0      0.0     12.6      0.0
0
## 6           0           0           0      0.0      0.0     12.6
0
##   SMQ020Yes.2 SMQ020Yes.3
## 1           0           0
## 2           0           0
## 3           0           0
## 4           0           0
## 5           0           0
## 6           0           0
```

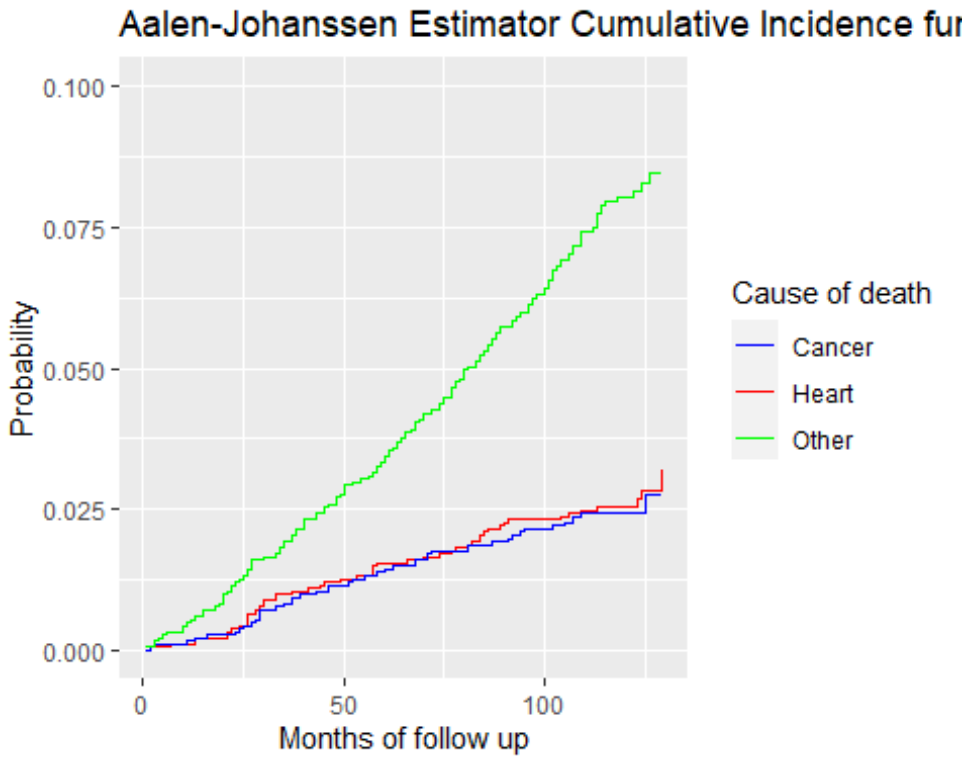
Non-parametric estimators for cumulative incidence

Aalen-Johanssen Estimator

```r
ci <- Cuminc(time = "time", status = "status", data = data_short)

head(ci)
```

```
##   time      Surv         CI.1         CI.2         CI.3       seSurv
## 1    1 0.9994472 0.0000000000 0.0000000000 0.0005527916 0.0005526388
## 2    2 0.9983416 0.0005527916 0.0005527916 0.0005527916 0.0009566689
## 3    3 0.9966833 0.0005527916 0.0011055832 0.0016583748 0.0013518099
## 4    4 0.9961305 0.0005527916 0.0011055832 0.0022111664 0.0014597167
## 5    5 0.9955777 0.0005527916 0.0011055832 0.0027639580 0.0015600697
## 6    6 0.9950249 0.0005527916 0.0011055832 0.0033167496 0.0016542443
##         seCI.1       seCI.2       seCI.3
## 1 0.0000000000 0.0000000000 0.0005526388
## 2 0.0005526388 0.0005526388 0.0005526388
## 3 0.0005526388 0.0007813331 0.0009566689
## 4 0.0005526388 0.0007813331 0.0011043602
## 5 0.0005526388 0.0007813331 0.0012343702
## 6 0.0005526388 0.0007813331 0.0013518099
```

```r
# Plot
ggplot(ci) +
      geom_step(aes(x = time, y = CI.1, color = 'Heart')) +
      geom_step(aes(x = time, y = CI.2, color = 'Cancer')) +
  geom_step(aes(x = time, y = CI.3, color = 'Other')) +
  labs(title = 'Aalen-Johanssen Estimator Cumulative Incidence function') +
xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="Cause of death",
    values=c(Heart="red", Cancer="blue", Other = "green"))
```



1: Cox model

```
# Stratified hazards model
c1 <- coxph(Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +  RIDAGEYR.3  +
RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3  + BMXBMI.1 + BMXBMI.2
+ BMXBMI.3  + LBDLDL.1 + LBDLDL.2 + LBDLDL.3 + DIQ010Borderline.1 +
DIQ010Borderline.2 + DIQ010Borderline.3 + DIQ010Yes.1 + DIQ010Yes.2 +
DIQ010Yes.3 + BPXSY1.1 + BPXSY1.2 + BPXSY1.3 + SMQ020Yes.1 + SMQ020Yes.2 +
SMQ020Yes.3 + strata(trans), data = data_long)


summary(c1)

## Call:
## coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
##     RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
##     BMXBMI.1 + BMXBMI.2 + BMXBMI.3 + LBDLDL.1 + LBDLDL.2 + LBDLDL.3 +
##     DIQ010Borderline.1 + DIQ010Borderline.2 + DIQ010Borderline.3 +
##     DIQ010Yes.1 + DIQ010Yes.2 + DIQ010Yes.3 + BPXSY1.1 + BPXSY1.2 +
##     BPXSY1.3 + SMQ020Yes.1 + SMQ020Yes.2 + SMQ020Yes.3 + strata(trans),
##     data = data_long)
##
##   n= 5427, number of events= 241
##
##                          coef exp(coef)  se(coef)      z Pr(>|z|)
## RIDAGEYR.1           0.725583  2.065935  0.116254  6.241 4.34e-10 ***
## RIDAGEYR.2           0.837456  2.310481  0.124960  6.702 2.06e-11 ***
## RIDAGEYR.3           0.825470  2.282955  0.068883 11.984  < 2e-16 ***
## RIAGENDRFemale.1    -0.723129  0.485231  0.325143 -2.224  0.02615 *
## RIAGENDRFemale.2    -0.585039  0.557084  0.325258 -1.799  0.07207 .
## RIAGENDRFemale.3    -0.016164  0.983966  0.173985 -0.093  0.92598
## BMXBMI.1            -0.015148  0.984966  0.027072 -0.560  0.57578
## BMXBMI.2             0.034787  1.035399  0.015870  2.192  0.02838 *
## BMXBMI.3            -0.016408  0.983726  0.015058 -1.090  0.27585
## LBDLDL.1            -0.016031  0.984097  0.041231 -0.389  0.69742
## LBDLDL.2            -0.065582  0.936522  0.044421 -1.476  0.13984
## LBDLDL.3            -0.063351  0.938614  0.024706 -2.564  0.01034 *
## DIQ010Borderline.1 -0.449747  0.637789  1.020466 -0.441  0.65941
## DIQ010Borderline.2  0.255683  1.291343  0.734145  0.348  0.72764
## DIQ010Borderline.3 -0.222766  0.800302  0.593135 -0.376  0.70723
## DIQ010Yes.1         -0.013510  0.986581  0.431107 -0.031  0.97500
## DIQ010Yes.2          0.289745  1.336087  0.388163  0.746  0.45539
## DIQ010Yes.3          0.617177  1.853687  0.206199  2.993  0.00276 **
## BPXSY1.1             0.158713  1.172001  0.064290  2.469  0.01356 *
## BPXSY1.2            -0.028800  0.971611  0.075208 -0.383  0.70177
## BPXSY1.3            -0.001786  0.998215  0.040503 -0.044  0.96482
## SMQ020Yes.1          0.239402  1.270490  0.309425  0.774  0.43911
## SMQ020Yes.2          0.586168  1.797088  0.334536  1.752  0.07974 .
## SMQ020Yes.3          0.321161  1.378727  0.177801  1.806  0.07087 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR.1             2.0659     0.4840   1.64498    2.5946
## RIDAGEYR.2             2.3105     0.4328   1.80857    2.9517
## RIDAGEYR.3             2.2830     0.4380   1.99464    2.6129
## RIAGENDRFemale.1       0.4852     2.0609   0.25656    0.9177
## RIAGENDRFemale.2       0.5571     1.7951   0.29448    1.0539
## RIAGENDRFemale.3       0.9840     1.0163   0.69965    1.3838
## BMXBMI.1               0.9850     1.0153   0.93407    1.0386
## BMXBMI.2               1.0354     0.9658   1.00369    1.0681
## BMXBMI.3               0.9837     1.0165   0.95512    1.0132
## LBDLDL.1               0.9841     1.0162   0.90770    1.0669
## LBDLDL.2               0.9365     1.0678   0.85843    1.0217
## LBDLDL.3               0.9386     1.0654   0.89425    0.9852
## DIQ010Borderline.1     0.6378     1.5679   0.08631    4.7130
## DIQ010Borderline.2     1.2913     0.7744   0.30629    5.4444
## DIQ010Borderline.3     0.8003     1.2495   0.25025    2.5594
## DIQ010Yes.1            0.9866     1.0136   0.42381    2.2966
## DIQ010Yes.2            1.3361     0.7485   0.62435    2.8592
## DIQ010Yes.3            1.8537     0.5395   1.23743    2.7769
## BPXSY1.1               1.1720     0.8532   1.03325    1.3294
## BPXSY1.2               0.9716     1.0292   0.83845    1.1259
## BPXSY1.3               0.9982     1.0018   0.92204    1.0807
## SMQ020Yes.1            1.2705     0.7871   0.69277    2.3300
## SMQ020Yes.2            1.7971     0.5565   0.93285    3.4620
## SMQ020Yes.3            1.3787     0.7253   0.97305    1.9535
##
## Concordance= 0.86  (se = 0.012 )
## Likelihood ratio test= 486.8  on 24 df,   p=<2e-16
## Wald test            = 336.1  on 24 df,   p=<2e-16
## Score (logrank) test = 500.7  on 24 df,   p=<2e-16
```

Using just coxph()

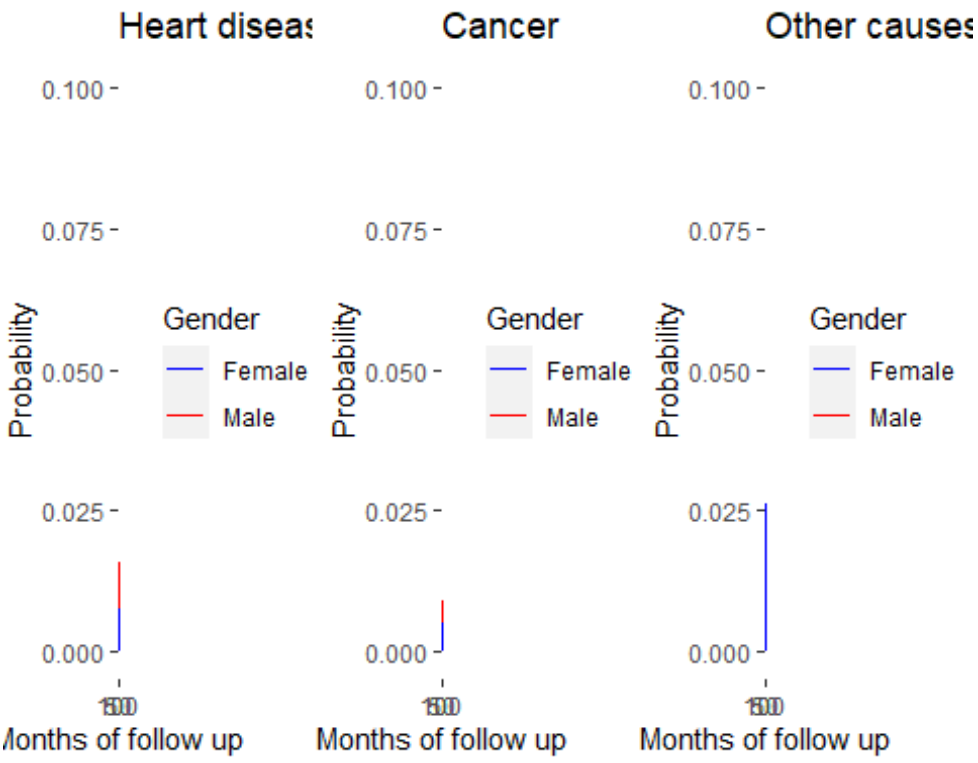
```
# Alternative method using just coxph()

coxph(Surv(time,status==1)~  RIDAGEYR + RIAGENDR + BMXBMI + LBDLDL + DIQ010 +
BPXSY1 + SMQ020, data = data_short)

## Call:
## coxph(formula = Surv(time, status == 1) ~ RIDAGEYR + RIAGENDR +
##     BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data = data_short)
##
##                       coef exp(coef) se(coef)      z        p
## RIDAGEYR           0.72558   2.06594  0.11625  6.241 4.34e-10
## RIAGENDRFemale    -0.72313   0.48523  0.32514 -2.224   0.0261
## BMXBMI            -0.01515   0.98497  0.02707 -0.560   0.5758
## LBDLDL            -0.01603   0.98410  0.04123 -0.389   0.6974
## DIQ010Borderline -0.44975   0.63779  1.02047 -0.441   0.6594
## DIQ010Yes         -0.01351   0.98658  0.43111 -0.031   0.9750
```

```
## BPXSY1            0.15871   1.17200  0.06429  2.469   0.0136
## SMQ020Yes         0.23940   1.27049  0.30942  0.774   0.4391
##
## Likelihood ratio test=99.89  on 8 df, p=< 2.2e-16
## n= 1809, number of events= 49

# Gives same results
```

Get predicted Cumulative incidence using mstate package

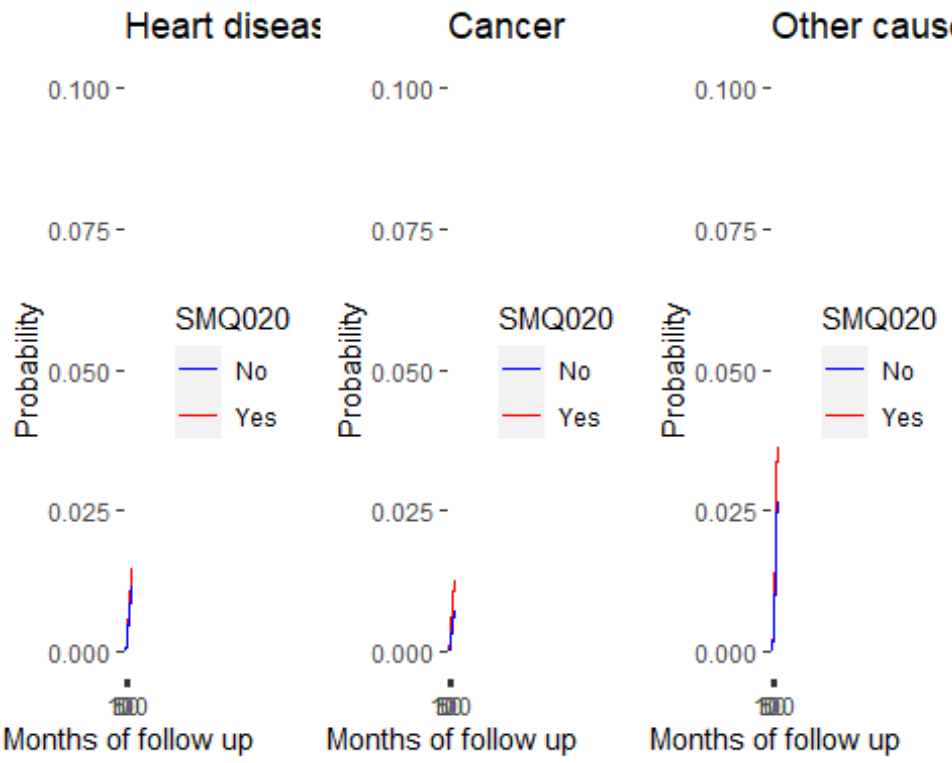
```
par(mfrow = c(1, 3))

c1 <- coxph(Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +  RIDAGEYR.3  +
RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3  + BMXBMI.1 + BMXBMI.2
+ BMXBMI.3  + LBDLDL.1 + LBDLDL.2 + LBDLDL.3 + DIQ010Borderline.1 +
DIQ010Borderline.2 + DIQ010Borderline.3 + DIQ010Yes.1 + DIQ010Yes.2 +
DIQ010Yes.3 + BPXSY1.1 + BPXSY1.2 + BPXSY1.3 + SMQ020Yes.1 + SMQ020Yes.2 +
SMQ020Yes.3 + strata(trans), data = data_long, method = 'breslow')


# New data for individual with average covariate values and factors

# Compare incidence of each outcome for males and females, both BMI 28.8, LDL
115md/dL, aged 48, systolic blood pressure 124mmHg, with no diabetes and not
smoked more than 100 cigarettes in lifetime
Male <- data.frame(RIDAGEYR.1 = c(4.8,0,0), RIDAGEYR.2 = c(0,4.8,0),
RIDAGEYR.3 = c(0,0,4.8), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(28.8, 0,0 ), BMXBMI.2 =
c( 0,28.8,0 ), BMXBMI.3 = c( 0,0,28.8), LBDLDL.1 = c(11.5,0 ,0 ), LBDLDL.2 =
c( 0,11.5,0 ), LBDLDL.3 = c( 0,0 ,11.5), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(12.4,0,0), BPXSY1.2 =c(0,12.4,0), BPXSY1.3 = c(0,0,12.4),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))

Female <- data.frame(RIDAGEYR.1 = c(4.8,0,0), RIDAGEYR.2 = c(0,4.8,0),
RIDAGEYR.3 = c(0,0,4.8), RIAGENDRFemale.1 = c(1,0,0), RIAGENDRFemale.2 =
c(0,1,0), RIAGENDRFemale.3 = c(0,0,1), BMXBMI.1 = c(28.8, 0,0 ), BMXBMI.2 =
c( 0,28.8,0 ), BMXBMI.3 = c( 0,0,28.8), LBDLDL.1 = c(11.5,0 ,0 ), LBDLDL.2 =
c( 0,11.5,0 ), LBDLDL.3 = c( 0,0 ,11.5), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(12.4,0,0), BPXSY1.2 =c(0,12.4,0), BPXSY1.3 = c(0,0,12.4),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))


# Estimated cumulative hazards for all event times
msf.Male <- msfit(c1, Male, trans = tmat)
```

```r
msf.Female <- msfit(c1, Female, trans = tmat)



# Caluculates Cumulative Incidence
pt.Male <- probtrans(msf.Male, 0)[[1]]


pt.Female <- probtrans(msf.Female, 0)[[1]]
#pt.Female


# Plot

# Heart disease
plot11 <- ggplot(NULL, aes(x = time, y = pstate2)) +
      geom_step(data = pt.Male, aes(color = 'Male')) +
      geom_step(data = pt.Female, aes(color = 'Female')) + labs(title =
'Heart disease') + xlab('Months of follow up') + ylab('Probability') +
ylim(0,0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))

# Cancer

plot12 <- ggplot(NULL, aes(x = time, y = pstate3)) +
      geom_step(data = pt.Male, aes(color = 'Male')) +
      geom_step(data = pt.Female, aes(color = 'Female')) + labs(title =
'Cancer') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))


# Other causes

plot13 <- ggplot(NULL, aes(x = time, y = pstate4)) +
      geom_step(data = pt.Male, aes(color = 'Male')) +
      geom_step(data = pt.Female, aes(color = 'Female')) + labs(title =
'Other causes') + xlab('Months of follow up') + ylab('Probability') +
ylim(0,0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))

grid.arrange(plot11, plot12, plot13, nrow=1, ncol=3)
```



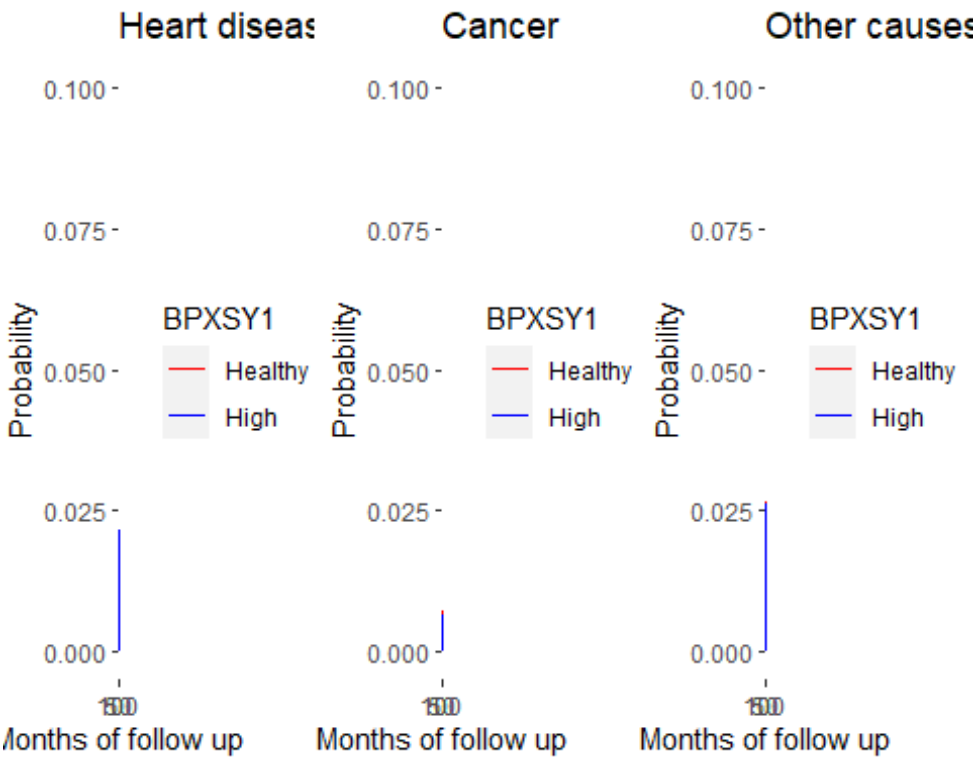

