Overall survival

University of Waterloo Health Data Science Lab

March 2018

```
# 4-year survival analysis

set.seed(4375) # "HDSL"

admissions_df <- read.csv("../../../MIMIC/sentiment_data/patient_df.csv")
sapsii_df <- read.csv("../../../MIMIC/sentiment_data/sapsii_df.csv")
notes_df <- read_csv("../../../MIMIC/sentiment_data/notes_df_sntmnt.csv")
```

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

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   row_id = col_double(),
##   subject_id = col_double(),
##   hadm_id = col_double(),
##   chartdate = col_datetime(format = ""),
##   charttime = col_datetime(format = ""),
##   storetime = col_datetime(format = ""),
##   category = col_character(),
##   description = col_character(),
##   cgid = col_double(),
##   iserror = col_logical(),
##   text = col_character(),
##   polarity = col_double(),
##   subjectivity = col_double()
## )
```

```
## Warning: 244 parsing failures.
##    row     col           expected actual                                                file
##  59631 iserror 1/0/T/F/TRUE/FALSE    1.0 '../../../MIMIC/sentiment_data/notes_df_sntmnt.csv'
## 210653 iserror 1/0/T/F/TRUE/FALSE    1.0 '../../../MIMIC/sentiment_data/notes_df_sntmnt.csv'
## 271374 iserror 1/0/T/F/TRUE/FALSE    1.0 '../../../MIMIC/sentiment_data/notes_df_sntmnt.csv'
## 271376 iserror 1/0/T/F/TRUE/FALSE    1.0 '../../../MIMIC/sentiment_data/notes_df_sntmnt.csv'
## 271377 iserror 1/0/T/F/TRUE/FALSE    1.0 '../../../MIMIC/sentiment_data/notes_df_sntmnt.csv'
## ...... ....... .................. ...... ....................................................
## See problems(...) for more details.
```

```
notes_df$text <- NULL

notes_df <- notes_df %>% left_join(admissions_df, by="hadm_id")

# Get the difference between the time of death and the time that the note was taken.
# We don't want to look at notes close to death (~12hrs)

deathtime.POSIX <- as.POSIXlt(strptime(notes_df$dod, "%Y-%m-%d %H:%M:%S"))
charttime.POSIX <- as.POSIXlt(strptime(notes_df$charttime, "%Y-%m-%d %H:%M:%S"))
notes_df$death_chart_diffhrs <- difftime(deathtime.POSIX,
                                         charttime.POSIX,
```

```
                                        units="hours")

setDT(notes_df)

notes_df <- notes_df[is.na(death_chart_diffhrs) |
                       death_chart_diffhrs > 12]

notes_df <- notes_df[is.na(iserror)]

analysis_df <- admissions_df

setDT(analysis_df)
analysis_df[dod != "", surv_time := difftime(as.Date(dod), as.Date(admittime), units="days")]
# Remove patients with data entry error: death time recorded before admission.
analysis_df <- analysis_df[is.na(surv_time) | surv_time >= 0]

analysis_df[dod != "", surv_event := 1]

# set discharge time and alive event for patients that have not died.
analysis_df[dod == "" & dbsource == "metavision",
            surv_time := 90]
analysis_df[dod == "" & dbsource == "carevue",
            surv_time := 365.242 * 4]
analysis_df[dod == "", surv_event := 0]

# Ensure that we are looking at a 4-year survival model
# (with censoring at 90 days for metavision)
analysis_df[surv_time > 365.242 * 4, surv_event := 0]
analysis_df[surv_time > 365.242 * 4, surv_time := 365.242 * 4]


# A few hundred rows with unknown dbsource
analysis_df <- analysis_df[!is.na(dbsource)]

head(analysis_df[, c("dod", "surv_time", "surv_event", "dbsource"), with=FALSE],
     n=15L)
```

```
##                    dod     surv_time surv_event dbsource
##  1: 2102-06-14 00:00:00  237.000 days          1  carevue
##  2:                     1460.968 days          0  carevue
##  3: 2104-08-20 00:00:00   13.000 days          1  carevue
##  4:                     1460.968 days          0  carevue
##  5:                     1460.968 days          0  carevue
##  6: 2135-02-08 00:00:00  150.000 days          1  carevue
##  7:                     1460.968 days          0  carevue
##  8:                     1460.968 days          0  carevue
##  9:                     1460.968 days          0  carevue
## 10:                     1460.968 days          0  carevue
## 11: 2133-09-30 00:00:00  273.000 days          1  carevue
## 12: 2199-11-22 00:00:00 1460.968 days          0  carevue
## 13:                     1460.968 days          0  carevue
## 14: 2104-01-08 00:00:00    6.000 days          1  carevue
## 15: 2198-09-06 00:00:00  296.000 days          1  carevue
```

