Section 1 Assignment

Instructions (read carefully):

- Each student must submit one assignment. Members of the same group can submit the same assignment.
- Each member of the same group will receive the same grade.
- Please put the name of each group member on the first page.
- Assignments must be done using RMarkdown.
- Submissions must include the .pdf file and the reproducible .rmd file used to do the homework. R code for all applied questions must be provided and be executable in the .rmd file.
- This assignment is due electronically through CANVAS on Friday Jan 28 2022, unless otherwise noted.

Grader Notes:

1) please look for **consistency**. For example, a few points should be deducted if, in one questions, students refer to "observations" in the data, and in another, they refer to "participants" or "subjects" or something else.

2) Students should follow good figure practices. By this I mean low data-ink ratio, no frivolous colors (e.g., pink plot background, just because), no frivolous symbols (e.g., cat emojis for scatter plot points), etc. If this becomes a problem with the homeworks, I will try to do a session on visual analytics.

3) Generally, the document should be neat and orderly. Think formal publication. Extraneous and unnecessary information (e.g., lengthy warning messages, console output that's not relevant to the question) should be penalized.

4) Of course, you should TELL all this to students so that they know what to expect.

Question 1) Using the language of "censoring" and/or "truncation" (left, right, and/or interval), explain why a prospective cohort study is often seen as higher quality than a retrospective cohort study.

The key to correctly answering this question is to recognize that prospective studies are less likely to suffer from left truncation. The reason is that in a prospective study, the start time is determined, and eligible individuals are identified as a result of this. On the other hand, with a retrospective study, the converse is usually the case: individuals are identified, and then the start time is determined. The problem with this is that individuals would have to make it at least to the point where they can be identified for inclusion in the retrospective study.

Question 2) Using Figure 1, draw the line diagram for for ID = 0 that would result if this individual was left truncated.

Draw a line that experiences an event before age 46, and thus does not get included in the study.

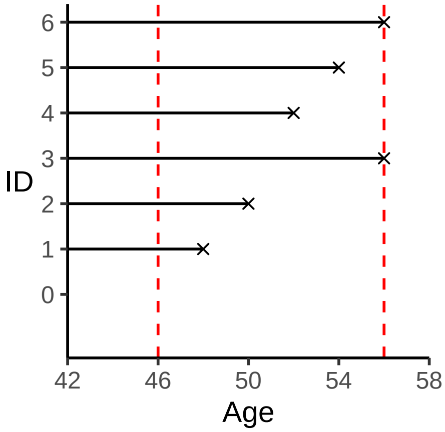

Figure 1: Line diagram with six hypothetical individuals who enter into a study at age 46 and who exit at age 56.

Question 3) For a randomly selected subset of 25 observations out of the N = 100 observations in the "2021_12_30-section1_cohort.csv" data, fit a line plot for the time-to-event outcome using ggplot. With this line plot, explore different ggplot themes. Pick your two favorite themes and compare them to the classic theme (i.e., `theme_classic()`). Title each plot with the name of the theme used. Plot each of the three themes in a grid with one row and three columns. Save the plot as a pdf or png file to directory. Include this plot in your homework output and provide an informative caption with the plot.

