Insights into Sepsis Survival

HealthData Explorers: Alberto Barbesin, Nicolae Bologa, Eleonora Marcassa

Contents

The Dataset

Sepsis is a life-threatening condition caused by an exaggerated reaction of the body to an infection, that leads to organ failure or even death. Since *sepsis* can kill a patient even in just one hour, survival prediction is an urgent priority among the medical community: even if laboratory tests and hospital analyses can provide insightful information about the patient, in fact, they might not come in time to allow medical doctors to recognize an immediate death risk and treat it properly. The features of patients that has been recorded at the hospital admission are:

- `sex`,

- age,
- septic episode number.

We considered a cohort of 110,204 patient admissions.

For the **primary cohort**, the records represent patients with potential sepsis preconditions (according to the pre-Sepsis-3 definition); for the **study cohort**, they represent only patient admissions defined by the new Sepsis-3 definition.

```
# survival_primary_cohort
(survival_primary_cohort <- read_csv("primary_cohort.csv") %>%
  rename(age = age_years,
         sex = sex_0male_1female,
         count_episode = episode_number,
         hospital_outcome = hospital_outcome_1alive_0dead) %>%
  mutate(sex_cat = ifelse(sex == 0, "male", "female"), hospital_outcome_cat = ifelse(hospital_outcome ==
```

```
## Rows: 110204 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): age_years, sex_0male_1female, episode_number, hospital_outcome_1ali...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 110,204 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1    21     1             1                1 female  alive
##  2    20     1             1                1 female  alive
##  3    21     1             1                1 female  alive
##  4    77     0             1                1 male    alive
##  5    72     0             1                1 male    alive
##  6    83     0             1                1 male    alive
##  7    74     0             1                1 male    alive
##  8    74     1             1                1 female  alive
##  9    69     0             1                1 male    alive
## 10    53     1             1                1 female  alive
## # i 110,194 more rows
```

```
# survival_study_cohort
(survival_study_cohort <- read_csv("study_cohort.csv") %>%
  rename(age = age_years,
         sex = sex_0male_1female,
         count_episode = episode_number,
         hospital_outcome = hospital_outcome_1alive_0dead) %>%
    mutate(sex_cat = ifelse(sex == 0, "male", "female"), hospital_outcome_cat = ifelse(hospital_outcome
```

```
## Rows: 19051 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): age_years, sex_0male_1female, episode_number, hospital_outcome_1ali...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 19,051 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1     7     1             1                1 female  alive
##  2    17     0             2                1 male    alive
##  3    70     0             1                1 male    alive
##  4    76     0             1                1 male    alive
##  5     8     0             1                1 male    alive
##  6    41     0             2                1 male    alive
##  7    60     0             1                0 male    dead
##  8    89     1             1                0 female  dead
##  9    76     0             3                0 male    dead
## 10    81     1             1                1 female  alive
## # i 19,041 more rows
```

```
# survival_validation_cohort
(survival_validation_cohort <- read_csv("validation_cohort.csv") %>%
  rename(age = age_years,
         sex = sex_0male_1female,
         count_episode = episode_number,
         hospital_outcome = hospital_outcome_1alive_0dead) %>%
    mutate(sex_cat = ifelse(sex == 0, "male", "female"), hospital_outcome_cat = ifelse(hospital_outcome
```

```
## Rows: 137 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): age_years, sex_0male_1female, episode_number, hospital_outcome_1ali...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 137 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1    20     0             1                1 male    alive
##  2    22     0             1                1 male    alive
##  3    26     1             2                0 female  dead
##  4    33     1             1                1 female  alive
##  5    33     0             1                1 male    alive
##  6    33     0             2                0 male    dead
##  7    35     0             1                1 male    alive
##  8    35     1             1                1 female  alive
##  9    36     0             1                1 male    alive
## 10    36     1             1                1 female  alive
## # i 127 more rows
```

A first glance to all the three datasets reveals that the variables `sex` and `hospital_outcome` are binary variables. In order to carry out the analysis, it can be helpful converting them in categorical variables.

Exploratory Analysis

Survival primary cohort

Structure and summary of data and some exploratory plots

```
str(survival_primary_cohort)
```

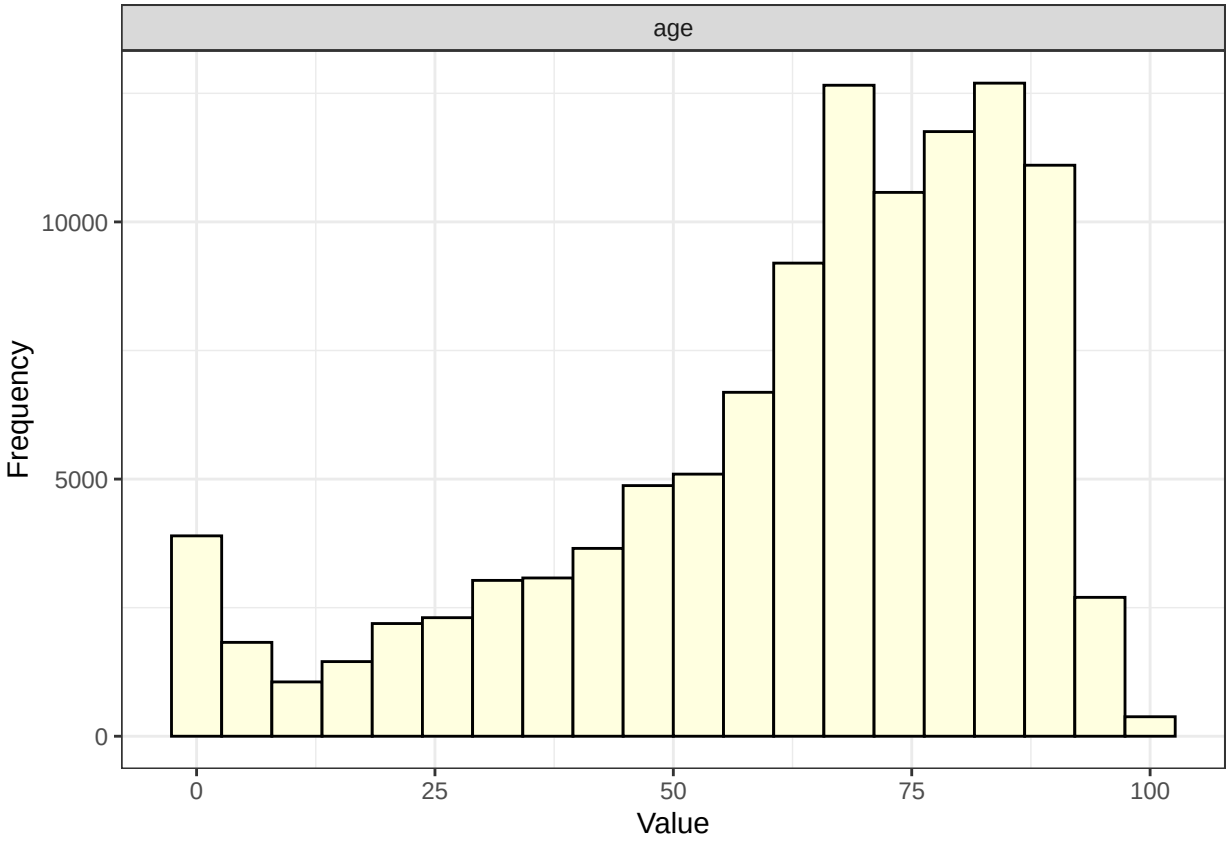
```
## tibble [110,204 x 6] (S3: tbl_df/tbl/data.frame)
##  $ age                 : num [1:110204] 21 20 21 77 72 83 74 74 69 53 ...
##  $ sex                 : num [1:110204] 1 1 1 0 0 0 0 1 0 1 ...
##  $ count_episode       : num [1:110204] 1 1 1 1 1 1 1 1 1 1 ...
##  $ hospital_outcome    : num [1:110204] 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex_cat             : chr [1:110204] "female" "female" "female" "male" ...
##  $ hospital_outcome_cat: chr [1:110204] "alive" "alive" "alive" "alive" ...
```

```
summary(survival_primary_cohort[!(colnames(survival_primary_cohort) %in% c("sex_cat", "hospital_outcome_
```

```
##       age              sex         count_episode   hospital_outcome
##  Min.   :  0.00   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.: 51.00   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.0000
##  Median : 68.00   Median :0.0000   Median :1.000   Median :1.0000
##  Mean   : 62.74   Mean   :0.4739   Mean   :1.349   Mean   :0.9265
##  3rd Qu.: 81.00   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :100.00   Max.   :1.0000   Max.   :5.000   Max.   :1.0000
```

Variables `age`,`count hospital` and `hospital outcome` are *quite symmetrical*. We can't say the same for the other ones.

```
survival_primary_cohort %>%
  select(age) %>%
  gather(key = "cols", value = "value") %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 20, fill = "lightyellow", col = "black") +
  facet_wrap(. ~ cols, ncol = 1) +
  labs(x = "Value", y = "Frequency") +
  theme_bw()
```

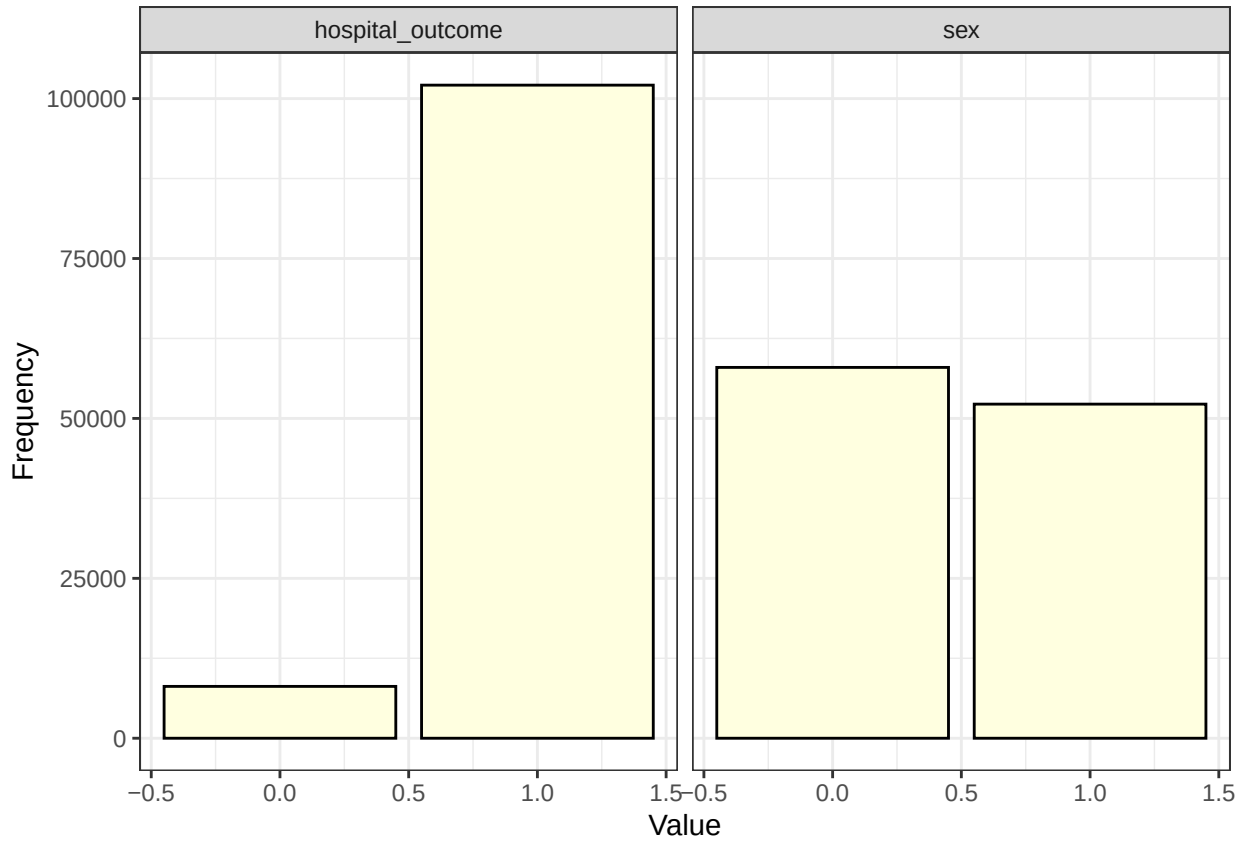


This very first graph highlights the fact the data collected belong mostly to older patients. As a matter of fact, the distribution is *left skewed*.

```
survival_primary_cohort %>%
  group_by(sex_cat) %>%
  summarise(n = n(), mean= n/length(survival_primary_cohort$sex_cat))
```

```
## # A tibble: 2 x 3
##   sex_cat     n  mean
##   <chr>   <int> <dbl>
## 1 female  52231 0.474
## 2 male    57973 0.526
```

```
survival_primary_cohort %>%
  select(sex, hospital_outcome) %>%
  gather(cols, value) %>%
  ggplot(aes(x = value)) +
  geom_bar(fill = "lightyellow", col = "black") +
  facet_wrap(.~ cols, ncol = 2) +
  labs(x = "Value", y = "Frequency") +
  theme_bw()
```



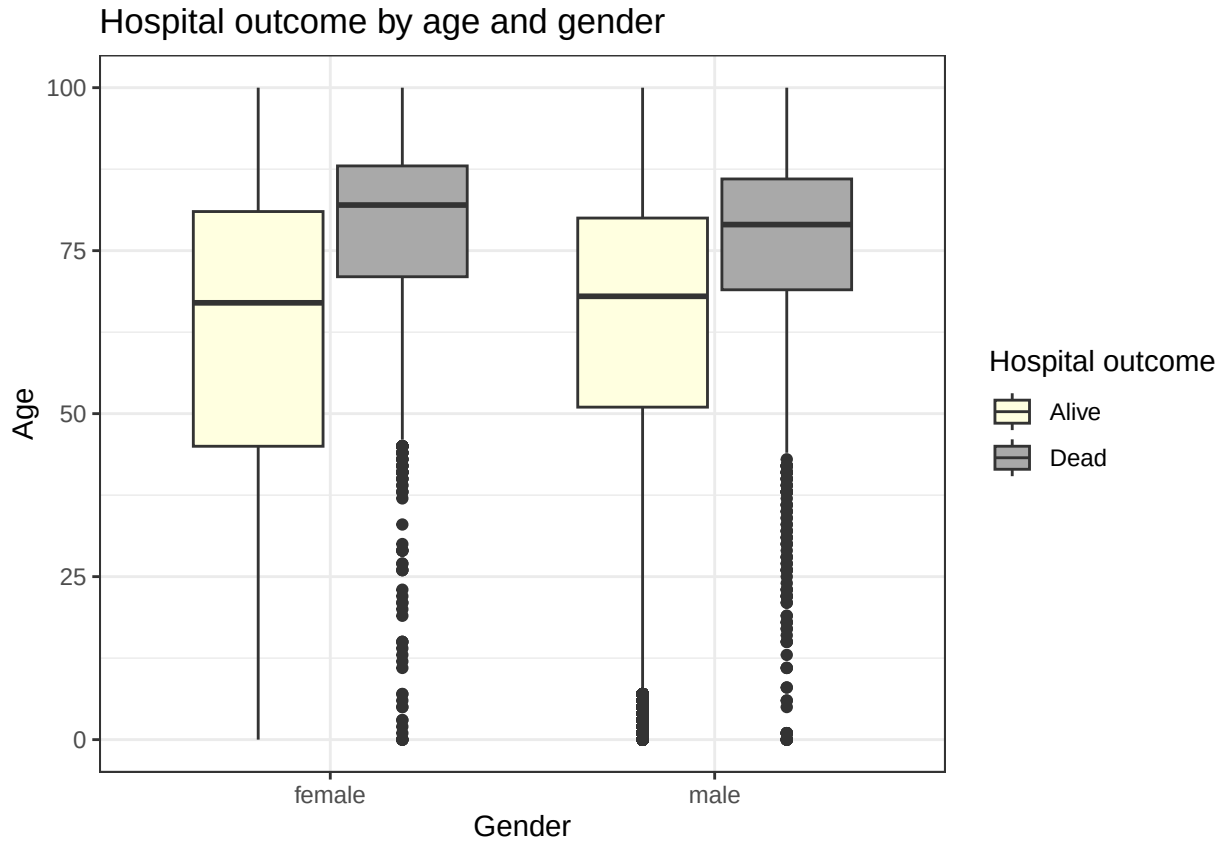
The graph representing `hospital_outcome` reveals a low mortality rate among patients affected by sepsis, with the majority surviving.
Furthermore, analysis of the `sex` variable indicates a higher proportion of male patients compared to females, as showed in the accompanying table.

```
survival_primary_cohort %>%
  group_by(sex) %>%
  summarise(m_age = mean(age),
            sd_age = sd(age),
            freq = n()) %>%
  arrange(desc(sex))
```

```
## # A tibble: 2 x 4
##     sex m_age sd_age  freq
##   <dbl> <dbl>  <dbl> <int>
## 1     1  62.2   25.2 52231
## 2     0  63.3   23.1 57973
```

```
spc_by_age <- survival_primary_cohort %>%
  group_by(age) %>%
  summarise(m_sex = mean(sex),
            sd_sex = sd(sex),
            freq = n()) %>%
  arrange(desc(freq))
```

```
survival_primary_cohort %>%
  ggplot(aes(x = sex_cat, y = age, fill = hospital_outcome_cat)) +
  geom_boxplot(show.legend = TRUE) +
  labs(title = "Hospital outcome by age and gender",
       x = "Gender",
       y = "Age",
       fill = "Hospital outcome")+
  scale_fill_manual(values = c("alive" = "lightyellow", "dead" = "darkgrey"),
                    labels = c("Alive", "Dead")) +
  theme_bw()
```



From a very first glance at the graph, an asymmetry in the data can be noticed by looking at the location of the median. Besides the fact that it seems to be a quite strong relationship between `age` and `death`. In addition, *females* are likely to live two years longer than *males*.

The dataset is splitted into two parts, in order to better distinguish *survived* and *dead* people.

```
(dead_survival_primary_cohort <- survival_primary_cohort%>%
  filter(hospital_outcome_cat=="dead" & hospital_outcome==0))
```

```
## # A tibble: 8,105 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1    72     0             1                0 male    dead
##  2    63     0             1                0 male    dead
```

```
##  3    89     1             2                0 female  dead
##  4    80     0             3                0 male    dead
##  5    62     1             3                0 female  dead
##  6    56     1             3                0 female  dead
##  7    63     0             1                0 male    dead
##  8    60     0             1                0 male    dead
##  9    89     1             1                0 female  dead
## 10    61     0             1                0 male    dead
## # i 8,095 more rows
```

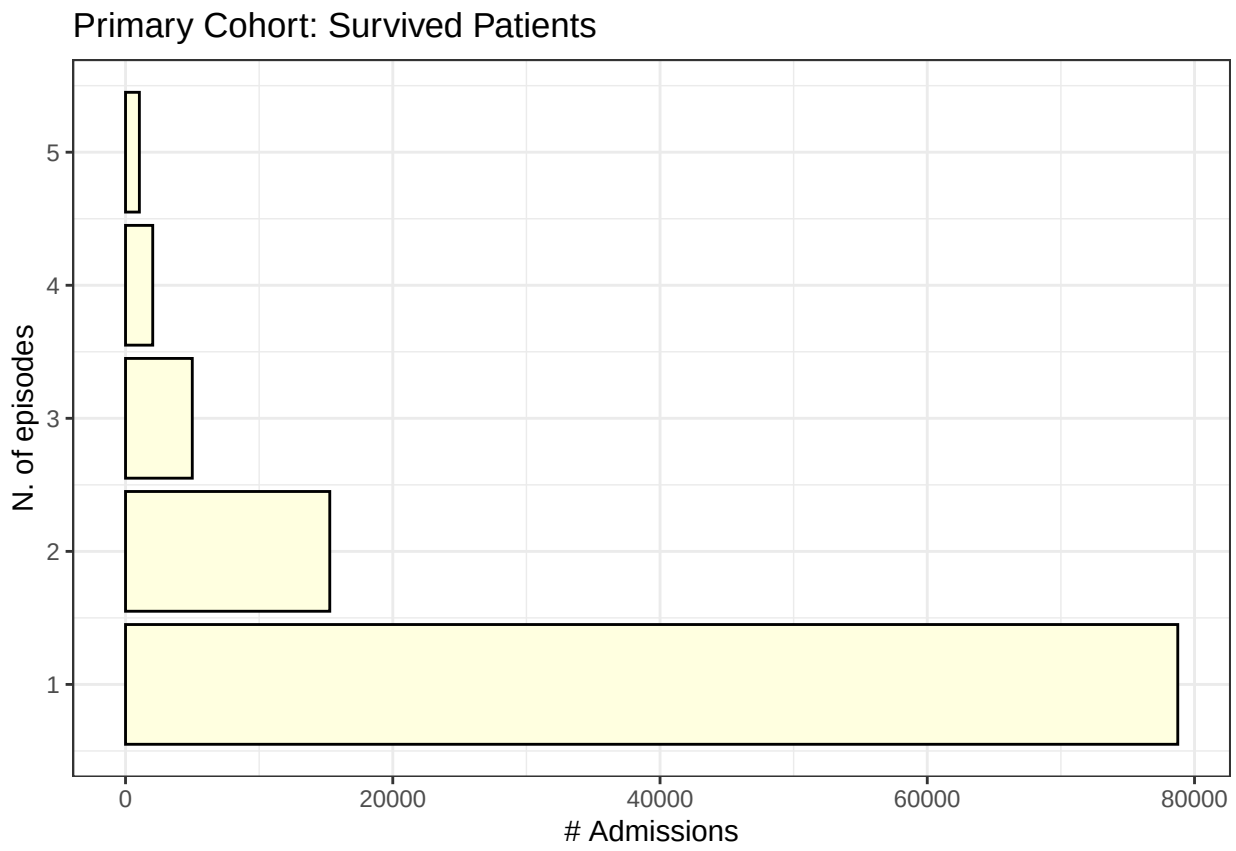
```
(survived_survival_primary_cohort <- survival_primary_cohort%>%
   filter(hospital_outcome_cat=="alive" & hospital_outcome==1))
```

```
## # A tibble: 102,099 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1    21     1             1                1 female  alive
##  2    20     1             1                1 female  alive
##  3    21     1             1                1 female  alive
##  4    77     0             1                1 male    alive
##  5    72     0             1                1 male    alive
##  6    83     0             1                1 male    alive
##  7    74     0             1                1 male    alive
##  8    74     1             1                1 female  alive
##  9    69     0             1                1 male    alive
## 10    53     1             1                1 female  alive
## # i 102,089 more rows
```

```
(n_episodes <- survived_survival_primary_cohort%>%
group_by(count_episode)%>%
summarise(n = n()))
```

```
## # A tibble: 5 x 2
##   count_episode     n
##           <dbl> <int>
## 1             1 78747
## 2             2 15285
## 3             3  4996
## 4             4  2036
## 5             5  1035
```

```
ggplot(n_episodes, aes(x = count_episode, y = n)) +
  geom_bar(stat = "identity", fill = "lightyellow", color = "black") +
  labs(y = "# Admissions", x = "N. of episodes", title = "Primary Cohort: Survived Patients") +
  theme_bw() +
  coord_flip()
```



Thanks to a combined reading of the previously presented tables and this last graph, it can be understood that the majority of the people had just one episode of *sepsis*. As a matter of fact, this one and only episode involve about 80,000 people. This number fastly decrease: "just" 15,000 people showed two episodes of *sepsis*. The greater the number of episodes the lower is the number of people who presented the infection: it gets lower than 5,000.