```
### Smoking

# Compare incidence of each outcome for those who have smoked more than 100
cigarettes in life with those who haven't, both BMI 23,male, LDL 90md/dL,
aged 45, systolic blood pressure 110mmHg, with no diabetes
Smoke <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c(
0,23,0 ), BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0
), LBDLDL.3 = c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(11,0,0), BPXSY1.2 =c(0,11,0), BPXSY1.3 = c(0,0,11),
SMQ020Yes.1 = c(1,0,0), SMQ020Yes.2 = c(0,1,0), SMQ020Yes.3 = c(0,0,1), trans
= c(1, 2, 3), strata = c(1, 2, 3))

Not_smoked  <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c(
0,23,0 ), BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0
), LBDLDL.3 = c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(11,0,0), BPXSY1.2 =c(0,11,0), BPXSY1.3 = c(0,0,11),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))
```

```r
# Estimated cumulative hazards for all event times
msf.Smoke <- msfit(c1, Smoke, trans = tmat)
msf.Not_smoked <- msfit(c1, Not_smoked, trans = tmat)


# Caluculates Cumulative Incidence
pt.Smoke <- probtrans(msf.Smoke, 0)[[1]]

pt.Not_smoked <- probtrans(msf.Not_smoked, 0)[[1]]

# Plot

# Heart disease
plot21 <- ggplot(NULL, aes(x = time, y = pstate2)) +
      geom_step(data = pt.Smoke, aes(color = 'Yes')) +
      geom_step(data = pt.Not_smoked, aes(color = 'No')) + labs(title =
'Heart disease') + xlab('Months of follow up') + ylab('Probability') +
ylim(0,0.1) +
scale_colour_manual(name="SMQ020",
    values=c(Yes="red", No="blue"))

# Cancer

plot22 <- ggplot(NULL, aes(x = time, y = pstate3)) +
      geom_step(data = pt.Smoke, aes(color = 'Yes')) +
      geom_step(data = pt.Not_smoked, aes(color = 'No')) + labs(title =
'Cancer') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="SMQ020",
    values=c(Yes="red", No="blue"))


# Other causes

plot23 <- ggplot(NULL, aes(x = time, y = pstate4)) +
      geom_step(data = pt.Smoke, aes(color = 'Yes')) +
      geom_step(data = pt.Not_smoked, aes(color = 'No')) + labs(title =
'Other causes') + xlab('Months of follow up') + ylab('Probability') +
ylim(0,0.1) +
scale_colour_manual(name="SMQ020",
    values=c(Yes="red", No="blue"))

grid.arrange(plot21, plot22, plot23, nrow=1, ncol=3)
```



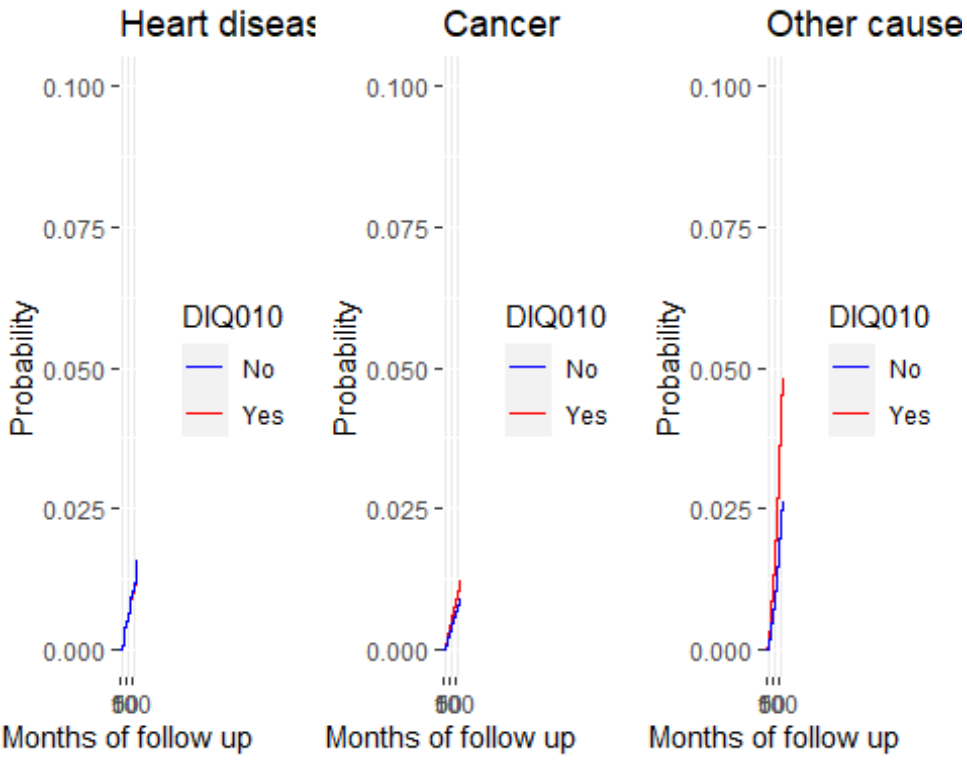

```
### High blood pressure

# Compare incidence of each outcome for those with systolic bloof pressure
110mmHh (healthy) with 150 mmHg (high), both BMI 23, male, LDL 90md/dL, aged
45, systolic blood pressure 110mmHg, with no diabetes and not smoked more
than 100 cigarettes in lifetime
Healthy <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c(
0,23,0 ), BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0
), LBDLDL.3 = c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(11,0,0), BPXSY1.2 =c(0,11,0), BPXSY1.3 = c(0,0,11),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))

High <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c(
0,23,0 ), BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0
), LBDLDL.3 = c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(15,0,0), BPXSY1.2 =c(0,15,0), BPXSY1.3 = c(0,0,15),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
```

```
= c(1, 2, 3), strata = c(1, 2, 3))


# Estimated cumulative hazards for all event times
msf.Healthy <- msfit(c1, Healthy, trans = tmat)
msf.High <- msfit(c1, High, trans = tmat)


# Caluculates Cumulative Incidence
pt.Healthy <- probtrans(msf.Healthy, 0)[[1]]

pt.High <- probtrans(msf.High, 0)[[1]]

# Plot

# Heart disease
plot31 <- ggplot(NULL, aes(x = time, y = pstate2)) +
      geom_step(data = pt.Healthy, aes(color = 'Healthy')) +
      geom_step(data = pt.High, aes(color = 'High')) + labs(title = 'Heart
disease') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="BPXSY1",
    values=c(Healthy="red", High="blue"))

# Cancer
plot32 <- ggplot(NULL, aes(x = time, y = pstate3)) +
      geom_step(data = pt.Healthy, aes(color = 'Healthy')) +
      geom_step(data = pt.High, aes(color = 'High')) + labs(title = 'Cancer')
+ xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="BPXSY1",
    values=c(Healthy="red", High="blue"))



# Other causes
plot33 <- ggplot(NULL, aes(x = time, y = pstate4)) +
      geom_step(data = pt.Healthy, aes(color = 'Healthy')) +
      geom_step(data = pt.High, aes(color = 'High')) + labs(title = 'Other
causes') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="BPXSY1",
    values=c(Healthy="red", High="blue"))

grid.arrange(plot31, plot32, plot33, nrow=1, ncol=3)
```




```
### Diabetic

# Compare incidence of each outcome for those with and without diabetes, both
BMI 23, male,  LDL 90md/dL, aged 45, systolic blood pressure 110mmHg, who
have not smoked more than 100 cigarettes in lifetime
Diabetic <- data.frame(RIDAGEYR.1 = c(4.8,0,0), RIDAGEYR.2 = c(0,4.8,0),
RIDAGEYR.3 = c(0,0,4.8), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(28.8, 0,0 ), BMXBMI.2 =
c( 0,28.8,0 ), BMXBMI.3 = c( 0,0,28.8), LBDLDL.1 = c(11.5,0 ,0 ), LBDLDL.2 =
c( 0,11.5,0 ), LBDLDL.3 = c( 0,0 ,11.5), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(1 ,0 ,0 ), DIQ010Yes.2 = c(0 ,1 ,0 ), DIQ010Yes.3 = c(0 ,0
,1), BPXSY1.1 = c(12.4,0,0), BPXSY1.2 =c(0,12.4,0), BPXSY1.3 = c(0,0,12.4),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))

Not_diabetic <- data.frame(RIDAGEYR.1 = c(4.8,0,0), RIDAGEYR.2 = c(0,4.8,0),
RIDAGEYR.3 = c(0,0,4.8), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(28.8, 0,0 ), BMXBMI.2 =
c( 0,28.8,0 ), BMXBMI.3 = c( 0,0,28.8), LBDLDL.1 = c(11.5,0 ,0 ), LBDLDL.2 =
c( 0,11.5,0 ), LBDLDL.3 = c( 0,0 ,11.5), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(12.4,0,0), BPXSY1.2 =c(0,12.4,0), BPXSY1.3 = c(0,0,12.4),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))
```

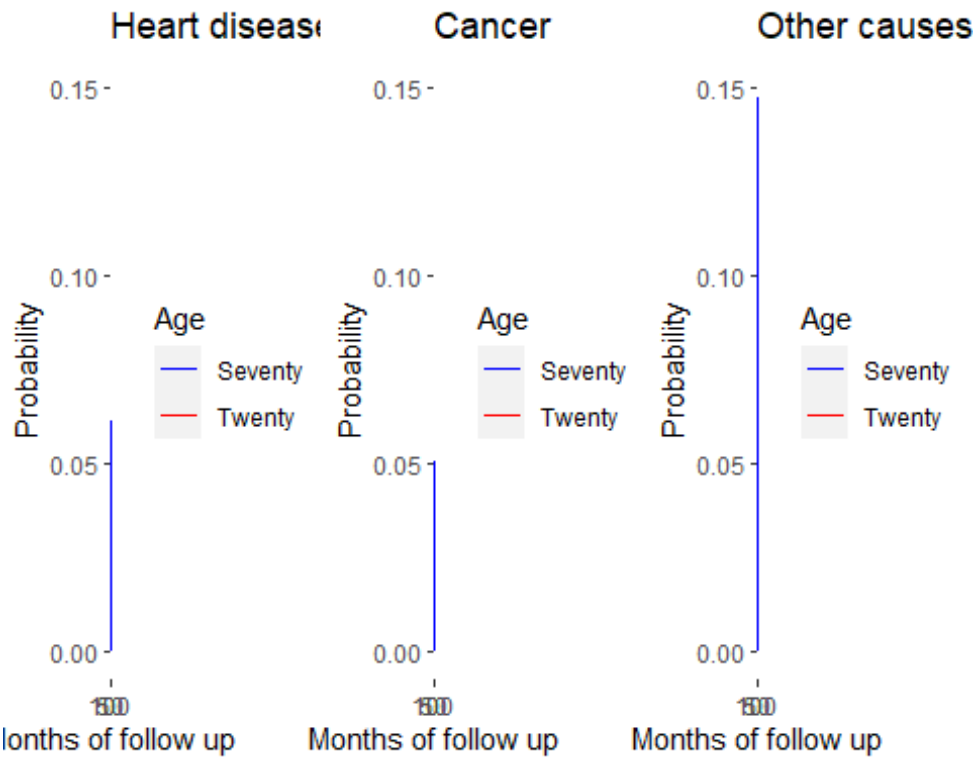
```r
# Estimated cumulative hazards for all event times
msf.diabetic <- msfit(c1, Diabetic, trans = tmat)
msf.not_diabetic <- msfit(c1, Not_diabetic, trans = tmat)


# Caluculates Cumulative Incidence
pt.diabetic <- probtrans(msf.diabetic, 0)[[1]]

pt.not_diabetic <- probtrans(msf.not_diabetic, 0)[[1]]

# Plot

# Heart disease
plot41 <- ggplot(NULL, aes(x = time, y = pstate2)) +
      geom_step(data = pt.diabetic, aes(color = 'Yes')) +
      geom_step(data = pt.not_diabetic, aes(color = 'No')) + labs(title =
'Heart disease') + xlab('Months of follow up') + ylab('Probability') +
ylim(0,0.1) +
scale_colour_manual(name="DIQ010",
    values=c(Yes="red", No="blue"))

# Cancer
plot42<- ggplot(NULL, aes(x = time, y = pstate3)) +
      geom_step(data = pt.diabetic, aes(color = 'Yes')) +
      geom_step(data = pt.not_diabetic, aes(color = 'No')) + labs(title =
'Cancer') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="DIQ010",
    values=c(Yes="red", No="blue"))

plot43<- ggplot(NULL, aes(x = time, y = pstate4)) +
      geom_step(data = pt.diabetic, aes(color = 'Yes')) +
      geom_step(data = pt.not_diabetic, aes(color = 'No')) + labs(title =
'Other causes') + xlab('Months of follow up') + ylab('Probability') +
ylim(0,0.1) +
scale_colour_manual(name="DIQ010",
    values=c(Yes="red", No="blue"))

grid.arrange(plot41, plot42, plot43, nrow=1, ncol=3)
```




```
### Age

# Compare incidence of each outcome for a 20 and 70 year old, both BMI 23,
male, LDL 90md/dL, systolic blood pressure 110mmHg, with no diabetes and not
smoked more than 100 cigarettes in lifetime
Young <- data.frame(RIDAGEYR.1 = c(2,0,0), RIDAGEYR.2 = c(0,2,0), RIDAGEYR.3
= c(0,0,2), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 = c(0,0,0),
RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c( 0,23,0 ),
BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0 ), LBDLDL.3
= c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ), DIQ010Borderline.2 = c(0 ,0
,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ), DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2
= c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0 ), BPXSY1.1 = c(11,0,0), BPXSY1.2
=c(0,11,0), BPXSY1.3 = c(0,0,11), SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 =
c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans = c(1, 2, 3), strata = c(1, 2, 3))

Old <-  data.frame(RIDAGEYR.1 = c(7,0,0), RIDAGEYR.2 = c(0,7,0), RIDAGEYR.3 =
c(0,0,7), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 = c(0,0,0),
RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c( 0,23,0 ),
BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0 ), LBDLDL.3
= c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ), DIQ010Borderline.2 = c(0 ,0
,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ), DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2
= c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0 ), BPXSY1.1 = c(11,0,0), BPXSY1.2
=c(0,11,0), BPXSY1.3 = c(0,0,11), SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 =
c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans = c(1, 2, 3), strata = c(1, 2, 3))
```

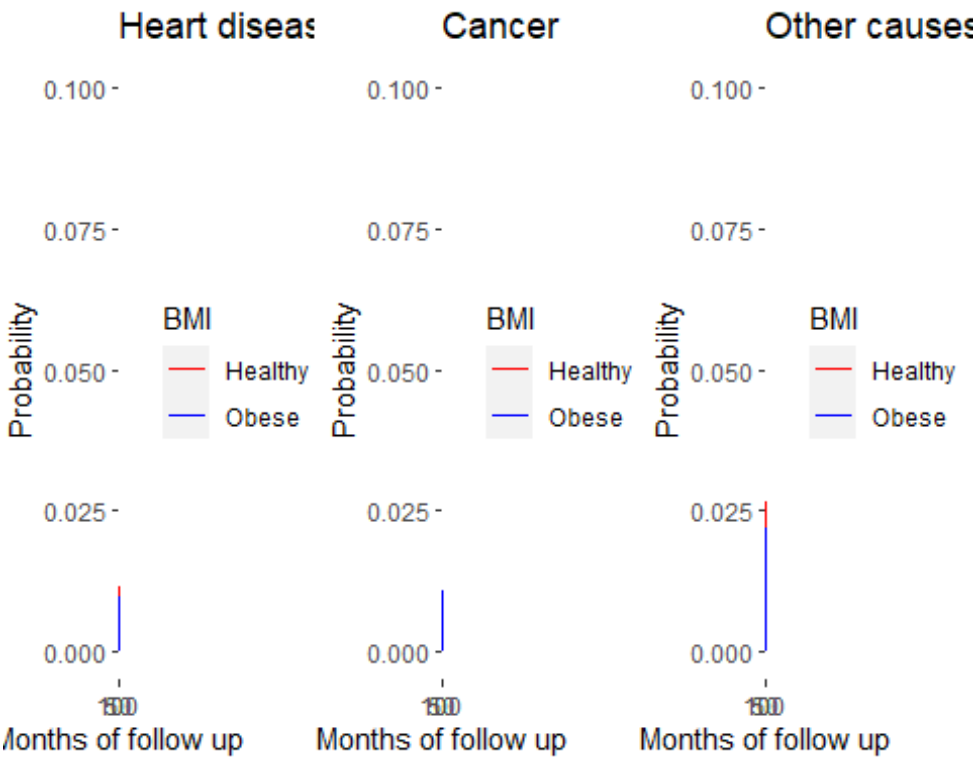
```r
# Estimated cumulative hazards for all event times
msf.young <- msfit(c1, Young, trans = tmat)
msf.old <- msfit(c1, Old, trans = tmat)


# Caluculates Cumulative Incidence
pt.young <- probtrans(msf.young, 0)[[1]]

pt.old <- probtrans(msf.old, 0)[[1]]

# Plot

# Heart disease
plot51 <- ggplot(NULL, aes(x = time, y = pstate2)) +
      geom_step(data = pt.young, aes(color = 'Twenty')) +
      geom_step(data = pt.old, aes(color = 'Seventy')) + labs(title = 'Heart
disease') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.15)
+
scale_colour_manual(name="Age",
    values=c(Twenty="red", Seventy="blue"))

# Cancer
plot52 <- ggplot(NULL, aes(x = time, y = pstate3)) +
      geom_step(data = pt.young, aes(color = 'Twenty')) +
      geom_step(data = pt.old, aes(color = 'Seventy')) + labs(title =
'Cancer') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.15)
+
scale_colour_manual(name="Age",
    values=c(Twenty="red", Seventy="blue"))


# Other causes
plot53 <- ggplot(NULL, aes(x = time, y = pstate4)) +
      geom_step(data = pt.young, aes(color = 'Twenty')) +
      geom_step(data = pt.old, aes(color = 'Seventy')) + labs(title = 'Other
causes') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.15) +
scale_colour_manual(name="Age",
    values=c(Twenty="red", Seventy="blue"))

grid.arrange(plot51, plot52, plot53, nrow=1, ncol=3)
```

```
## Warning: Removed 14 row(s) containing missing values (geom_path).
```



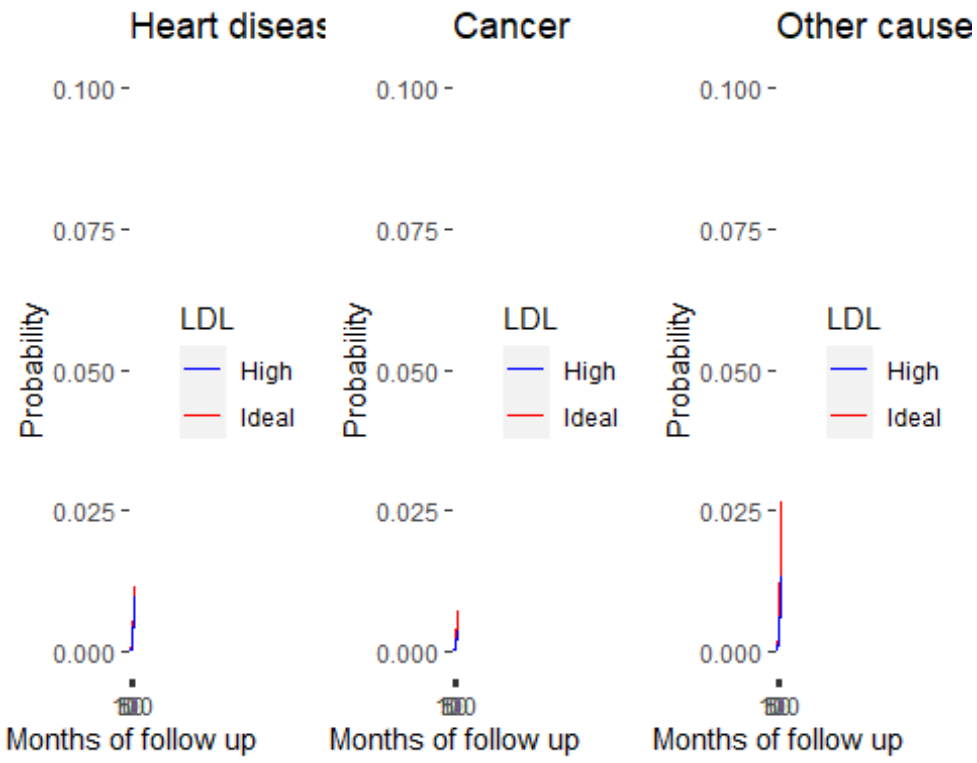