This code chunk imports the data, creates the plots, and saves the grid plot to a directory:

```r
library(tidyverse)
library(ggplot2)
library(gridExtra)

## import the dataset
cohort <- read_csv("../../data/2021_12_30-section1_cohort.csv")

## randomly select 25 observations (don't forget to set the seed)
set.seed(123)
sample_size <- 25
cohort <- cohort %>% sample_n(., sample_size)

## rank randomly selected IDs by stop time
cohort <- cohort %>% mutate(new_ID=rank(stop,ties="random"))

p1 <- ggplot(cohort) +
  geom_linerange(aes(y=new_ID,xmin=start,xmax=stop)) +
  geom_point(aes(y=new_ID,x=stop,shape=factor(outcome))) +
  scale_y_continuous(expand=c(0,0),
                     breaks=seq(1,sample_size,1)) +
  scale_x_continuous(expand=c(0,0.05)) +
  scale_shape_manual(name = "Outcome",values = c(1, 17, 4)) +
  ggtitle("Classic Theme") +
  theme_classic() +
  theme(axis.title.y = element_text(angle=0,vjust=.5))
```

```
p2 <- ggplot(cohort) +
  geom_linerange(aes(y=new_ID,xmin=start,xmax=stop)) +
  geom_point(aes(y=new_ID,x=stop,shape=factor(outcome))) +
  scale_y_continuous(expand=c(0,0),
                     breaks=seq(1,sample_size,1)) +
  scale_x_continuous(expand=c(0,0.05)) +
  scale_shape_manual(name = "Outcome",values = c(1, 17, 4)) +
  ggtitle("Minimal Theme") +
  theme_minimal() +
  theme(axis.title.y = element_text(angle=0,vjust=.5))

p3 <- ggplot(cohort) +
  geom_linerange(aes(y=new_ID,xmin=start,xmax=stop)) +
  geom_point(aes(y=new_ID,x=stop,shape=factor(outcome))) +
  scale_y_continuous(expand=c(0,0),
                     breaks=seq(1,sample_size,1)) +
  scale_x_continuous(expand=c(0,0.05)) +
  scale_shape_manual(name = "Outcome",values = c(1, 17, 4)) +
  ggtitle("Linedraw Theme") +
  theme_linedraw() +
  theme(axis.title.y = element_text(angle=0,vjust=.5))

ggsave("./question3_plot.pdf",grid.arrange(p1,p2,p3,nrow=1), width=30, height=10, units="cm")
```

This code chunk plots the figure in the homework pdf. The caption code is not shown here, but examples of how to create captions are provided in the notes (e.g., page 21):

```
knitr::include_graphics("./question3_plot.pdf")
```

The key to this question is to be able to save the figure, and present it using the "‘include_graphics‘" code chunk, which provides many options as to how the figure can be presented. Students should also use the ", ‘fig.show=’hide’‘" in the first code chunk to avoid presenting the figure in the first code chunk.

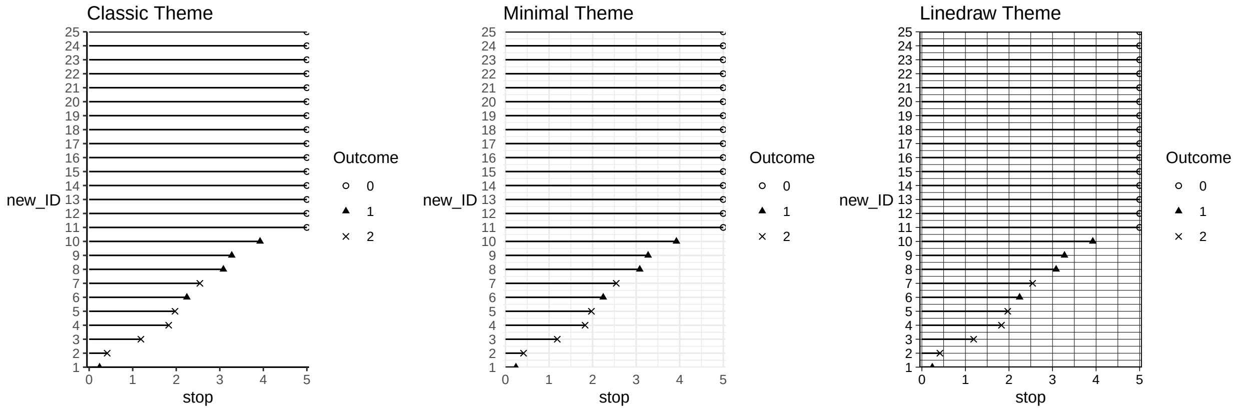

Figure 2: Line plot for 25 randomly selected observations in the Section 1 Cohort dataset, comparing three different themes from ggplot.