Keep in mind that these number represent all the people who got affected by the disease but survived.

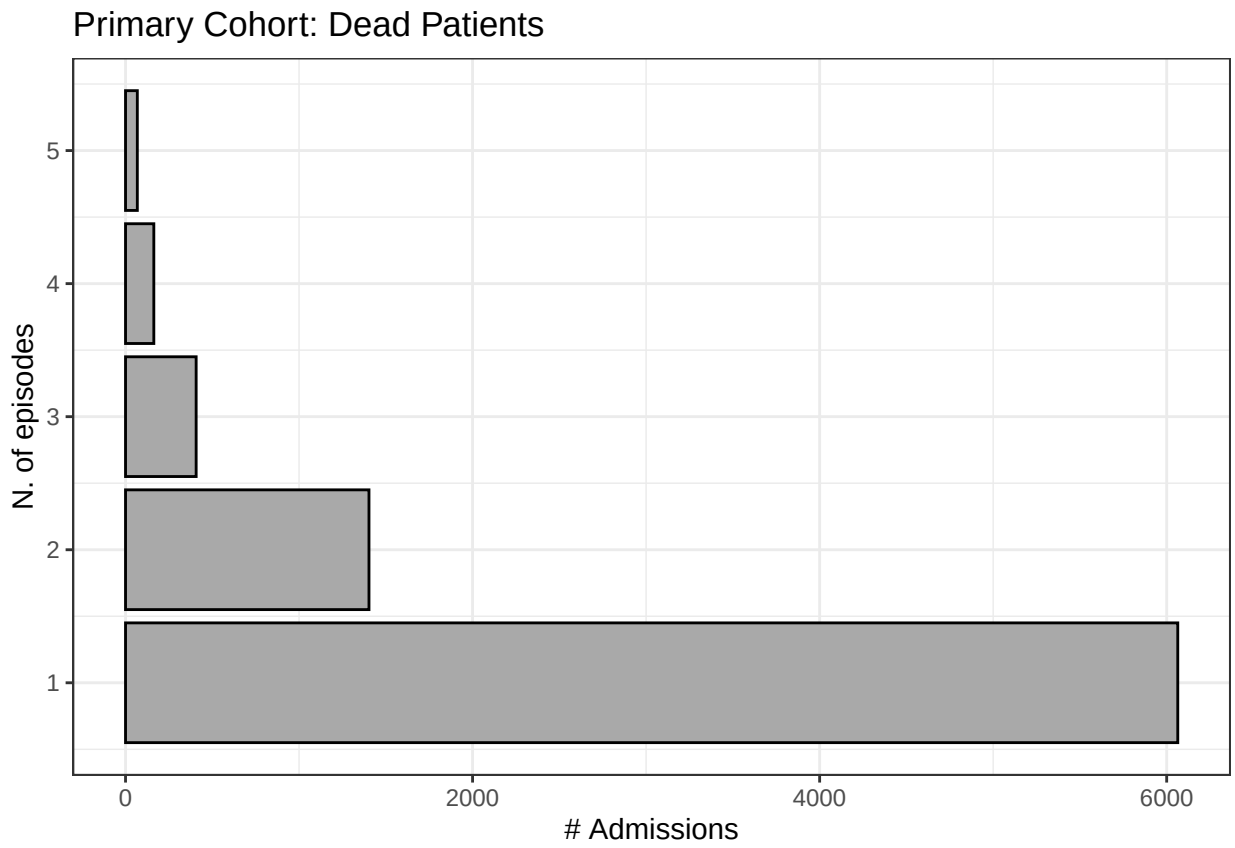
While the following results are about people who got the infection but died.

```
(n_episodes <- dead_survival_primary_cohort%>%
group_by(count_episode)%>%
summarise(n = n()))
```

```
## # A tibble: 5 x 2
##   count_episode     n
##           <dbl> <int>
## 1             1  6064
## 2             2  1403
## 3             3   407
## 4             4   163
## 5             5    68
```

```
ggplot(data = n_episodes, aes(x = count_episode, y = n)) +
  geom_bar(stat = "identity", fill = "darkgrey", color = "black") +
  labs(y = "# Admissions", x = "N. of episodes", title = "Primary Cohort: Dead Patients") +
```

```
  theme_bw() +
  coord_flip()
```



In this case, the number of dead people who showed just one case of *sepsis* is around 6,000. Then the number decreases to c.a. 1,500 when the episodes increase at 2. As before, the greater the number of episodes the lower the number of infected patients who died.

Probability of getting *sepsis*

We are interested in determining whether men with *sepsis* tend to be older than women, and vice versa. Additionally, we are exploring whether there is a gender and age-related difference in survival likelihood.

```
male = survival_primary_cohort %>% filter(sex == 0)
female = survival_primary_cohort %>% filter(sex == 1)

m_mean_age = mean(male$age)
m_sd_age = sd(male$age)
f_mean_age = mean(female$age)
f_sd_age = sd(female$age)

# males' age average
m_mean_age
```

```
## [1] 63.25243
```

```r
# males' age standard deviation
m_sd_age
```

```
## [1] 23.14269
```

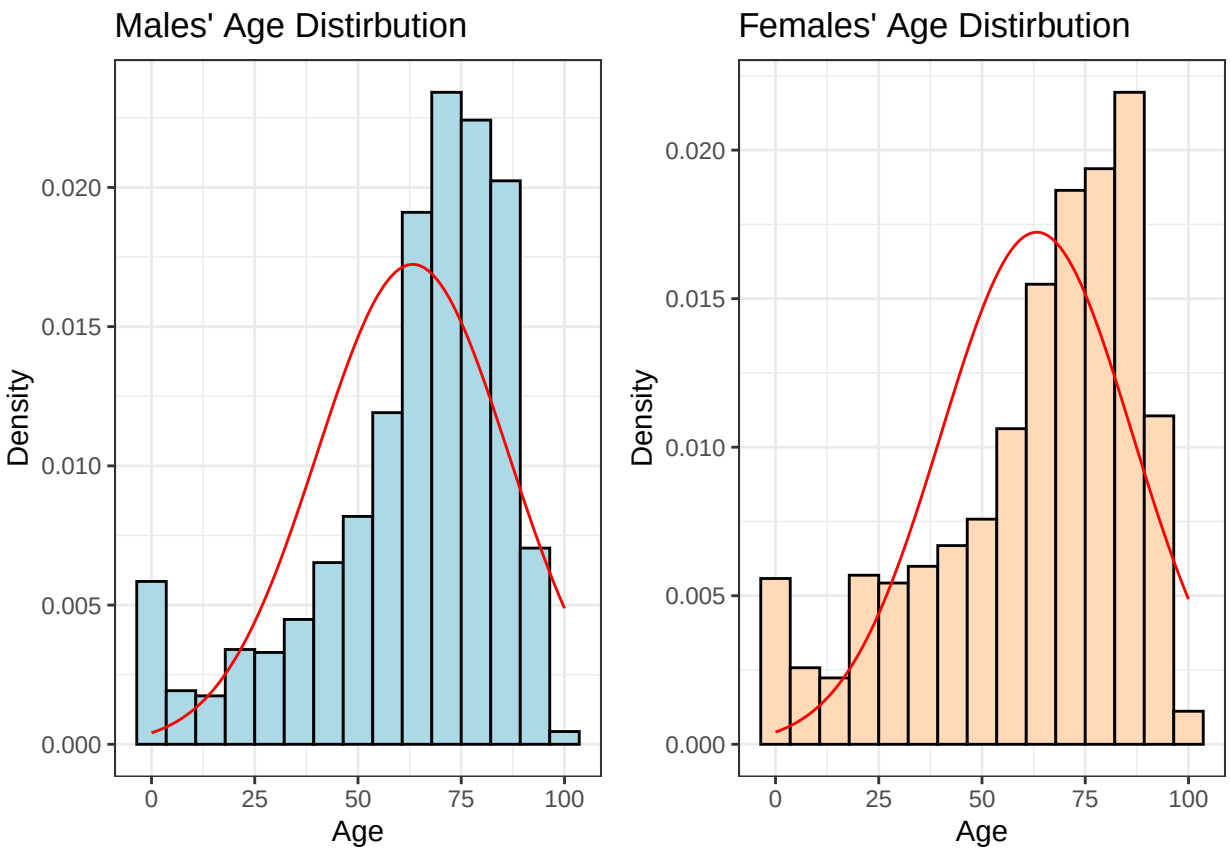
```r
# females' age average
f_mean_age
```

```
## [1] 62.16123
```

```r
# females' age standard deviation
f_sd_age
```

```
## [1] 25.16188
```

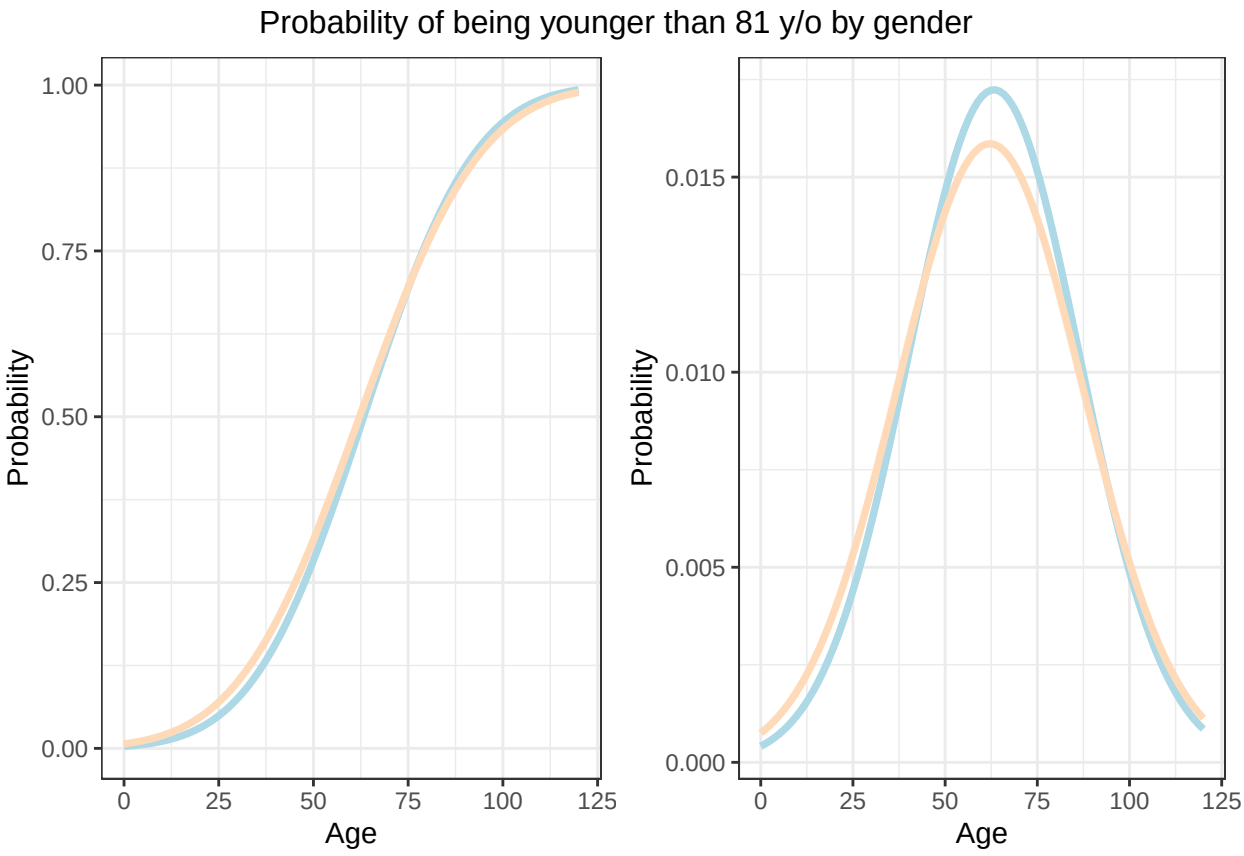
```r
plot_male <- ggplot(male) +
  geom_histogram(aes(x = age,
                     y = after_stat(density)),
                 bins = 15,
                 fill = "lightblue",
                 color = "black") +
  stat_function(fun = dnorm,
                args = list (mean = m_mean_age, sd = m_sd_age),
                color = "red") +
  labs(title = "Males' Age Distirbution", x = "Age", y = "Density") +
  theme_bw()

plot_female <- ggplot(female) +
  geom_histogram(aes(x = age,
                     y = after_stat(density)),
                 bins = 15,
                 fill = "#FFDAB9",
                 color = "black") +
  stat_function(fun = dnorm,
                args = list (mean = m_mean_age, sd = m_sd_age),
                color = "red") +
  labs(title = "Females' Age Distirbution", x = "Age", y = "Density") +
  theme_bw()

gridExtra::grid.arrange(grobs = list(plot_male, plot_female), nrow = 1, ncol = 2)
```



Building upon our initial observations, individuals affected by *sepsis* tend to be older, as evidenced by the continued presence of a *left-skewed* distribution. Now, let's look at whether men and women under 81, which is younger than the third quantile, are more likely to get *sepsis* according.

```
# probability males are younger than 81
prob_males <- pnorm(quantile(survival_primary_cohort$age, probs = 0.75),
                    m_mean_age,
                    m_sd_age,
                    lower.tail = T)

# probability females are younger than 81
prob_females <- pnorm(quantile(survival_primary_cohort$age, probs = 0.75),
                      f_mean_age,
                      f_sd_age,
                      lower.tail = T)

probs_df <- data.frame(age = seq(0,120, by = 1),
                       prob_males = pnorm(seq(0,120,by=1),
                                          m_mean_age,
                                          m_sd_age,
                                          lower.tail = T),
                       prob_females = pnorm(seq(0,120,by=1),
                                            f_mean_age,
                                            f_sd_age,
                                            lower.tail = T))
```

```r
pnorm_plot <- ggplot(probs_df, aes(x = age)) +
  stat_function(fun = pnorm,
                args = list(mean = m_mean_age, sd = m_sd_age),
                color = "lightblue",
                linewidth = 1.3) +
  stat_function(fun = pnorm,
                args = list(mean = f_mean_age, sd = f_sd_age),
                color = "#FFDAB9",
                linewidth = 1.3) +
  labs(x = "Age", y="Probability") +
  theme_bw()

dnorm_plot <- ggplot(probs_df, aes(x = age)) +
  stat_function(fun = dnorm,
                args = list(mean = m_mean_age, sd = m_sd_age),
                color = "lightblue",
                linewidth = 1.3) +
  stat_function(fun = dnorm,
                args = list(mean = f_mean_age, sd = f_sd_age),
                color = "#FFDAB9",
                linewidth = 1.3) +
  labs(x = "Age", y="Probability") +
  theme_bw()

combined_plots <- gridExtra::grid.arrange(grobs = list(pnorm_plot, dnorm_plot), ncol = 2, top = "Probabi
```

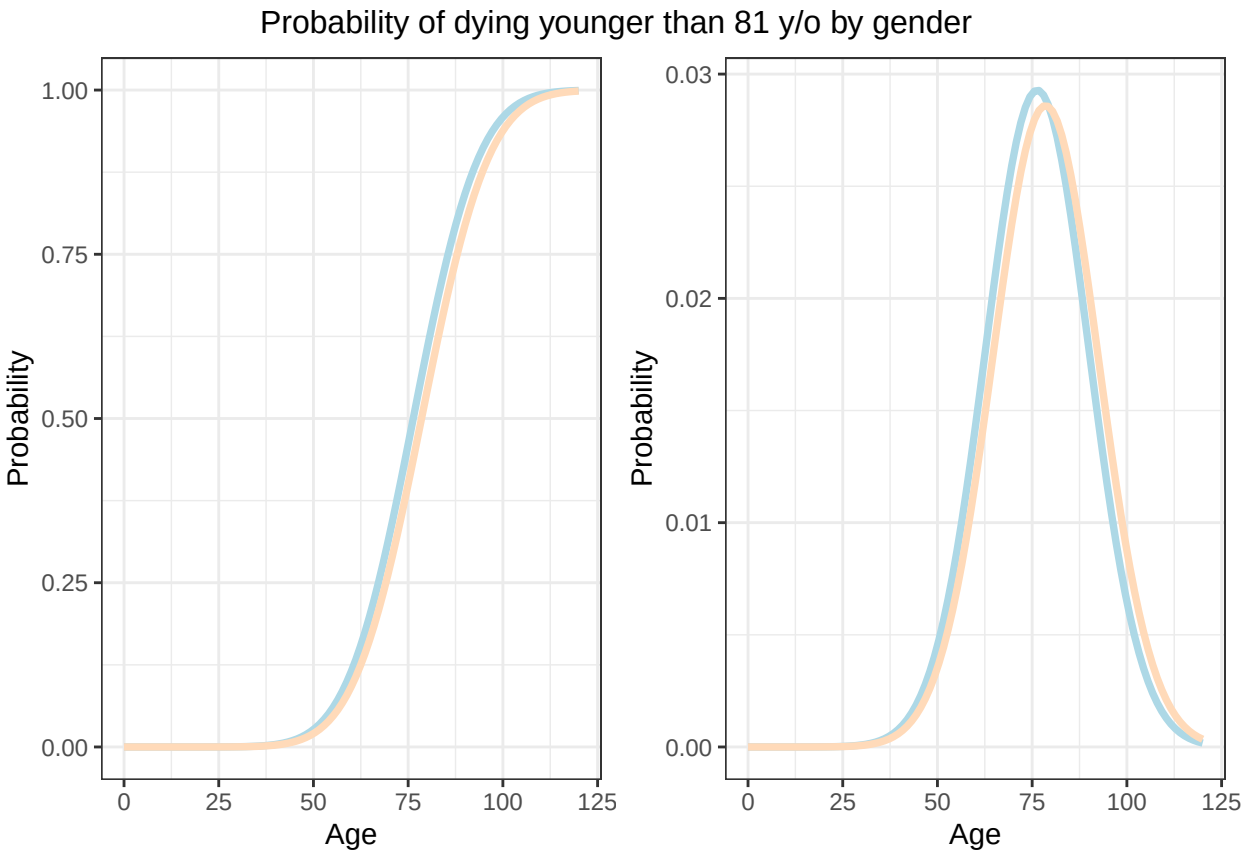


Once computed the probability of getting the *sepsis* infection, it is possible to affirm that the 77,30% of women and the 77,84% of men younger than the age third quantile get the infection at least one time. So quite similar probabilities for both women and men.

Do they have the same probability of death if they are younger than the third quantile?

```
male_dead = survival_primary_cohort %>%
  filter(sex == 0 & hospital_outcome == 0)

female_dead = survival_primary_cohort %>%
  filter(sex == 1 & hospital_outcome == 0)

md_mean_age = mean(male_dead$age)
md_sd_age = sd(male_dead$age)
fd_mean_age = mean(female_dead$age)
fd_sd_age = sd(female_dead$age)

# probability males who died of sepsis are younger than 81
prob_males <- pnorm(quantile(survival_primary_cohort$age, probs = 0.75),
                    md_mean_age,
                    md_sd_age,
                    lower.tail = T)

# probability females who died of sepsis are younger than 81
prob_females <- pnorm(quantile(survival_primary_cohort$age, probs = 0.75),
                      fd_mean_age,
                      fd_sd_age,
                      lower.tail = T)

probs_df <- data.frame(age = seq(0, 120, by = 1),
                       prob_males = pnorm(seq(0, 120, by = 1),
                                          md_mean_age,
                                          md_sd_age,
                                          lower.tail = TRUE),
                       prob_females = pnorm(seq(0, 120, by = 1),
                                            fd_mean_age,
                                            fd_sd_age,
                                            lower.tail = TRUE))

pnorm_plot <- ggplot(data = probs_df, aes(x = age)) +
  stat_function(fun = pnorm,
                args = list(mean = md_mean_age, sd = md_sd_age),
                color = "lightblue",
                linewidth = 1.3) +
  stat_function(fun = pnorm,
                args = list(mean = fd_mean_age, sd = fd_sd_age),
                color = "#FFDAB9",
                linewidth= 1.3) +
  labs(x = "Age", y = "Probability") +
  theme_bw()

dnorm_plot <- ggplot(data = probs_df, aes(x = age)) +
  stat_function(fun = dnorm,
                args = list(mean = md_mean_age,
```

```
                            sd = md_sd_age),
                color = "lightblue",
                linewidth = 1.3) +
  stat_function(fun = dnorm,
                args = list(mean = fd_mean_age, sd = fd_sd_age),
                color = "#FFDAB9",
                linewidth = 1.3) +
  labs(x = "Age", y = "Probability") +
  theme_bw()

combined_plots <- gridExtra::grid.arrange(grobs = list(pnorm_plot, dnorm_plot), ncol = 2, top = "Probabi
```



Relying upon this probabilities, men younger than 81 are more likely to get *sepsis* before in life time than women. As a matter of fact, women who get infected before the third quantile age are the 57.03% of the total feminine population; while men are the 63.39% of the total masculine population.