```
### BMI

# Compare incidence of each outcome for BMI 23 (healthy) vs BMI 35 (obese),
both male, LDL 90md/dL, aged 45, systolic blood pressure 110mmHg, with no
diabetes and not smoked more than 100 cigarettes in lifetime
Healthy <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c(
0,23,0 ), BMXBMI.3 = c( 0,0, 23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0
), LBDLDL.3 = c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(11,0,0), BPXSY1.2 =c(0,11,0), BPXSY1.3 = c(0,0,11),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))

Obese <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(35, 0,0 ), BMXBMI.2 = c(
0,35,0 ), BMXBMI.3 = c( 0,0,35), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0
), LBDLDL.3 = c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(11,0,0), BPXSY1.2 =c(0,11,0), BPXSY1.3 = c(0,0,11),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))
```

```r
# Estimated cumulative hazards for all event times
msf.healthy <- msfit(c1, Healthy, trans = tmat)
msf.obese<- msfit(c1, Obese, trans = tmat)


# Caluculates Cumulative Incidence
pt.healthy <- probtrans(msf.healthy, 0)[[1]]

pt.obese <- probtrans(msf.obese, 0)[[1]]

# Plot

# Heart disease
plot61 <- ggplot(NULL, aes(x = time, y = pstate2)) +
      geom_step(data = pt.healthy, aes(color = 'Healthy')) +
      geom_step(data = pt.obese, aes(color = 'Obese')) + labs(title = 'Heart
disease') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="BMI",
    values=c(Healthy="red", Obese="blue"))

plot62 <- ggplot(NULL, aes(x = time, y = pstate3)) +
      geom_step(data = pt.healthy, aes(color = 'Healthy')) +
      geom_step(data = pt.obese, aes(color = 'Obese')) + labs(title =
'Cancer') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="BMI",
    values=c(Healthy="red", Obese="blue"))

plot63 <- ggplot(NULL, aes(x = time, y = pstate4)) +
      geom_step(data = pt.healthy, aes(color = 'Healthy')) +
      geom_step(data = pt.obese, aes(color = 'Obese')) + labs(title = 'Other
causes') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="BMI",
    values=c(Healthy="red", Obese="blue"))

grid.arrange(plot61, plot62, plot63, nrow=1, ncol=3)
```




```
### LDL

# Compare incidence of each outcome for individial with LDL 90mg/dL vs.
200mg/dL, both BMI 23, male, aged 45, systolic blood pressure 110mmHg, with
no diabetes and not smoked more than 100 cigarettes in lifetime
Ideal <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c(
0,23,0 ), BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(9,0 ,0 ), LBDLDL.2 = c( 0,9,0
), LBDLDL.3 = c( 0,0 ,9), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(11,0,0), BPXSY1.2 =c(0,11,0), BPXSY1.3 = c(0,0,11),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
= c(1, 2, 3), strata = c(1, 2, 3))


High <- data.frame(RIDAGEYR.1 = c(4.5,0,0), RIDAGEYR.2 = c(0,4.5,0),
RIDAGEYR.3 = c(0,0,4.5), RIAGENDRFemale.1 = c(0,0,0), RIAGENDRFemale.2 =
c(0,0,0), RIAGENDRFemale.3 = c(0,0,0), BMXBMI.1 = c(23, 0,0 ), BMXBMI.2 = c(
0,23,0 ), BMXBMI.3 = c( 0,0,23), LBDLDL.1 = c(20,0 ,0 ), LBDLDL.2 = c( 0,20,0
), LBDLDL.3 = c( 0,0 ,20), DIQ010Borderline.1 = c(0 ,0 ,0 ),
DIQ010Borderline.2 = c(0 ,0 ,0 ), DIQ010Borderline.3 = c(0 ,0 ,0 ),
DIQ010Yes.1 =c(0 ,0 ,0 ), DIQ010Yes.2 = c(0 ,0 ,0 ), DIQ010Yes.3 = c(0 ,0 ,0
), BPXSY1.1 = c(11,0,0), BPXSY1.2 =c(0,11,0), BPXSY1.3 = c(0,0,11),
SMQ020Yes.1 = c(0,0,0), SMQ020Yes.2 = c(0,0,0), SMQ020Yes.3 = c(0,0,0), trans
```

```
= c(1, 2, 3), strata = c(1, 2, 3))


# Estimated cumulative hazards for all event times
msf.ideal <- msfit(c1, Ideal, trans = tmat)
msf.vhigh <- msfit(c1, High, trans = tmat)


# Caluculates Cumulative Incidence
pt.ideal <- probtrans(msf.ideal, 0)[[1]]

pt.vhigh <- probtrans(msf.vhigh, 0)[[1]]

# Plot

# Heart disease
plot71 <- ggplot(NULL, aes(x = time, y = pstate2)) +
      geom_step(data = pt.ideal, aes(color = 'Ideal')) +
      geom_step(data = pt.vhigh, aes(color = 'High')) + labs(title = 'Heart
disease') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="LDL",
    values=c(Ideal="red", High="blue"))

plot72<- ggplot(NULL, aes(x = time, y = pstate3)) +
      geom_step(data = pt.ideal, aes(color = 'Ideal')) +
      geom_step(data = pt.vhigh, aes(color = 'High')) + labs(title =
'Cancer') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="LDL",
    values=c(Ideal="red", High="blue"))

plot73 <- ggplot(NULL, aes(x = time, y = pstate4)) +
      geom_step(data = pt.ideal, aes(color = 'Ideal')) +
      geom_step(data = pt.vhigh, aes(color = 'High')) + labs(title = 'Other
causes') + xlab('Months of follow up') + ylab('Probability') + ylim(0,0.1) +
scale_colour_manual(name="LDL",
    values=c(Ideal="red", High="blue"))

grid.arrange(plot71, plot72, plot73, nrow=1, ncol=3)
```



2: Fine and Gray Model

The code below implementing the Fine-Gray model on the NHANES mortality linked dataset was adapted from a tutorial by L.Scrucca et al. on finding estimates using the `crr()` function in R [5].

```
# Remove individuals with missing covariate data:
data_complete <- na.omit(data_short) %>% select(-c(stat1, stat2, stat3))
summary(data_complete)
```

```
##       SEQN           time       status          cause        RIAGENDR
##  Min.   :   2   Min.   :  1.0   0:1568   cancer    :  46   Male  :884
##  1st Qu.: 886   1st Qu.:112.0   1:  49   event-free:1568   Female:925
##  Median :1697   Median :119.0   2:  46   heart     :  49
##  Mean   :1697   Mean   :112.8   3: 146   other     : 146
##  3rd Qu.:2528   3rd Qu.:126.0
##  Max.   :3351   Max.   :132.0
##     RIDAGEYR         BMXBMI           LBDLDL             DIQ010
##  Min.   :2.000   Min.   : 15.68   Min.   : 2.20   No        :1611
##  1st Qu.:3.200   1st Qu.: 24.20   1st Qu.: 8.80   Borderline:  30
##  Median :4.600   Median : 27.64   Median :11.30   Yes       : 168
##  Mean   :4.795   Mean   : 28.76   Mean   :11.49
##  3rd Qu.:6.300   3rd Qu.: 31.89   3rd Qu.:13.80
##  Max.   :8.500   Max.   :130.21   Max.   :32.80
##      BPXSY1      SMQ020
```

```
##  Min.   : 8.00   No :923
##  1st Qu.:11.20   Yes:886
##  Median :12.00
##  Mean   :12.41
##  3rd Qu.:13.40
##  Max.   :22.40
```

```
ftime <- data_complete$time
fstatus <- data_complete$status
```

```
# Convert covariate to 0/1 numerical variable:
Gender <- as.numeric(data_complete$RIAGENDR)-1
```

```
# Diabetes indicator variables
DIQ010.Yes <- ifelse(data_complete$DIQ010 == 'Yes', 1, 0)
DIQ010.Borderline <- ifelse(data_complete$DIQ010 == 'Borderline', 1, 0)
```

```
# Check
sum(DIQ010.Yes)
```

```
## [1] 168
```

```
sum(DIQ010.Borderline)
```

```
## [1] 30
```

```
# Smoking indicator variables
SMQ020.Yes <- ifelse(data_complete$SMQ020 == 'Yes', 1, 0)
```

```
sum(SMQ020.Yes )
```

```
## [1] 886
```

```
# Create matrix of covariates - row per individual:
x = data.frame(RIDAGEYR = data_complete$RIDAGEYR, RIAGENDR.Female = Gender,
BMXBMI = data_complete$BMXBMI, LBDLDL = data_complete$LBDLDL,
DIQ010.Borderline, DIQ010.Yes, BPXSY1 = data_complete$BPXSY1, SMQ020.Yes)
```

```
head(data_complete)
```

```
##   SEQN time status      cause RIAGENDR RIDAGEYR BMXBMI LBDLDL DIQ010
BPXSY1
## 1    5  123      0 event-free     Male      7.3  30.63    9.8     No
13.0
## 2   12  123      0 event-free     Male      3.8  25.61    9.1     No
12.6
## 3   13  119      0 event-free     Male      7.1  21.65   11.0     No
15.4
```

```
## 4    3  129      0 event-free     Male      7.0  24.74    7.5    Yes
13.8
## 5   18  126      0 event-free   Female      3.3  25.70    8.6     No
11.4
## 6   20  129      0 event-free   Female      2.2  25.26   11.2     No
10.8
##   SMQ020
## 1     No
## 2     No
## 3    Yes
## 4     No
## 5     No
## 6     No
```

```
head(x)
```

```
##   RIDAGEYR RIAGENDR.Female BMXBMI LBDLDL DIQ010.Borderline DIQ010.Yes
BPXSY1
## 1      7.3               0  30.63    9.8                0          0
13.0
## 2      3.8               0  25.61    9.1                0          0
12.6
## 3      7.1               0  21.65   11.0                0          0
15.4
## 4      7.0               0  24.74    7.5                0          1
13.8
## 5      3.3               1  25.70    8.6                0          0
11.4
## 6      2.2               1  25.26   11.2                0          0
10.8
##   SMQ020.Yes
## 1          0
## 2          0
## 3          1
## 4          0
## 5          0
## 6          0
```

```
### Heart disease deaths
mod1 <- crr(ftime, fstatus, x)
summary(mod1)
```

```
## Competing Risks Regression
##
## Call:
## crr(ftime = ftime, fstatus = fstatus, cov1 = x)
##
##                       coef exp(coef) se(coef)      z p-value
## RIDAGEYR           0.62674     1.872   0.1192  5.258 1.5e-07
## RIAGENDR.Female   -0.66495     0.514   0.3012 -2.208 2.7e-02
## BMXBMI            -0.01206     0.988   0.0309 -0.390 7.0e-01
```

```
## LBDLDL            -0.00527     0.995   0.0463 -0.114 9.1e-01
## DIQ010.Borderline -0.39465     0.674   1.0092 -0.391 7.0e-01
## DIQ010.Yes        -0.12004     0.887   0.4924 -0.244 8.1e-01
## BPXSY1             0.15725     1.170   0.0765  2.057 4.0e-02
## SMQ020.Yes         0.20656     1.229   0.2931  0.705 4.8e-01
##
##                   exp(coef) exp(-coef)   2.5% 97.5%
## RIDAGEYR              1.872      0.534 1.4816 2.364
## RIAGENDR.Female       0.514      1.944 0.2850 0.928
## BMXBMI                0.988      1.012 0.9299 1.050
## LBDLDL                0.995      1.005 0.9085 1.089
## DIQ010.Borderline     0.674      1.484 0.0932 4.871
## DIQ010.Yes            0.887      1.128 0.3379 2.328
## BPXSY1                1.170      0.854 1.0074 1.359
## SMQ020.Yes            1.229      0.813 0.6922 2.184
##
## Num. cases = 1809
## Pseudo Log-likelihood = -321
## Pseudo likelihood ratio test = 84.6  on 8 df,
```

```
## Obtain overall p-value for factors with more than 2 levels
library(aod)
```

```
## Warning: package 'aod' was built under R version 3.6.3
```

```
##
## Attaching package: 'aod'
```

```
## The following object is masked from 'package:survival':
##
##     rats
```

```
wald.test(mod1$var, mod1$coef, Terms = 5:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.21, df = 2, P(> X2) = 0.9
```

```
# Indicates diabetes factor is not statistically significant
```

```
### Cancer deaths
```

```r
mod2 <- crr(ftime, fstatus, x, failcode = 2)
summary(mod2)
```

```
## Competing Risks Regression
##
## Call:
## crr(ftime = ftime, fstatus = fstatus, cov1 = x, failcode = 2)
##
##                       coef exp(coef) se(coef)      z p-value
## RIDAGEYR            0.7162     2.047   0.0925  7.746 9.5e-15
## RIAGENDR.Female    -0.5090     0.601   0.3180 -1.600 1.1e-01
## BMXBMI              0.0355     1.036   0.0114  3.104 1.9e-03
## LBDLDL             -0.0534     0.948   0.0354 -1.509 1.3e-01
## DIQ010.Borderline   0.3413     1.407   0.7363  0.464 6.4e-01
## DIQ010.Yes          0.1956     1.216   0.3899  0.502 6.2e-01
## BPXSY1             -0.0287     0.972   0.0710 -0.404 6.9e-01
## SMQ020.Yes          0.5683     1.765   0.3230  1.759 7.9e-02
##
##                    exp(coef) exp(-coef)  2.5% 97.5%
## RIDAGEYR               2.047      0.489 1.707  2.45
## RIAGENDR.Female        0.601      1.664 0.322  1.12
## BMXBMI                 1.036      0.965 1.013  1.06
## LBDLDL                 0.948      1.055 0.885  1.02
## DIQ010.Borderline      1.407      0.711 0.332  5.96
## DIQ010.Yes             1.216      0.822 0.566  2.61
## BPXSY1                 0.972      1.029 0.846  1.12
## SMQ020.Yes             1.765      0.567 0.937  3.32
##
## Num. cases = 1809
## Pseudo Log-likelihood = -302
## Pseudo likelihood ratio test = 80.7  on 8 df,
```

```r
## Obtain overall p-value for factors with more than 2 levels

wald.test(mod2$var, mod1$coef, Terms = 5:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.34, df = 2, P(> X2) = 0.84
```

```r
# Indicates diabetes factor is not statistically significant
```

```r
### Other deaths
mod3 <- crr(ftime, fstatus, x, failcode = 3)
summary(mod3)
```

```
## Competing Risks Regression
##
## Call:
## crr(ftime = ftime, fstatus = fstatus, cov1 = x, failcode = 3)
##
##                        coef exp(coef) se(coef)      z p-value
## RIDAGEYR            0.75741     2.133   0.0774  9.785  0.0000
## RIAGENDR.Female     0.06825     1.071   0.1748  0.391  0.7000
## BMXBMI             -0.01205     0.988   0.0152 -0.793  0.4300
## LBDLDL             -0.05175     0.950   0.0264 -1.959  0.0500
## DIQ010.Borderline  -0.23401     0.791   0.6046 -0.387  0.7000
## DIQ010.Yes          0.61819     1.856   0.2099  2.945  0.0032
## BPXSY1             -0.00961     0.990   0.0425 -0.226  0.8200
## SMQ020.Yes          0.26531     1.304   0.1756  1.511  0.1300
##
##                    exp(coef) exp(-coef)  2.5% 97.5%
## RIDAGEYR               2.133      0.469 1.833  2.48
## RIAGENDR.Female        1.071      0.934 0.760  1.51
## BMXBMI                 0.988      1.012 0.959  1.02
## LBDLDL                 0.950      1.053 0.902  1.00
## DIQ010.Borderline      0.791      1.264 0.242  2.59
## DIQ010.Yes             1.856      0.539 1.230  2.80
## BPXSY1                 0.990      1.010 0.911  1.08
## SMQ020.Yes             1.304      0.767 0.924  1.84
##
## Num. cases = 1809
## Pseudo Log-likelihood = -955
## Pseudo likelihood ratio test = 258  on 8 df,
```

```r
## Obtain overall p-value for factors with more than 2 levels

wald.test(mod3$var, mod1$coef, Terms = 5:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.66, df = 2, P(> X2) = 0.72
```

```r
# Indicates diabetes factor is not statistically significant
```

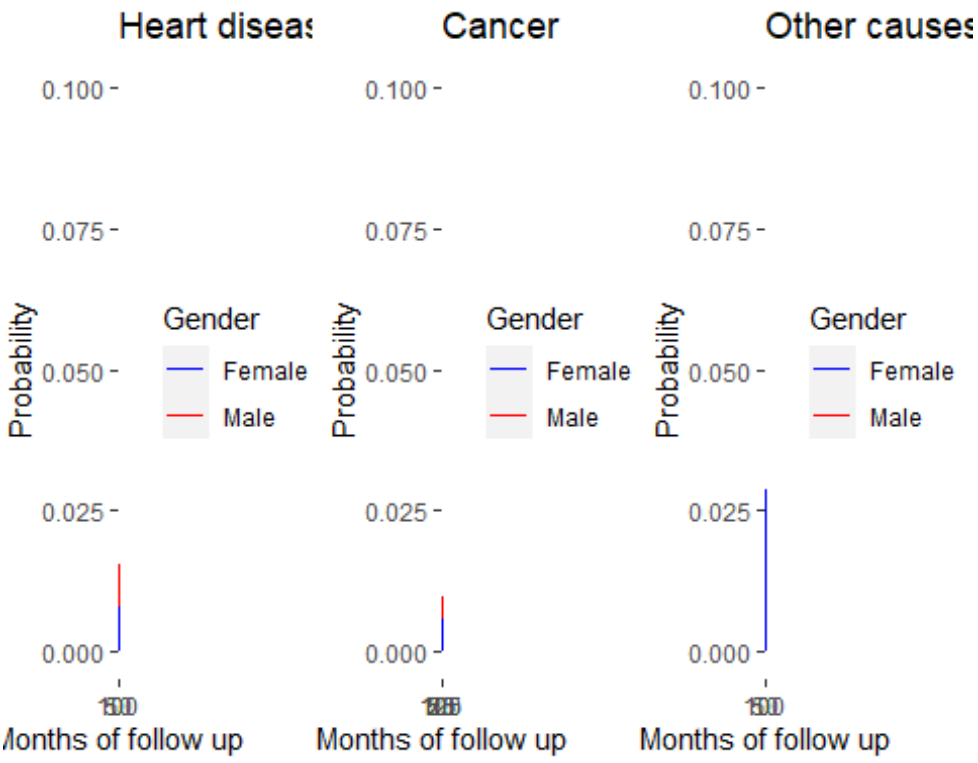
```
crr(ftime, fstatus, DIQ010.Yes)
```

```
## convergence:  TRUE
## coefficients:
## DIQ010.Yes1
##      0.4921
## standard errors:
## [1] 0.4067
## two-sided p-values:
## DIQ010.Yes1
##        0.23
```

Cumulative incidence function prediction

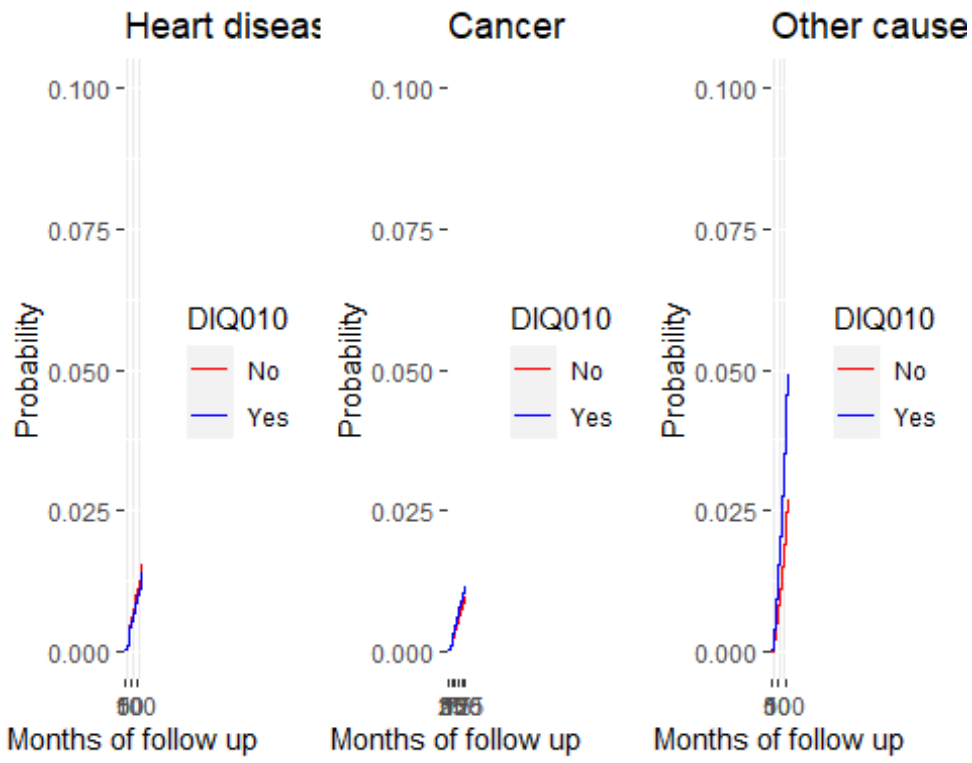
```
# Create matrix to estimate at:
x_ = data.frame(Gender, data_complete$RIDAGEYR, data_complete$BMXBMI,
data_complete$LBDLDL, DIQ010.Borderline, DIQ010.Yes, data_complete$BPXSY1,
SMQ020.Yes)