```r
# have to make surv_time numeric for the Surv() function in the survival package
analysis_df[, surv_time := as.numeric(surv_time)]

notes_df <- notes_df %>%
  group_by(hadm_id) %>%
  summarize(mean_polarity = mean(polarity),
            mean_subjectivity = mean(subjectivity))

# joining notes by admission id.
setDT(notes_df)
setDT(analysis_df)
analysis_df <- analysis_df %>% left_join(notes_df, by="hadm_id")


setDT(sapsii_df)
# joining SAPS II scores by ICU stay ID.
analysis_df <- analysis_df %>% left_join(sapsii_df, by="icustay_id")

analysis_df_w_pol <- analysis_df %>%
  filter(!is.na(mean_polarity))
```

Mean polarity and subjectivity quartiles

```r
output.folder <- "../../../MIMIC/"

gg_color_hue <- function(n) {
  hues = seq(15, 375, length = n + 1)
  hcl(h = hues, l = 65, c = 100)[1:n]
}

# 4 Groups
# Polarity
analysis_df_w_pol$Quartile <- factor(ntile(analysis_df_w_pol$mean_polarity, 4))
quartiles <- levels(analysis_df_w_pol$Quartile)
quartiles
```

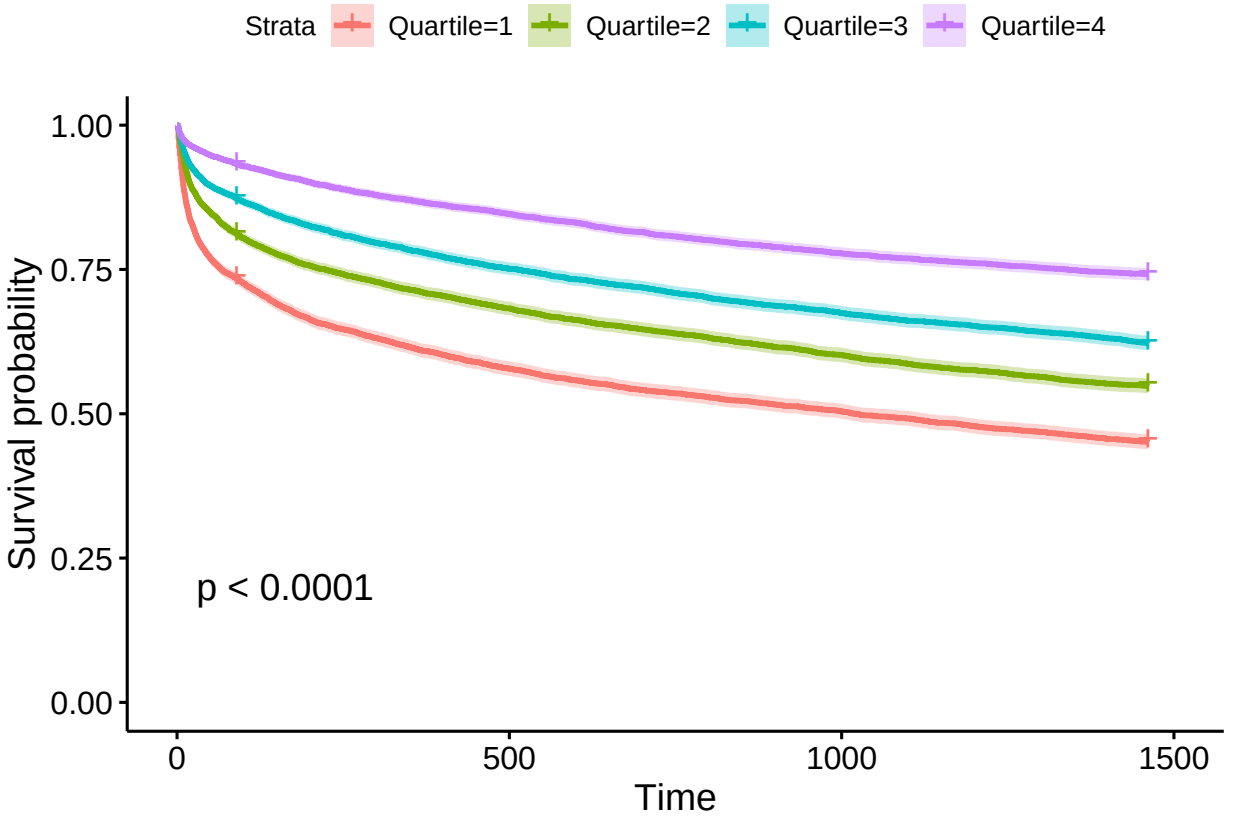
```
## [1] "1" "2" "3" "4"
```