Survival study cohort

```
survival_study_cohort
```

```
## # A tibble: 19,051 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
```

```
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1     7     1             1                1 female  alive
##  2    17     0             2                1 male    alive
##  3    70     0             1                1 male    alive
##  4    76     0             1                1 male    alive
##  5     8     0             1                1 male    alive
##  6    41     0             2                1 male    alive
##  7    60     0             1                0 male    dead
##  8    89     1             1                0 female  dead
##  9    76     0             3                0 male    dead
## 10    81     1             1                1 female  alive
## # i 19,041 more rows
```

```
str(survival_study_cohort)
```

```
## tibble [19,051 x 6] (S3: tbl_df/tbl/data.frame)
##  $ age                 : num [1:19051] 7 17 70 76 8 41 60 89 76 81 ...
##  $ sex                 : num [1:19051] 1 0 0 0 0 0 0 1 0 1 ...
##  $ count_episode       : num [1:19051] 1 2 1 1 1 2 1 1 3 1 ...
##  $ hospital_outcome    : num [1:19051] 1 1 1 1 1 1 0 0 0 1 ...
##  $ sex_cat             : chr [1:19051] "female" "male" "male" "male" ...
##  $ hospital_outcome_cat: chr [1:19051] "alive" "alive" "alive" "alive" ...
```

```
summary(survival_study_cohort)
```

```
##       age             sex         count_episode   hospital_outcome
##  Min.   :  0.0   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.: 65.0   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.0000
##  Median : 77.0   Median :0.0000   Median :1.000   Median :1.0000
##  Mean   : 72.5   Mean   :0.4486   Mean   :1.396   Mean   :0.8107
##  3rd Qu.: 85.0   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :100.0   Max.   :1.0000   Max.   :5.000   Max.   :1.0000
##    sex_cat          hospital_outcome_cat
##  Length:19051       Length:19051
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

As already did for the previous dataset, the following one is split into two too: *survived* and *dead* people.

```
(survived_survival_study_cohort <- survival_study_cohort %>%
   filter(hospital_outcome_cat=="alive" & hospital_outcome==1))
```

```
## # A tibble: 15,445 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1     7     1             1                1 female  alive
##  2    17     0             2                1 male    alive
##  3    70     0             1                1 male    alive
##  4    76     0             1                1 male    alive
```

```
##  5     8     0             1                1 male    alive
##  6    41     0             2                1 male    alive
##  7    81     1             1                1 female  alive
##  8    55     0             1                1 male    alive
##  9    33     1             2                1 female  alive
## 10    48     0             1                1 male    alive
## # i 15,435 more rows
```

```
(dead_survival_study_cohort <- survival_study_cohort %>%
   filter(hospital_outcome==0 & hospital_outcome_cat =="dead"))
```

```
## # A tibble: 3,606 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1    60     0             1                0 male    dead
##  2    89     1             1                0 female  dead
##  3    76     0             3                0 male    dead
##  4    66     1             1                0 female  dead
##  5    63     0             2                0 male    dead
##  6    73     0             2                0 male    dead
##  7    66     0             1                0 male    dead
##  8    79     1             1                0 female  dead
##  9    87     1             1                0 female  dead
## 10    59     0             1                0 male    dead
## # i 3,596 more rows
```

```
(n_episodes <- survived_survival_study_cohort %>%
   group_by(count_episode) %>%
   summarise(n = n()))
```

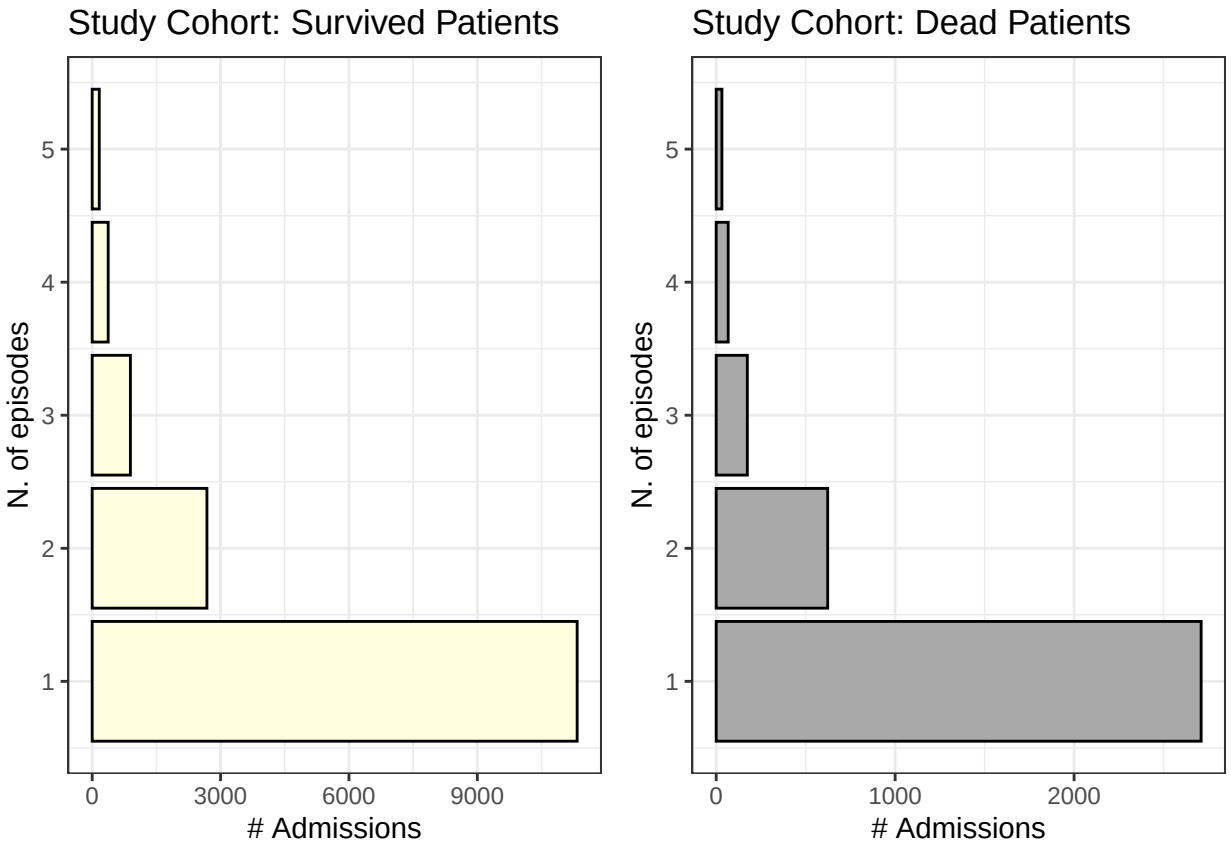
```
## # A tibble: 5 x 2
##   count_episode     n
##           <dbl> <int>
## 1             1 11332
## 2             2  2681
## 3             3   893
## 4             4   374
## 5             5   165
```

```
plot_survived <- ggplot(data = n_episodes, aes(x = count_episode, y = n)) +
  geom_bar(stat = "identity", fill = "lightyellow", color = "black") +
  labs(title = "Study Cohort: Survived Patients",
       y = "# Admissions",
       x = "N. of episodes") +
  theme_bw() +
  coord_flip()

(n_episodes <- dead_survival_study_cohort %>%
    group_by(count_episode) %>%
    summarise(n = n()))
```

```
## # A tibble: 5 x 2
```

```
##   count_episode     n
##           <dbl> <int>
## 1             1  2710
## 2             2   623
## 3             3   174
## 4             4    67
## 5             5    32
```

```
plot_dead <- ggplot( n_episodes, aes(x = count_episode, y = n)) +
  geom_bar(stat = "identity", color = "black", fill = "darkgrey") +
  labs(title = "Study Cohort: Dead Patients",
       y = "# Admissions",
       x = "N. of episodes") +
  theme_bw() +
  coord_flip()

combined_plots <- gridExtra::grid.arrange(grobs = list(plot_survived, plot_dead), ncol = 2)
```



By looking at the plots, it is possible to notice that even in the `survival_study_cohort`, distinguished in *survived* and *dead*, the behavior of the data is the same as the one previously seen: the greater the `N. of episodes`, the lower is the `# Admissions`.

```
(numerical_data_study <- survival_study_cohort[, sapply(survival_study_cohort, is.numeric)])
```

```
## # A tibble: 19,051 x 4
```

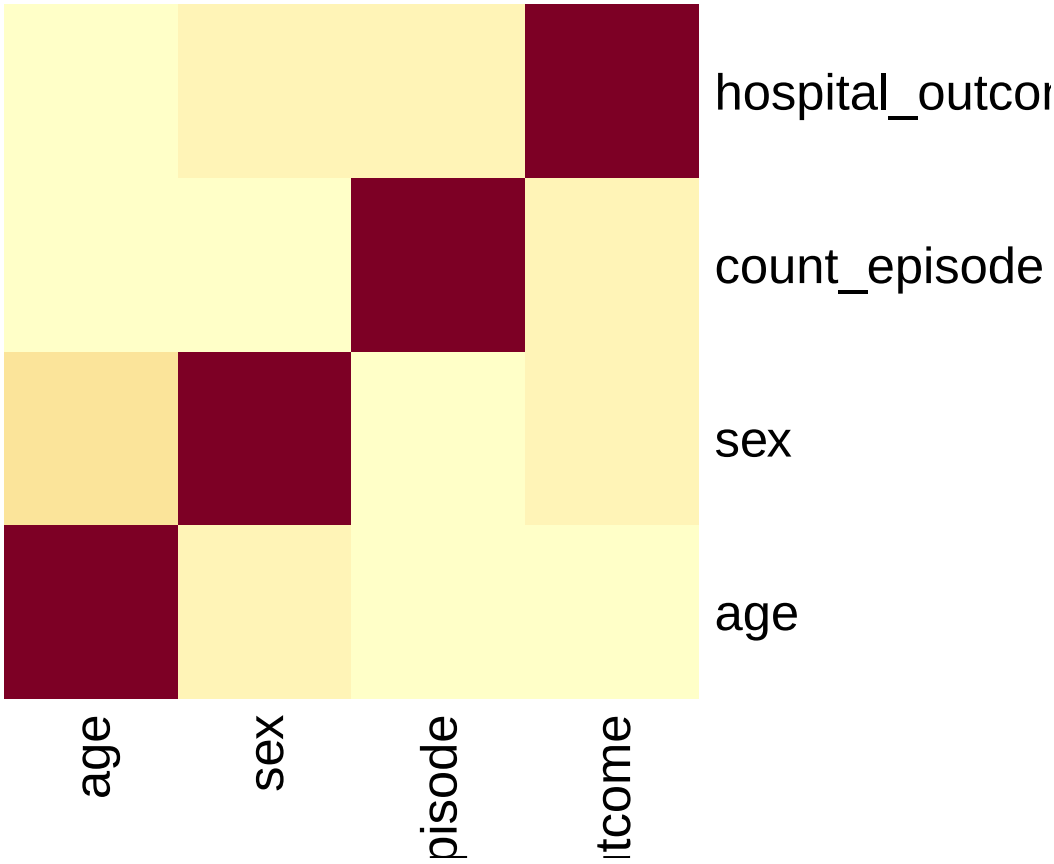
```
##      age   sex count_episode hospital_outcome
##    <dbl> <dbl>         <dbl>            <dbl>
##  1     7     1             1                1
##  2    17     0             2                1
##  3    70     0             1                1
##  4    76     0             1                1
##  5     8     0             1                1
##  6    41     0             2                1
##  7    60     0             1                0
##  8    89     1             1                0
##  9    76     0             3                0
## 10    81     1             1                1
## # i 19,041 more rows
```

```
cor(numerical_data_study)
```

```
##                          age         sex count_episode hospital_outcome
## age               1.00000000  0.06393699   -0.06829214      -0.12617415
## sex               0.06393699  1.00000000   -0.03964150       0.01524892
## count_episode    -0.06829214 -0.03964150    1.00000000       0.02203592
## hospital_outcome -0.12617415  0.01524892    0.02203592       1.00000000
```

While computing the correlation between variables in the dataset, it can be concluded that there isn't a strong relationship.

```
corr_matrix_study = cor(numerical_data_study)
heatmap(corr_matrix_study,
        Colv = NA,
        Rowv = NA,
        scale="column")
```



Survival validation

```
str(survival_validation_cohort)
```

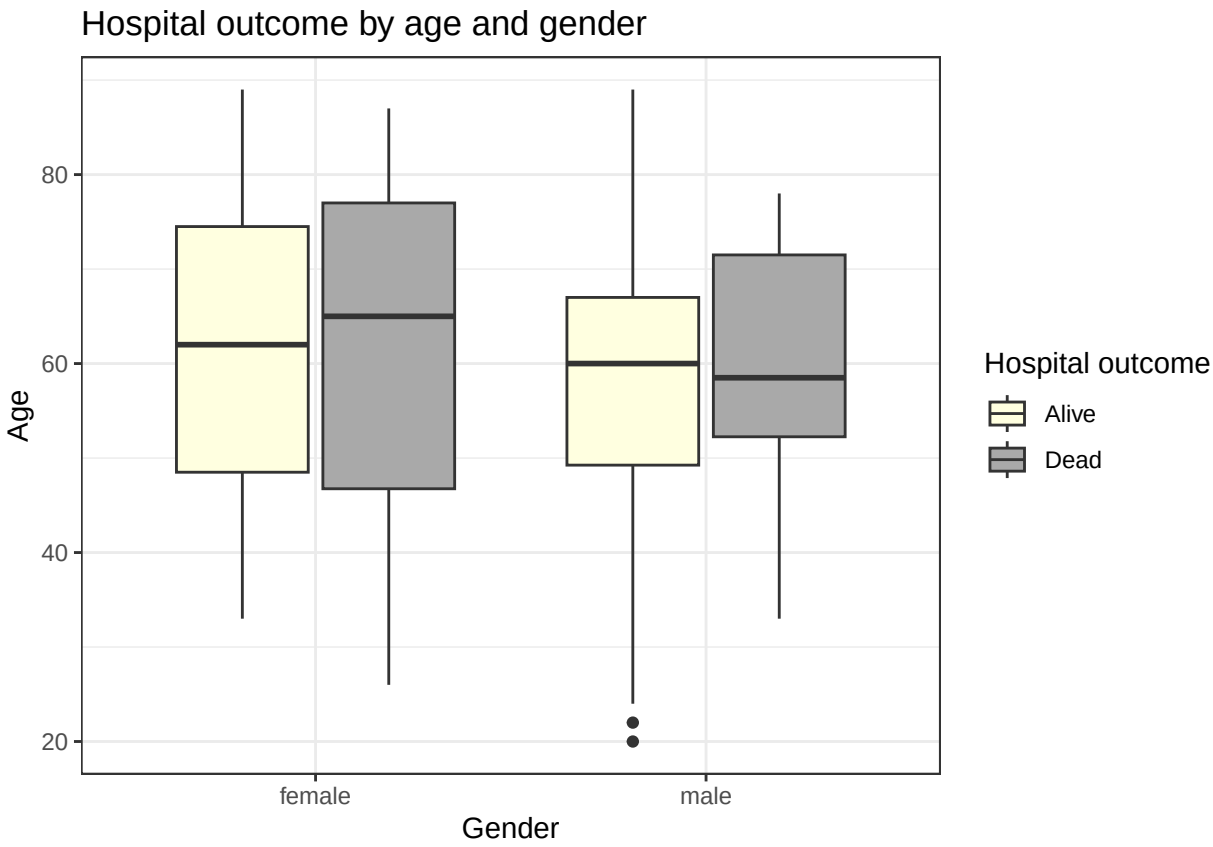
```
## tibble [137 x 6] (S3: tbl_df/tbl/data.frame)
##  $ age                 : num [1:137] 20 22 26 33 33 33 35 35 36 36 ...
##  $ sex                 : num [1:137] 0 0 1 1 0 0 0 1 0 1 ...
##  $ count_episode       : num [1:137] 1 1 2 1 1 2 1 1 1 1 ...
##  $ hospital_outcome    : num [1:137] 1 1 0 1 1 0 1 1 1 1 ...
##  $ sex_cat             : chr [1:137] "male" "male" "female" "female" ...
##  $ hospital_outcome_cat: chr [1:137] "alive" "alive" "dead" "alive" ...
```

```
summary(survival_validation_cohort[!(colnames(survival_primary_cohort) %in%
                                       c("sex_cat", "hospital_outcome_cat"))])
```

```
##       age              sex         count_episode   hospital_outcome
##  Min.   :20.00   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:50.00   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.0000
##  Median :60.00   Median :0.0000   Median :1.000   Median :1.0000
##  Mean   :59.54   Mean   :0.3431   Mean   :1.161   Mean   :0.8248
##  3rd Qu.:72.00   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :89.00   Max.   :1.0000   Max.   :2.000   Max.   :1.0000
```

Here we can see that `age` is strongly symmetrical. While, the `count_episode` seems to be just quite symmetrical.

```
(survival_validation_cohort %>%
  ggplot(aes(x = sex_cat,
             y = age,
             fill = hospital_outcome_cat)) +
  geom_boxplot(show.legend = TRUE) +
  labs(title = "Hospital outcome by age and gender",
       x = "Gender",
       y = "Age",
       fill = "Hospital outcome")+
  scale_fill_manual(values = c("alive" = "lightyellow",
                               "dead" = "darkgrey"),
                    labels = c("Alive",
                               "Dead")) +
  theme_bw())
```



As already did for the previous datasets, the `survival_validation_cohort` is split into two parts too: *survived* and *dead* people.

```
(survived_survival_study_cohort <- survival_study_cohort %>%
  filter(hospital_outcome_cat=="alive" & hospital_outcome==1))
```

```
## # A tibble: 15,445 x 6
```

```
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1     7     1             1                1 female  alive
##  2    17     0             2                1 male    alive
##  3    70     0             1                1 male    alive
##  4    76     0             1                1 male    alive
##  5     8     0             1                1 male    alive
##  6    41     0             2                1 male    alive
##  7    81     1             1                1 female  alive
##  8    55     0             1                1 male    alive
##  9    33     1             2                1 female  alive
## 10    48     0             1                1 male    alive
## # i 15,435 more rows
```

```
(dead_survival_study_cohort <- survival_study_cohort %>%
  filter(hospital_outcome==0 & hospital_outcome_cat =="dead"))
```

```
## # A tibble: 3,606 x 6
##      age   sex count_episode hospital_outcome sex_cat hospital_outcome_cat
##    <dbl> <dbl>         <dbl>            <dbl> <chr>   <chr>
##  1    60     0             1                0 male    dead
##  2    89     1             1                0 female  dead
##  3    76     0             3                0 male    dead
##  4    66     1             1                0 female  dead
##  5    63     0             2                0 male    dead
##  6    73     0             2                0 male    dead
##  7    66     0             1                0 male    dead
##  8    79     1             1                0 female  dead
##  9    87     1             1                0 female  dead
## 10    59     0             1                0 male    dead
## # i 3,596 more rows
```

```
(n_episodes <- survived_survival_study_cohort %>%
   group_by(count_episode) %>%
   summarise(n = n()))
```

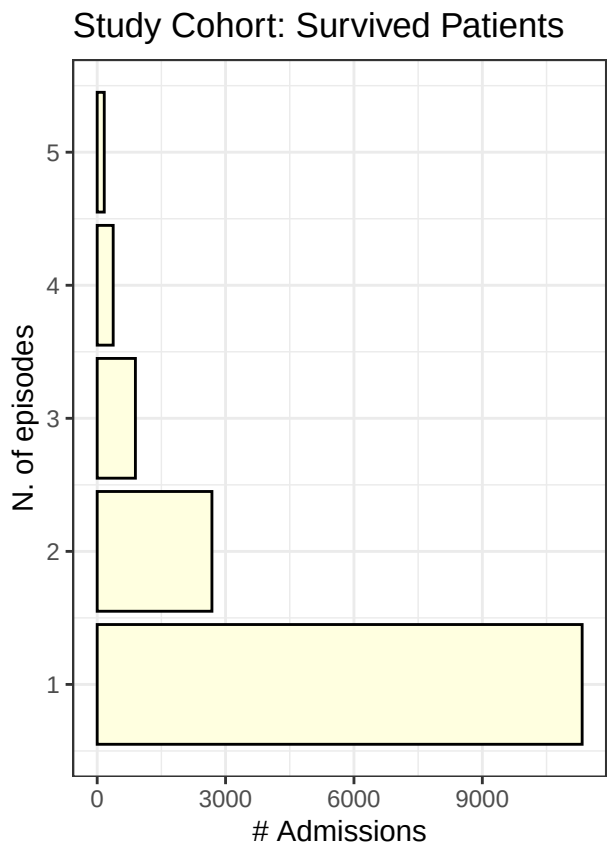
```
## # A tibble: 5 x 2
##   count_episode     n
##           <dbl> <int>
## 1             1 11332
## 2             2  2681
## 3             3   893
## 4             4   374
## 5             5   165
```

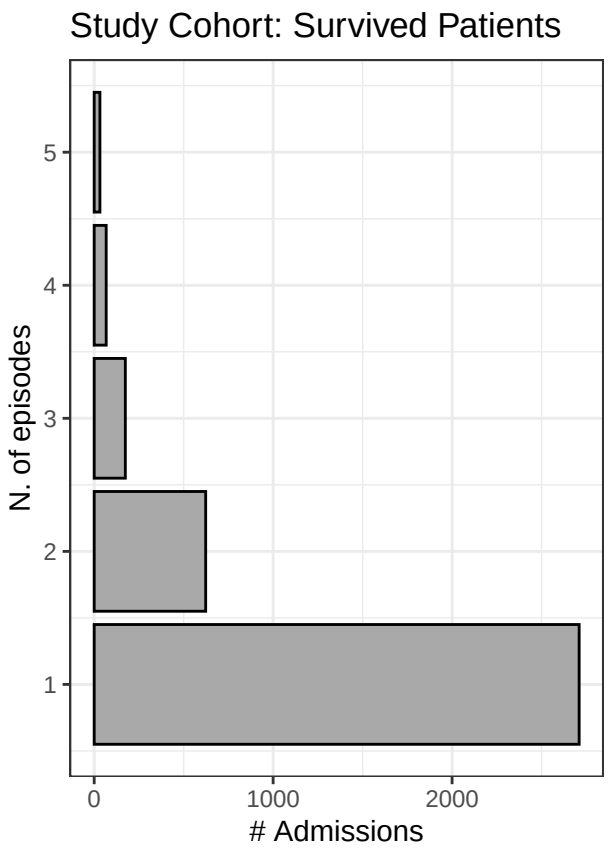
```
plot_survived <- ggplot(data = n_episodes,
                        aes(x = count_episode,
                            y = n)) +
  geom_bar(stat = "identity",
           fill = "lightyellow",
           color = "black") +
  labs(title = "Study Cohort: Survived Patients",
```