#### GENDER
# Set of covariates to predict for:
# Heart
pred_genderH <- predict(mod1,rbind(c(4.8,0,28.8,11.5,0,0,12.4,0),
                                   c(4.8,1,28.8,11.5,0,0,12.4,0)))

#pred_genderH

gender_heart <- data.frame(time = pred_genderH[,1], m = pred_genderH[,2], f =
pred_genderH[,3])


# Cancer
pred_genderC <- predict(mod2,rbind(c(4.8,0,28.8,11.5,0,0,12.4,0),
                                   c(4.8,1,28.8,11.5,0,0,12.4,0)))

gender_cancer <- data.frame(time = pred_genderC[,1], m = pred_genderC[,2], f
= pred_genderC[,3])


# Other causes
pred_genderO <- predict(mod3,rbind(c(4.8,0,28.8,11.5,0,0,12.4,0),
                                   c(4.8,1,28.8,11.5,0,0,12.4,0)))

gender_other <- data.frame(time = pred_genderO[,1], m = pred_genderO[,2], f =
pred_genderO[,3])
```

```r
x <- ggplot(gender_heart) +
      geom_step( aes(x = time, y = m, color = 'Male')) +
      geom_step(aes(x = time, y = f, color = 'Female')) +
  labs(title = 'Heart disease') + xlab('Months of follow up')+
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))

y <- ggplot(gender_cancer) +
      geom_step( aes(x = time, y = m, color = 'Male')) +
      geom_step(aes(x = time, y = f, color = 'Female')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))

z <- ggplot(gender_other) +
      geom_step( aes(x = time, y = m, color = 'Male')) +
      geom_step(aes(x = time, y = f, color = 'Female')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))


grid.arrange(x, y, z, nrow=1, ncol=3)
```



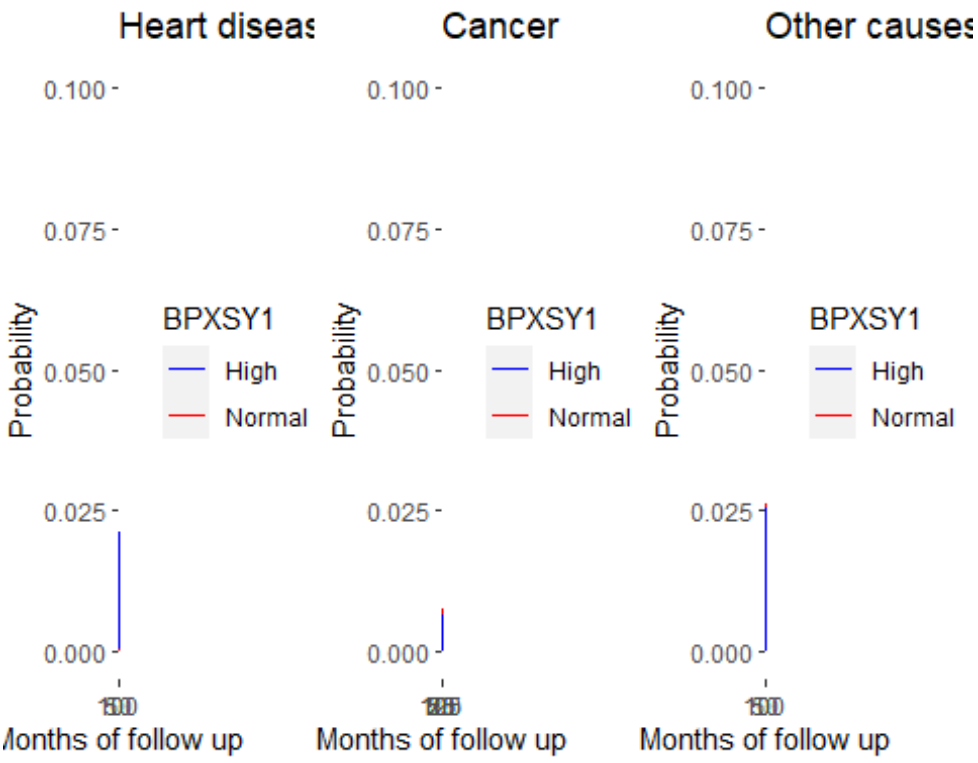

```
#### DIABETES
# Set of covariates to predict for:
# Heart
pred_dbH <- predict(mod1,rbind(c(4.8,0,28.8,11.5,0,0,12.4,0),
                               c(4.8,0,28.8,11.5,0,1,12.4,0)))


db_heart <- data.frame(time = pred_dbH[,1], m = pred_dbH[,2], f =
pred_dbH[,3])


# Cancer
pred_dbC <- predict(mod2,rbind(c(4.8,0,28.8,11.5,0,0,12.4,0),
                               c(4.8,0,28.8,11.5,0,1,12.4,0)))

db_cancer <- data.frame(time = pred_dbC[,1], m = pred_dbC[,2], f =
pred_dbC[,3])


# Other causes
pred_dbO <- predict(mod3,rbind(c(4.8,0,28.8,11.5,0,0,12.4,0),
                               c(4.8,0,28.8,11.5,0,1,12.4,0)))

db_other <- data.frame(time = pred_dbO[,1], m = pred_dbO[,2], f =
pred_dbO[,3])
```

```r
# PLot
x <- ggplot(db_heart) +
      geom_step( aes(x = time, y = m, color = 'No')) +
      geom_step(aes(x = time, y = f, color = 'Yes')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="DIQ010",
    values=c(No="red", Yes="blue"))

y <- ggplot(db_cancer) +
      geom_step( aes(x = time, y = m, color = 'No')) +
      geom_step(aes(x = time, y = f, color = 'Yes')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="DIQ010",
    values=c(No="red", Yes="blue"))

z <- ggplot(db_other) +
      geom_step( aes(x = time, y = m, color = 'No')) +
      geom_step(aes(x = time, y = f, color = 'Yes')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="DIQ010",
    values=c(No="red", Yes="blue"))


grid.arrange(x, y, z, nrow=1, ncol=3)
```



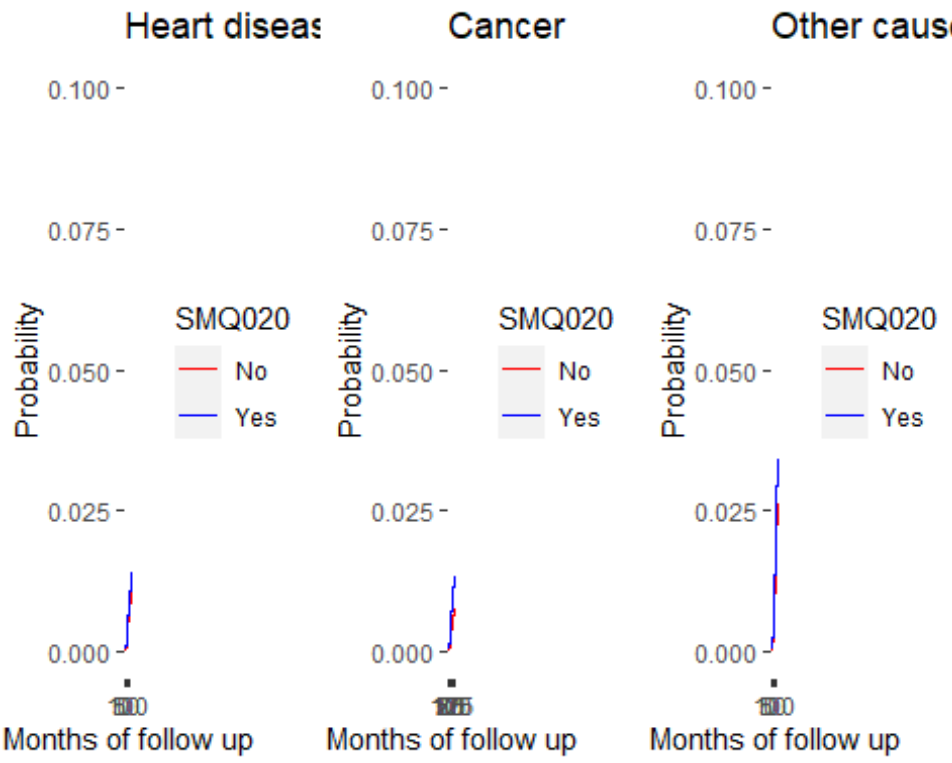

```
#### HIGH BLOOD PRESSURE
# Set of covariates to predict for:
# Heart
pred_highbpH <- predict(mod1,rbind(c(4.5,0,23,9,0,0,11,0),
                                   c(4.5,0,23,9,0,0,15,0)))


highbp_heart <- data.frame(time = pred_highbpH[,1],
                           m = pred_highbpH[,2], f = pred_highbpH[,3])


# Cancer
pred_highbpC <- predict(mod2,rbind(c(4.5,0,23,9,0,0,11,0),
                                   c(4.5,0,23,9,0,0,15,0)))

highbp_cancer <- data.frame(time = pred_highbpC[,1], m = pred_highbpC[,2], f
= pred_highbpC[,3])


# Other causes
pred_highbpO <- predict(mod3,rbind(c(4.5,0,23,9,0,0,11,0),
                                   c(4.5,0,23,9,0,0,15,0)))

highbp_other <- data.frame(time = pred_highbpO[,1], m = pred_highbpO[,2], f =
pred_highbpO[,3])
```

```r
# PLot
x <- ggplot(highbp_heart) +
      geom_step( aes(x = time, y = m, color = 'Normal')) +
      geom_step(aes(x = time, y = f, color = 'High')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="BPXSY1",
    values=c(Normal="red", High="blue"))

y <- ggplot(highbp_cancer) +
      geom_step( aes(x = time, y = m, color = 'Normal')) +
      geom_step(aes(x = time, y = f, color = 'High')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="BPXSY1",
    values=c(Normal="red", High="blue"))

z <- ggplot(highbp_other) +
      geom_step( aes(x = time, y = m, color = 'Normal')) +
      geom_step(aes(x = time, y = f, color = 'High')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="BPXSY1",
    values=c(Normal="red", High="blue"))


grid.arrange(x, y, z, nrow=1, ncol=3)
```




```
#### SMOKING
# Set of covariates to predict for:
# Heart
pred_smokeH <- predict(mod1,rbind(c(4.5,0,23,9,0,0,11,0),
                                  c(4.5,0,23,9,0,0,11,1)))


smoke_heart <- data.frame(time = pred_smokeH[,1],
                          m = pred_smokeH[,2], f = pred_smokeH[,3])


# Cancer
pred_smokeC <- predict(mod2,rbind(c(4.5,0,23,9,0,0,11,0),
                                  c(4.5,0,23,9,0,0,11,1)))

smoke_cancer <- data.frame(time = pred_smokeC[,1], m = pred_smokeC[,2], f =
pred_smokeC[,3])


# Other causes
pred_smokeO <- predict(mod3,rbind(c(4.5,0,23,9,0,0,11,0),
                                  c(4.5,0,23,9,0,0,11,1)))

smoke_other <- data.frame(time = pred_smokeO[,1], m = pred_smokeO[,2], f =
pred_smokeO[,3])
```

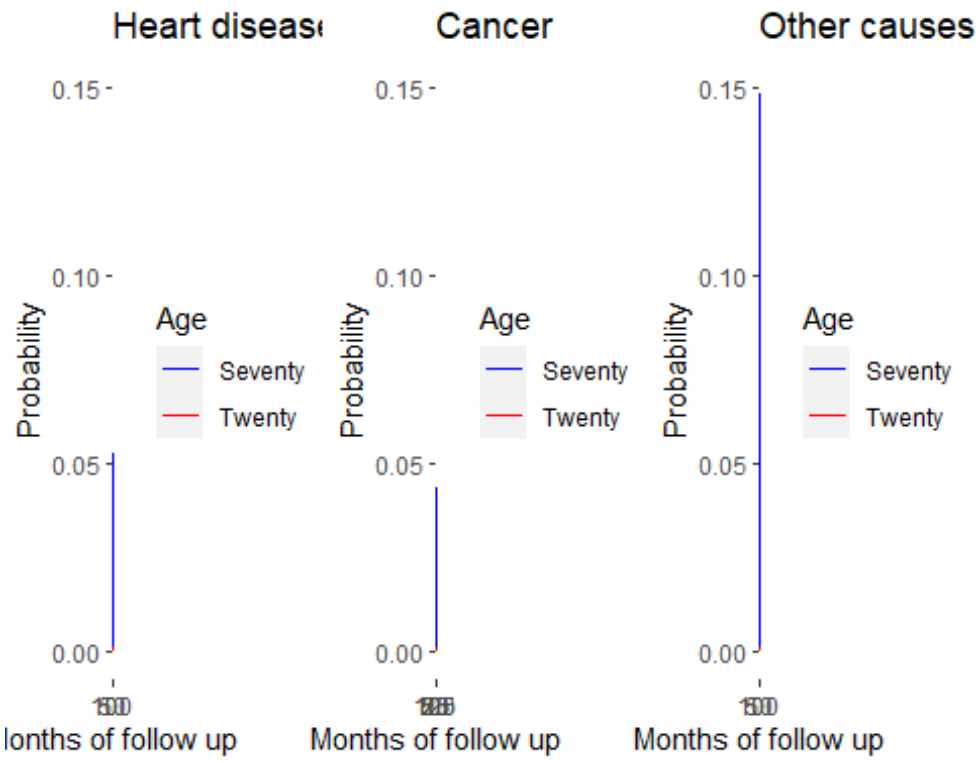
```r
# PLot
x <- ggplot(smoke_heart) +
      geom_step( aes(x = time, y = m, color = 'No')) +
      geom_step(aes(x = time, y = f, color = 'Yes')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="SMQ020",
    values=c(No="red", Yes="blue"))

y <- ggplot(smoke_cancer) +
      geom_step( aes(x = time, y = m, color = 'No')) +
      geom_step(aes(x = time, y = f, color = 'Yes')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="SMQ020",
    values=c(No="red", Yes="blue"))

z <- ggplot(smoke_other) +
      geom_step( aes(x = time, y = m, color = 'No')) +
      geom_step(aes(x = time, y = f, color = 'Yes')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="SMQ020",
    values=c(No="red", Yes="blue"))



grid.arrange(x, y, z, nrow=1, ncol=3)
```




```
#### AGE
# Set of covariates to predict for (comparing 20 and 70 year olds with sam
BMI, LDL, who don't smoke, have diabetes or high blood pressure:
# Heart
pred_ageH <- predict(mod1,rbind(c(2,0,23,9,0,0,11,0),
                                c(7,0,23,9,0,0,11,0)))


age_heart <- data.frame(time = pred_ageH[,1],
                        m = pred_ageH[,2], f = pred_ageH[,3])


# Cancer
pred_ageC <- predict(mod2,rbind(c(2,0,23,9,0,0,11,0),
                                c(7,0,23,9,0,0,11,0)))

age_cancer <- data.frame(time = pred_ageC[,1], m = pred_ageC[,2], f =
pred_ageC[,3])


# Other causes
pred_ageO <- predict(mod3,rbind(c(2,0,23,9,0,0,11,0),
                                c(7,0,23,9,0,0,11,0)))

age_other <- data.frame(time = pred_ageO[,1], m = pred_ageO[,2], f =
pred_ageO[,3])
```

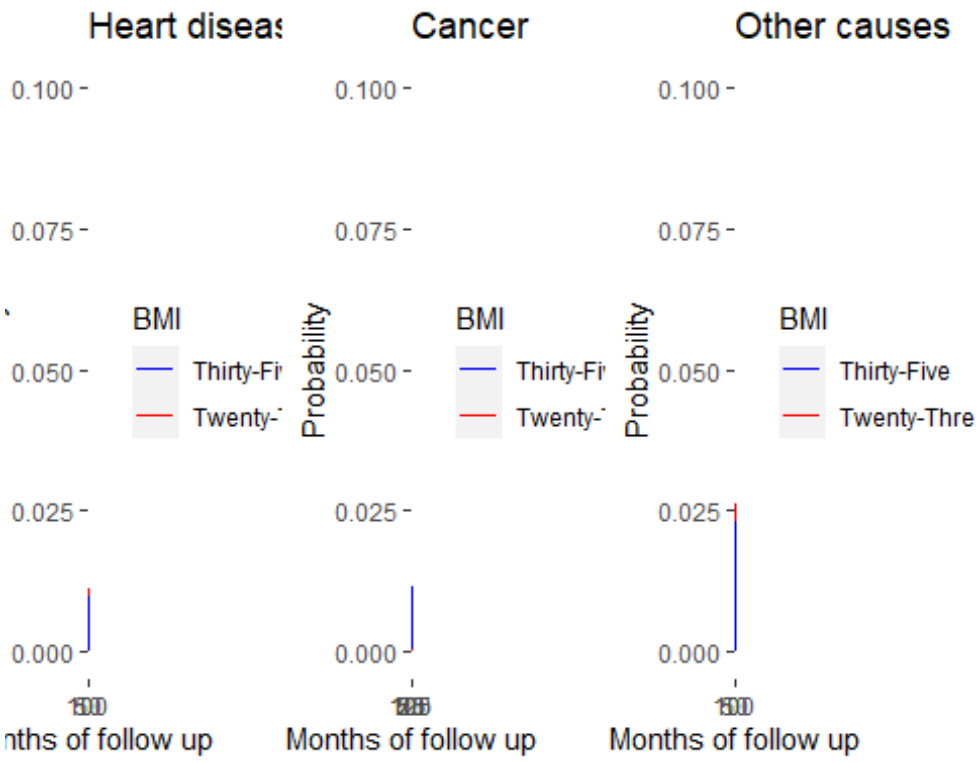
```r
# PLot
x <- ggplot(age_heart) +
      geom_step( aes(x = time, y = m, color = 'Twenty')) +
      geom_step(aes(x = time, y = f, color = 'Seventy')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.15) +
scale_colour_manual(name="Age",
    values=c(Twenty="red", Seventy="blue"))

y <- ggplot(age_cancer) +
      geom_step( aes(x = time, y = m, color = 'Twenty')) +
      geom_step(aes(x = time, y = f, color = 'Seventy')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.15) +
scale_colour_manual(name="Age",
    values=c(Twenty="red", Seventy="blue"))

z <- ggplot(age_other) +
      geom_step( aes(x = time, y = m, color = 'Twenty')) +
      geom_step(aes(x = time, y = f, color = 'Seventy')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.15) +
scale_colour_manual(name="Age",
    values=c(Twenty="red", Seventy="blue"))

grid.arrange(x, y, z, nrow=1, ncol=3)
```

```
## Warning: Removed 6 row(s) containing missing values (geom_path).
```



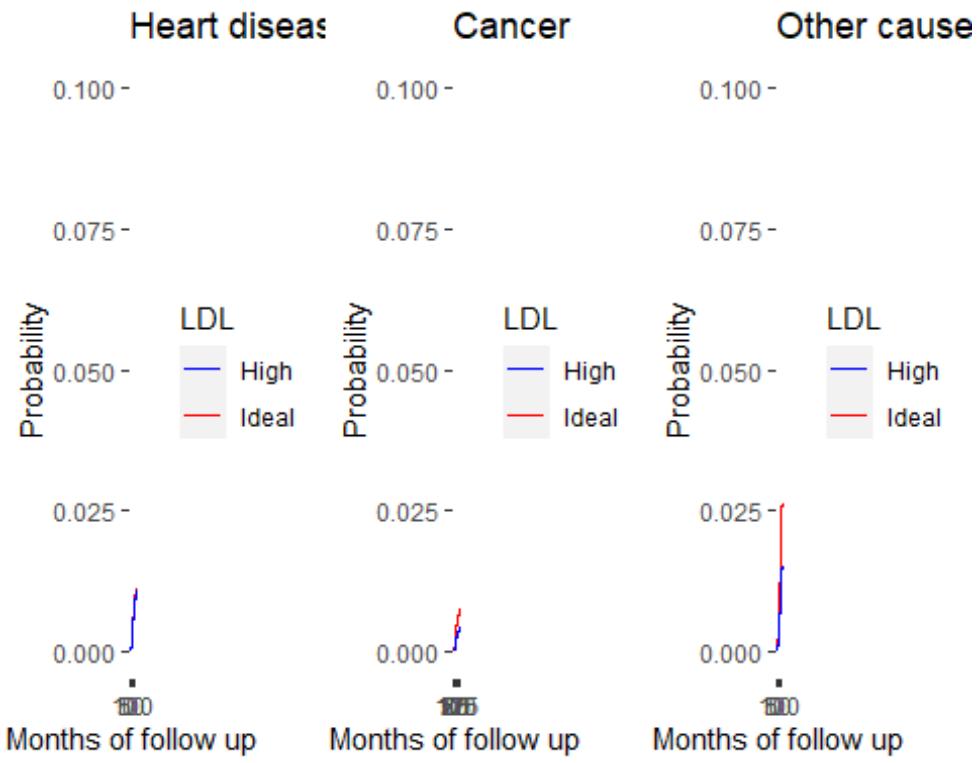