Question 4) Please do a basic exploratory analysis of the "2021_12_30-section1_cohort.csv" dataset. No more than 1/2 page of results. Provide results for the exposure, the confounder, and the outcome.

This should be an easy question to answer. I am looking for basic descriptives for the exposure, the confounder, and the outcome, in tabular and/or graphical form.

```
cohort <- read_csv("../../data/2021_12_30-section1_cohort.csv")


thm <- theme_classic() +
  theme(
    legend.position = "top",
    legend.background = element_rect(fill = "transparent", colour = NA),
    legend.key = element_rect(fill = "transparent", colour = NA)
  )
theme_set(thm)


p1 <- ggplot(cohort) +
  scale_y_continuous(expand=c(0,0)) +
  scale_x_continuous(expand=c(0,0)) +
  theme(text = element_text(size=20),
        axis.text.x = element_text(color="black"),
        axis.text.y = element_text(color="black")) +
  geom_bar(aes(exposure))
p2 <- ggplot(cohort) +
  scale_y_continuous(expand=c(0,0)) +
  scale_x_continuous(expand=c(0,0)) +
  theme(text = element_text(size=20),
        axis.text.x = element_text(color="black"),
        axis.text.y = element_text(color="black")) +
  geom_bar(aes(confounder))
p3 <- ggplot(cohort) +
  scale_y_continuous(expand=c(0,0)) +
  scale_x_continuous(expand=c(0,0)) +
  theme(text = element_text(size=20),
        axis.text.x = element_text(color="black"),
        axis.text.y = element_text(color="black")) +
  geom_bar(aes(outcome))

plot_eda <- grid.arrange(p1,p2,p3,nrow=1)
```

```
ggsave("./question4_plot.pdf",grid.arrange(p1,p2,p3,nrow=1), width=30, height=10, units="cm")
```

```
cohort %>%
  group_by(outcome) %>%
  summarize_at(vars(exposure,confounder,stop),mean)
```

```
## # A tibble: 3 x 4
##   outcome exposure confounder  stop
##     <dbl>    <dbl>      <dbl> <dbl>
## 1       0    0.239      0.435  5
## 2       1    0.464      0.429  2.06
## 3       2    0          0.462  2.02
```

This code chunk plots the figure in the homework pdf. Be sure to look for an informative caption describing the figure:

```r
knitr::include_graphics("./question4_plot.pdf")
```

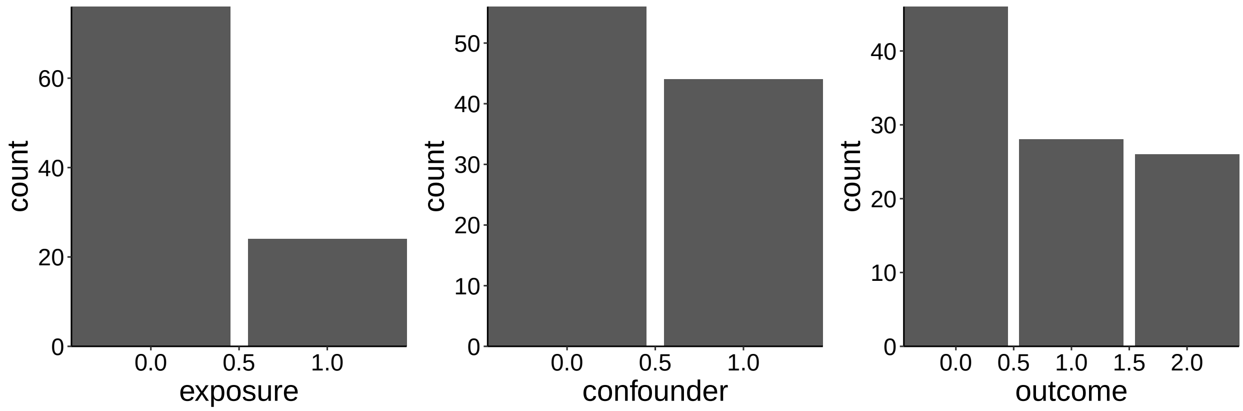

Figure 3: Bar plots showing the proportions of the exposure, confounder, and the outcome in 100 observations from the Section 1 Cohort data.