```
       y = "# Admissions",
       x = "N. of episodes") +
  theme_bw() +
  coord_flip()

(n_episodes <- dead_survival_study_cohort %>%
    group_by(count_episode) %>%
    summarise(n = n()))
```

```
## # A tibble: 5 x 2
##   count_episode     n
##           <dbl> <int>
## 1             1  2710
## 2             2   623
## 3             3   174
## 4             4    67
## 5             5    32
```

```
plot_dead<- ggplot(data = n_episodes,
                   aes(x = count_episode, y = n)) +
  geom_bar(stat = "identity",
           fill = "darkgrey",
           color = "black") +
  labs(title = "Study Cohort: Survived Patients",
       y = "# Admissions",
       x = "N. of episodes") +
  theme_bw() +
  coord_flip()

combined_plots <- gridExtra::grid.arrange(grobs = list(plot_survived, plot_dead), ncol = 2)
```






```
(numerical_data_study <- survival_study_cohort[, sapply(survival_study_cohort, is.numeric)])
```

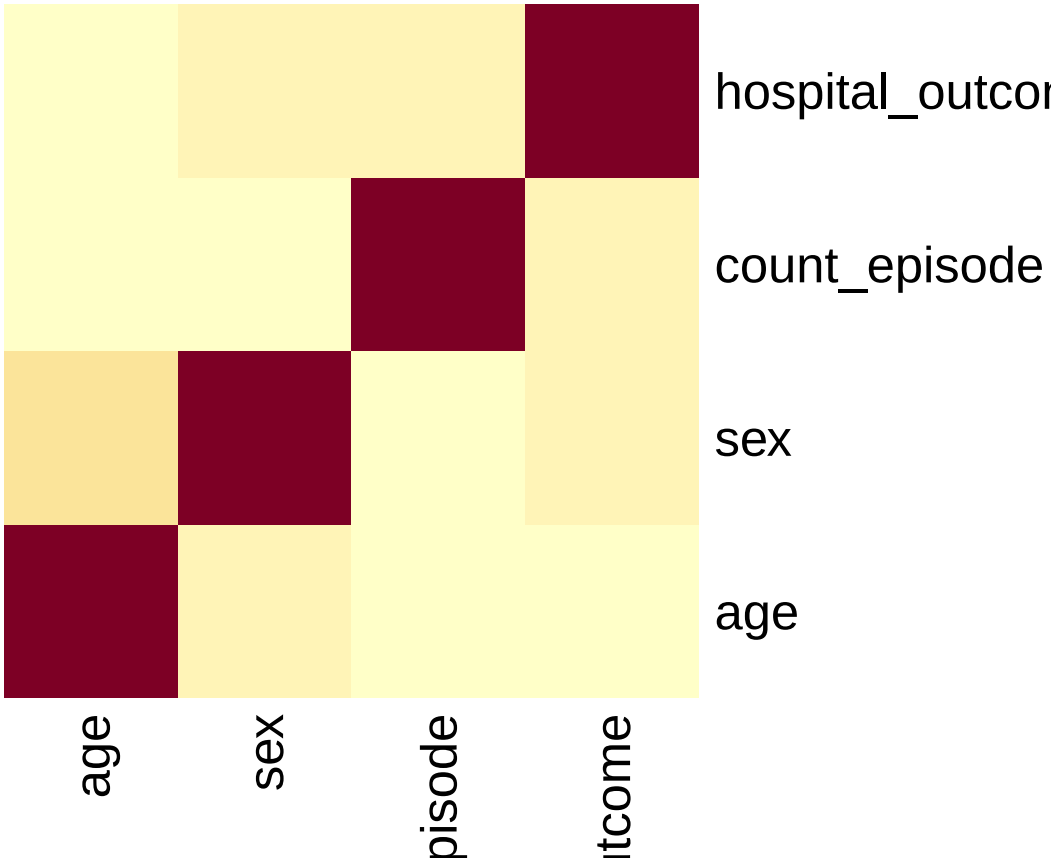
```
## # A tibble: 19,051 x 4
##      age   sex count_episode hospital_outcome
##    <dbl> <dbl>         <dbl>            <dbl>
##  1     7     1             1                1
##  2    17     0             2                1
##  3    70     0             1                1
##  4    76     0             1                1
##  5     8     0             1                1
##  6    41     0             2                1
##  7    60     0             1                0
##  8    89     1             1                0
##  9    76     0             3                0
## 10    81     1             1                1
## # i 19,041 more rows
```

```
cor(numerical_data_study)
```

```
##                             age         sex count_episode hospital_outcome
## age                1.00000000  0.06393699   -0.06829214      -0.12617415
## sex                0.06393699  1.00000000   -0.03964150       0.01524892
## count_episode     -0.06829214 -0.03964150    1.00000000       0.02203592
## hospital_outcome -0.12617415  0.01524892    0.02203592       1.00000000
```

The computed correlation shows that there isn't a strong relationship between the variables.

```
corr_matrix_study = cor(numerical_data_study)
heatmap(corr_matrix_study,
        Colv = NA,
        Rowv = NA,
        scale="column")
```



Sampling methods

```
table(survival_primary_cohort$hospital_outcome)
```

```
##
##      0      1
##   8105 102099
```

```
survival_primary_cohort <- survival_primary_cohort %>%
  select(age,
         sex,
         count_episode,
         hospital_outcome)
```

```r
survival_primary_cohort$hospital_outcome <- as.factor(survival_primary_cohort$hospital_outcome)
survival_primary_cohort$sex <- as.factor(survival_primary_cohort$sex)
survival_primary_cohort$count_episode <- as.factor(survival_primary_cohort$count_episode)
```

Logistic Regression

We want to fit a model of ***logistic regression*** of this type:

$$p(x) = \Pr(Y = 0|x) = \frac{e^{\beta_0+\beta_k x}}{1 + e^{\beta_0+\beta_k x}}$$

Adjusting the Loss Function

We increase the weights of the minority class for the loss function. Then when we fit the *logistic regression* model, we specify the resulting weights in the parameter *weights*.

- `table(target_variable)` calculates the frequency of each class in the binary target variable.
- `prop.table(class_freq)` calculates the proportions of each class frequency relative to the total number of observations.
- `1 / prop.table(class_freq)` calculates the inverse of the proportions to create class weights. The less frequent class (minority class) will have a higher weight, while the more frequent class (majority class) will have a lower weight.

```r
class_freq <- table(survival_primary_cohort$hospital_outcome) # calculates the frequency of each class

# Calculate class weights
class_weights <- 1 / prop.table(class_freq)
# calculates the inverse of the proportions to create class weights.
(class_weights <- round(class_weights, digits = 5))
```

```
##
##        0        1
## 13.59704  1.07938
```

```r
survival_primary_cohort
```

```
## # A tibble: 110,204 x 4
##      age sex   count_episode hospital_outcome
##    <dbl> <fct> <fct>         <fct>
##  1    21 1     1             1
##  2    20 1     1             1
##  3    21 1     1             1
##  4    77 0     1             1
##  5    72 0     1             1
##  6    83 0     1             1
##  7    74 0     1             1
##  8    74 1     1             1
##  9    69 0     1             1
## 10    53 1     1             1
## # i 110,194 more rows
```

Computing the optimal treshold

```
log_survival_cohort <- glm(hospital_outcome ~ .,
                           data = survival_primary_cohort,
                           family ="binomial")
summary(log_survival_cohort)
```

```
##
## Call:
## glm(formula = hospital_outcome ~ ., family = "binomial", data = survival_primary_cohort)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     5.6168241  0.0652935  86.024  < 2e-16 ***
## age            -0.0443361  0.0008285 -53.514  < 2e-16 ***
## sex1            0.1767691  0.0237553   7.441 9.98e-14 ***
## count_episode2 -0.1052041  0.0314822  -3.342 0.000833 ***
## count_episode3 -0.0255946  0.0541160  -0.473 0.636243
## count_episode4 -0.0741647  0.0838599  -0.884 0.376487
## count_episode5  0.0776840  0.1276565   0.609 0.542830
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 57904  on 110203  degrees of freedom
## Residual deviance: 53580  on 110197  degrees of freedom
## AIC: 53594
##
## Number of Fisher Scoring iterations: 6
```

It is possible to see that `count_episode` and `age` have a *negative relationship* with the target variable (`hospital outcome`), meaning that to an increase in the number of episodes of *sepsis* or in the age - or both of them - leads to a decrease in the survival.
Instead, `sex` has a positive relationship with the target variable, so that males has an higher probability of death in comparison with women. In addition, `age` and `sex` are *strongly significant*. In the `count_episode` variable, just `count_episode2` is strongly significant, whose alpha value is approximately 0.001. While the other `count_episode` variables aren't that significant.
On the other hand, we can say that to an increase of the number of episodes there's a higher probability of dying while comparing men and women together.

Note that we used the `count_episode` variables as categorical as we interested in discovering the correlation between the number of episodes and the negative hospital outcome. This is due to the fact that it is not an only mortal hilliness, but it can bring to the failure of any organ in the human body. Moreover, we do not have any data that help us know which organs fail most or how many episodes manifest before an organ failure. Indeed, working on a numerical `count_episode` variable would lead us to take conclusions that may be different from the reality.

Let's test the data by creating a training and a test set. The train set includes randomly the 80% of the observation in the `survival_primary_cohort` dataset. While the test set includes the remaining random 20% of the same dataset.

```r
n.sample <- nrow(survival_primary_cohort)
size = round(0.8 * n.sample)
set.seed(123)
idx = sample(n.sample, size)
train_data_primary = survival_primary_cohort[idx,]
test_data_primary = survival_primary_cohort[-idx,]
```

```r
train_data_primary
```

```
## # A tibble: 88,163 x 4
##      age sex   count_episode hospital_outcome
##    <dbl> <fct> <fct>         <fct>
##  1    87 1     1             1
##  2    82 1     2             1
##  3    80 0     3             1
##  4    71 1     1             1
##  5    78 1     1             1
##  6     9 0     1             1
##  7    92 0     1             1
##  8    60 0     1             1
##  9    10 1     4             1
## 10    25 1     1             1
## # i 88,153 more rows
```

```r
log_survival_primary <- glm(hospital_outcome~ .,
                            data = train_data_primary,
                            family ="binomial")

summary(log_survival_primary)
```

```
##
## Call:
## glm(formula = hospital_outcome ~ ., family = "binomial", data = train_data_primary)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     5.6258511  0.0729507  77.119  < 2e-16 ***
## age            -0.0445753  0.0009254 -48.169  < 2e-16 ***
## sex1            0.1701160  0.0264224   6.438 1.21e-10 ***
## count_episode2 -0.0989568  0.0351186  -2.818  0.00484 **
## count_episode3 -0.0045092  0.0606881  -0.074  0.94077
## count_episode4 -0.0890533  0.0929524  -0.958  0.33804
## count_episode5  0.0420376  0.1409372   0.298  0.76550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46682  on 88162  degrees of freedom
## Residual deviance: 43176  on 88156  degrees of freedom
## AIC: 43190
##
## Number of Fisher Scoring iterations: 6
```

It is possible to observe that the results are quite the same of the previous *logistic regression* analysis.

In order to see if the model is making a good prediction of a positive `hospital_outcome`, we might want to calculate the *misclasification rate*.

```
table(test_data_primary$hospital_outcome)
```

```
##
##     0     1
##  1550 20491
```

```
pred_primary <- predict(log_survival_primary,
                        test_data_primary,
                        type="response")

class_primary <- ifelse(pred_primary >= 0.85, 1,0)
(mr_primary <- mean(class_primary != test_data_primary$hospital_outcome))
```

```
## [1] 0.1236786
```

A *misclassification rate* of the 12.37% indicates a relatively low level of error, implying that the model has a pretty strong predictive capacity.

```
# confusion matrix
(cf_primary <- table(predicted = class_primary,
                     actual = test_data_primary$hospital_outcome))
```

```
##          actual
## predicted     0     1
##         0   269  1445
##         1  1281 19046
```

```
# computing sensitivity
(Se <- cf_primary[2, 2] / sum(cf_primary[, 2]))
```

```
## [1] 0.9294812
```

```
# computing specificity
(Spe <- cf_primary[1, 1] / sum(cf_primary[, 1]))
```

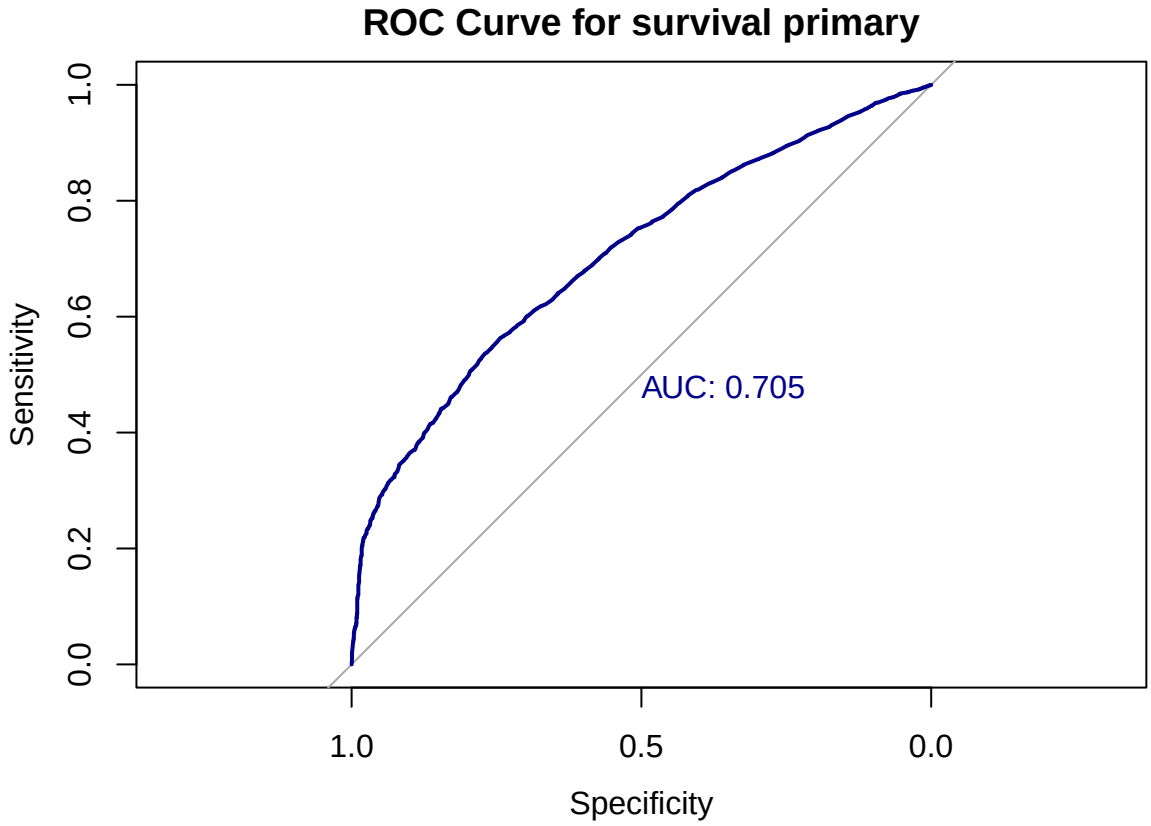
```
## [1] 0.1735484
```

The *confusion matrix* suggest us that the model identifies well the positive outcomes. However, the low value of TN, in conjunction of the 1281 FP, suggests the model struggling in predicting a negative outcome. This influences the result in the computation of the **ROC Curve**.

```
roc_curve <- roc(test_data_primary$hospital_outcome~pred_primary, plot = T, col="darkblue", print.auc =
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



As a matter of fact, an **Area Under the Curve** of 0.705 highlights that the model has a good discriminant capacity, but it can be surely improved - take into account that an **AUC** $= 0.5$ means that the model makes random predictions.

```
(optimal_threshold <- coords(roc_curve, "best", best.method = "closest.topleft"))
```

```
##   threshold specificity sensitivity
## 1 0.9174576   0.6845161   0.6110976
```

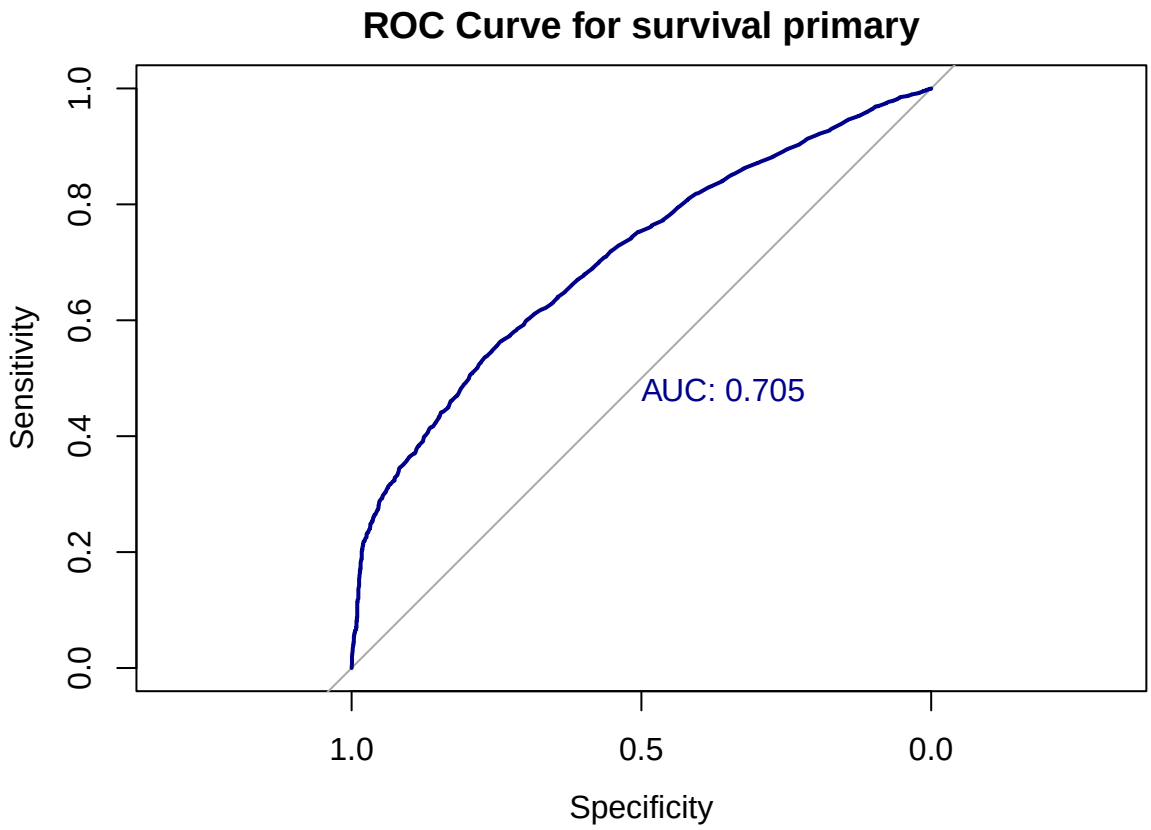
```
opt_pred_primary <- predict(log_survival_primary, test_data_primary, type="response")
opt_class_primary <- ifelse(pred_primary >= optimal_threshold$threshold, 1 , 0)
(mr_primary <- mean(opt_class_primary != test_data_primary$hospital_outcome))
```

```
## [1] 0.3837394
```

```
roc_curve <- roc(test_data_primary$hospital_outcome~opt_pred_primary,
                 plot = T,
                 col="darkblue",
                 print.auc = T,
                 main = "ROC Curve for survival primary")
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Even with the optima threshold the results do not vary. This can be due to the small amount of observations present in the dataset.

Logistic regression: study cohort

Logistic regression is now performed on the `survival_study_cohort`.
First, we select the variables of our interest, transform them into factorial and then split the dataset into *training* and *test* set.