```
#### BMI
# Set of covariates to predict for (comparing people with BMI 23 and BMI 35
with same age, LDL, who don't smoke, have diabetes or high blood pressure:
# Heart
pred_bmiH <- predict(mod1,rbind(c(4.5,0,23,9,0,0,11,0),
                                c(4.5,0,35,9,0,0,11,0)))


bmi_heart <- data.frame(time = pred_bmiH[,1],
                        m = pred_bmiH[,2], f = pred_bmiH[,3])


# Cancer
pred_bmiC <- predict(mod2,rbind(c(4.5,0,23,9,0,0,11,0),
                                c(4.5,0,35,9,0,0,11,0)))

bmi_cancer <- data.frame(time = pred_bmiC[,1], m = pred_bmiC[,2], f =
pred_bmiC[,3])


# Other causes
pred_bmiO <- predict(mod3,rbind(c(4.5,0,23,9,0,0,11,0),
                                c(4.5,0,35,9,0,0,11,0)))

bmi_other <- data.frame(time = pred_bmiO[,1], m = pred_bmiO[,2], f =
pred_bmiO[,3])
```

```r
# PLot
x <- ggplot(bmi_heart) +
      geom_step( aes(x = time, y = m, color = 'Twenty-Three')) +
      geom_step(aes(x = time, y = f, color = 'Thirty-Five')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="BMI",
    values=c('Twenty-Three'="red", 'Thirty-Five'="blue"))

y <- ggplot(bmi_cancer) +
      geom_step( aes(x = time, y = m, color = 'Twenty-Three')) +
      geom_step(aes(x = time, y = f, color = 'Thirty-Five')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="BMI",
    values=c('Twenty-Three'="red", 'Thirty-Five'="blue"))

z <- ggplot(bmi_other) +
      geom_step( aes(x = time, y = m, color = 'Twenty-Three')) +
      geom_step(aes(x = time, y = f, color = 'Thirty-Five')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="BMI",
    values=c('Twenty-Three'="red", 'Thirty-Five'="blue"))

grid.arrange(x, y, z, nrow=1, ncol=3)
```




```
#### LDL
# Set of covariates to predict for (comparing people with BMI 23 and BMI 35
with same age, BMI, who don't smoke, have diabetes or high blood pressure:
# Heart
pred_ldlH <- predict(mod1,rbind(c(4.5,0,23,9,0,0,11,0),
                                c(4.5,0,23,20,0,0,11,0)))


ldl_heart <- data.frame(time = pred_ldlH[,1],
                        m = pred_ldlH[,2], f = pred_ldlH[,3])


# Cancer
pred_ldlC <- predict(mod2,rbind(c(4.5,0,23,9,0,0,11,0),
                                c(4.5,0,23,20,0,0,11,0)))

ldl_cancer <- data.frame(time = pred_ldlC[,1], m = pred_ldlC[,2], f =
pred_ldlC[,3])


# Other causes
pred_ldlO <- predict(mod3,rbind(c(4.5,0,23,9,0,0,11,0),
                                c(4.5,0,23,20,0,0,11,0)))

ldl_other <- data.frame(time = pred_ldlO[,1], m = pred_ldlO[,2], f =
pred_ldlO[,3])
```

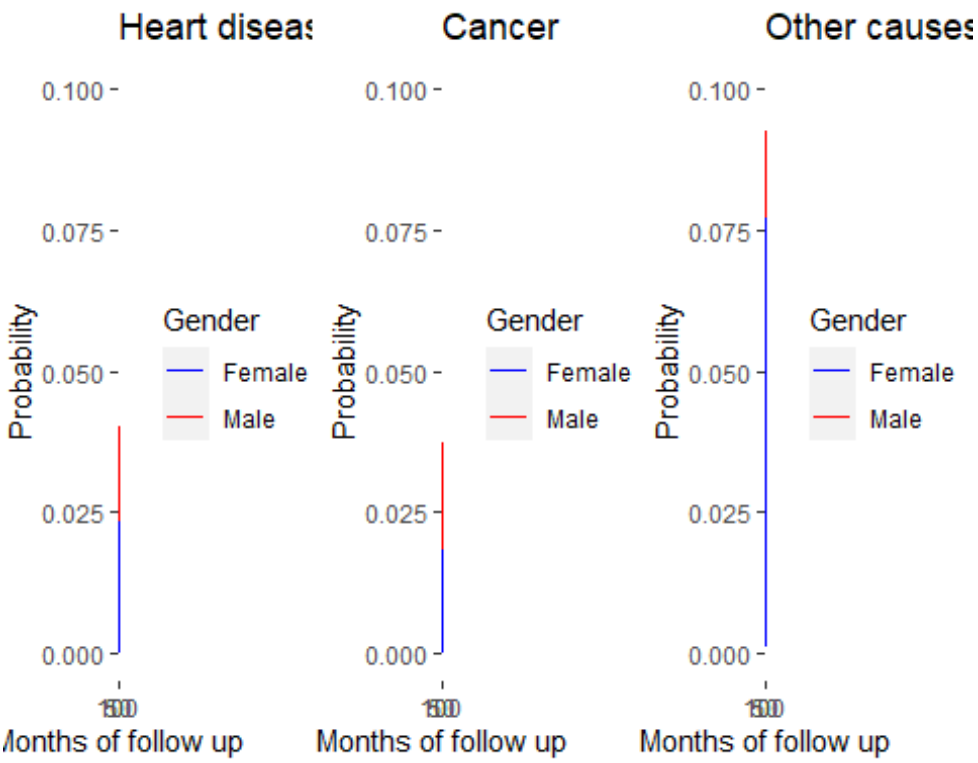
```r
# PLot
x <- ggplot(ldl_heart) +
      geom_step( aes(x = time, y = m, color = 'Ideal')) +
      geom_step(aes(x = time, y = f, color = 'High')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="LDL",
    values=c('Ideal'="red", 'High'="blue"))

y <- ggplot(ldl_cancer) +
      geom_step( aes(x = time, y = m, color = 'Ideal')) +
      geom_step(aes(x = time, y = f, color = 'High')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="LDL",
    values=c('Ideal'="red", 'High'="blue"))


z <- ggplot(ldl_other) +
      geom_step( aes(x = time, y = m, color = 'Ideal')) +
      geom_step(aes(x = time, y = f, color = 'High')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="LDL",
    values=c('Ideal'="red", 'High'="blue"))

grid.arrange(x, y, z, nrow=1, ncol=3)
```



We can compare the Fine-Gray cumulative hazard estimates to the non-parametric cumulative hazard estimates to assess fit

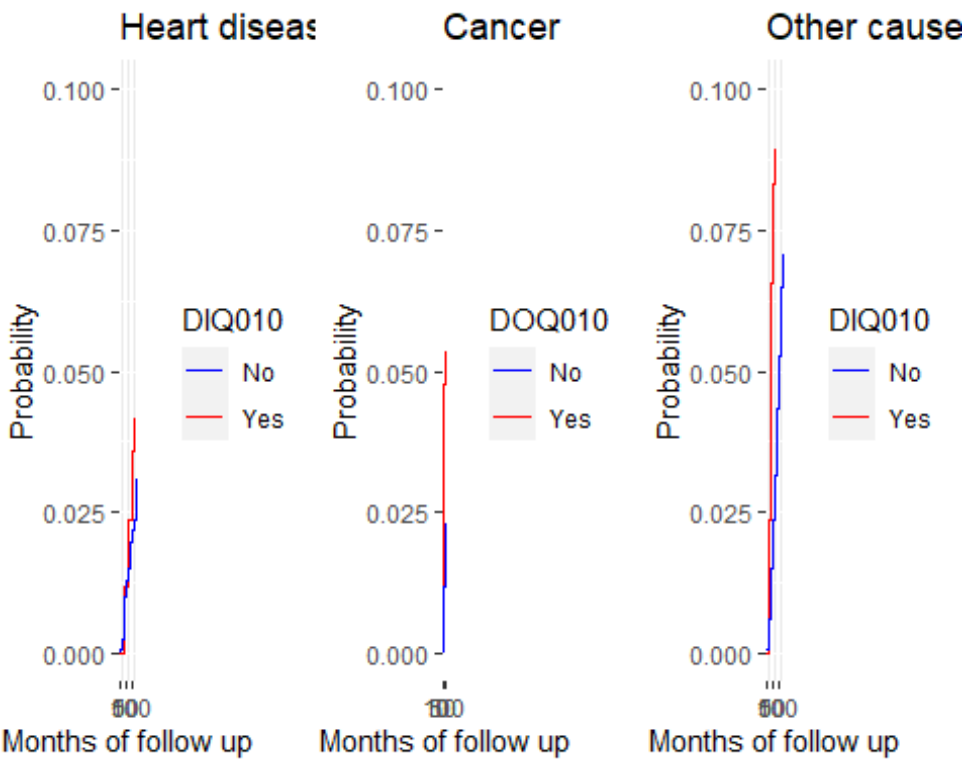
```
## GENDER
ci <- Cuminc(data_complete$time, as.numeric(data_complete$status),
             group = data_complete$RIAGENDR)

ci.m <- ci[ci$group == "Male", ]
ci.f <- ci[ci$group == "Female", ]

#Plot

x <- ggplot(data = NULL, aes(x= time, y = CI.2)) +
      geom_step(data = ci.m, aes(color = 'Male')) +
      geom_step(data = ci.f, aes(color = 'Female')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))

y <- ggplot(data = NULL, aes(x= time, y = CI.3)) +
      geom_step(data = ci.m, aes(color = 'Male')) +
      geom_step(data = ci.f, aes(color = 'Female')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))
```

```
z <- ggplot(data = NULL, aes(x= time, y = CI.4)) +
      geom_step(data = ci.m, aes( color = 'Male')) +
      geom_step(data = ci.f, aes( color = 'Female')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="Gender",
    values=c(Male="red", Female="blue"))


grid.arrange(x, y, z, nrow=1, ncol=3)
```



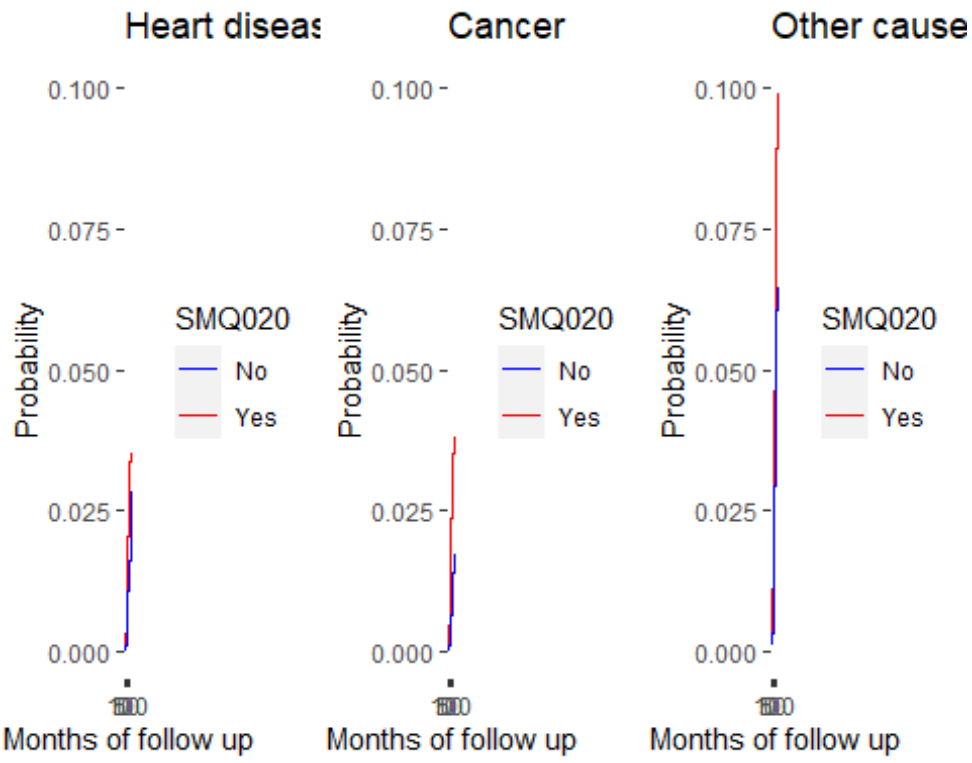

```
## DIABETES
ci <- Cuminc(data_complete$time, as.numeric
             (data_complete$status),
             group = data_complete$DIQ010)

ci.yes <- ci[ci$group == "Yes", ]
ci.no <- ci[ci$group == "No", ]

#Plot

x <- ggplot(data = NULL, aes(x= time, y = CI.2)) +
      geom_step(data = ci.yes, aes(color = 'Yes')) +
      geom_step(data = ci.no, aes(color = 'No')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
```

```
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="DIQ010",
    values=c(Yes="red", No="blue"))

y <- ggplot(data = NULL, aes(x= time, y = CI.3)) +
      geom_step(data = ci.yes, aes(color = 'Yes')) +
      geom_step(data = ci.no, aes(color = 'No')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="DOQ010",
    values=c(Yes="red", No="blue"))

z <- ggplot(data = NULL, aes(x= time, y = CI.4)) +
      geom_step(data = ci.yes, aes(color = 'Yes')) +
      geom_step(data = ci.no, aes(color = 'No')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="DIQ010",
    values=c(Yes="red", No="blue"))

grid.arrange(x, y, z, nrow=1, ncol=3)

## Warning: Removed 22 row(s) containing missing values (geom_path).
```




```
## SMOKE
ci <- Cuminc(data_complete$time, as.numeric
             (data_complete$status),
```

```
             group = data_complete$SMQ020)

ci.yes <- ci[ci$group == "Yes", ]
ci.no <- ci[ci$group == "No", ]


#Plot

x <- ggplot(data = NULL, aes(x= time, y = CI.2)) +
      geom_step(data = ci.yes, aes(color = 'Yes')) +
      geom_step(data = ci.no, aes(color = 'No')) +
  labs(title = 'Heart disease') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="SMQ020",
    values=c(Yes="red", No="blue"))

y <- ggplot(data = NULL, aes(x= time, y = CI.3)) +
      geom_step(data = ci.yes, aes(color = 'Yes')) +
      geom_step(data = ci.no, aes(color = 'No')) +
  labs(title = 'Cancer') + xlab('Months of follow up') + ylab('Probability')
+ ylim(0, 0.1) +
scale_colour_manual(name="SMQ020",
    values=c(Yes="red", No="blue"))

z <- ggplot(data = NULL, aes(x= time, y = CI.4)) +
      geom_step(data = ci.yes, aes(color = 'Yes')) +
      geom_step(data = ci.no, aes(color = 'No')) +
  labs(title = 'Other causes') + xlab('Months of follow up') +
ylab('Probability') + ylim(0, 0.1) +
scale_colour_manual(name="SMQ020",
    values=c(Yes="red", No="blue"))

grid.arrange(x, y, z, nrow=1, ncol=3)

## Warning: Removed 3 row(s) containing missing values (geom_path).
```



Diagnostics - Proportionality assumption

Fine-Gray model Schoenfeld residuals

```
### Fine-Gray

# Schoenfeld residual plots with smoother:
# If the residuals do not have a constant mean across time the proportional
hazards assumption is violated.


# cause 1


datafgcoxph <- data_complete %>% mutate(status = factor(status))
data_h <- finegray(Surv(time, status) ~ ., id = SEQN, data=datafgcoxph)

summary(data_h)

##       SEQN             cause        RIAGENDR       RIDAGEYR
##  Min.   :   2   cancer     : 267   Male  :1400   Min.   :2.000
##  1st Qu.: 905   event-free:1568   Female:1329   1st Qu.:3.700
##  Median :1716   heart      :  49                 Median :5.800
##  Mean   :1701   other      : 845                 Mean   :5.562
##  3rd Qu.:2500                                    3rd Qu.:7.400
##  Max.   :3351                                    Max.   :8.500
```

```
##      BMXBMI           LBDLDL              DIQ010         BPXSY1      SMQ020
##  Min.   : 15.68   Min.   : 2.20   No        :2289   Min.   : 8.0   No
:1273
##  1st Qu.: 24.24   1st Qu.: 8.60   Borderline:  50   1st Qu.:11.2
Yes:1456
##  Median : 27.64   Median :10.70   Yes       : 390   Median :12.4
##  Mean   : 28.72   Mean   :11.13                     Mean   :12.8
##  3rd Qu.: 32.26   3rd Qu.:13.50                     3rd Qu.:14.0
##  Max.   :130.21   Max.   :32.80                     Max.   :22.4
##     fgstart          fgstop         fgstatus           fgwt
##  Min.   :  0.00   Min.   :  2.0   Min.   :0.00000   Min.   :0.1569
##  1st Qu.:  0.00   1st Qu.:113.0   1st Qu.:0.00000   1st Qu.:0.8336
##  Median :  0.00   Median :121.0   Median :0.00000   Median :1.0000
##  Mean   : 40.08   Mean   :118.2   Mean   :0.01796   Mean   :0.8494
##  3rd Qu.:112.00   3rd Qu.:126.0   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :128.00   Max.   :132.0   Max.   :1.00000   Max.   :1.0000
```

```
fgc1 <- coxph(Surv(fgstart, fgstop, fgstatus) ~ RIDAGEYR + RIAGENDR + BMXBMI
+ LBDLDL + DIQ010 + BPXSY1 + SMQ020,
                    weight=fgwt, data=data_h)
summary(fgc1)
```

```
## Call:
## coxph(formula = Surv(fgstart, fgstop, fgstatus) ~ RIDAGEYR +
##     RIAGENDR + BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data = data_h,
##     weights = fgwt)
##
##   n= 2729, number of events= 49
##
##                       coef exp(coef)  se(coef)      z Pr(>|z|)
## RIDAGEYR          0.626873  1.871748  0.109983  5.700  1.2e-08 ***
## RIAGENDRFemale   -0.664887  0.514332  0.325184 -2.045   0.0409 *
## BMXBMI           -0.012095  0.987978  0.026008 -0.465   0.6419
## LBDLDL           -0.005211  0.994803  0.040911 -0.127   0.8986
## DIQ010Borderline -0.394023  0.674338  1.018432 -0.387   0.6988
## DIQ010Yes        -0.120747  0.886258  0.428112 -0.282   0.7779
## BPXSY1            0.157602  1.170700  0.063463  2.483   0.0130 *
## SMQ020Yes         0.205952  1.228695  0.309440  0.666   0.5057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR            1.8717     0.5343   1.50879    2.3220
## RIAGENDRFemale      0.5143     1.9443   0.27192    0.9728
## BMXBMI              0.9880     1.0122   0.93888    1.0396
## LBDLDL              0.9948     1.0052   0.91815    1.0779
## DIQ010Borderline    0.6743     1.4829   0.09162    4.9633
## DIQ010Yes           0.8863     1.1283   0.38296    2.0510
## BPXSY1              1.1707     0.8542   1.03377    1.3258
## SMQ020Yes           1.2287     0.8139   0.66996    2.2534
```

```
##
## Concordance= 0.841  (se = 0.027 )
## Likelihood ratio test= 84.64  on 8 df,   p=6e-15
## Wald test            = 64.19  on 8 df,   p=7e-11
## Score (logrank) test = 86.91  on 8 df,   p=2e-15
```

```
# cause 2
data_c <- finegray(Surv(time, status) ~ ., id = SEQN, etype = '2',
data=datafgcoxph)

fgc2 <- coxph(Surv(fgstart, fgstop, fgstatus) ~ RIDAGEYR + RIAGENDR + BMXBMI
+ LBDLDL + DIQ010 + BPXSY1 + SMQ020,
                    weight=fgwt, data=data_c)
summary(fgc2)
```

```
## Call:
## coxph(formula = Surv(fgstart, fgstop, fgstatus) ~ RIDAGEYR +
##     RIAGENDR + BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data = data_c,
##     weights = fgwt)
##
##   n= 2347, number of events= 46
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## RIDAGEYR          0.71639   2.04703  0.11548  6.204 5.52e-10 ***
## RIAGENDRFemale   -0.50914   0.60101  0.32473 -1.568   0.1169
## BMXBMI            0.03547   1.03610  0.01520  2.334   0.0196 *
## LBDLDL           -0.05338   0.94802  0.04310 -1.238   0.2155
## DIQ010Borderline  0.34218   1.40802  0.73293  0.467   0.6406
## DIQ010Yes         0.19574   1.21620  0.38544  0.508   0.6116
## BPXSY1           -0.02869   0.97171  0.07326 -0.392   0.6953
## SMQ020Yes         0.56829   1.76524  0.33380  1.702   0.0887 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR            2.0470     0.4885    1.6324     2.567
## RIAGENDRFemale      0.6010     1.6639    0.3180     1.136
## BMXBMI              1.0361     0.9652    1.0057     1.067
## LBDLDL              0.9480     1.0548    0.8712     1.032
## DIQ010Borderline    1.4080     0.7102    0.3348     5.922
## DIQ010Yes           1.2162     0.8222    0.5714     2.589
## BPXSY1              0.9717     1.0291    0.8417     1.122
## SMQ020Yes           1.7652     0.5665    0.9176     3.396
##
## Concordance= 0.866  (se = 0.018 )
## Likelihood ratio test= 80.68  on 8 df,   p=4e-14
## Wald test            = 58.54  on 8 df,   p=9e-10
## Score (logrank) test = 80.82  on 8 df,   p=3e-14
```