Question 5) Describe, in words, the interpretation of the CDF:

$$F(t) = P(T \leq t)$$

AND the survival function:

$$S(t) = P(T > t)$$

if T represents age at death from all causes, and t represents 64 years of age.

The interpretation for the CDF in the given context is the probability of the death from any cause occurring at time 64 years of age or before.

The interpretation of the survival function in the given context is the probability of death from any cause occurring after 64 years of age. (note: not at or after 64 years, but after 64 years).

To get full points for this question, the students must: 1) interpret F(t) and S(t) in the specified context (i.e., all cause mortality, 64 years of age); and 2) they must correctly (and precisely) interpret the equations.

Question 6) Using the first five observations from the synthetic data in Table 1 of the course notes, write out (but do not solve for) the terms for the Kaplan-Meier estimator $\hat{S}(t) = \prod_{k \in t_k \leq t}(1 - d_k/n_k)$. Assume that the total population at risk includes all 10 observations in Table 1.

$$\hat{S}(t) = \prod_{k \in t_k \leq t} (1 - d_k/n_k) = \underset{\text{for } t\text{=0.03}}{\left(1 - \frac{1}{10}\right)} \times \underset{\text{for } t\text{=0.49}}{\left(1 - \frac{1}{9}\right)} \times \underset{\text{for } t\text{=0.65}}{\left(1 - \frac{1}{8}\right)} \times \underset{\text{for } t\text{=2.61}}{\left(1 - \frac{1}{6}\right)}$$

The key here is to recognize that the risk-set for the last term excludes the censored observation occuring in ID = 4.

Question 7) Please explain the difference between the `Surv()` function, and the `survfit()` function in the survival package.

There are several ways to answer this question, but they should all lead to some version of the followign: the ‘Surv()‘ function converts a start time, a stop time, and an event indicator, to a time-to-event outcome that can be used in subsequent analyses. The ‘survfit()‘ function fits the Kaplan-Meier to time-to-event outcomes created by the ‘Surv()‘ function.

```
library(survival)


head(with(cohort,Surv(time=start, time2=stop, event=factor(outcome))))
```

```
## [1] (0,5.000000+]  (0,5.000000+]  (0,2.094849:1] (0,5.000000+]  (0,2.076077:1]
## [6] (0,4.162218:1]
```

Question 8) Refer to Figure 3 in the Section 1 course notes. Note that the dashed blue line in Figure 3 is from the Kaplan-Meier estimator, while the solid black line is from the simple calculation shown in the equations above the Figure (on page 10). Why don't these figures align exactly?

The reason these two curves don't align is because the blue curve is "less discrete" than the black curve. The time-scale used for the black curve is year. Consequently, all events that occur in the third year are classified as events at the beginning of the year (and thus the curve jumps to its end of year 2 risk at the beginning of the year). In contrast, the KM estimator treats each event separately, and thus each distinct jump occurs exactly when the event takes place.