```
study_cohort <- survival_study_cohort %>%
  select(age,
         sex,
         count_episode,
         hospital_outcome)

study_cohort$age <- as.numeric(study_cohort$age)
study_cohort$count_episode <- as.factor(study_cohort$count_episode)
study_cohort$sex <- as.factor(study_cohort$sex)
study_cohort$hospital_outcome <- as.factor(study_cohort$hospital_outcome)

set.seed(123)
idx <- sample(nrow(study_cohort)*0.75)
```

```r
train_study <- study_cohort[idx,]
test_study <- study_cohort[-idx,]
```

```r
log_study <- glm(hospital_outcome ~ .,
                 data = train_study,
                 family = "binomial" )
summary(log_study)
```

```
##
## Call:
## glm(formula = hospital_outcome ~ ., family = "binomial", data = train_study)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.885661   0.116488  24.772  < 2e-16 ***
## age            -0.019662   0.001477 -13.310  < 2e-16 ***
## sex1            0.156121   0.044319   3.523 0.000427 ***
## count_episode2  0.038138   0.058029   0.657 0.511042
## count_episode3  0.201056   0.101095   1.989 0.046725 *
## count_episode4  0.202771   0.152090   1.333 0.182457
## count_episode5  0.148974   0.224920   0.662 0.507752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13553  on 14287  degrees of freedom
## Residual deviance: 13328  on 14281  degrees of freedom
## AIC: 13342
##
## Number of Fisher Scoring iterations: 4
```

After performing the *logistic regression* on the study cohort, we can notice that there is a *positive relationship* between the intercept and `sex` covariate, and a *negative relationship* between the intercept and `age`. Covariates `age` and `sex` are strongly significant as previously assessed with the primary cohort. There is also a slight positive relationship with `count_episode3` on a significance level of 0.05.

We proceed by computing the ***misclassification rate*** in order to evaluate the overall performance of the model.

```r
prediction_study <- predict(log_study,
                            newdata = test_study,
                            type = "response")
class_pred_study <- ifelse(prediction_study > 0.75,1,0)
(mr_study <- mean(ifelse(prediction_study > 0.75,1,0)!= test_study$hospital_outcome))
```

```
## [1] 0.2221289
```

In this case, the rate is greater compared to the primary cohort, with a value corresponding to 22.21%, meaning that the models' classification accuracy is quite lower.

Next step is to create a confusion matrix in order to compute *sensitivity* and *specificity*.

```r
(conf_mat_study <- table(actual = test_study$hospital_outcome,
                         predicted = class_pred_study))
```

```
##       predicted
## actual    0    1
##      0   56  951
##      1  107 3649
```

```r
(sens_study <- conf_mat_study[2,2]/sum(conf_mat_study[,2]))
```

```
## [1] 0.7932609
```

```r
(spec_study <- conf_mat_study[1,1]/sum(conf_mat_study[,1]))
```

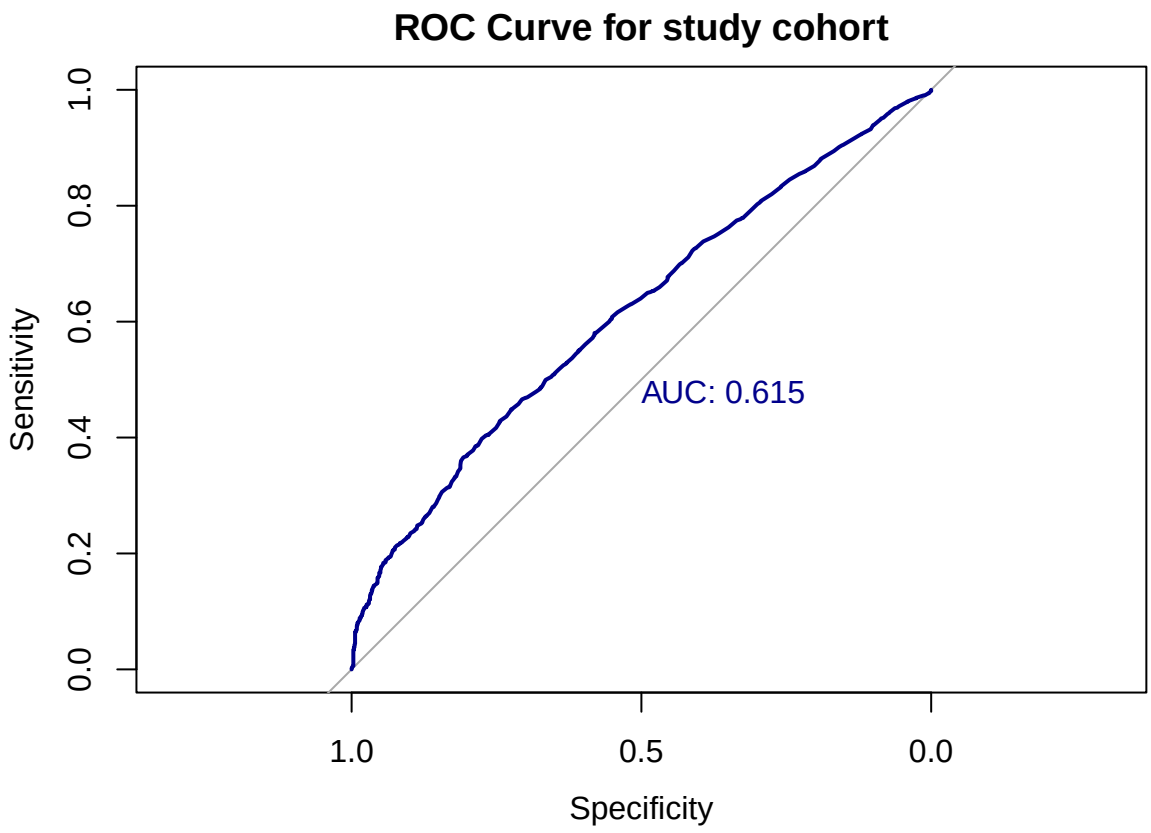
```
## [1] 0.3435583
```

The outcomes correspond to a *sensitivity* value of 0.79, meaning that the rate of TP is almost 79%, and a *specificity* value of 0.34, so the rate of TN is just 34%. This is due to a *positively imbalanced* dataset.

```r
(roc_curve_study <- roc(test_study$hospital_outcome~prediction_study,
                        plot = T,
                        col="darkblue",
                        fill = "lightblue",
                        print.auc = T,
                        main = "ROC Curve for study cohort"))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.formula(formula = test_study$hospital_outcome ~ prediction_study,     plot = T, col = "darkblue"
##
## Data: prediction_study in 1007 controls (test_study$hospital_outcome 0) < 3756 cases (test_study$hosp
## Area under the curve: 0.6148
```

The **Area Under the Curve** has a value of 0.615, meaning that the model's performance is slightly above the average (case of random classification). The classification performance on the `survival_study_cohort` seems to be lower than previously computed on the primary cohort. The model still require improvement for better accuracy.

The validation cohort

```
str(survival_validation_cohort)
```

```
## tibble [137 x 6] (S3: tbl_df/tbl/data.frame)
##  $ age                 : num [1:137] 20 22 26 33 33 33 35 35 36 36 ...
##  $ sex                 : num [1:137] 0 0 1 1 0 0 0 1 0 1 ...
##  $ count_episode       : num [1:137] 1 1 2 1 1 2 1 1 1 1 ...
##  $ hospital_outcome    : num [1:137] 1 1 0 1 1 0 1 1 1 1 ...
##  $ sex_cat             : chr [1:137] "male" "male" "female" "female" ...
##  $ hospital_outcome_cat: chr [1:137] "alive" "alive" "dead" "alive" ...
```

```
survival_validation_cohort <- survival_validation_cohort %>%
  select(age,
         sex,
         count_episode,
         hospital_outcome)
```

```
survival_validation_cohort$sex <- as.factor(survival_validation_cohort$sex)
survival_validation_cohort$count_episode <- as.factor(survival_validation_cohort$count_episode)
survival_validation_cohort$hospital_outcome <- as.factor(survival_validation_cohort$hospital_outcome)
```

```
head(predictions <- predict(log_survival_primary,
                            newdata = survival_validation_cohort,
                            type="response"), 20)
```

```
##         1         2         3         4         5         6         7         8
## 0.9912884 0.9904838 0.9894190 0.9869391 0.9845544 0.9829752 0.9831385 0.9857386
##         9        10        11        12        13        14        15        16
## 0.9823834 0.9850982 0.9837312 0.9790171 0.9814467 0.9771053 0.9776925 0.9739150
##        17        18        19        20
## 0.9769215 0.9758948 0.9722775 0.9702936
```

```
class_pred_validation <- ifelse(predictions >= 0.85, 1,0)
(mr_primary <- mean(class_pred_validation != survival_validation_cohort$hospital_outcome))
```

```
## [1] 0.189781
```

```
class_pred_validation
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
##   1   1   1   0   0   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
```

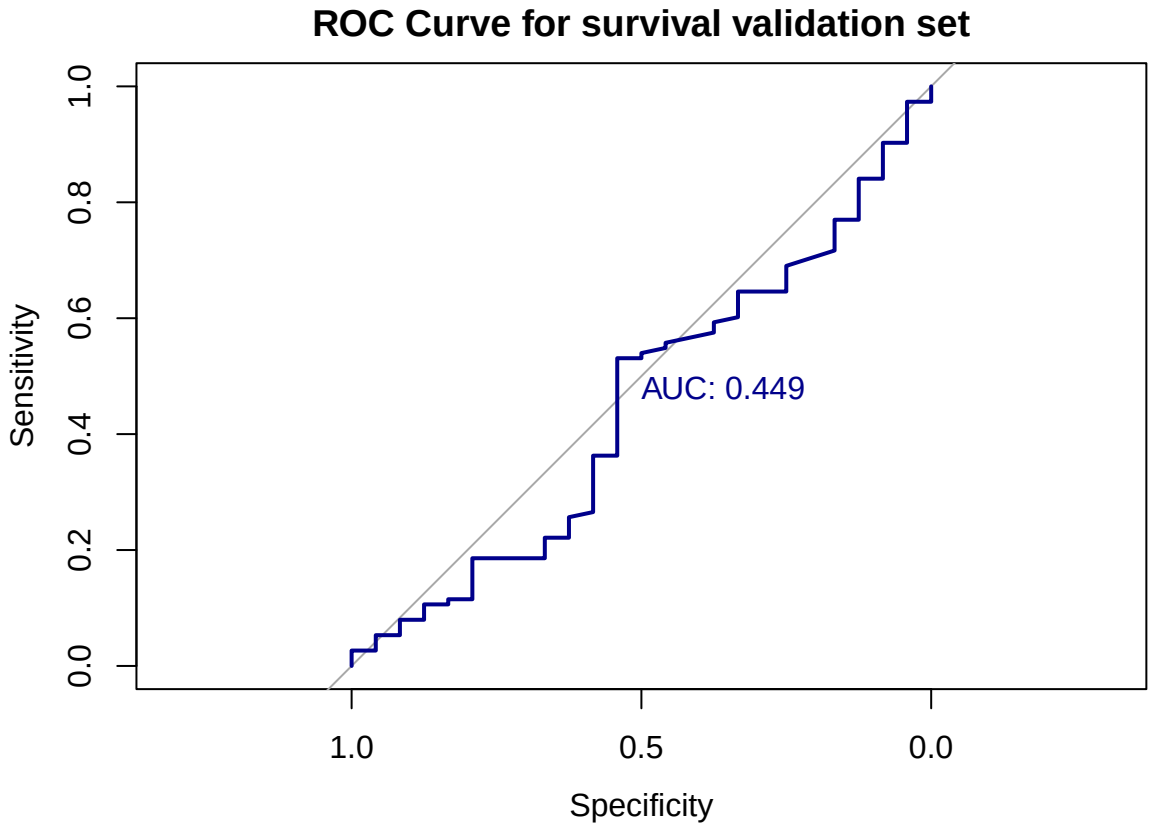
```
(confusion_matrix <- table(predicted = class_pred_validation,
                           actual = survival_validation_cohort$hospital_outcome))
```

```
##          actual
## predicted   0   1
##         0   0   2
##         1  24 111
```

```r
roc_curve <- roc(survival_validation_cohort$hospital_outcome~predictions,
                 plot = T,
                 col="darkblue",
                 print.auc = T,
                 main = "ROC Curve for survival validation set")
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls > cases
```



As previously noticed, it is evident that there is a scarcity of data in the latest **ROC Curve** computation. Indeed, the **AUC** is below 0.5, meaning that the model is generating random predictions.
This kind of difficulty can be avoid by taking into account all the negative outcome from the dataset and just a bounded selection of the positive ones, potentially yielding more favorable results. On the other hand, this approach would fail to represent reality, which is the primary aim.
Alternative models can be taken into account in order to achieve more coherent results.

Survival Analysis

In conducting this survival analysis, the aim is to model the the time until the infection from *sepsis* occurs. From the data source, it is known are the start and the of the study. However, what isn't known are the recovery and the end of the recovery dates.
Indeed, take into account the following information:

Start of the study: 2010-01-01
End of the study: 2011-12-31
The study lasts **2 years**

```
set.seed(123)

# Defining the beginning and the ending dates
start_date <- as.Date("2010-01-01")
end_date <- as.Date("2011-12-31")

# Creating a sequence of dates from start_date to end_date
all_dates <- seq(start_date, end_date, by = "day")

# N. observations
n_obs <- nrow(survival_primary_cohort)

# Random generation of recovering date
start_dates <- sample(all_dates, n_obs, replace = TRUE)

# Generate a random interval of days for the duration of hospitalization (between 1 and 30 days)
recovery_duration <- sample(1:30, n_obs, replace = TRUE)

# Calculate the discharge dates by adding the interval of days to the admission date.
end_dates <- start_dates + recovery_duration

# Ensure that the discharge dates do not exceed the overall end date.
end_dates <- pmin(end_dates, end_date)

dates <- data.frame(start_date = start_dates,
                    end_date = end_dates)
head(dates, 10)
```

```
##    start_date   end_date
## 1  2011-02-19 2011-03-13
## 2  2011-04-08 2011-04-22
## 3  2010-06-28 2010-07-08
## 4  2011-06-10 2011-06-11
## 5  2010-07-14 2010-07-18
## 6  2010-04-28 2010-05-11
## 7  2010-10-26 2010-11-19
## 8  2010-08-17 2010-09-06
## 9  2010-09-01 2010-09-04
## 10 2010-01-14 2010-01-31
```

```
survival_primary_cohort <- cbind(survival_primary_cohort,dates)
head(survival_primary_cohort)
```

```
##   age sex count_episode hospital_outcome start_date   end_date
## 1  21   1             1                1 2011-02-19 2011-03-13
## 2  20   1             1                1 2011-04-08 2011-04-22
## 3  21   1             1                1 2010-06-28 2010-07-08
## 4  77   0             1                1 2011-06-10 2011-06-11
## 5  72   0             1                1 2010-07-14 2010-07-18
## 6  83   0             1                1 2010-04-28 2010-05-11
```

```
survival_primary_cohort$duration <- difftime(survival_primary_cohort$end_date,
                                             survival_primary_cohort$start_date, units="days")

survival_primary_cohort$duration <- as.numeric(survival_primary_cohort$duration)

head(survival_primary_cohort)
```

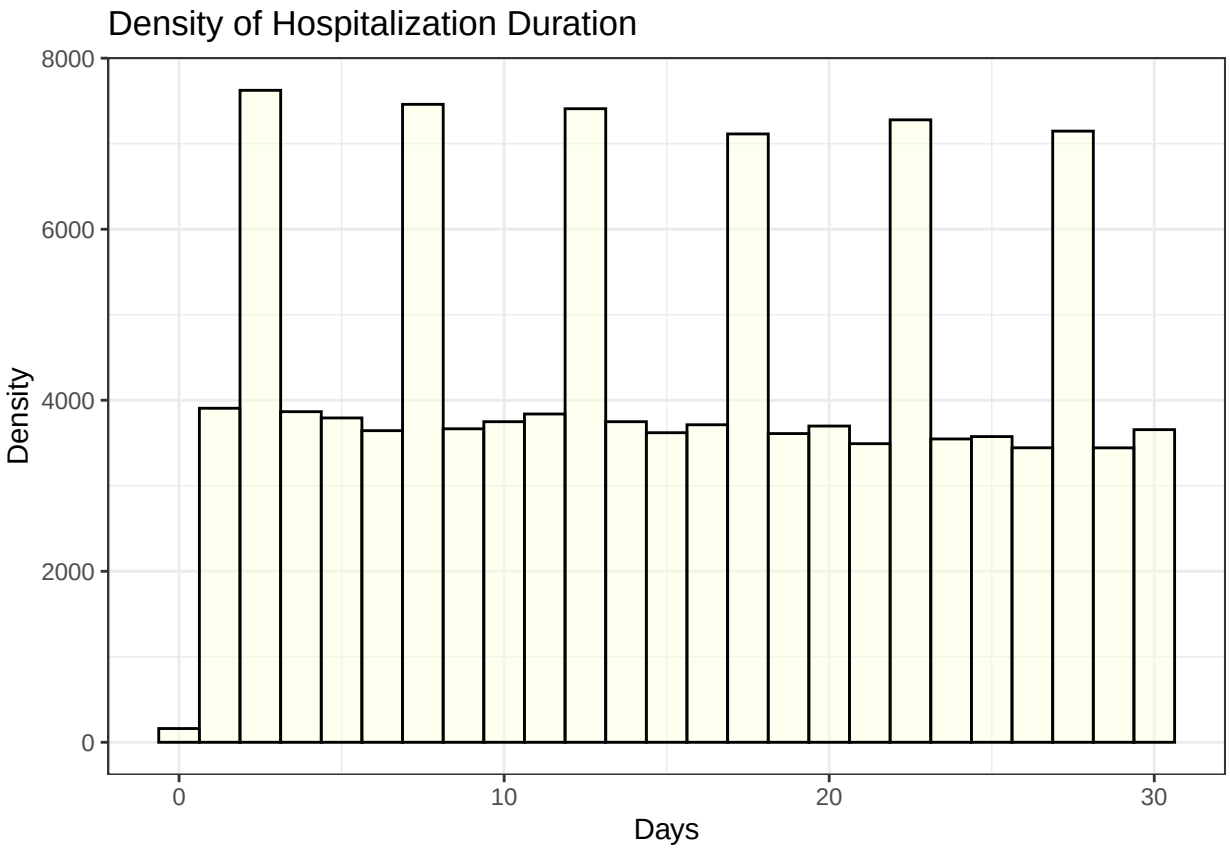
```
##   age sex count_episode hospital_outcome start_date   end_date duration
## 1  21   1             1                1 2011-02-19 2011-03-13       22
## 2  20   1             1                1 2011-04-08 2011-04-22       14
## 3  21   1             1                1 2010-06-28 2010-07-08       10
## 4  77   0             1                1 2011-06-10 2011-06-11        1
## 5  72   0             1                1 2010-07-14 2010-07-18        4
## 6  83   0             1                1 2010-04-28 2010-05-11       13
```

```
str(survival_primary_cohort)
```

```
## 'data.frame':    110204 obs. of  7 variables:
##  $ age             : num  21 20 21 77 72 83 74 74 69 53 ...
##  $ sex             : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 2 1 2 ...
##  $ count_episode   : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ hospital_outcome: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ start_date      : Date, format: "2011-02-19" "2011-04-08" ...
##  $ end_date        : Date, format: "2011-03-13" "2011-04-22" ...
##  $ duration        : num  22 14 10 1 4 13 24 20 3 17 ...
```

Let's plot the randomly sampled recovery dates just found, in order to better understand the distribution of the hospital stay durations.

```
ggplot(data = survival_primary_cohort) +
  geom_histogram(aes(x = duration),
                 bins = 25,
                 fill = "lightyellow",
                 alpha = 0.5,
                 color = "black") +
  labs(x = "Days",
       y = "Density",
       title = "Density of Hospitalization Duration") +
  theme_bw()
```



As we expected, the distribution shows that the data aren't evenly spread out. There are some noticeable peaks in the middle, but the rest seems to be fairly consistent.

Censoring

Let's get deep down into this *survival analysis* considering **censored** data. From now on, ***right censoring*** is taken into account, considering subjects who exit the study before an event occurs or when the study concludes before the event manifests. A subject may be censored due to:

- Loss to follow-up
- Withdrawal from study
- No event by end of fixed study period

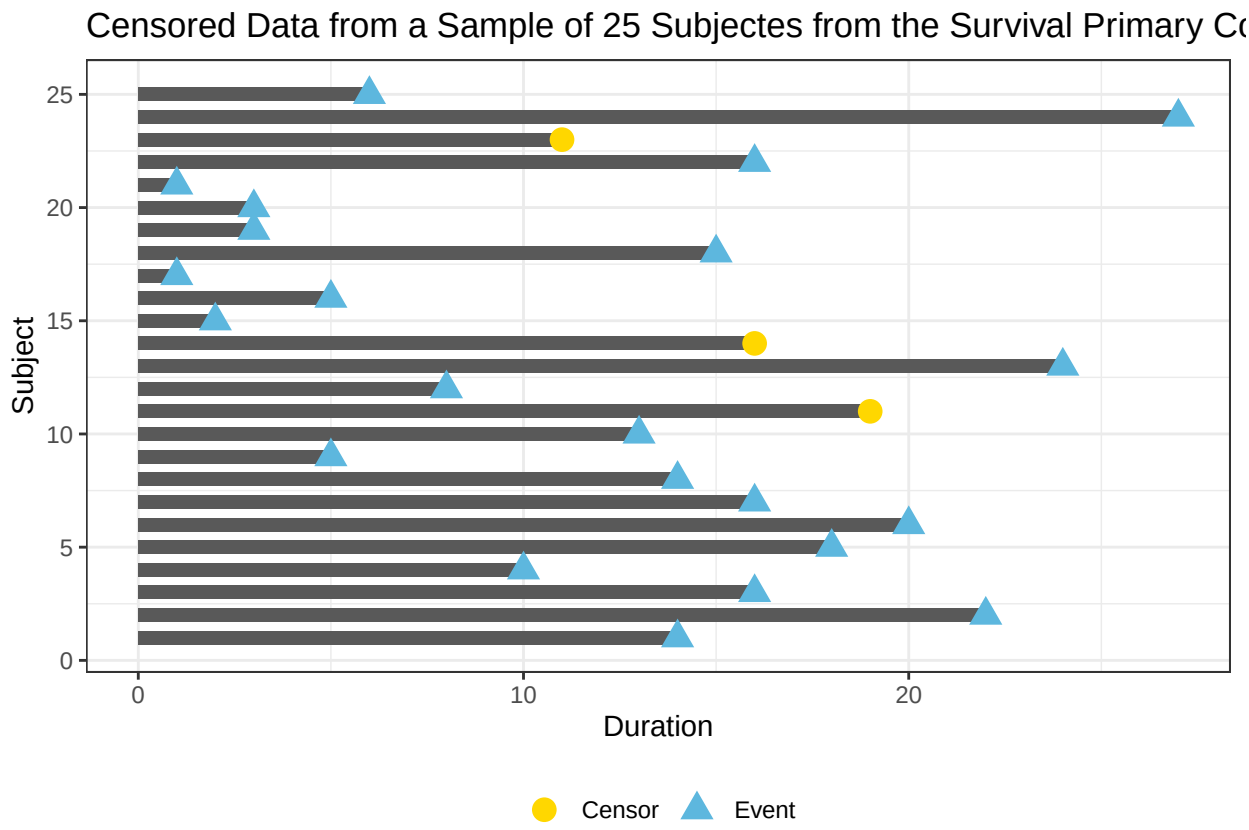
In particular, referring to the main topic of this research, we analyze patients who manifested the infection by *sepsis* during the two years of data collection. Let's assume that those patients who have had *sepsis* more than one and then passed away are **censored**.