```r
# cause 3
data_o <- finegray(Surv(time, status) ~ ., id = SEQN, etype = '3',
data=datafgcoxph)

fgc3 <- coxph(Surv(fgstart, fgstop, fgstatus) ~ RIDAGEYR + RIAGENDR + BMXBMI
+ LBDLDL + DIQ010 + BPXSY1 + SMQ020,
                    weight=fgwt, data=data_o)
summary(fgc3)
```

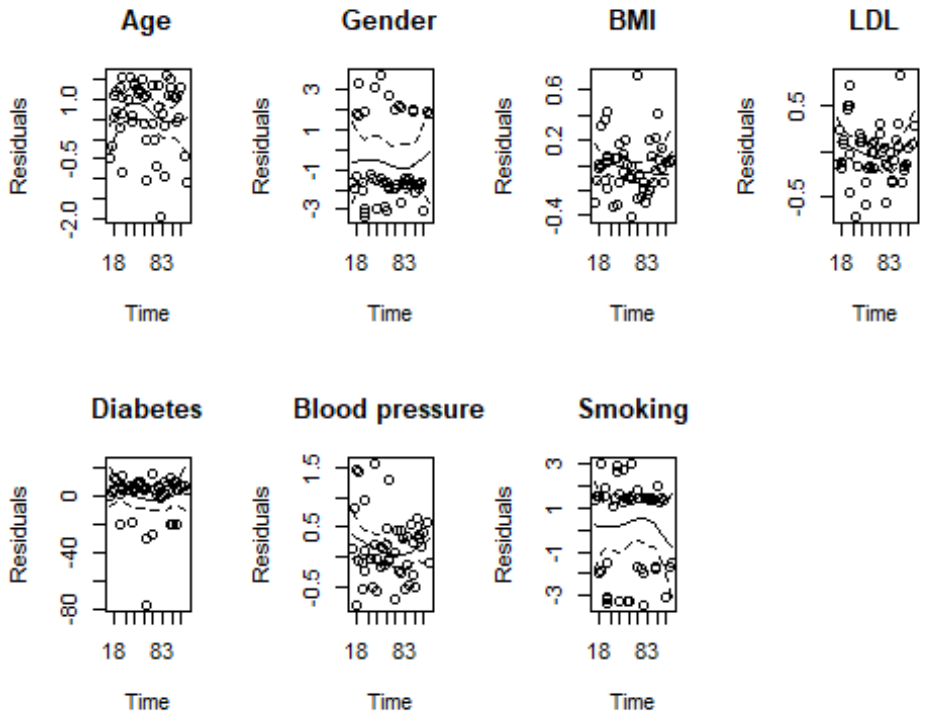
```
## Call:
## coxph(formula = Surv(fgstart, fgstop, fgstatus) ~ RIDAGEYR +
##     RIAGENDR + BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data = data_o,
##     weights = fgwt)
##
##   n= 2704, number of events= 146
##
##                       coef exp(coef)  se(coef)      z Pr(>|z|)
## RIDAGEYR          0.758016  2.134039  0.066871 11.335  < 2e-16 ***
## RIAGENDRFemale    0.068597  1.071005  0.173695  0.395  0.69290
## BMXBMI           -0.012056  0.988016  0.014527 -0.830  0.40660
## LBDLDL           -0.051822  0.949497  0.024206 -2.141  0.03228 *
## DIQ010Borderline -0.234133  0.791257  0.593929 -0.394  0.69343
## DIQ010Yes         0.618663  1.856445  0.203956  3.033  0.00242 **
## BPXSY1           -0.009654  0.990392  0.039339 -0.245  0.80614
## SMQ020Yes         0.265223  1.303722  0.177002  1.498  0.13402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR            2.1340     0.4686    1.8719    2.4329
## RIAGENDRFemale      1.0710     0.9337    0.7620    1.5054
## BMXBMI              0.9880     1.0121    0.9603    1.0166
## LBDLDL              0.9495     1.0532    0.9055    0.9956
## DIQ010Borderline    0.7913     1.2638    0.2470    2.5344
## DIQ010Yes           1.8564     0.5387    1.2447    2.7688
## BPXSY1              0.9904     1.0097    0.9169    1.0698
## SMQ020Yes           1.3037     0.7670    0.9216    1.8444
##
## Concordance= 0.841  (se = 0.016 )
## Likelihood ratio test= 258.2  on 8 df,   p=<2e-16
## Wald test            = 182.3  on 8 df,   p=<2e-16
## Score (logrank) test = 267.8  on 8 df,   p=<2e-16
```

```r
temp <- cox.zph(fgc1)
temp2 <- cox.zph(fgc2)
temp3 <- cox.zph(fgc3)
print(temp)
```

```
##            chisq df    p
## RIDAGEYR 0.99516  1 0.32
```

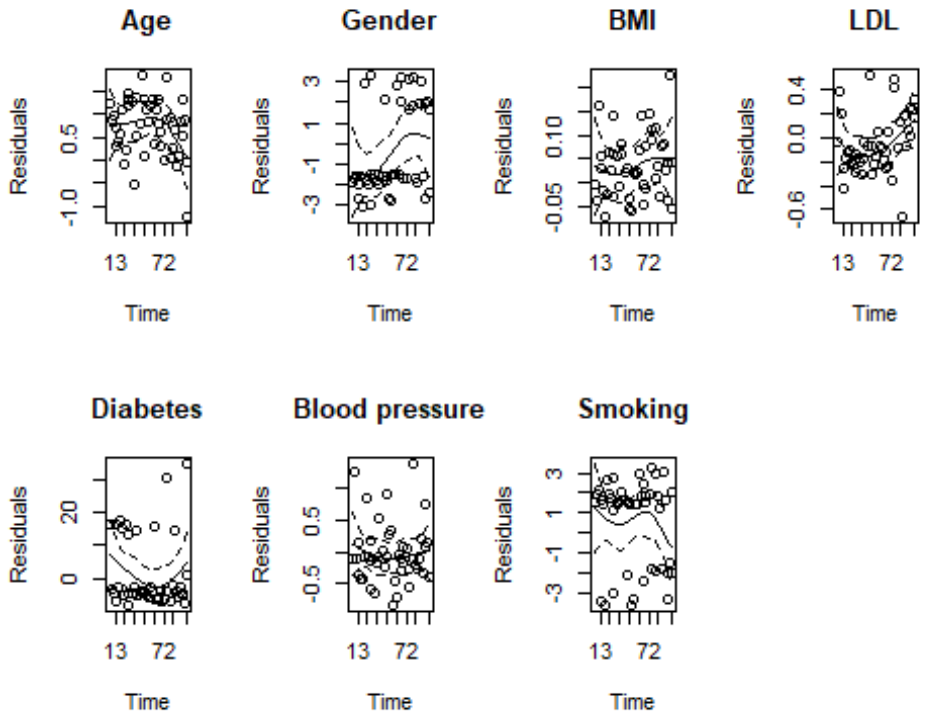
```
## RIAGENDR 0.00621  1 0.94
## BMXBMI   0.15490  1 0.69
## LBDLDL   0.24365  1 0.62
## DIQ010   0.64381  2 0.72
## BPXSY1   1.06043  1 0.30
## SMQ020   0.05982  1 0.81
## GLOBAL   2.61681  8 0.96
```

```
# plot curves

par(mfrow = c(2,4))

names = c('Age', 'Gender', 'BMI', 'LDL', 'Diabetes', 'Blood pressure',
'Smoking')

for(j in 1:7)
  plot(temp[j], resid = T, se = T, main = names[j],
                  xlab = 'Time',
                  ylab = 'Residuals')

par(mfrow = c(2,4))
```

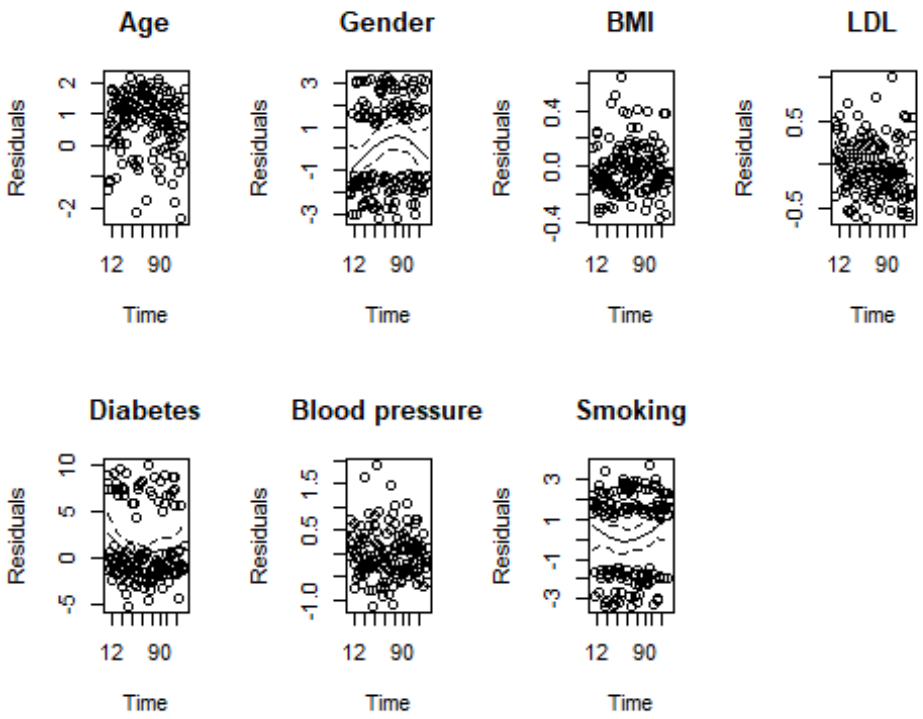



```
for(j in 1:7)
  plot(temp2[j], resid = T, se = T, main = names[j],
                  xlab = 'Time',
                  ylab = 'Residuals')
```

```r
par(mfrow = c(2,4))
```



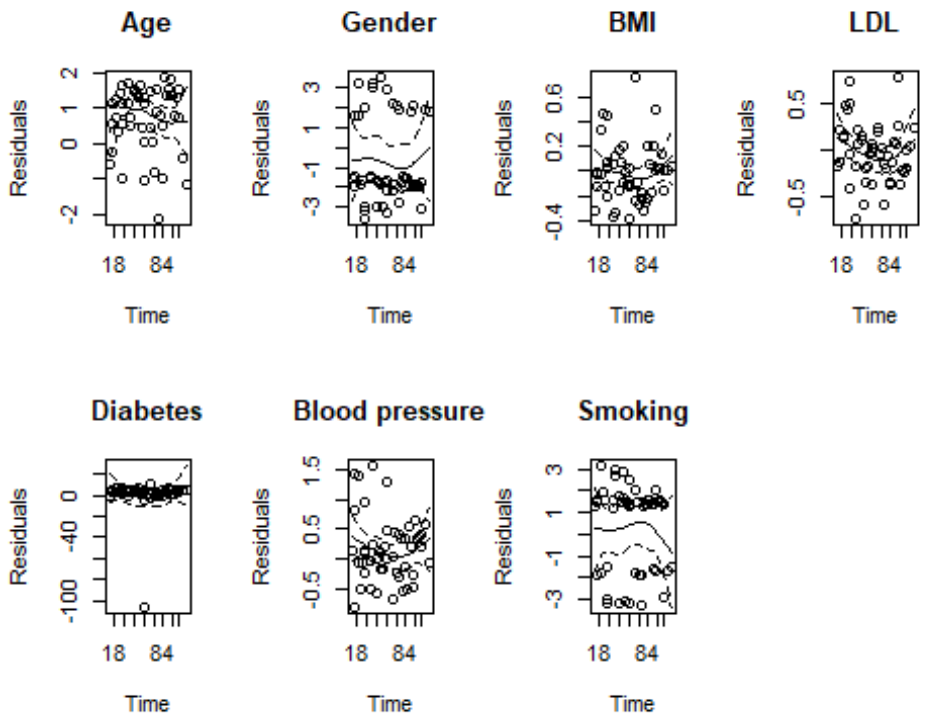

```r
for(j in 1:7)
  plot(temp3[j], resid = T, se = T, main = names[j],
                 xlab = 'Time',
                 ylab = 'Residuals')
```



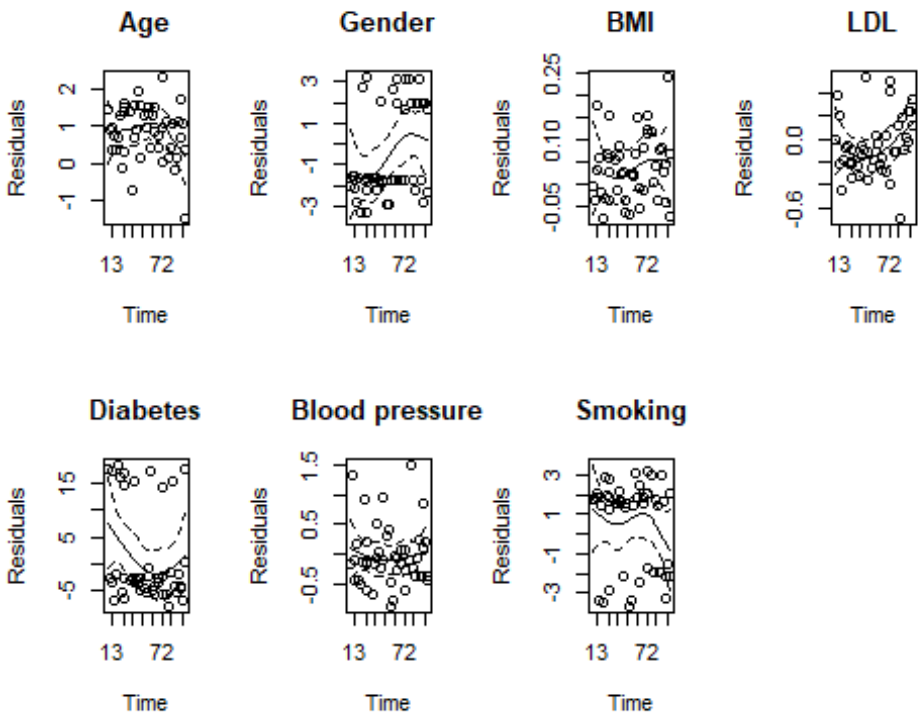
Cox model Schoenfeld residuals

```
# Proportional hazards assumption
c1 <- coxph(Surv(time,status==1)~  RIDAGEYR + RIAGENDR + BMXBMI + LBDLDL +
DIQ010 + BPXSY1 + SMQ020, data = data_short)

temp <- cox.zph(c1)
print(temp)

##             chisq df    p
## RIDAGEYR 0.14115  1 0.71
## RIAGENDR 0.00345  1 0.95
## BMXBMI   0.06981  1 0.79
## LBDLDL   0.75282  1 0.39
## DIQ010   0.82555  2 0.66
## BPXSY1   0.72865  1 0.39
## SMQ020   0.06817  1 0.79
## GLOBAL   2.29184  8 0.97

# plot curves

par(mfrow = c(2,4))

names = c('Age', 'Gender', 'BMI', 'LDL', 'Diabetes', 'Blood pressure',
'Smoking')

for(j in 1:7)
```

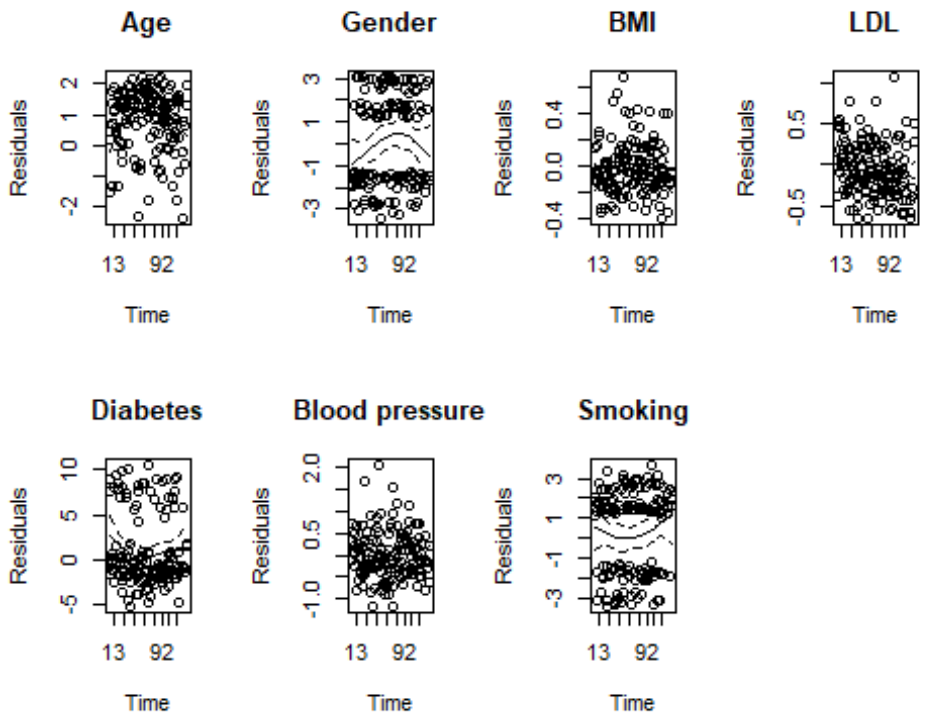
```
  plot(temp[j], resid = T, se = T, main = names[j],
                 xlab = 'Time',
                 ylab = 'Residuals')

c2 <- coxph(Surv(time,status==2)~  RIDAGEYR + RIAGENDR + BMXBMI + LBDLDL +
DIQ010 + BPXSY1 + SMQ020, data = data_short)
temp2 <- cox.zph(c2)

# plot curves
par(mfrow = c(2,4))
```




```
for(j in 1:7)
  plot(temp2[j], resid = T, se = T, main = names[j],
                 xlab = 'Time',
                 ylab = 'Residuals')

c3 <- coxph(Surv(time,status==3)~  RIDAGEYR + RIAGENDR + BMXBMI + LBDLDL +
DIQ010 + BPXSY1 + SMQ020, data = data_short)
temp3 <- cox.zph(c3)
par(mfrow = c(2,4))
```




```
for(j in 1:7)
  plot(temp3[j], resid = T, se = T, main = names[j],
                 xlab = 'Time',
                 ylab = 'Residuals')
```



To obtain the Schoenfeld residuals to check the proportionality assumptions for the Cox and Fine-Gray models above, the `cox.zph` funtion was used from the `survival` package [6].

Simpler models with fewer covariates

```
## Try fitting models with just age, gender and one other covariate to check
if significant...