```r
# Map quartiles to specific colors
quartile.colors <- gg_color_hue(4)
names(quartile.colors) <- quartiles


surv.obj <- survfit(Surv(time = surv_time, event = surv_event) ~ Quartile, data=analysis_df_w_pol)

pol.ggsurv.plot <- GGally::ggsurv(surv.obj, CI=T, cens.col = 2, lty.ci = 3) +
  xlab("Time (days)") +
  theme(text=element_text(size=8), axis.line.x = element_blank(),
        axis.title.x = element_blank(), axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  scale_y_continuous(breaks = seq(0.4, 1, by = 0.2), limits = c(0.4, 1)) +
  guides(linetype = FALSE) +
  scale_colour_manual(name="Polarity      \nQuartile",
```

```
                        values = quartile.colors)
```

```
## Scale for 'colour' is already present. Adding another scale for
## 'colour', which will replace the existing scale.
```

```
# Log-rank test
ggsurvplot(
   surv.obj,                     # survfit object with calculated statistics.
   pval = TRUE,              # show p-value of log-rank test.
   conf.int = TRUE,          # show confidence intervals for
   xlim = c(0,1500)         # present narrower X axis
)
```




```
# Subjectivity
analysis_df_w_pol$Quartile <- factor(ntile(analysis_df_w_pol$mean_subjectivity, 4))
levels(analysis_df_w_pol$Quartile)
```

```
## [1] "1" "2" "3" "4"
```

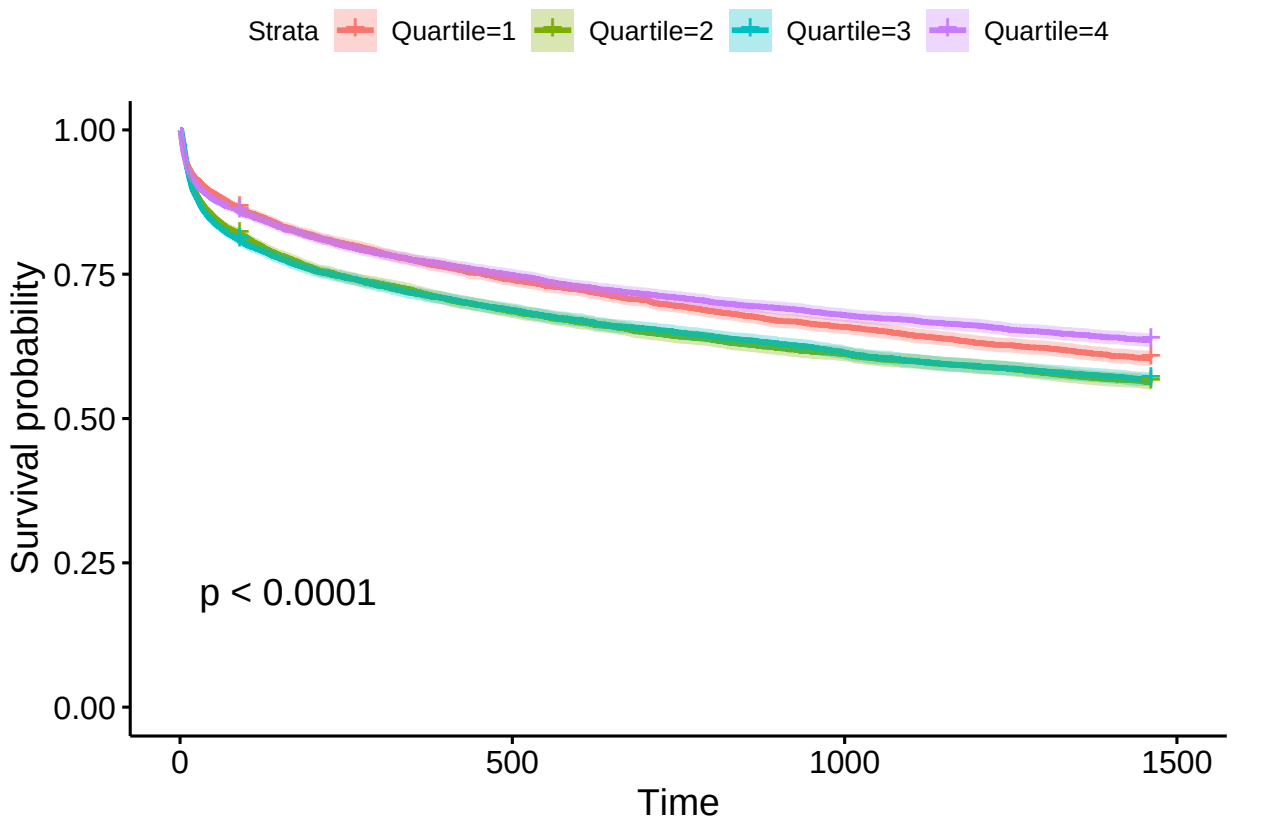
```
surv.obj <- survfit(Surv(time = surv_time, event = surv_event) ~ Quartile, data=analysis_df_w_pol)
```