Question 9) Fit the `survfit()` function to the “2021_12_30-section1_cohort.csv” data. Before you fit, be sure to re-code the outcome so that any non-zero event counts as an event (i.e., re-code outcome=2 to outcome=1). Examine the R object that you get from this fit. How many elements are in this object? What are the first six elements (describe them briefly, don't just provide their element names). Is there enough information in this object for you to determine the median survival time for the outcome? If so, what is the median survival time.

```
library(tidyverse)
library(survival)


cohort <- read_csv("../../data/2021_12_30-section1_cohort.csv") %>% mutate(outcome = as.numeric(outcome


surv_obj <- survfit(Surv(time=start,time2=stop,event=outcome)~1,data=cohort)
```

```r
names(surv_obj)
```

```
##  [1] "n"         "time"      "n.risk"    "n.event"   "n.censor"  "surv"
##  [7] "std.err"   "cumhaz"    "std.chaz"  "n.enter"   "type"      "logse"
## [13] "conf.int"  "conf.type" "lower"     "upper"     "call"
```

```r
str(surv_obj)
```

```
## List of 17
##  $ n        : int 100
##  $ time     : num [1:55] 0.0375 0.0526 0.1184 0.2061 0.2393 ...
##  $ n.risk   : num [1:55] 100 99 98 97 96 95 94 93 92 91 ...
##  $ n.event  : num [1:55] 1 1 1 1 1 1 1 1 1 1 ...
##  $ n.censor : num [1:55] 0 0 0 0 0 0 0 0 0 0 ...
##  $ surv     : num [1:55] 0.99 0.98 0.97 0.96 0.95 0.94 0.93 0.92 0.91 0.9 ...
##  $ std.err  : num [1:55] 0.0101 0.0143 0.0176 0.0204 0.0229 ...
##  $ cumhaz   : num [1:55] 0.01 0.0201 0.0303 0.0406 0.051 ...
##  $ std.chaz : num [1:55] 0.01 0.0142 0.0175 0.0203 0.0228 ...
##  $ n.enter  : num [1:55] 1.00e+02 1.91e-313 1.49e+195 3.17e-120 8.75e-154 ...
##  $ type     : chr "counting"
##  $ logse    : logi TRUE
##  $ conf.int : num 0.95
##  $ conf.type: chr "log"
##  $ lower    : num [1:55] 0.971 0.953 0.937 0.922 0.908 ...
##  $ upper    : num [1:55] 1 1 1 0.999 0.994 ...
##  $ call     : language survfit(formula = Surv(time = start, time2 = stop, event = outcome) ~ 1,
##  - attr(*, "class")= chr "survfit"
```

```r
med_time <- tibble(time=surv_obj$time,risk=surv_obj$surv) %>% filter(round(risk,2)==0.50)


med_time
```

```
## # A tibble: 1 x 2
##    time  risk
##   <dbl> <dbl>
## 1  4.16 0.500
```

There are 17 elements in the object created by ‘survfit‘. The first six elements are: n, time, n.risk, n.event, n.censor, surv. These are the original sample size, the unique event times, the number at risk at each event time, the number of events at each event time, the number of censored observations at each event time, and the estimated survival probability at each event time. There is, in fact, enough information to evaluate median survival time in this cohort, which is the time at which 50 percent of the sample survives. The median survival time is 4.16.

Some important elements to answer this question: the median survival time should be reported to 1 or 2 decimal places. Any more is unwarranted, and should probably result in loss of points.

Question 10) Using the fit from Question 6, plot the cumulative distribution function (not the survival function) using the KM estimator. Interpret the curve assuming that the outcome is death from any cause and the time-scale is year on study.

```
plot_dat <- tibble(time = surv_obj$time,
                   risk = 1 - surv_obj$surv)


p1 <- ggplot(plot_dat) +
  scale_y_continuous(expand=c(0,0), limits=c(0,1)) +
  scale_x_continuous(expand=c(0,0)) +
  ylab("Risk of Death") +
  xlab("Year on Study") +
  theme_classic() +
  theme(axis.text.x = element_text(size=12,color="black"),
        axis.text.y = element_text(size=12,color="black")) +
  geom_step(aes(x=time,y=risk))


ggsave("./question9_plot.pdf",p1, height=10, width=10, units="cm")
```

This figure shows a steady increase in the risk of death from the start to the end of study. By year 5 on study, just over 50% of the sample died, and by the 5th year on study, 54% of the sample died.

```
knitr::include_graphics("./question9_plot.pdf")
```