#########################################################################
########### COX MODEL

cox_mod <- coxph(Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +  RIDAGEYR.3
+ RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3  + BMXBMI.1 +
BMXBMI.2 + BMXBMI.3  + LBDLDL.1 + LBDLDL.2 + LBDLDL.3 + DIQ010Borderline.1 +
DIQ010Borderline.2 + DIQ010Borderline.3 + DIQ010Yes.1 + DIQ010Yes.2 +
DIQ010Yes.3 + BPXSY1.1 + BPXSY1.2 + BPXSY1.3 + SMQ020Yes.1 + SMQ020Yes.2 +
SMQ020Yes.3 + strata(trans), data = data_long)

summary(cox_mod)

## Call:
## coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
##     RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
##     BMXBMI.1 + BMXBMI.2 + BMXBMI.3 + LBDLDL.1 + LBDLDL.2 + LBDLDL.3 +
##     DIQ010Borderline.1 + DIQ010Borderline.2 + DIQ010Borderline.3 +
##     DIQ010Yes.1 + DIQ010Yes.2 + DIQ010Yes.3 + BPXSY1.1 + BPXSY1.2 +
```

```
##     BPXSY1.3 + SMQ020Yes.1 + SMQ020Yes.2 + SMQ020Yes.3 + strata(trans),
##     data = data_long)
##
##   n= 5427, number of events= 241
##
##                          coef exp(coef)  se(coef)      z Pr(>|z|)
## RIDAGEYR.1           0.725583  2.065935  0.116254  6.241 4.34e-10 ***
## RIDAGEYR.2           0.837456  2.310481  0.124960  6.702 2.06e-11 ***
## RIDAGEYR.3           0.825470  2.282955  0.068883 11.984  < 2e-16 ***
## RIAGENDRFemale.1    -0.723129  0.485231  0.325143 -2.224  0.02615 *
## RIAGENDRFemale.2    -0.585039  0.557084  0.325258 -1.799  0.07207 .
## RIAGENDRFemale.3    -0.016164  0.983966  0.173985 -0.093  0.92598
## BMXBMI.1            -0.015148  0.984966  0.027072 -0.560  0.57578
## BMXBMI.2             0.034787  1.035399  0.015870  2.192  0.02838 *
## BMXBMI.3            -0.016408  0.983726  0.015058 -1.090  0.27585
## LBDLDL.1            -0.016031  0.984097  0.041231 -0.389  0.69742
## LBDLDL.2            -0.065582  0.936522  0.044421 -1.476  0.13984
## LBDLDL.3            -0.063351  0.938614  0.024706 -2.564  0.01034 *
## DIQ010Borderline.1 -0.449747  0.637789  1.020466 -0.441  0.65941
## DIQ010Borderline.2  0.255683  1.291343  0.734145  0.348  0.72764
## DIQ010Borderline.3 -0.222766  0.800302  0.593135 -0.376  0.70723
## DIQ010Yes.1         -0.013510  0.986581  0.431107 -0.031  0.97500
## DIQ010Yes.2          0.289745  1.336087  0.388163  0.746  0.45539
## DIQ010Yes.3          0.617177  1.853687  0.206199  2.993  0.00276 **
## BPXSY1.1             0.158713  1.172001  0.064290  2.469  0.01356 *
## BPXSY1.2            -0.028800  0.971611  0.075208 -0.383  0.70177
## BPXSY1.3            -0.001786  0.998215  0.040503 -0.044  0.96482
## SMQ020Yes.1          0.239402  1.270490  0.309425  0.774  0.43911
## SMQ020Yes.2          0.586168  1.797088  0.334536  1.752  0.07974 .
## SMQ020Yes.3          0.321161  1.378727  0.177801  1.806  0.07087 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR.1             2.0659     0.4840   1.64498    2.5946
## RIDAGEYR.2             2.3105     0.4328   1.80857    2.9517
## RIDAGEYR.3             2.2830     0.4380   1.99464    2.6129
## RIAGENDRFemale.1       0.4852     2.0609   0.25656    0.9177
## RIAGENDRFemale.2       0.5571     1.7951   0.29448    1.0539
## RIAGENDRFemale.3       0.9840     1.0163   0.69965    1.3838
## BMXBMI.1               0.9850     1.0153   0.93407    1.0386
## BMXBMI.2               1.0354     0.9658   1.00369    1.0681
## BMXBMI.3               0.9837     1.0165   0.95512    1.0132
## LBDLDL.1               0.9841     1.0162   0.90770    1.0669
## LBDLDL.2               0.9365     1.0678   0.85843    1.0217
## LBDLDL.3               0.9386     1.0654   0.89425    0.9852
## DIQ010Borderline.1     0.6378     1.5679   0.08631    4.7130
## DIQ010Borderline.2     1.2913     0.7744   0.30629    5.4444
## DIQ010Borderline.3     0.8003     1.2495   0.25025    2.5594
## DIQ010Yes.1            0.9866     1.0136   0.42381    2.2966
```

```
## DIQ010Yes.2           1.3361     0.7485   0.62435    2.8592
## DIQ010Yes.3           1.8537     0.5395   1.23743    2.7769
## BPXSY1.1              1.1720     0.8532   1.03325    1.3294
## BPXSY1.2              0.9716     1.0292   0.83845    1.1259
## BPXSY1.3              0.9982     1.0018   0.92204    1.0807
## SMQ020Yes.1           1.2705     0.7871   0.69277    2.3300
## SMQ020Yes.2           1.7971     0.5565   0.93285    3.4620
## SMQ020Yes.3           1.3787     0.7253   0.97305    1.9535
##
## Concordance= 0.86  (se = 0.012 )
## Likelihood ratio test= 486.8  on 24 df,   p=<2e-16
## Wald test            = 336.1  on 24 df,   p=<2e-16
## Score (logrank) test = 500.7  on 24 df,   p=<2e-16
```

```
##### BMI
```

```
c_bmi <- coxph(Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +  RIDAGEYR.3  +
RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3  + BMXBMI.1 + BMXBMI.2
+ BMXBMI.3 + strata(trans), data = data_long)

summary(c_bmi)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
##     RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
##     BMXBMI.1 + BMXBMI.2 + BMXBMI.3 + strata(trans), data = data_long)
##
##   n= 5427, number of events= 241
##
##                       coef exp(coef)  se(coef)      z Pr(>|z|)
## RIDAGEYR.1        0.811076  2.250328  0.108272  7.491 6.83e-14 ***
## RIDAGEYR.2        0.857524  2.357316  0.117400  7.304 2.79e-13 ***
## RIDAGEYR.3        0.850256  2.340245  0.063902 13.306  < 2e-16 ***
## RIAGENDRFemale.1 -0.759204  0.468039  0.310555 -2.445  0.01450 *
## RIAGENDRFemale.2 -0.710692  0.491304  0.314624 -2.259  0.02389 *
## RIAGENDRFemale.3 -0.057136  0.944466  0.166250 -0.344  0.73109
## BMXBMI.1         -0.013066  0.987019  0.026037 -0.502  0.61580
## BMXBMI.2          0.035665  1.036308  0.013684  2.606  0.00915 **
## BMXBMI.3         -0.004164  0.995845  0.014161 -0.294  0.76873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR.1          2.2503     0.4444    1.8201    2.7823
## RIDAGEYR.2          2.3573     0.4242    1.8728    2.9672
## RIDAGEYR.3          2.3402     0.4273    2.0648    2.6525
## RIAGENDRFemale.1    0.4680     2.1366    0.2546    0.8603
## RIAGENDRFemale.2    0.4913     2.0354    0.2652    0.9102
## RIAGENDRFemale.3    0.9445     1.0588    0.6818    1.3083
## BMXBMI.1            0.9870     1.0132    0.9379    1.0387
```

```
## BMXBMI.2            1.0363     0.9650    1.0089    1.0645
## BMXBMI.3            0.9958     1.0042    0.9686    1.0239
##
## Concordance= 0.852  (se = 0.012 )
## Likelihood ratio test= 451.7  on 9 df,   p=<2e-16
## Wald test            = 308.7  on 9 df,   p=<2e-16
## Score (logrank) test = 445.4  on 9 df,   p=<2e-16
```

```
# BMI not significant - only on cancer at 10% level
```

```
##### LDL

c_ldl <- coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
    RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
LBDLDL.1 + LBDLDL.2 + LBDLDL.3 + strata(trans), data = data_long)

summary(c_ldl)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
##     RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
##     LBDLDL.1 + LBDLDL.2 + LBDLDL.3 + strata(trans), data = data_long)
##
##   n= 5427, number of events= 241
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## RIDAGEYR.1         0.81790   2.26573  0.10830  7.552 4.29e-14 ***
## RIDAGEYR.2         0.78962   2.20256  0.10931  7.223 5.07e-13 ***
## RIDAGEYR.3         0.83364   2.30168  0.06258 13.321  < 2e-16 ***
## RIAGENDRFemale.1 -0.76833   0.46379  0.31010 -2.478  0.01322 *
## RIAGENDRFemale.2 -0.68083   0.50620  0.31473 -2.163  0.03053 *
## RIAGENDRFemale.3 -0.06202   0.93987  0.16588 -0.374  0.70850
## LBDLDL.1         -0.00873   0.99131  0.04058 -0.215  0.82967
## LBDLDL.2         -0.08565   0.91791  0.04381 -1.955  0.05059 .
## LBDLDL.3         -0.07567   0.92712  0.02444 -3.096  0.00196 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR.1          2.2657     0.4414    1.8324    2.8015
## RIDAGEYR.2          2.2026     0.4540    1.7778    2.7288
## RIDAGEYR.3          2.3017     0.4345    2.0360    2.6020
## RIAGENDRFemale.1    0.4638     2.1562    0.2526    0.8517
## RIAGENDRFemale.2    0.5062     1.9755    0.2732    0.9380
## RIAGENDRFemale.3    0.9399     1.0640    0.6790    1.3009
## LBDLDL.1            0.9913     1.0088    0.9155    1.0734
```

```
## LBDLDL.2            0.9179     1.0894    0.8424    1.0002
## LBDLDL.3            0.9271     1.0786    0.8838    0.9726
##
## Concordance= 0.853  (se = 0.012 )
## Likelihood ratio test= 461.1  on 9 df,   p=<2e-16
## Wald test            = 326.4  on 9 df,   p=<2e-16
## Score (logrank) test = 470.2  on 9 df,   p=<2e-16
```

```
# LDL coefficients similar to in full model
```

```
#### Diabetes
```

```
c_diabetes <- coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
    RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
    DIQ010Yes.1 + DIQ010Yes.2 + DIQ010Yes.3 + DIQ010Borderline.1 +
    DIQ010Borderline.2 + DIQ010Borderline.3 + strata(trans),
    data = data_long)
```

```
summary(c_diabetes)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
##     RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
##     DIQ010Yes.1 + DIQ010Yes.2 + DIQ010Yes.3 + DIQ010Borderline.1 +
##     DIQ010Borderline.2 + DIQ010Borderline.3 + strata(trans),
##     data = data_long)
##
##   n= 5427, number of events= 241
##
##                        coef exp(coef) se(coef)      z Pr(>|z|)
## RIDAGEYR.1          0.82284   2.27696  0.10818  7.606 2.83e-14 ***
## RIDAGEYR.2          0.80790   2.24320  0.11252  7.180 6.98e-13 ***
## RIDAGEYR.3          0.85089   2.34173  0.06515 13.060  < 2e-16 ***
## RIAGENDRFemale.1   -0.77378   0.46127  0.31205 -2.480 0.013149 *
## RIAGENDRFemale.2   -0.72347   0.48507  0.31700 -2.282 0.022476 *
## RIAGENDRFemale.3   -0.12289   0.88436  0.16734 -0.734 0.462717
## DIQ010Yes.1         0.06121   1.06312  0.41185  0.149 0.881858
## DIQ010Yes.2         0.46543   1.59269  0.37703  1.234 0.217031
## DIQ010Yes.3         0.64158   1.89948  0.19462  3.297 0.000979 ***
## DIQ010Borderline.1 -0.43084   0.64996  1.01331 -0.425 0.670705
## DIQ010Borderline.2  0.41711   1.51757  0.72872  0.572 0.567064
## DIQ010Borderline.3 -0.29621   0.74363  0.58617 -0.505 0.613326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR.1            2.2770     0.4392    1.8419    2.8147
```

```
## RIDAGEYR.2            2.2432     0.4458    1.7992    2.7967
## RIDAGEYR.3            2.3417     0.4270    2.0610    2.6607
## RIAGENDRFemale.1      0.4613     2.1679    0.2502    0.8503
## RIAGENDRFemale.2      0.4851     2.0616    0.2606    0.9029
## RIAGENDRFemale.3      0.8844     1.1308    0.6371    1.2276
## DIQ010Yes.1           1.0631     0.9406    0.4743    2.3831
## DIQ010Yes.2           1.5927     0.6279    0.7607    3.3347
## DIQ010Yes.3           1.8995     0.5265    1.2971    2.7816
## DIQ010Borderline.1    0.6500     1.5386    0.0892    4.7361
## DIQ010Borderline.2    1.5176     0.6590    0.3638    6.3305
## DIQ010Borderline.3    0.7436     1.3447    0.2357    2.3459
##
## Concordance= 0.854  (se = 0.012 )
## Likelihood ratio test= 459.4  on 12 df,   p=<2e-16
## Wald test            = 308.2  on 12 df,   p=<2e-16
## Score (logrank) test = 452.1  on 12 df,   p=<2e-16
```

```
# Similar to full model
```

```
##### Smoking
```

```
c_smoke <- coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
    RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
    SMQ020Yes.1 + SMQ020Yes.2 + SMQ020Yes.3 + strata(trans),
    data = data_long)

summary(c_smoke)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
##     RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
##     SMQ020Yes.1 + SMQ020Yes.2 + SMQ020Yes.3 + strata(trans),
##     data = data_long)
##
##   n= 5427, number of events= 241
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## RIDAGEYR.1         0.82013   2.27079  0.10859  7.552 4.28e-14 ***
## RIDAGEYR.2         0.81363   2.25608  0.11260  7.226 4.98e-13 ***
## RIDAGEYR.3         0.85570   2.35302  0.06389 13.393  < 2e-16 ***
## RIAGENDRFemale.1 -0.70379   0.49471  0.32116 -2.191   0.0284 *
## RIAGENDRFemale.2 -0.53340   0.58661  0.32524 -1.640   0.1010
## RIAGENDRFemale.3  0.02196   1.02220  0.17249  0.127   0.8987
```

```
## SMQ020Yes.1       0.23535   1.26536  0.30696  0.767   0.4433
## SMQ020Yes.2       0.55842   1.74791  0.33118  1.686   0.0918 .
## SMQ020Yes.3       0.30069   1.35080  0.17593  1.709   0.0874 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR.1          2.2708     0.4404    1.8354    2.8094
## RIDAGEYR.2          2.2561     0.4432    1.8093    2.8132
## RIDAGEYR.3          2.3530     0.4250    2.0761    2.6669
## RIAGENDRFemale.1    0.4947     2.0214    0.2636    0.9284
## RIAGENDRFemale.2    0.5866     1.7047    0.3101    1.1097
## RIAGENDRFemale.3    1.0222     0.9783    0.7290    1.4334
## SMQ020Yes.1         1.2654     0.7903    0.6933    2.3094
## SMQ020Yes.2         1.7479     0.5721    0.9133    3.3452
## SMQ020Yes.3         1.3508     0.7403    0.9568    1.9070
##
## Concordance= 0.853  (se = 0.011 )
## Likelihood ratio test= 453.8  on 9 df,   p=<2e-16
## Wald test            = 306.9  on 9 df,   p=<2e-16
## Score (logrank) test = 441.9  on 9 df,   p=<2e-16
```

```
# Similar to in full model
```

```
### Blood pressure

c_bp <- coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
    RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
    BPXSY1.1 + BPXSY1.2 + BPXSY1.3 + strata(trans),
    data = data_long)


summary(c_bp)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ RIDAGEYR.1 + RIDAGEYR.2 +
##     RIDAGEYR.3 + RIAGENDRFemale.1 + RIAGENDRFemale.2 + RIAGENDRFemale.3 +
##     BPXSY1.1 + BPXSY1.2 + BPXSY1.3 + strata(trans), data = data_long)
##
##   n= 5427, number of events= 241
##
##                       coef exp(coef)  se(coef)      z Pr(>|z|)
## RIDAGEYR.1        0.742939  2.102104  0.113223  6.562 5.32e-11 ***
## RIDAGEYR.2        0.824635  2.281048  0.115684  7.128 1.02e-12 ***
## RIDAGEYR.3        0.852217  2.344840  0.066355 12.843  < 2e-16 ***
## RIAGENDRFemale.1 -0.810510  0.444631  0.310943 -2.607  0.00914 **
```

```
## RIAGENDRFemale.2 -0.677193  0.508041  0.314832 -2.151  0.03148 *
## RIAGENDRFemale.3 -0.061166  0.940667  0.165979 -0.369  0.71249
## BPXSY1.1          0.154790  1.167413  0.063895  2.423  0.01541 *
## BPXSY1.2         -0.022239  0.978006  0.074277 -0.299  0.76463
## BPXSY1.3          0.002569  1.002572  0.040369  0.064  0.94926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIDAGEYR.1          2.1021     0.4757    1.6838    2.6244
## RIDAGEYR.2          2.2810     0.4384    1.8183    2.8616
## RIDAGEYR.3          2.3448     0.4265    2.0589    2.6705
## RIAGENDRFemale.1    0.4446     2.2491    0.2417    0.8179
## RIAGENDRFemale.2    0.5080     1.9683    0.2741    0.9416
## RIAGENDRFemale.3    0.9407     1.0631    0.6794    1.3023
## BPXSY1.1            1.1674     0.8566    1.0300    1.3232
## BPXSY1.2            0.9780     1.0225    0.8455    1.1313
## BPXSY1.3            1.0026     0.9974    0.9263    1.0851
##
## Concordance= 0.851  (se = 0.012 )
## Likelihood ratio test= 452.8  on 9 df,   p=<2e-16
## Wald test            = 312.6  on 9 df,   p=<2e-16
## Score (logrank) test = 445.4  on 9 df,   p=<2e-16
```

Alternative approach to apply Fine-Gray model

To calculate the weights for the Fine-Gray model the `finegray()` function was used below [7].

```
### Fitting Fine - Gray model using coxph():


heart_fgdat <- finegray(Surv(time, cause) ~ ., data=data_complete,
etype="heart")

summary(heart_fgdat)
```

```
##       SEQN        status     RIAGENDR       RIDAGEYR         BMXBMI
##  Min.   :   2.0   0:2731   Male  :2288   Min.   :2.000   Min.   : 15.68
##  1st Qu.: 930.8   1:  49   Female:2178   1st Qu.:3.600   1st Qu.: 24.15
##  Median :1732.5   2:  46                 Median :5.500   Median : 27.34
##  Mean   :1704.4   3:1640                 Mean   :5.485   Mean   : 28.44
##  3rd Qu.:2456.0                          3rd Qu.:7.400   3rd Qu.: 31.60
##  Max.   :3351.0                          Max.   :8.500   Max.   :130.21
##      LBDLDL             DIQ010         BPXSY1     SMQ020        fgstart
##  Min.   : 2.20   No        :3748   Min.   : 8.0   No :2161   Min.   :
0.00
##  1st Qu.: 8.70   Borderline:  70   1st Qu.:11.2   Yes:2305   1st Qu.:
0.00
##  Median :11.00   Yes       : 648   Median :12.4              Median :
```

```
55.00
##  Mean   :11.28                     Mean   :12.8              Mean   :
58.94
##  3rd Qu.:13.80                     3rd Qu.:14.0              3rd
Qu.:125.00
##  Max.   :32.80                     Max.   :22.4              Max.
:125.00
##      fgstop         fgstatus            fgwt
##  Min.   :  2.0   Min.   :0.00000   Min.   :0.9705
##  1st Qu.:105.0   1st Qu.:0.00000   1st Qu.:0.9955
##  Median :125.0   Median :0.00000   Median :0.9968
##  Mean   :110.4   Mean   :0.01097   Mean   :0.9959
##  3rd Qu.:132.0   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :132.0   Max.   :1.00000   Max.   :1.0000

dim(heart_fgdat)

## [1] 4466   13

summary(data_complete)

##       SEQN           time       status          cause        RIAGENDR
##  Min.   :   2   Min.   :  1.0   0:1568   cancer    :  46   Male  :884
##  1st Qu.: 886   1st Qu.:112.0   1:  49   event-free:1568   Female:925
##  Median :1697   Median :119.0   2:  46   heart     :  49
##  Mean   :1697   Mean   :112.8   3: 146   other     : 146
##  3rd Qu.:2528   3rd Qu.:126.0
##  Max.   :3351   Max.   :132.0
##     RIDAGEYR         BMXBMI           LBDLDL             DIQ010
##  Min.   :2.000   Min.   : 15.68   Min.   : 2.20   No        :1611
##  1st Qu.:3.200   1st Qu.: 24.20   1st Qu.: 8.80   Borderline:  30
##  Median :4.600   Median : 27.64   Median :11.30   Yes       : 168
##  Mean   :4.795   Mean   : 28.76   Mean   :11.49
##  3rd Qu.:6.300   3rd Qu.: 31.89   3rd Qu.:13.80
##  Max.   :8.500   Max.   :130.21   Max.   :32.80
##      BPXSY1      SMQ020
##  Min.   : 8.00   No :923
##  1st Qu.:11.20   Yes:886
##  Median :12.00
##  Mean   :12.41
##  3rd Qu.:13.40
##  Max.   :22.40

fgfit_heart <- coxph(Surv(fgstart, fgstop, fgstatus) ~ RIAGENDR + RIDAGEYR +
BMXBMI+ LBDLDL + DIQ010+ BPXSY1 + SMQ020, data=heart_fgdat, weight= fgwt)

summary(fgfit_heart)

## Call:
## coxph(formula = Surv(fgstart, fgstop, fgstatus) ~ RIAGENDR +
##     RIDAGEYR + BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data =
```

```
heart_fgdat,
##     weights = fgwt)
##
##   n= 4466, number of events= 49
##
##                       coef exp(coef)  se(coef)      z Pr(>|z|)
## RIAGENDRFemale   -0.697692  0.497733  0.325603 -2.143   0.0321 *
## RIDAGEYR          0.646692  1.909214  0.110150  5.871 4.33e-09 ***
## BMXBMI           -0.013180  0.986907  0.026374 -0.500   0.6173
## LBDLDL           -0.004947  0.995065  0.040763 -0.121   0.9034
## DIQ010Borderline -0.409373  0.664067  1.018300 -0.402   0.6877
## DIQ010Yes        -0.104478  0.900795  0.428157 -0.244   0.8072
## BPXSY1            0.157059  1.170065  0.063567  2.471   0.0135 *
## SMQ020Yes         0.230119  1.258750  0.309363  0.744   0.4570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## RIAGENDRFemale      0.4977     2.0091   0.26293    0.9422
## RIDAGEYR            1.9092     0.5238   1.53849    2.3693
## BMXBMI              0.9869     1.0133   0.93719    1.0393
## LBDLDL              0.9951     1.0050   0.91866    1.0778
## DIQ010Borderline    0.6641     1.5059   0.09025    4.8864
## DIQ010Yes           0.9008     1.1101   0.38920    2.0848
## BPXSY1              1.1701     0.8547   1.03300    1.3253
## SMQ020Yes           1.2587     0.7944   0.68645    2.3082
##
## Concordance= 0.833  (se = 0.03 )
## Likelihood ratio test= 88.54  on 8 df,   p=9e-16
## Wald test            = 66.71  on 8 df,   p=2e-11
## Score (logrank) test = 91.16  on 8 df,   p=3e-16
```

3: Pseudo-value Approach

The code used below to implement the pseudo-value approach was adapted from a tutorial by Klein et al. on producing pseudo-value estimates using the `pseudo` package in R [8].

Select grid of 5 time points