```
# Creating a survival object
Surv(survival_primary_cohort$duration, survival_primary_cohort$hospital_outcome)[1:10]
```

```
##  [1] 22:1 14:1 10:1  1:1  4:1 13:1 24:1 20:1  3:1 17:1
```

```r
# Switching values of hospital_outcome variable
survival_primary_cohort <-
  survival_primary_cohort %>%
  mutate(hospital_outcome=dplyr::recode(hospital_outcome,
                                        "0"=1,
                                        "1"=0))
```

```r
set.seed(192837465)
data <- survival_primary_cohort %>%
  mutate(censor = ifelse(count_episode != 1 & hospital_outcome == 1,
                         "Censor",
                         "Event"),
         duration = as.numeric(duration)) %>%
  select(duration, censor, count_episode) %>%
  sample_n(size = 25, replace = FALSE) %>%
  mutate(n = row_number())

ggplot(data, aes(n, duration)) +
    geom_bar(stat = "identity", width = 0.6) +
    geom_point(aes(color = censor, shape = censor),
               size = 4) +
    coord_flip() +
    theme_bw() +
    theme(legend.title = element_blank(),
          legend.position = "bottom") +
    scale_color_manual(values = c("Censor" = "gold", "Event" = "#5DB7DE")) +
    labs(y = "Duration",
         x = "Subject",
         title = "Censored Data from a Sample of 25 Subjectes from the Survival Primary Cohort Dataset"
```



The dataset under study comprises 110,204 observations. Therefore, in order to best visualize the above graph, a random sample of twenty-five observations was chosen.
Indeed, in the study time considered, it is possible to observe that only three subjects out of twenty-five are deliberated as *censored*, *i.e.* the subjects died once they have contracted the infection more than a single time. Let's compute the proportion of those who are event-free 20 days of hospital stay:

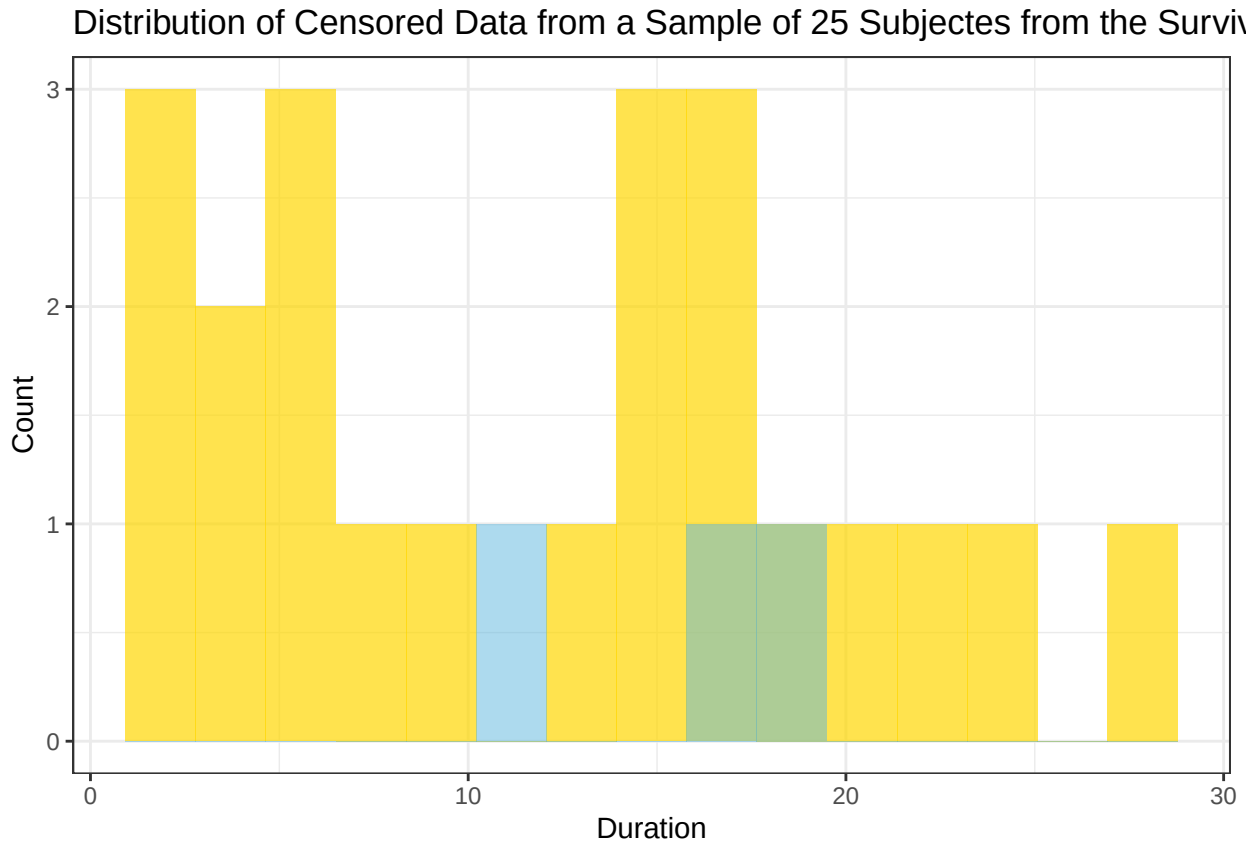
- Subjects 11, 14 and 23 were **censored before 20 days** of hospital stay, meaning that we are aware of the fact that they had more than one episode of *sepsis* and died before the 20^{th} day.
- Subjects 2, 13 and 24 were **event free at 20 days**.
- Subject 7 had the **event on the 20** th day.
- All the other seventeen subjects had the **event before 20 days** and they survived. Moreover, we know that they had one or more episode of *sepsis*, but don't know how many of them precisely.

Of course this graph is not representative of the whole dataset; however is a good representation to understand how to compute the proportion mentioned before. Let's see what happens if the number of subjects increases to fifty.

As the distribution of the follow-up might be skewed as we are considering a sample of the whole data, it could be interesting seeing how the number of *censored patients* differs from those with the event. In order to do that, let's plot an histogram:

```
ggplot(data, aes(duration, fill = censor)) +
  geom_histogram(data = subset(data, censor == "Event"),
```

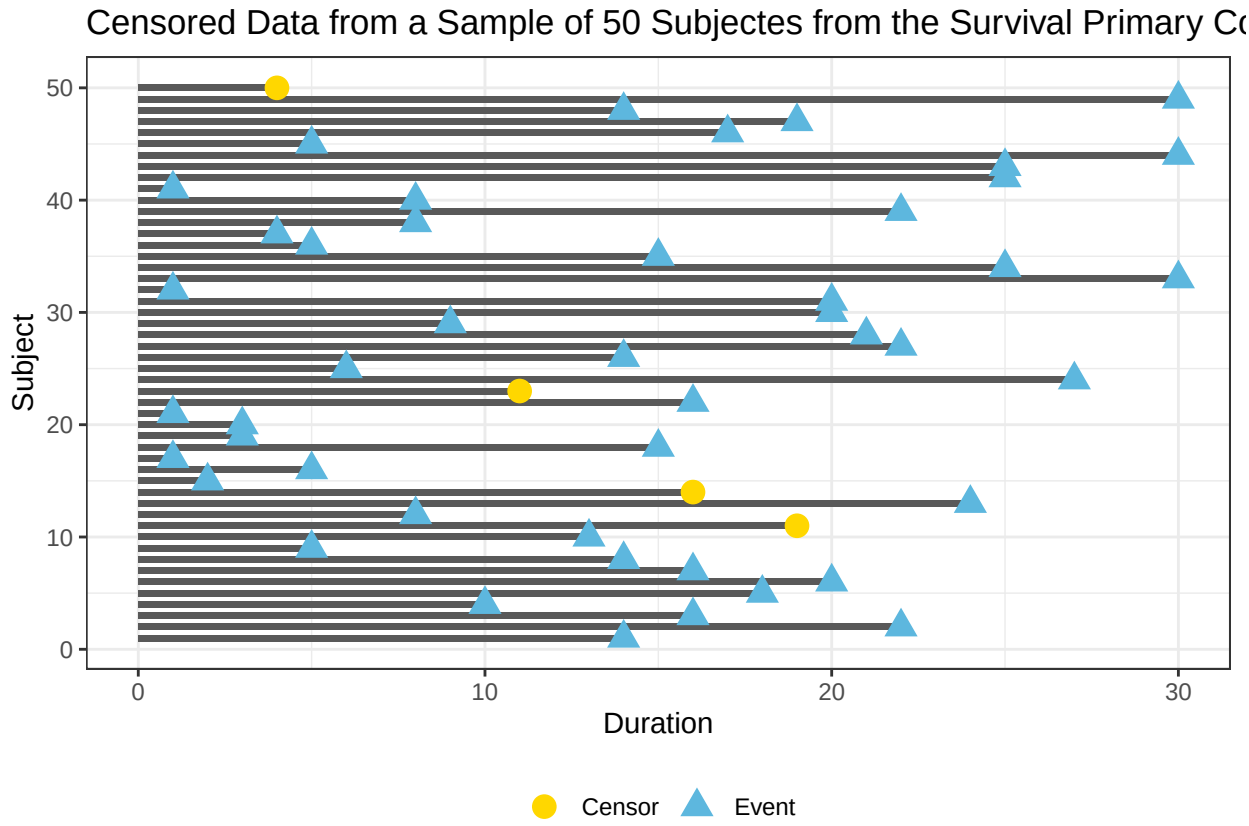
```
                aes(x = duration),
                bins = 15, alpha = 0.7, fill = "gold") +
  geom_histogram(data = subset(data, censor == "Censor"),
                aes(x = duration),
                bins = 15, alpha = 0.5, fill = "#5DB7DE") +
  theme_bw() +
  labs(x = "Duration",
       y = "Count",
       fill = "Censor",
       title = "Distribution of Censored Data from a Sample of 25 Subjectes from the Survival Primary C
```



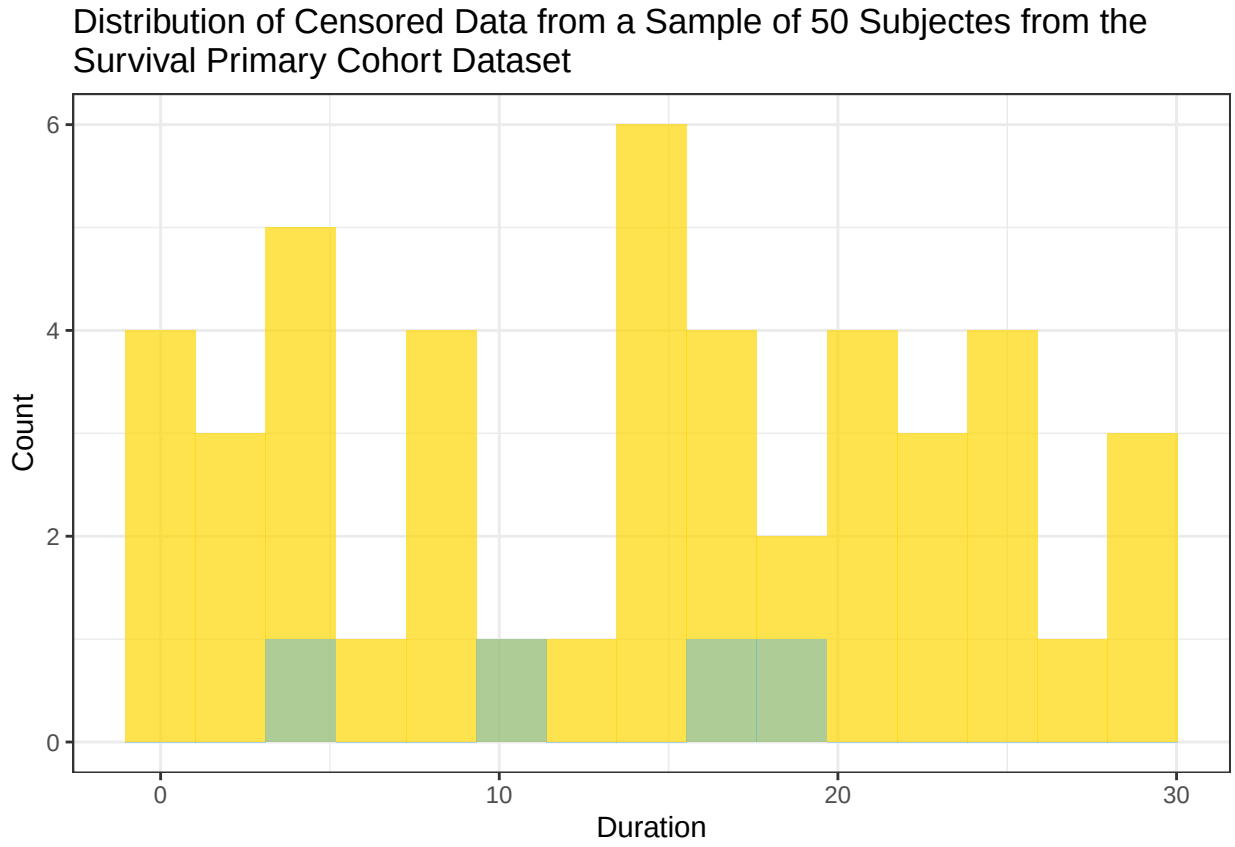
The distribution doesn't present any particular skweness in this case, indeed as the general distribution.

```
set.seed(192837465)
data <- survival_primary_cohort %>%
  mutate(censor = ifelse(count_episode != 1 & hospital_outcome == 1,
                         "Censor",
                         "Event"),
         duration = as.numeric(duration)) %>%
  select(duration, censor, count_episode) %>%
  sample_n(size = 50, replace = FALSE) %>%
  mutate(n = row_number())

ggplot(data, aes(n, duration)) +
    geom_bar(stat = "identity", width = 0.6) +
```

```
    geom_point(aes(color = censor, shape = censor),
               size = 4) +
    coord_flip() +
    theme_bw() +
    theme(legend.title = element_blank(),
          legend.position = "bottom") +
    scale_color_manual(values = c("Censor" = "gold",
                                  "Event" = "#5DB7DE")) +
    labs(y = "Duration",
         x = "Subject",
         title = "Censored Data from a Sample of 50 Subjectes from the Survival Primary Cohort Dataset")
```




```
ggplot(data, aes(duration, fill = censor)) +
  geom_histogram(data = subset(data, censor == "Event"),
                 aes(x = duration),
                 bins = 15, alpha = 0.7, fill = "gold") +
  geom_histogram(data = subset(data, censor == "Censor"),
                 aes(x = duration),
                 bins = 15, alpha = 0.5, fill = "#5DB7DE") +
  theme_bw() +
  labs(x = "Duration",
       y = "Count",
       fill = "Censor",
       title = "Distribution of Censored Data from a Sample of 50 Subjectes from the \nSurvival Primary
```



Very briefly, it is possible to notice that the number of **censored subjects** increased by one, but still all of them passed away before the 20 th day of hospitalization.
Additionally, subjects 6, 30 and 31 experienced the event on the 20^{th} day of hospitalization. At the same time, the number of those subject who were event free before the same day increased to twelve. Among all the other individuals, the infection occurred before 20 days of hospital stay.
However, the exact timing of the event occurrence remains unknown. What is certain is that they survived and that the event occurred at least once.

Regarding the density distribution of this sample, a very slight positive skewness is evident for individuals with the event.

Keep in mind that this sample is too small to be a good representation of the whole dataset too. Still, it is a good way to visualize censored data.

Empirical Distribution Function

The ***Empirical Distribution Function*** is a non-parametric approach used to estimate the underlying distribution of the data.
The ***EDF*** is defined as $\hat{F}(t) = 1 - \hat{S}(t)$, where $\hat{S}(t)$ is the *empirical survival function.*

Let's step back a little bit, in order to define the latest:

$$\hat{S}(t) = \frac{\text{Number of individuals with survival time} \geq t}{\text{Number of individuals in the dataset}}$$

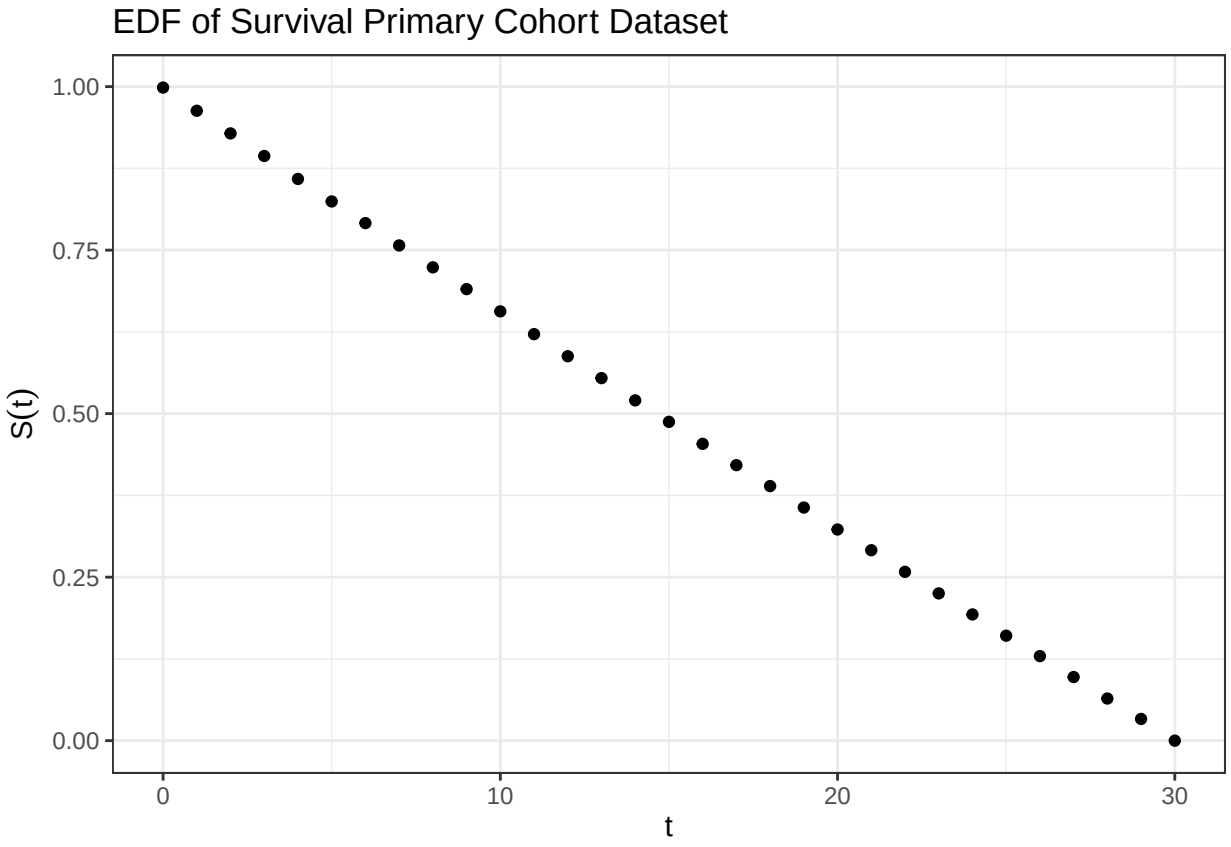
where $t \in T$, $T = [0, 800]$.

Therefore, the **EDF** can be re-write as:

$$\hat{F}(t) = 1 - \frac{\text{Number of individual with survival time} \geq t}{\text{Number of individuals in the dataset}}$$

Let's now compute our result.