```
sub.ggsurv.plot <- GGally::ggsurv(surv.obj, CI=T, cens.col = 2, lty.ci = 3) +
  xlab("Time (days)") +
  theme(text=element_text(size=8)) +
  scale_y_continuous(breaks = seq(0.4, 1, by = 0.2), limits = c(0.4, 1)) +
  guides(linetype = FALSE) +
  scale_colour_manual(name="Subjectivity\nQuartile",
```

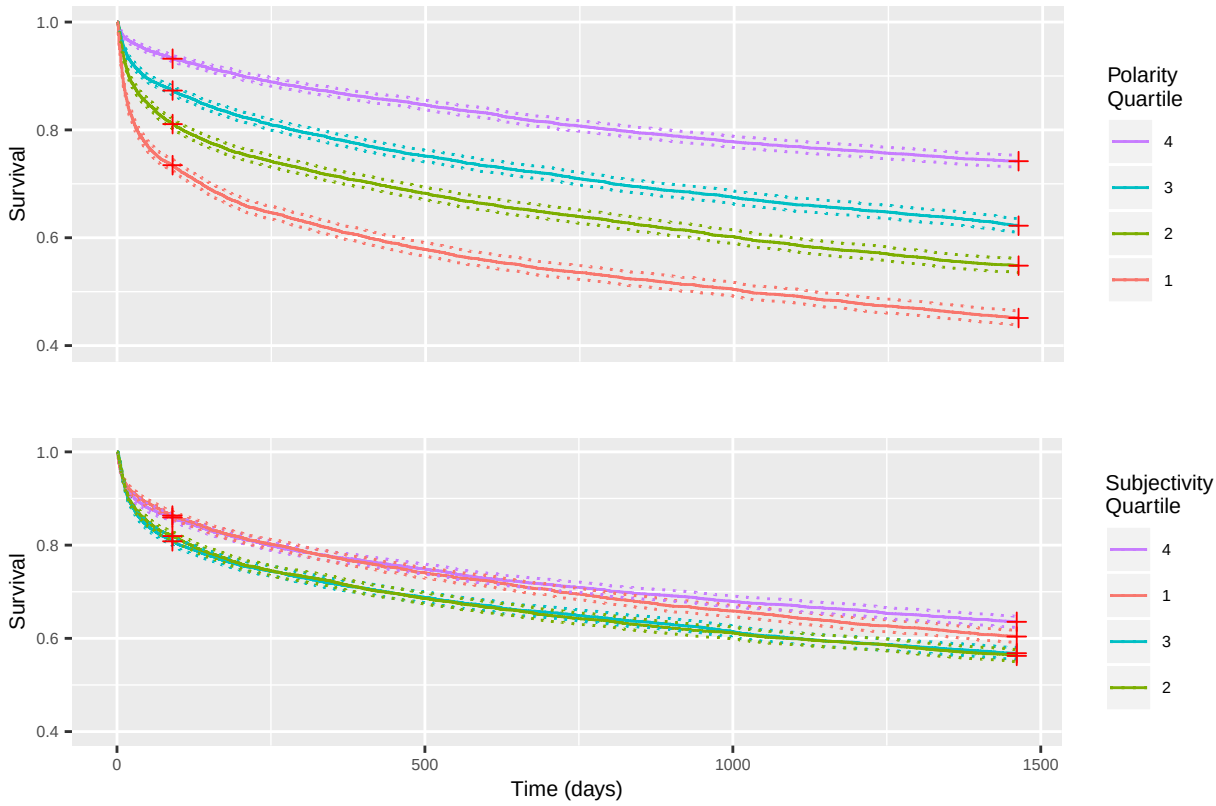
```
                      values = quartile.colors)
```

```
## Scale for 'colour' is already present. Adding another scale for
## 'colour', which will replace the existing scale.
```

```
# Log-rank test
ggsurvplot(
   surv.obj,                     # survfit object with calculated statistics.
   #risk.table = TRUE,        # show risk table.
   pval = TRUE,              # show p-value of log-rank test.
   conf.int = TRUE,          # show confidence intervals for
                             # point estimaes of survival curves.
   xlim = c(0,1500)         # present narrower X axis, but nsot affect
)
```




```
ggsurv.plot <- grid.arrange(pol.ggsurv.plot, sub.ggsurv.plot)
```




```
ggsurv.plot
```

```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[layout]
```

```
ggsave(file.path(output.folder, "KM_curves.tiff"), plot = ggsurv.plot,
       width=15, height=10, units="cm", dpi=300)
```