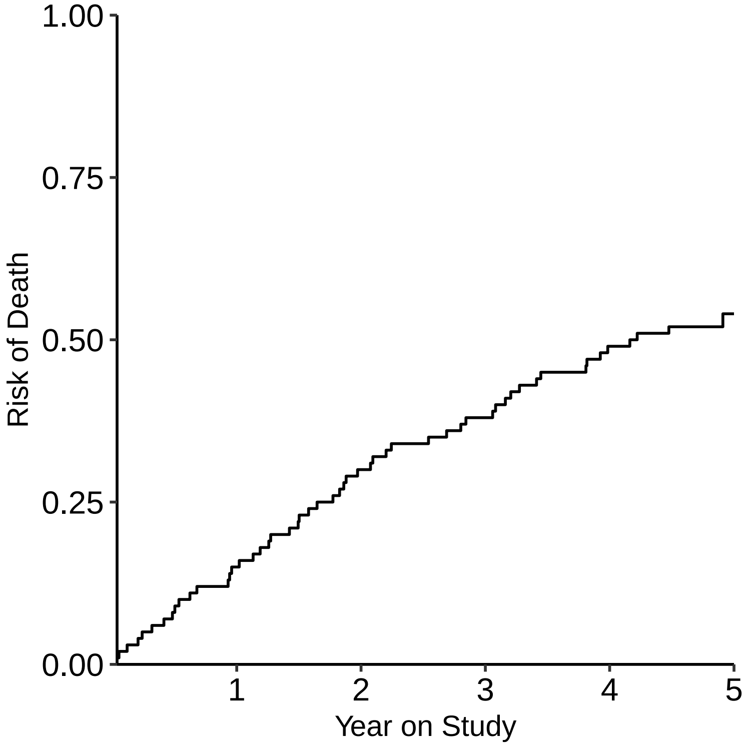

Figure 4: Five year cumulative risk of death among 100 observations in the Section 1 Cohort dataset.

Question 11) Referring to Figure 6 of the section 1 course notes, why is the cumulative risk represented by the dashed line higher than the cumulative risks represented by the solid black line, even though they are the same events?

The dashed line in Figure 6 represents the cumulative risk from the KM estimator, which treats the competing event (outcome = 2) as a censored observation. Practically, this means that the risks for individuals who experience outcome = 2 is re-distributed as risk for outcome = 1. This works in theory when individuals experiencing outcome = 2 are still at risk for experiencing outcome = 1 (i.e., when outcome = 1 is a censoring event), but not when outcome = 2 is a competing event.

Question 12) What is the main problem with using the cause-specific risk to understand the causal effects of exposures on outcomes of interest?

Cause-specific risks for an event of interest (e.g., outcome = 1) capture the risk that we'd observe if we were able to prevent a competing risk (e.g., outcome = 2) from occurring. More often than not, there is no way to prevent competing risks from occurring (e.g., when the competing risk is some form of mortality, either all-cause or cause-specific). As such, in the presence of competing risks, one can never actually observe cause specific risks because they require interventions on competing risks that are not usually possible.

Question 13) Provide a single plot of the cause-specific and sub-distribution risk for "outcome = 1" in the "2021_12_30-section1_cohort.csv" using the Kaplan-Meier, Aalen-Johansen, and Gray's CIF estimators.