```
data <- survival_primary_cohort %>%
  mutate(censor = ifelse(count_episode != 1 & hospital_outcome == 1, 1, 0),
         duration = duration) %>%
  select(duration, censor)

# EDF
ggplot(survival_primary_cohort) +
  cmstatr::stat_esf(aes(x = duration)) +
  xlab(expression(t)) + ylab(expression(S(t))) +
  labs(title = "EDF of Survival Primary Cohort Dataset") +
  theme_bw()
```

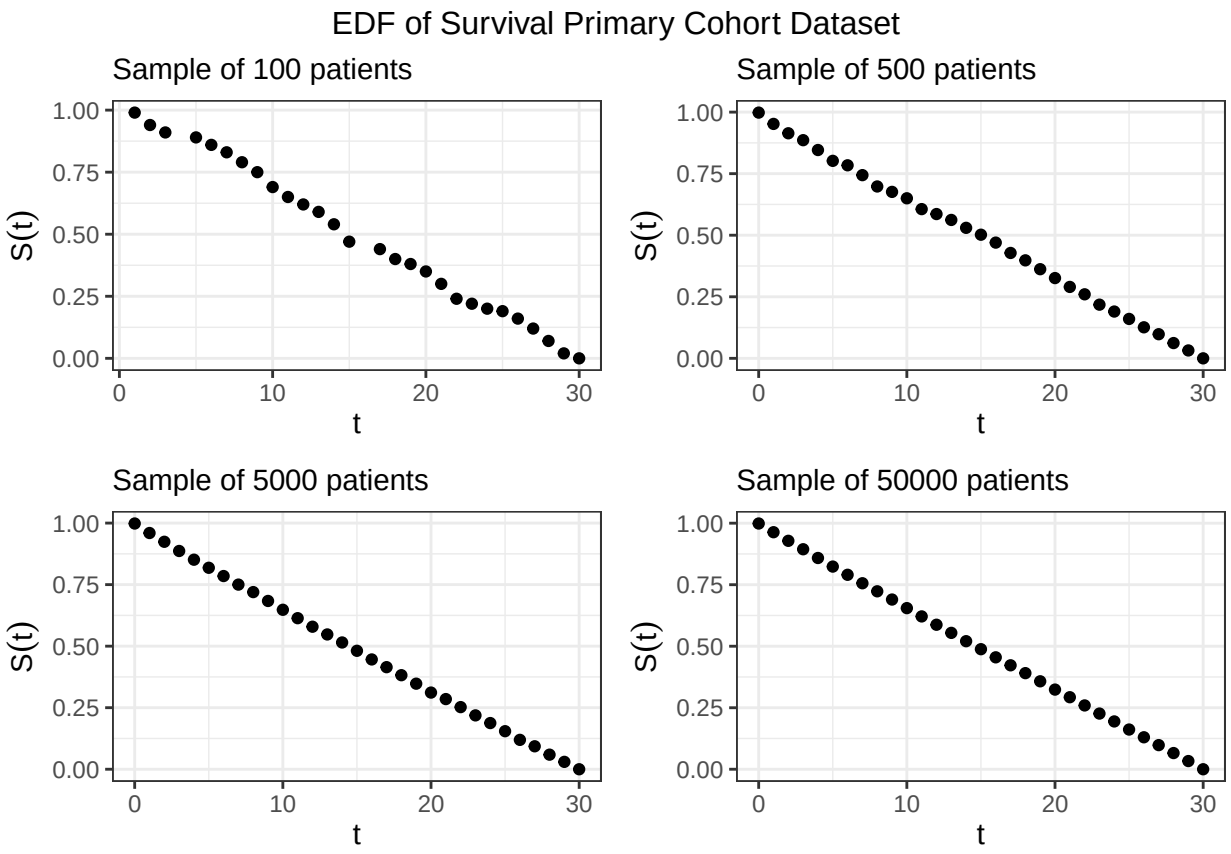


It appears that the data follow the path of straight line, as it is considering all the 110204. Indeed, the graph tells us the fact that the patients have the pretty same probability to be alive before time t. This is due to the low variance on the dataset.

In order to have a better glance of how the **EDF** works, we consider some samples (100; 500; 5,000; 50,000).

```
# Sampling
sample_sizes <- c(100, 500, 5000, 50000)
```

```
plots <- lapply(sample_sizes, function(size) {
  data <- survival_primary_cohort %>%
    sample_n(size = size, replace = FALSE)

  ggplot(data) +
    cmstatr::stat_esf(aes(x = duration)) +
    xlab(expression(t)) + ylab(expression(S(t))) +
    labs(subtitle = paste("Sample of", size, "patients")) +
    theme_bw() +
    theme(plot.title = element_text(size = 12))
})

gridExtra::grid.arrange(grobs = plots, nrow = 2, ncol = 2,
                        top = "EDF of Survival Primary Cohort Dataset")
```



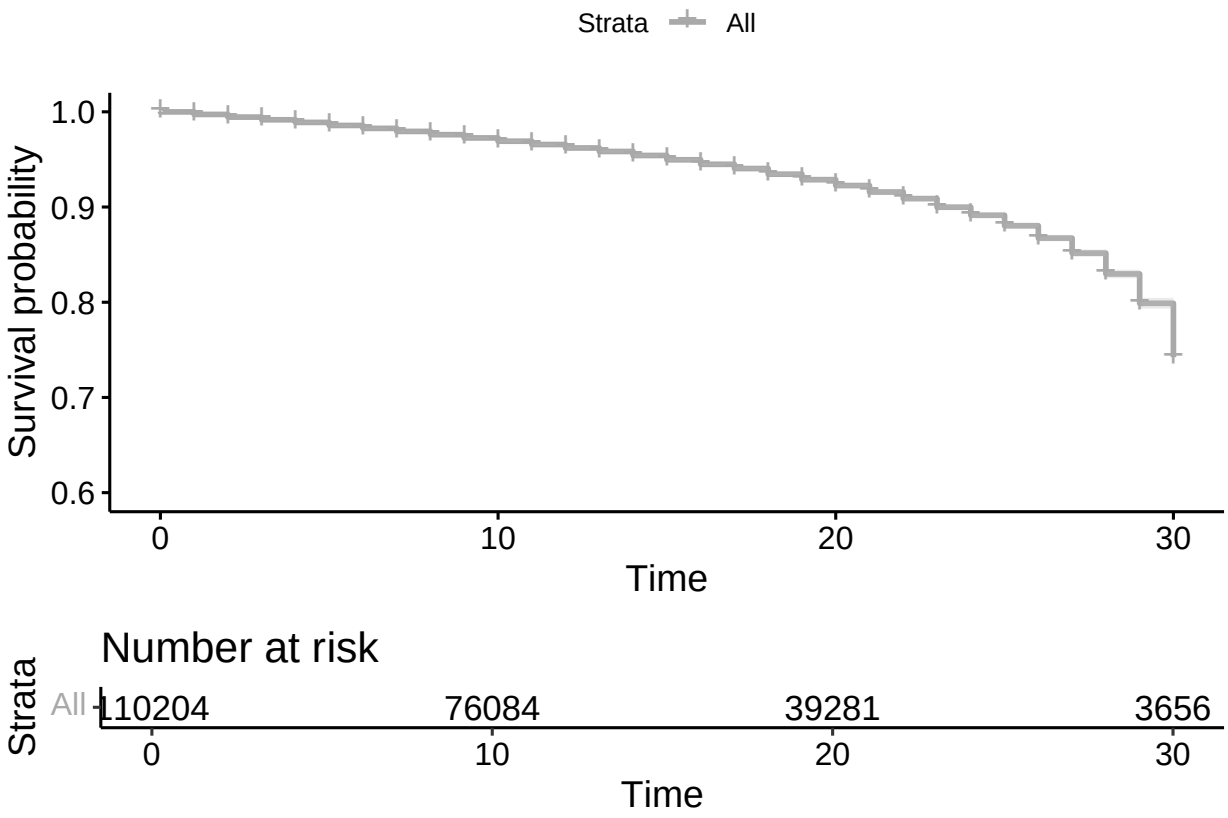
By examining each sample, we can conclude that the smaller the sample size, the higher the variance in the probability of being alive after time t.

```
s <- Surv(survival_primary_cohort$duration,
          survival_primary_cohort$hospital_outcome)
sfit <- survfit(s~1)
summary(sfit)
```

```
## Call: survfit(formula = s ~ 1)
##
```

```
##  time n.risk n.event survival  std.err lower 95% CI upper 95% CI
##     0 110204      11    1.000 3.01e-05        1.000        1.000
##     1 110043     290    0.997 1.57e-04        0.997        0.998
##     2 106137     300    0.994 2.26e-04        0.994        0.995
##     3 102321     292    0.992 2.80e-04        0.991        0.992
##     4  98513     271    0.989 3.24e-04        0.988        0.990
##     5  94647     306    0.986 3.71e-04        0.985        0.986
##     6  90854     283    0.983 4.13e-04        0.982        0.983
##     7  87210     284    0.979 4.53e-04        0.979        0.980
##     8  83443     291    0.976 4.93e-04        0.975        0.977
##     9  79750     279    0.973 5.32e-04        0.972        0.974
##    10  76084     270    0.969 5.70e-04        0.968        0.970
##    11  72335     259    0.966 6.08e-04        0.964        0.967
##    12  68496     259    0.962 6.46e-04        0.961        0.963
##    13  64777     247    0.958 6.85e-04        0.957        0.960
##    14  61087     267    0.954 7.28e-04        0.953        0.956
##    15  57338     273    0.950 7.75e-04        0.948        0.951
##    16  53718     267    0.945 8.23e-04        0.943        0.947
##    17  50005     237    0.940 8.69e-04        0.939        0.942
##    18  46425     293    0.934 9.30e-04        0.933        0.936
##    19  42891     264    0.929 9.90e-04        0.927        0.931
##    20  39281     261    0.923 1.05e-03        0.920        0.925
##    21  35583     261    0.916 1.13e-03        0.914        0.918
##    22  32091     244    0.909 1.20e-03        0.906        0.911
##    23  28448     282    0.900 1.31e-03        0.897        0.902
##    24  24812     233    0.891 1.41e-03        0.889        0.894
##    25  21265     263    0.880 1.54e-03        0.877        0.883
##    26  17690     261    0.867 1.72e-03        0.864        0.871
##    27  14246     259    0.852 1.95e-03        0.848        0.855
##    28  10721     275    0.830 2.30e-03        0.825        0.834
##    29   7099     264    0.799 2.89e-03        0.793        0.805
##    30   3656     259    0.742 4.33e-03        0.734        0.751
```

```
ggsurvplot(fit = sfit,
           data = survival_primary_cohort,
           risk.table = TRUE,
           conf.int = TRUE,
           palette = "darkgrey",
           ylim = c(0.6,1))
```



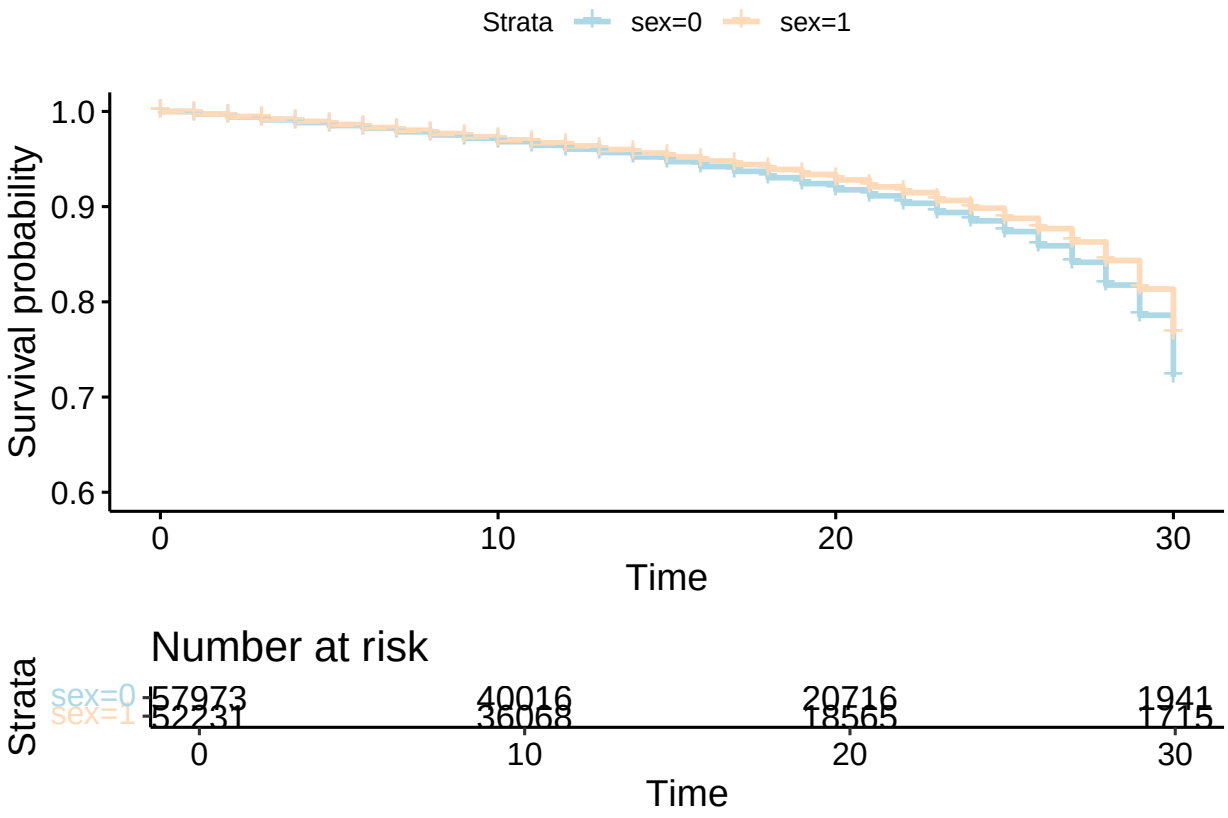
Looking at the graph, it is possible to notice that the *survival probability* is constant till the 10^{th} day, as the curve seems to follow a horizontal trajectory. Subsequently, there is a noticeable decline after this point, the slope becoming extremely steeper after the 20^{th}-day.

```
sfit <- survfit(Surv (duration,hospital_outcome ) ~ sex,
                data = survival_primary_cohort)
summary(sfit)
```

```
## Call: survfit(formula = Surv(duration, hospital_outcome) ~ sex, data = survival_primary_cohort)
##
##                 sex=0
##  time n.risk n.event survival  std.err lower 95% CI upper 95% CI
##     0  57973       4    1.000 3.45e-05        1.000        1.000
##     1  57894     163    0.997 2.23e-04        0.997        0.998
##     2  55826     165    0.994 3.19e-04        0.994        0.995
##     3  53854     159    0.991 3.94e-04        0.990        0.992
##     4  51850     155    0.988 4.59e-04        0.987        0.989
##     5  49828     159    0.985 5.21e-04        0.984        0.986
##     6  47839     138    0.982 5.73e-04        0.981        0.983
##     7  45930     169    0.979 6.35e-04        0.977        0.980
##     8  43962     152    0.975 6.90e-04        0.974        0.977
##     9  42009     149    0.972 7.43e-04        0.970        0.973
##    10  40016     147    0.968 7.97e-04        0.967        0.970
##    11  38058     152    0.964 8.53e-04        0.963        0.966
##    12  36061     148    0.960 9.09e-04        0.959        0.962
```

```
##    13  34067     127    0.957 9.60e-04        0.955        0.959
##    14  32165     155    0.952 1.02e-03        0.950        0.954
##    15  30239     157    0.947 1.09e-03        0.945        0.949
##    16  28319     155    0.942 1.16e-03        0.940        0.944
##    17  26361     141    0.937 1.23e-03        0.935        0.939
##    18  24455     174    0.930 1.32e-03        0.928        0.933
##    19  22583     151    0.924 1.41e-03        0.921        0.927
##    20  20716     146    0.918 1.50e-03        0.915        0.921
##    21  18786     130    0.911 1.59e-03        0.908        0.914
##    22  16946     145    0.904 1.70e-03        0.900        0.907
##    23  15022     161    0.894 1.85e-03        0.890        0.897
##    24  13137     129    0.885 1.98e-03        0.881        0.889
##    25  11238     143    0.874 2.17e-03        0.870        0.878
##    26   9304     159    0.859 2.43e-03        0.854        0.864
##    27   7472     151    0.842 2.76e-03        0.836        0.847
##    28   5639     160    0.818 3.27e-03        0.811        0.824
##    29   3716     144    0.786 4.07e-03        0.778        0.794
##    30   1941     160    0.721 6.17e-03        0.709        0.733
##
##                sex=1
##  time n.risk n.event survival  std.err lower 95% CI upper 95% CI
##     0  52231       7    1.000 5.07e-05        1.000        1.000
##     1  52149     127    0.997 2.22e-04        0.997        0.998
##     2  50311     135    0.995 3.19e-04        0.994        0.995
##     3  48467     133    0.992 3.96e-04        0.991        0.993
##     4  46663     116    0.990 4.57e-04        0.989        0.990
##     5  44819     147    0.986 5.28e-04        0.985        0.987
##     6  43015     145    0.983 5.94e-04        0.982        0.984
##     7  41280     115    0.980 6.45e-04        0.979        0.982
##     8  39481     139    0.977 7.06e-04        0.975        0.978
##     9  37741     130    0.973 7.63e-04        0.972        0.975
##    10  36068     123    0.970 8.17e-04        0.969        0.972
##    11  34277     107    0.967 8.65e-04        0.965        0.969
##    12  32435     111    0.964 9.17e-04        0.962        0.966
##    13  30710     120    0.960 9.76e-04        0.958        0.962
##    14  28922     112    0.956 1.03e-03        0.954        0.958
##    15  27099     116    0.952 1.10e-03        0.950        0.954
##    16  25399     112    0.948 1.16e-03        0.946        0.950
##    17  23644      96    0.944 1.22e-03        0.942        0.947
##    18  21970     119    0.939 1.30e-03        0.936        0.942
##    19  20308     113    0.934 1.38e-03        0.931        0.937
##    20  18565     115    0.928 1.48e-03        0.925        0.931
##    21  16797     131    0.921 1.60e-03        0.918        0.924
##    22  15145      99    0.915 1.70e-03        0.911        0.918
##    23  13426     121    0.907 1.84e-03        0.903        0.910
##    24  11675     104    0.898 1.99e-03        0.895        0.902
##    25  10027     120    0.888 2.19e-03        0.883        0.892
##    26   8386     102    0.877 2.41e-03        0.872        0.882
##    27   6774     108    0.863 2.72e-03        0.858        0.868
##    28   5082     115    0.843 3.21e-03        0.837        0.850
##    29   3383     120    0.813 4.10e-03        0.805        0.822
##    30   1715      99    0.767 5.99e-03        0.755        0.778
```

```r
colors <- c("lightblue", "#FFDAB9")

ggsurvplot(fit = sfit,
           data = survival_primary_cohort,
           risk.table = TRUE,
           palette = colors,
           ylim = c(0.6,1))
```



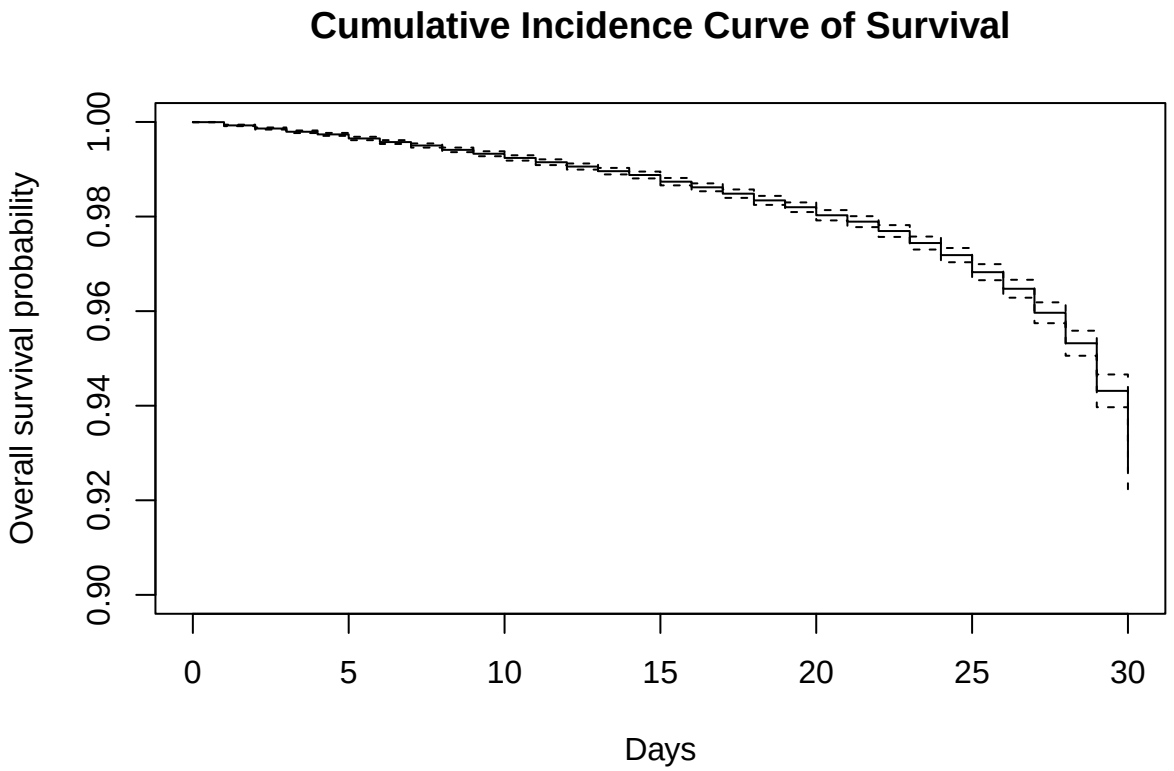
In particular, it is possible to conclude that *women* have a greater survival probability than *men*.

Kaplan-Meier

We now introduce the **Kaplan-Meier** method, which is a common tool used in the *survival analysis* in order to estimate cumulative survival over time.
In particular, **Kaplan-Meier** method is useful when dealing with censored survival data, where not all participants have experienced the event, *i.e.* had the *sepsis* infection at least one time and survived to it, within the two years of observation.
Keep in mind the the curve remains flat during time intervals where no events happen, and it decreases abruptly whenever there is a change in the `survival function` caused by an event occurrence.