```
library(pseudo)

## Warning: package 'pseudo' was built under R version 3.6.3

## Loading required package: KMsurv

## Loading required package: geepack

## Warning: package 'geepack' was built under R version 3.6.3
```

```r
data_pseudo <- data_complete[data_complete$status != 0, ]

data_pseudo <- data_pseudo %>%
  mutate(status = as.integer(status)-1)

skim(data_pseudo)
```

Data summary

Name	data_pseudo
Number of rows	241
Number of columns	11

Column type frequency:

factor	4
numeric	7

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cause	0	1	FALSE	3	oth: 146, hea: 49, can: 46, eve: 0
RIAGENDR	0	1	FALSE	2	Mal: 142, Fem: 99
DIQ010	0	1	FALSE	3	No: 183, Yes: 52, Bor: 6
SMQ020	0	1	FALSE	2	Yes: 151, No: 90

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
SEQN	0	1	1734.34	903.61	9.00	1019.00	1829.00	2500.00	3342.00	▆▆▆▇
time	0	1	62.75	34.01	1.00	33.00	63.00	91.00	129.00	▆▆▆▇
status	0	1	2.40	0.81	1.00	2.00	3.00	3.00	3.00	▃▁▃▁▇
RIDAGEYR	0	1	7.02	1.50	2.30	6.30	7.40	8.30	8.50	▁▁▁▃▇
BMXBMI	0	1	28.47	6.57	15.	24.13	27.26	32.32	52.02	▃▇▃▂▁

					92					
LBDLDL	0	1	10.65	3.70	2.20	8.20	10.00	12.90	22.40	
BPXSY1	0	1	13.64	2.45	8.80	11.80	13.40	15.00	22.40	

```r
summary(data_pseudo$status)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.402   3.000   3.000
```

```r
# Vector of 5-10 evenly spaced time points on the event scale - to find
pseudo-values at


# Quantiles
quantile(data_pseudo$time, probs = c(0.2,0.4,0.6,0.8,1) )
```

```
##  20%  40%  60%  80% 100%
##   27   50   77   97  129
```

```r
t_pts <- quantile(data_pseudo$time, probs = c(0.2,0.4,0.6,0.8,1) )

data_pseudo <- data_complete %>%
  mutate(status = as.integer(status)-1)

summary(data_pseudo)
```

```
##       SEQN           time           status              cause
RIAGENDR
##  Min.   :   2   Min.   :  1.0   Min.   :0.0000   cancer    :  46   Male
:884
##  1st Qu.: 886   1st Qu.:112.0   1st Qu.:0.0000   event-free:1568
Female:925
##  Median :1697   Median :119.0   Median :0.0000   heart     :  49
##  Mean   :1697   Mean   :112.8   Mean   :0.3201   other     : 146
##  3rd Qu.:2528   3rd Qu.:126.0   3rd Qu.:0.0000
##  Max.   :3351   Max.   :132.0   Max.   :3.0000
##     RIDAGEYR         BMXBMI           LBDLDL             DIQ010
##  Min.   :2.000   Min.   : 15.68   Min.   : 2.20   No        :1611
##  1st Qu.:3.200   1st Qu.: 24.20   1st Qu.: 8.80   Borderline:  30
##  Median :4.600   Median : 27.64   Median :11.30   Yes       : 168
##  Mean   :4.795   Mean   : 28.76   Mean   :11.49
##  3rd Qu.:6.300   3rd Qu.: 31.89   3rd Qu.:13.80
##  Max.   :8.500   Max.   :130.21   Max.   :32.80
##      BPXSY1      SMQ020
##  Min.   : 8.00   No :923
##  1st Qu.:11.20   Yes:886
##  Median :12.00
##  Mean   :12.41
```

```
##  3rd Qu.:13.40
##  Max.   :22.40
```

Estimate pseudovalues for each individual

```
pseudo <- pseudoci(time = data_pseudo$time, event = data_pseudo$status, tmax
= t_pts)

dim(pseudo$pseudo$cause1)

## [1] 1809    5

# HEart
b <- NULL
for(it in 1:length(pseudo$time)){
    b <- rbind(b,cbind(data_pseudo,pseudo = pseudo$pseudo$cause1[,it],
         tpseudo = pseudo$time[it],id=1:nrow(data_pseudo)))
}
b <- b[order(b$id),]

b$tpseudo <- factor(b$tpseudo)
skim(b$pseudo)
```

Data summary

Name	b$pseudo
Number of rows	9045
Number of columns	1

_____

Column type frequency:

numeric	1

_____

Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	0	1	0.02	0.15	-0.03	0	0	0	6.37	▇▁▁▁▁

```
# fit the model
library(geepack)


##### HEART
```

```r
fit_heart <- geese(pseudo ~ tpseudo + RIDAGEYR + RIAGENDR + BMXBMI + LBDLDL +
DIQ010 + BPXSY1 + SMQ020, data =b, id=id, jack = TRUE, scale.fix=TRUE,
family=gaussian,
    mean.link = "cloglog", corstr="independence")


#The results using the AJ variance estimate
h1 <- cbind(mean = round(fit_heart$beta,4), SD =
round(sqrt(diag(fit_heart$vbeta.ajs)),4),
    Z = round(fit_heart$beta/sqrt(diag(fit_heart$vbeta.ajs)),4),
    PVal = round(2-
2*pnorm(abs(fit_heart$beta/sqrt(diag(fit_heart$vbeta.ajs)))),4))

# Logit link
fit_heart2 <- geese(pseudo ~ as.factor(tpseudo) + RIDAGEYR + RIAGENDR +
BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data =b, id=id, jack = TRUE,
scale.fix=TRUE, family=gaussian,
    mean.link = "logit", corstr="independence")


#The results using the AJ variance estimate
h2 <- cbind(mean = round(fit_heart2$beta,4), SD =
round(sqrt(diag(fit_heart2$vbeta.ajs)),4),
    Z = round(fit_heart2$beta/sqrt(diag(fit_heart2$vbeta.ajs)),4),
    PVal = round(2-
2*pnorm(abs(fit_heart2$beta/sqrt(diag(fit_heart2$vbeta.ajs)))),4))







# One covariate
fit_heart_diabetes <- geese(pseudo ~ tpseudo + SMQ020, data =b, id=id, jack =
TRUE, scale.fix=TRUE, family=gaussian,
    mean.link = "cloglog", corstr="independence")


#The results using the AJ variance estimate
h1.diab <- cbind(mean = round(fit_heart_diabetes$beta,4), SD =
round(sqrt(diag(fit_heart_diabetes$vbeta.ajs)),4),
    Z =
round(fit_heart_diabetes$beta/sqrt(diag(fit_heart_diabetes$vbeta.ajs)),4),
    PVal = round(2-
2*pnorm(abs(fit_heart_diabetes$beta/sqrt(diag(fit_heart_diabetes$vbeta.ajs)))
),4))

h1.diab
```

```
##                mean     SD        Z   PVal
## (Intercept) -5.2698 0.3617 -14.5698 0.0000
## tpseudo50    0.7134 0.2158   3.3061 0.0009
## tpseudo77    1.0360 0.2396   4.3245 0.0000
## tpseudo97    1.3177 0.2545   5.1773 0.0000
## tpseudo129   1.5970 0.2786   5.7318 0.0000
## SMQ020Yes    0.4099 0.3120   1.3139 0.1889
```

```
#### Cancer

c <- NULL
for(it in 1:length(pseudo$time)){
    c <- rbind(c,cbind(data_pseudo,pseudo = pseudo$pseudo$cause2[,it],
         tpseudo = pseudo$time[it],id=1:nrow(data_pseudo)))
}
c <- c[order(c$id),]

skim(c)
```

Data summary

Name	c
Number of rows	9045
Number of columns	14

Column type frequency:

factor	4
numeric	10

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cause	0	1	FALSE	4	eve: 7840, oth: 730, hea: 245, can: 230
RIAGENDR	0	1	FALSE	2	Fem: 4625, Mal: 4420
DIQ010	0	1	FALSE	3	No: 8055, Yes: 840, Bor: 150
SMQ020	0	1	FALSE	2	No: 4615, Yes: 4430

Variable type: numeric

skim_vari	n_miss	complete_	mean	sd	p0	p25	p50	p75	p100	hist

able	ing	rate								
SEQN	0	1	1697.48	960.81	2.00	886.0	1697.00	2528.00	3351.00	
time	0	1	112.80	24.05	1.00	112.0	119.00	126.00	132.00	
status	0	1	0.32	0.87	0.00	0.0	0.00	0.00	3.00	
RIDAGEYR	0	1	4.79	1.88	2.00	3.2	4.60	6.30	8.50	
BMXBMI	0	1	28.76	7.02	15.68	24.2	27.64	31.89	130.21	
LBDLDL	0	1	11.49	3.67	2.20	8.8	11.30	13.80	32.80	
BPXSY1	0	1	12.41	1.98	8.00	11.2	12.00	13.40	22.40	
pseudo	0	1	0.02	0.13	-0.01	0.0	0.00	0.00	2.82	
tpseudo	0	1	76.00	35.58	27.00	50.0	77.00	97.00	129.00	
id	0	1	905.00	522.24	1.00	453.0	905.00	1357.00	1809.00	

```
fit_cancer <- geese(pseudo ~ as.factor(tpseudo) + RIDAGEYR + RIAGENDR +
BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data =c, id=id, jack = TRUE,
scale.fix=TRUE, family=gaussian,
    mean.link = "cloglog", corstr="independence")


#The results using the AJ variance estimate
c1 <- cbind(mean = round(fit_cancer$beta,3), SD =
round(sqrt(diag(fit_cancer$vbeta.ajs)),3),
    Z = round(fit_cancer$beta/sqrt(diag(fit_cancer$vbeta.ajs)),3),
    PVal = round(2-
2*pnorm(abs(fit_cancer$beta/sqrt(diag(fit_cancer$vbeta.ajs)))),3))

fit_cancer2 <- geese(pseudo ~ as.factor(tpseudo) + RIDAGEYR + RIAGENDR +
BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data =c, id=id, jack = TRUE,
scale.fix=TRUE, family=gaussian,
    mean.link = "logit", corstr="independence")

#The results using the AJ variance estimate
c2 <- cbind(mean = round(fit_cancer2$beta,3), SD =
round(sqrt(diag(fit_cancer2$vbeta.ajs)),3),
    Z = round(fit_cancer2$beta/sqrt(diag(fit_cancer2$vbeta.ajs)),3),
```

```
    PVal = round(2-
2*pnorm(abs(fit_cancer2$beta/sqrt(diag(fit_cancer2$vbeta.ajs)))),3))

#### OTHER CAUSES

d <- NULL
for(it in 1:length(pseudo$time)){
    d <- rbind(d,cbind(data_pseudo,pseudo = pseudo$pseudo$cause3[,it],
         tpseudo = pseudo$time[it],id=1:nrow(data_pseudo)))
}
d <- d[order(d$id),]

skim(d)
```

Data summary

Name	d
Number of rows	9045
Number of columns	14

Column type frequency:

factor	4
numeric	10

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cause	0	1	FALSE	4	eve: 7840, oth: 730, hea: 245, can: 230
RIAGENDR	0	1	FALSE	2	Fem: 4625, Mal: 4420
DIQ010	0	1	FALSE	3	No: 8055, Yes: 840, Bor: 150
SMQ020	0	1	FALSE	2	No: 4615, Yes: 4430

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
SEQN	0	1	1697.48	960.81	2.00	886.0	1697.00	2528.00	3351.00	▇▇▇▇▇
time	0	1	112.8	24.0	1.0	112	119.0	126.0	132.0	▁▁▁

			0	5	0	.0	0	0	0	_▁
status	0	1	0.32	0.87	0.00	0.0	0.00	0.00	3.00	▇_ _ _ _
RIDAGEYR	0	1	4.79	1.88	2.00	3.2	4.60	6.30	8.50	▆▆▅▃
BMXBMI	0	1	28.76	7.02	15.68	24.2	27.64	31.89	130.21	▇_ _ _ _
LBDLDL	0	1	11.49	3.67	2.20	8.8	11.30	13.80	32.80	▃▇▃_ _
BPXSY1	0	1	12.41	1.98	8.00	11.2	12.00	13.40	22.40	▃▇▃_ _
pseudo	0	1	0.05	0.22	-0.01	0.0	0.00	0.00	3.12	▇_ _ _ _
tpseudo	0	1	76.00	35.58	27.00	50.0	77.00	97.00	129.00	▇▇▇▇
id	0	1	905.00	522.24	1.00	453.0	905.00	1357.00	1809.00	▇▇▇▇

```
fit_other <- geese(pseudo ~ as.factor(tpseudo) + RIDAGEYR + RIAGENDR + BMXBMI
+ LBDLDL + DIQ010 + BPXSY1 + SMQ020, data = d, id=id, jack = TRUE,
scale.fix=TRUE, family=gaussian,
    mean.link = "cloglog", corstr="independence")


#The results using the AJ variance estimate
o1 <- cbind(mean = round(fit_other$beta,3), SD =
round(sqrt(diag(fit_other$vbeta.ajs)),3),
    Z = round(fit_other$beta/sqrt(diag(fit_other$vbeta.ajs)),3),
    PVal = round(2-
2*pnorm(abs(fit_other$beta/sqrt(diag(fit_other$vbeta.ajs)))),4))

fit_other2 <- geese(pseudo ~ as.factor(tpseudo) + RIDAGEYR + RIAGENDR +
BMXBMI + LBDLDL + DIQ010 + BPXSY1 + SMQ020, data = d, id=id, jack = TRUE,
scale.fix=TRUE, family=gaussian,
    mean.link = "logit", corstr="independence")

o2 <- cbind(mean = round(fit_other2$beta,3), SD =
round(sqrt(diag(fit_other2$vbeta.ajs)),3),
    Z = round(fit_other2$beta/sqrt(diag(fit_other2$vbeta.ajs)),3),
    PVal = round(2-
2*pnorm(abs(fit_other2$beta/sqrt(diag(fit_other2$vbeta.ajs)))),4))
```

Print pseudo-value approach results

```
th1 <- data.frame(covariate = h1[,0], estimate = h1[,1], se = h1[,2], p =
h1[,4])
```

```r
th2 <- data.frame(covariate = h2[,0], estimate = h2[,1], se = h2[,2], p =
h2[,4])

tc1 <- data.frame(covariate = c1[,0], estimate = c1[,1], se = c1[,2], p =
c1[,4])

tc2 <- data.frame(covariate = c2[,0], estimate = c2[,1], se = c2[,2], p =
c2[,4])

to1 <- data.frame(covariate = o1[,0], estimate = o1[,1], se = o1[,2], p =
o1[,4])

to2 <- data.frame(covariate = o2[,0], estimate = o2[,1], se = o2[,2], p =
o2[,4])

th1
```

```
##                  estimate     se      p
## (Intercept)      -19.7515 6.2286 0.0015
## tpseudo50          0.7387 0.4567 0.1058
## tpseudo77          0.9115 0.4687 0.0518
## tpseudo97          1.1296 0.4813 0.0189
## tpseudo129         1.4923 0.5191 0.0040
## RIDAGEYR           1.6550 0.6721 0.0138
## RIAGENDRFemale    -0.6559 0.5901 0.2664
## BMXBMI            -0.0249 0.0731 0.7336
## LBDLDL            -0.1344 0.1081 0.2137
## DIQ010Borderline  -0.4943 2.3072 0.8304
## DIQ010Yes         -4.3165 1.1364 0.0001
## BPXSY1             0.3710 0.1177 0.0016
## SMQ020Yes         -0.5897 0.5992 0.3250
```

```r
th2
```

```
##                        estimate     se      p
## (Intercept)           -22.4951 7.6549 0.0033
## as.factor(tpseudo)50    0.8462 0.4935 0.0864
## as.factor(tpseudo)77    1.0727 0.5035 0.0331
## as.factor(tpseudo)97    1.3597 0.5136 0.0081
## as.factor(tpseudo)129   1.7748 0.5614 0.0016
## RIDAGEYR                1.9003 0.7970 0.0171
## RIAGENDRFemale         -0.8760 0.6609 0.1850
## BMXBMI                 -0.0236 0.0820 0.7738
## LBDLDL                 -0.1444 0.1239 0.2438
## DIQ010Borderline       -0.7479 2.8246 0.7912
## DIQ010Yes              -5.0835 1.3187 0.0001
## BPXSY1                  0.4255 0.1566 0.0066
## SMQ020Yes              -0.7382 0.6920 0.2861
```

```
tc1

##                         estimate    se     p
## (Intercept)               -6.924 2.000 0.001
## as.factor(tpseudo)50       0.928 0.305 0.002
## as.factor(tpseudo)77       1.253 0.324 0.000
## as.factor(tpseudo)97       1.393 0.339 0.000
## as.factor(tpseudo)129      1.483 0.341 0.000
## RIDAGEYR                   0.489 0.138 0.000
## RIAGENDRFemale            -0.806 0.485 0.097
## BMXBMI                     0.007 0.035 0.833
## LBDLDL                    -0.068 0.043 0.109
## DIQ010Borderline          -0.151 0.906 0.868
## DIQ010Yes                  0.573 0.471 0.224
## BPXSY1                    -0.043 0.088 0.624
## SMQ020Yes                  0.614 0.469 0.191

tc2

##                         estimate    se     p
## (Intercept)               -7.118 2.076 0.001
## as.factor(tpseudo)50       0.966 0.314 0.002
## as.factor(tpseudo)77       1.314 0.331 0.000
## as.factor(tpseudo)97       1.453 0.346 0.000
## as.factor(tpseudo)129      1.552 0.347 0.000
## RIDAGEYR                   0.515 0.145 0.000
## RIAGENDRFemale            -0.857 0.501 0.087
## BMXBMI                     0.009 0.035 0.800
## LBDLDL                    -0.074 0.044 0.091
## DIQ010Borderline          -0.170 0.942 0.857
## DIQ010Yes                  0.604 0.495 0.222
## BPXSY1                    -0.044 0.092 0.629
## SMQ020Yes                  0.645 0.487 0.185

to1

##                         estimate    se      p
## (Intercept)              -14.139 2.000 0.0000
## as.factor(tpseudo)50       0.770 0.211 0.0003
## as.factor(tpseudo)77       1.343 0.248 0.0000
## as.factor(tpseudo)97       1.640 0.255 0.0000
## as.factor(tpseudo)129      1.889 0.259 0.0000
## RIDAGEYR                   1.449 0.187 0.0000
## RIAGENDRFemale             0.029 0.229 0.8981
## BMXBMI                     0.022 0.021 0.2838
## LBDLDL                     0.007 0.034 0.8252
## DIQ010Borderline          -0.316 0.874 0.7176
## DIQ010Yes                  0.775 0.259 0.0028
## BPXSY1                    -0.076 0.049 0.1242
## SMQ020Yes                  0.122 0.234 0.6010
```

```
to2

##                          estimate    se      p
## (Intercept)               -15.533 2.407 0.0000
## as.factor(tpseudo)50        0.848 0.228 0.0002
## as.factor(tpseudo)77        1.508 0.269 0.0000
## as.factor(tpseudo)97        1.865 0.276 0.0000
## as.factor(tpseudo)129       2.197 0.282 0.0000
## RIDAGEYR                    1.637 0.230 0.0000
## RIAGENDRFemale              0.001 0.274 0.9974
## BMXBMI                      0.021 0.025 0.3930
## LBDLDL                      0.007 0.040 0.8658
## DIQ010Borderline           -0.301 0.996 0.7628
## DIQ010Yes                   0.902 0.321 0.0049
## BPXSY1                     -0.083 0.059 0.1614
## SMQ020Yes                   0.131 0.280 0.6408
```

References

[1] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2005-2006. Available: https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2005.

[2] National Center for Health Statistics. Office of Analysis and Epidemiology, Public-use Linked Mortality File, 2015. Hyattsville, Maryland. (Available at the following address: https://www.cdc.gov/nchs/data-linkage/mortality-public.htm).

[3] C.J.Endres,"Introducing nhanesA",[Online], October 16 2018. Available: https://cran.r-project.org/web/packages/nhanesA/vignettes/Introducing_nhanesA.html

[4] H. Putter, "Tutorial in biostatistics: Competing risks and multi-state models Analyses using the mstate package", [Online], May 30, 2014. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.487.3943&rep=rep1&type=pdf

[5] L. Scrucca, A. Santucci, and F. Aversa, "Regression modeling of competing risk using R: an in depth guide for clinicians," Bone Marrow Transplantation, vol. 45, no. 9, pp. 1388-1395, 2010, doi: 10.1038/bmt.2009.359.

[6] T. Therneau, "A package for survival analysis in R",[Online], June 12 2020. Available: https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf

[7] T. Therneau, C. Crowson, and E. Atkinson, "Multi-state models and competing risks",[Online], June 12 2020. Available: https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf

[8] J. P. Klein, M. Gerster, P. K. Andersen, S. Tarima, and M. P. Perme, "SAS and R functions to compute pseudo-values for censored data regression," Computer Methods and Programs in Biomedicine, vol. 89, no. 3, pp. 289-300, 2007, doi: 10.1016/j.cmpb.2007.11.017.