There are many ways to do this. In my opinion, the best way is to fit the KM, AJ, and CIF estimators, obtain the risks and times for each, concatenate these into a single dataset with a variable that represents the estimator, and plot it in a single 'geom_step', like this

```
library(survival)
library(cmprsk)

## KM

cohort <- read_csv("../../data/2021_12_30-section1_cohort.csv")

cohort <- cohort %>% mutate(cs_outcome = as.numeric(outcome==1)) # note this is a numeric variable!

surv_model <- survfit(Surv(time=stop, event=cs_outcome) ~ 1, data=cohort)

plot_dat_km <- tibble(time=surv_model$time,
                      risk=1 - surv_model$surv,
                      Estimator="Kaplan-Meier")

## AJ

cohort <- read_csv("../../data/2021_12_30-section1_cohort.csv")

cohort <- cohort %>% mutate(outcome = factor(outcome, 0:2, labels=c("censor", "event", "competing risk"
aj_fit <- survfit(Surv(time=stop,event=outcome) ~ 1, data=cohort)

plot_dat_aj <- tibble(time=aj_fit$time,risk=aj_fit$pstate[,2], Estimator="Aalen-Johansen")

# Gray's CIF

cohort <- read_csv("../../data/2021_12_30-section1_cohort.csv")

gray_cif <- cuminc(cohort$stop, cohort$outcome, cencode=0) # note outcome is back to numeric!

plot_dat_cif <- tibble(time=gray_cif$`1 1`$time,
                       risk=gray_cif$`1 1`$est,
                       Estimator="Gray's CIF")

plot_dat <- rbind(plot_dat_km,
                  plot_dat_aj,
                  plot_dat_cif)

plot_obj <- ggplot(plot_dat) +
  scale_y_continuous(expand=c(0,0), limits=c(0,1)) +
  scale_x_continuous(expand=c(0,0)) +
  ylab("Cumulative Risk") +
  xlab("Time on Study") +
  scale_color_manual(values=c("#000000","#D55E00","#0072B2")) +
  geom_step(aes(x=time,y=risk,color=Estimator))

ggsave("question12_plot.pdf",plot=plot_obj)
```

```
knitr::include_graphics("./question12_plot.pdf")
```

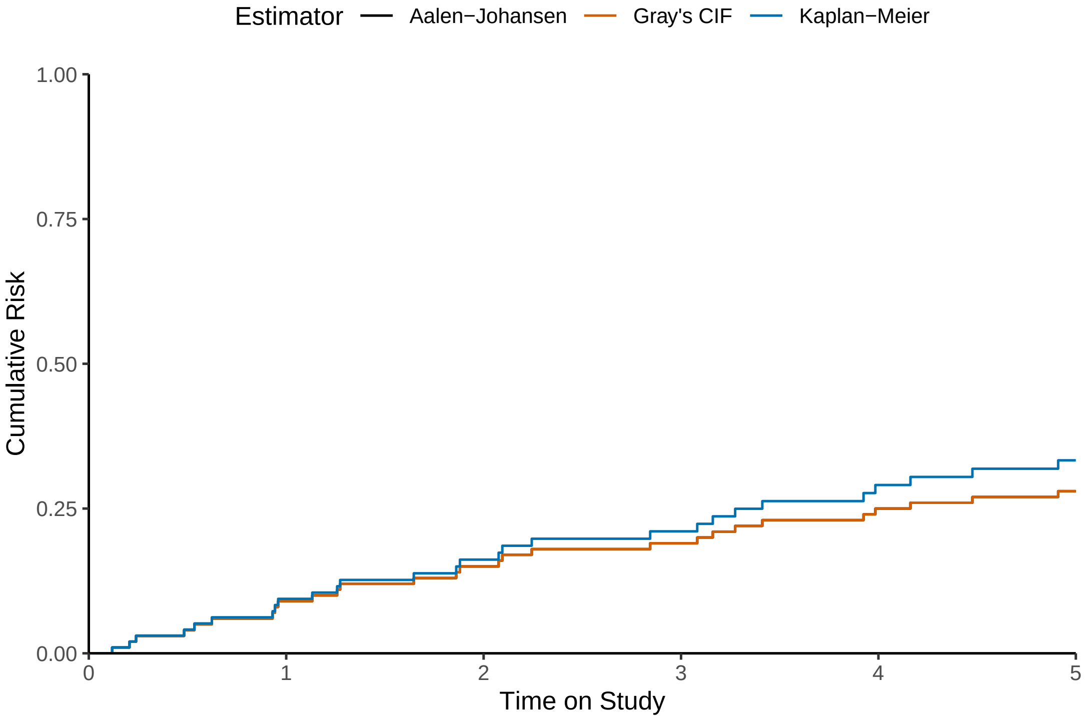

Figure 5: Comparison of the cumulative risk functions from the Kaplan-Meier (cause-specific), Aalen-Johansen (sub-distribution) and Gray's CIF (sub-distribution) estimators in 100 Observations from the Section 1 Data.