```r
data <- survival_primary_cohort %>%
  mutate(censor = ifelse(count_episode != 1 & hospital_outcome == 1, 1, 0),
         duration = as.numeric(duration)) %>%
  select(duration, censor)
```

```r
surv_fit <- survfit(Surv(data$duration, data$censor) ~ 1)

plot(surv_fit,
     xlab = "Days",
     ylab = "Overall survival probability",
     main = "Cumulative Incidence Curve of Survival",
     ylim = c(0.9,1))
```

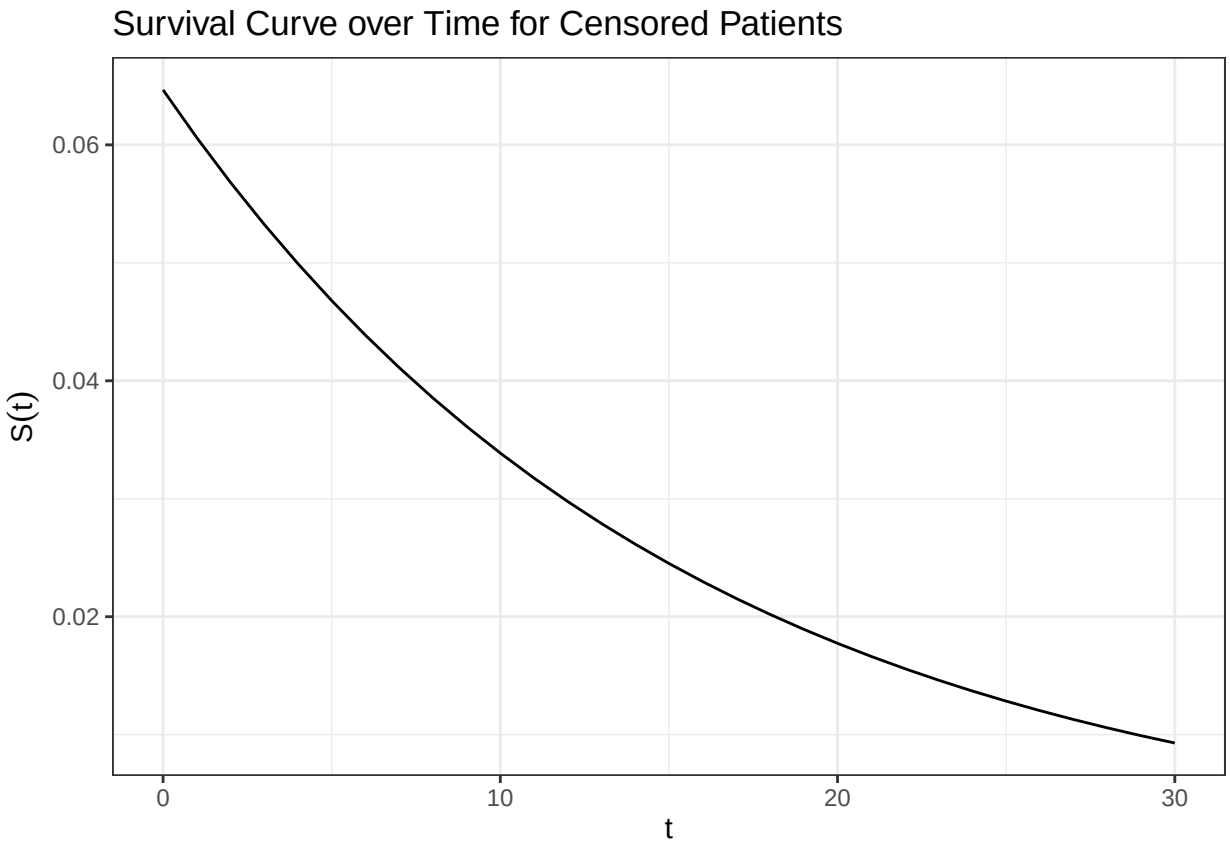


The graph above illustrates the "**Cumulative Incidence Curve of Survival**" and its evolution over time. It's evident that the curve maintains a relatively flat trajectory until the 10^{th} day, after which it gradually begins to decline until the 20^{th} day. Beyond this point, there is a sudden vertical drop in the curve. This pattern suggests that during periods of horizontal stability, the occurrence of events is minimal. On the contrary, when the curve experiences a **vertical decline**, there is a corresponding **decrease** in the **cumulative incidence** of survival probability.
It is possible to easily deduce that it doesn't behave differently from the **EDF** curve previously computed.

When discussing the survival probability of *censored* patients, it's evident that it decreases over time.
Indeed, the trend of the `survival curve` over time demonstrates the fluctuation in survival probability with duration.
Beginning a little above 0.06, it steadily declines over time, indicating a reduction in the likelihood of survival as the duration of recovery increases.

```r
data <- survival_primary_cohort %>%
  mutate(censor = ifelse(count_episode != 1 & hospital_outcome == 1, "Censor", "Event"),
         duration = as.numeric(duration))
```

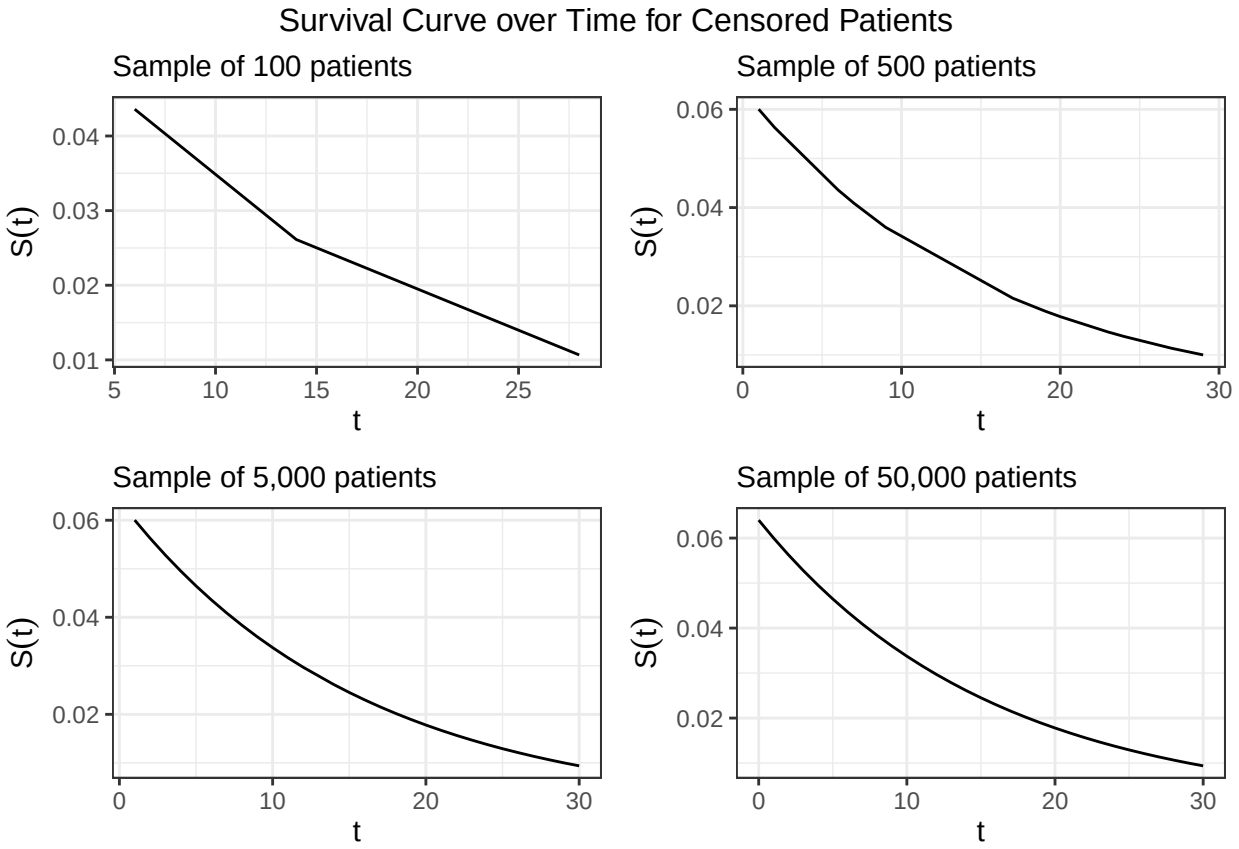
```r
data_0 <- data %>%
  dplyr::filter(censor == "Censor")

lambdahat <- mean(data_0$duration)^{-1}

data_0$par <- dexp(data_0$duration, rate = lambdahat)

ggplot(data_0) +
  geom_line(aes(x = duration, y = par)) +
  xlab(expression(t)) + ylab(expression(S(t))) +
  theme_bw() +
  labs(title = "Survival Curve over Time for Censored Patients")
```



We may wight sample data in order to see if there are changes if a lower number of patients is taken into account.

```r
data <- survival_primary_cohort %>%
  mutate(censor = ifelse(count_episode != 1 & hospital_outcome == 1, "Censor", "Event"),
         duration = as.numeric(duration))


calculate_lambda_hat <- function(data) {
  return(mean(data$duration)^{-1})
}
```

```r
create_survival_plots <- function(data, sample_sizes, titles) {
  plots <- list()

  for (i in seq_along(sample_sizes)) {
    sampled_data <- data %>%
      sample_n(size = sample_sizes[i], replace = FALSE) %>%
      dplyr::filter(censor == "Censor")

    lambdahat <- calculate_lambda_hat(sampled_data)

    p <- ggplot(sampled_data) +
      geom_line(aes(x = duration, y = dexp(duration, rate = lambdahat))) +
      xlab(expression(t)) + ylab(expression(S(t))) +
      theme_bw() +
      labs(subtitle = titles[i])

    plots[[i]] <- p
  }

  return(plots)
}

sample_sizes <- c(200, 500, 5000, 50000)
titles <- c("Sample of 100 patients", "Sample of 500 patients", "Sample of 5,000 patients", "Sample of 5

plots <- create_survival_plots(data, sample_sizes, titles)

gridExtra::grid.arrange(grobs = plots, ncol = 2, nrow = 2, top = "Survival Curve over Time for Censored
```



As expected, the fewer subjects included in each sample, the higher the starting point of the curve at time 0. This observation suggests that with smaller sample sizes, the initial survival probabilities appear to be greater. This is caused by the variability in smaller sample sizes, where single data points may have a greater influence on the overall trend.

Cox Model

The **Cox Proportional Hazard Model** is very useful to analyze the impact of several predictor variables on the event of interest.

```
(cox_model <- coxph(Surv(duration, hospital_outcome)~sex,
                    data = survival_primary_cohort))
```

```
## Call:
## coxph(formula = Surv(duration, hospital_outcome) ~ sex, data = survival_primary_cohort)
##
##          coef exp(coef) se(coef)      z        p
## sex1 -0.13951   0.86978  0.02238 -6.233 4.58e-10
##
## Likelihood ratio test=39.04  on 1 df, p=4.157e-10
## n= 110204, number of events= 8105
```

Let's investigate the variable `sex`, which is represented with 0 for **male** and 1 for **female**.

The coefficient for the variable **sex** in the **Cox Proportional Hazards Model** was estimated to be -0.14 ($p < 0.001$). This suggests that for every one-unit increase in the **sex** variable, the **log hazard ratio** decreases by 0.14 units.

Additionally, the exponential of the estimate 0.87 indicates that being female is associated with a lower hazard of experiencing the event compared to being male, given that it is less than 1. Moreover, the statistic of-6.23 indicates an extremely strong association between **sex** and outcome. A larger absolute z-value corresponds to a smaller p-value, underscoring stronger evidence against the null hypothesis ($H_0 = 0$) and a greater significance of the predictor variable in the model.

```
(model<- survdiff(Surv(duration, hospital_outcome)~sex,
                  data = survival_primary_cohort))
```

```
## Call:
## survdiff(formula = Surv(duration, hospital_outcome) ~ sex, data = survival_primary_cohort)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=0 57973     4548     4270      18.2      38.7
## sex=1 52231     3557     3835      20.2      38.7
##
##  Chisq= 38.7  on 1 degrees of freedom, p= 5e-10
```

```
tidy_model = tidy(model)
```

The **log-rank test** let us compare the survival distribution between the two groups of the variable **sex**. By testing the null hypothesis, we aim to determine if there is any difference in the survival distributions between these groups.

A total of $N = 57973$ males and $N = 52231$ females were observed in our analysis. The number of observed events (death) among males is *Observed* 4548, while among female is *Observed* 3557. Under the assumption of no difference in survival distributions, the expected number of events would be *Expected* 4270 among males and *Expected* 3835 among females.

This analysis highlights potential disparities in survival outcomes between males and females, based on observed and expected event counts.

```
x <- str(survival_primary_cohort)
```

```
## 'data.frame':    110204 obs. of  7 variables:
##  $ age             : num  21 20 21 77 72 83 74 74 69 53 ...
##  $ sex             : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 2 1 2 ...
##  $ count_episode   : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ hospital_outcome: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ start_date      : Date, format: "2011-02-19" "2011-04-08" ...
##  $ end_date        : Date, format: "2011-03-13" "2011-04-22" ...
##  $ duration        : num  22 14 10 1 4 13 24 20 3 17 ...
```

```
(full_model <- coxph(Surv(duration, hospital_outcome)~sex + count_episode + age,
                     data = survival_primary_cohort) )
```

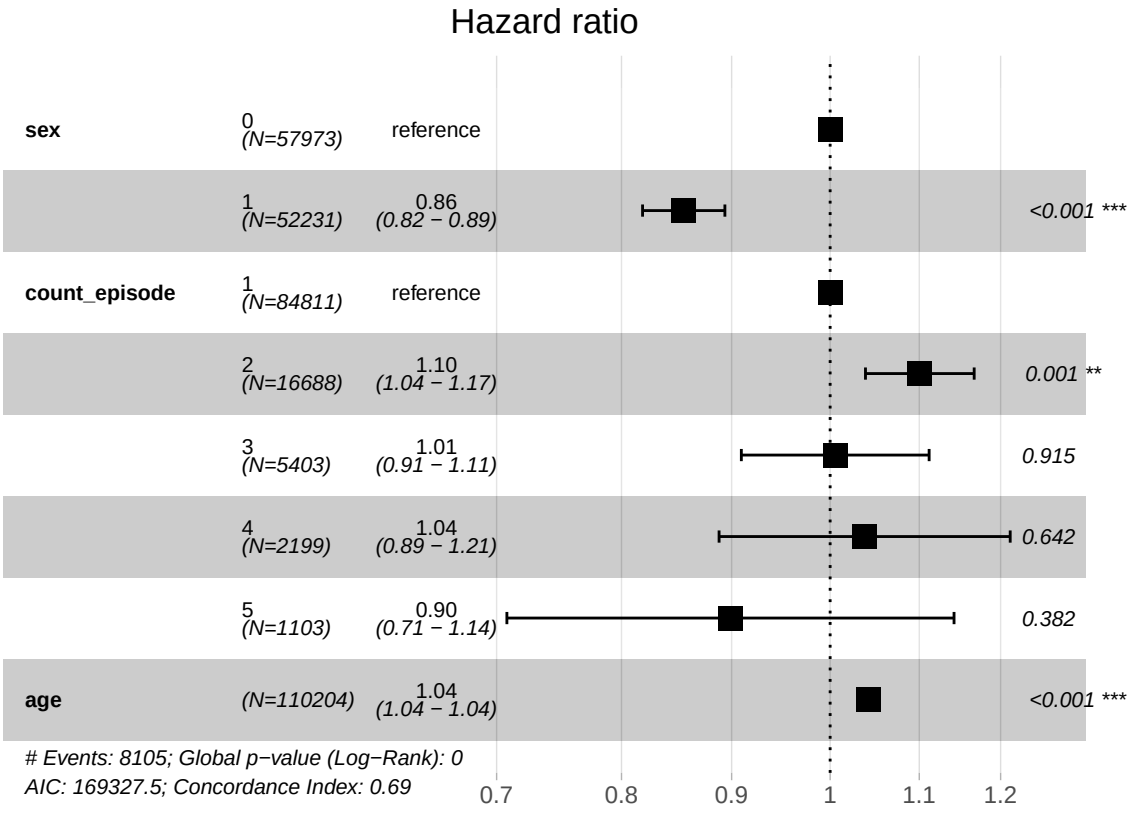
```
## Call:
## coxph(formula = Surv(duration, hospital_outcome) ~ sex + count_episode +
```

```
##     age, data = survival_primary_cohort)
##
##                       coef  exp(coef)   se(coef)      z       p
## sex1           -0.1565841  0.8550596  0.0224559 -6.973 3.1e-12
## count_episode2  0.0959012  1.1006503  0.0296515  3.234 0.00122
## count_episode3  0.0054379  1.0054528  0.0512300  0.106 0.91547
## count_episode4  0.0368880  1.0375768  0.0794200  0.464 0.64231
## count_episode5 -0.1066035  0.8988820  0.1219863 -0.874 0.38217
## age             0.0413704  1.0422380  0.0007874 52.540 < 2e-16
##
## Likelihood ratio test=4075  on 6 df, p=< 2.2e-16
## n= 110204, number of events= 8105
```

Here a **Cox proportional hazard model** is conducted for all variables in the dataset.

- `sex1`: the results are consistent with the previous Cox model. However, there's a different $coef$ value of -0.16. This suggests that for every one-unit increase in the `sex` variable, the **log hazard ratio** decreases by -0.16 units, which aligns with existing scientific evidence indicating that females tend to survive longer than males.
- `count_episode`: the variable includes 5 levels (1,2,3,4,5), each one indicating the episode's number of *sepsis*. All the count episodes have positive coefficients - except for `count_episode5`-. This indicates that for each unit increase in these count episode variables, the **log hazard ratio** increases. Nevertheless, `count_episode3`, `count_episode4` and `count_episode5` have large p-values, suggesting that these variables aren't significance for the model. So, just `count_episode2` results significant for the model.
- `age`: the $coef = 0.04$ suggest a **positive relationship** with the **log hazard ratio**. Specifically, the *log hazard ratio* increases by about 0.0414 for every one-unit increase in the age variable, which is expected given the higher likelihood of death among older individuals. In addition, its p-value ($p < 0.0001$) demonstrates strong significance for the analysis, as evidenced by its $z = 0.04$.

```
ggforest(full_model, data = survival_primary_cohort)
```



Each vertical lines represents the point estimates (*hazard ratio*) of the predictor variables in the model, determining the magnitude of the effect. The horizontal line represents the confidence interval around the point estimate.

In the figure, a dashed line at 1 represents the null value, helping in assessing the statistical significance of the effect size. If the confidence interval crosses this line, the effect is **not statistically significant**.

- `sex`: note male as a reference level. Females exhibit a **hazard ratio** of 0.85, with a **95% confidence interval** between 0.81 and 0.89. This suggests that patients receiving the treatment have a 0.15 **lower hazard** of the event compared to those who didn't receive it, with a 95% confidence that the **true hazard ratio** lies between 0.81 and 0.89.
- `count_episode`: note `count_episode1` as a reference level. The `count_episode2` has a **hazard ratio** of 1.11, with a 95% confidence interval between 1.04 and 1.17. Instead, the `count_episode3` and `count_episode4` have hazard ratios of 1.04, with confidence intervals ranging from 0.94 to 1.15 and 0.89 to 1.21, respectively. Additionally, `count_episode5` has a hazard ratio of 0.92, with a confidence interval between 0.72 and 1.16. Moreover, the `count_episode5` has a **hazard ratio** of 0.92, with a confidence interval between 0.72 and 1.16. So, briefly, `count_episode3` and `count_episode4` have a 0.04 **higher hazard** to the event compared to those that didn't receive the treatment, while `count_episode5` has a 0.08 **lower hazard** to the event. However, looking at the confidence intervals, the `count_episode4` and `count_episode5` appear more risky. In addition, looking at the p-values, we can't rely on `count_episode3`, `count_episode4` and `count_episode5` because they aren't statistically significant.
- `age`: it has an **hazard ratio** of 1.04, indicating that patients receiving treatment have a 0.04 **higher hazard** of mortality.

Finally, the **Concordance Index** is a measure of the discriminatory power of a survival model. Its value is 0.69 indicating a quite good predictive performance.

```
mv_fit <- coxph(Surv(duration, hospital_outcome) ~ sex + age,
                data = survival_primary_cohort)
cz <- cox.zph(mv_fit)
print(cz)
```

```
##        chisq df     p
## sex     4.71  1 0.030
## age     0.03  1 0.862
## GLOBAL  4.72  2 0.095
```

The evaluation of whether the assumptions underlying the *Cox Model*, particularly the proportional hazards assumption, hold for the predictor variables `sex` and `age` included in our model. This process helps ensure the validity of our *cox regression analysis*.

Looking at the p-values of `sex` and `age` (0.03 and 0.86), it is observed that they are greater than 0.05, suggesting no violation of the proportional hazard assumptions for these variables.
Additionally, for the `GLOBAL` assessment, the proportional hazard assumptions for the entire model are evaluated.

Looking at the p-value, which equals 0.09, it can be said that there is no violation of the proportional hazard assumptions for the entire model.

In the end, it has been decided to leave out `count_episode` variable as it is not statistically significant, in order to not fall into failing results.

Conclusion

In conclusion, the ***Cox Model***, presented similar results to the ***logistic regression***.

As a matter of fact, the variables `sex`, `age` and `count_episode2` are **strongly statistically significant**. This indicates their reliability in predicting whether an individual survives *sepsis*. Unfortunately, it cannot be said the same for `count_episode3`, `count_episode4` and `count_episode5`.

In addition, *females* have more probability to survive than *males*. People with two or more episodes have more probability to die, where those who are associated to variable `count_episode4` are exposed to a greater risk. The variable `age` has a positive relationship with death (`hospital_outcome` $== 1$), implying that older people have more chance to die.