An overview of time-to-event analysis for dental researchers

Tabitha K. Peter

October 11, 2022

Welcome to the tutorial "An overview of time-to-event analysis for dental researchers." The objective of this tutorial is to explain the foundational concepts of time-to-event analysis for an audience of dental researchers.

These foundational concepts are illustrated using an extended example from the dental research literature. See this link for the full text of this published work.

Set up

First, load the data set

```
source("R/R.R")
cmr <- read_csv("data/cmr.csv")
```

Once the data is loaded, take a look to see the contents of the data set. I notice that there are a lot of options for 'clinic' – this tells me I will need to combine some categories before analyzing the data.

```
# First look at the data
# See what changes we need to make
dplyr::glimpse(cmr) # notice: lots of options for 'clinic', and some categories are labeled as numbers!
## Rows: 1,002
## Columns: 23
## $ RecordID        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ No_Cases        <dbl> 2, NA, 1, 1, 1, 1, 2, NA, 4, NA, NA, 2, NA, 2, NA, 1, ~
## $ Age             <dbl> 67, NA, 67, 73, 81, 64, 76, NA, 62, NA, NA, 65, NA, 87~
## $ Gender          <chr> "F", NA, "F", "M", "F", "F", "F", NA, "F", NA, NA, "F"~
## $ CRA             <chr> "Not High Risk", NA, "High Risk", "Not High Risk", "Hi~
## $ Tooth_Number    <dbl> 13, 31, 2, 3, 13, 30, 11, 3, 3, 4, 13, 6, 9, 11, 4, 12~
## $ Tooth_Type      <chr> "P", "P", "P", "P", "P", "P", "A", "P", "P", "P", "P",~
## $ Jaw             <chr> "Mx", "Md", "Mx", "Mx", "Mx", "Md", "Mx", "Mx", "Mx", ~
## $ Repair_Material <chr> "Amal", "Amal", "Amal", "Amal", "Amal", "Amal", "RMGI"~
## $ Date_Repair     <date> 2017-12-04, 2018-05-14, 2016-07-26, 2017-04-06, 2015-~
## $ Surfaces        <chr> "Other", "Other", "Other", "L", "L", "B", "L", "B", "O~
## $ No_Surfaces     <dbl> 1, 1, 1, 2, 2, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, ~
## $ RCT             <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, ~
## $ Crown_Type      <chr> "PFM", "PFM", "PFM", "PFM", "PFM", "Other", "PFM", "PF~
## $ Clinic          <chr> "FGP", "FGP", "FAMD", "FAMD", "GMU", "FGP", "FAMD", "F~
## $ Provider_Type   <chr> "faculty", "faculty", "faculty", "faculty", "student",~
## $ Failure         <chr> "0", "0", "0", "0", "0", "0", NA, NA, NA, NA, "0", "0"~
## $ Last_exam       <date> 2019-06-24, 2019-06-24, 2019-05-15, 2019-02-22, 2018-~
## $ Failure_Date    <date> NA, NA, NA, NA, NA, NA, 2018-06-12, 2019-04-09, 2018-~
## $ Failure_Reason  <chr> NA, NA, NA, NA, NA, NA, "TE_or_Redo", "Caries_or_Repai~
## $ Status          <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ End_Date        <date> 2019-06-24, 2019-06-24, 2019-05-15, 2019-02-22, 2018-~
```

```
## $ Time            <dbl> 1.55236140, 1.11156742, 2.80082136, 1.88090349, 2.7022~
table(cmr$Clinic, useNA = "always") # hmmm... need to combine some categories
##
##   ADMS   CCST CODHDG   FAMD  FDDAU FEECLS    FGP    GMU  HDGEN HDPROS LTDCAR
##      5      6      3    440     80      1    181     27      1      2      1
##   OPER   PROS   SPEC   <NA>
##    112     65     78      0
```

I make some changes to the formatting of the data, so that I am ready to analyze the data. The goal of this analysis is to describe what factors influence the length of time that a crown margin repair lasts.

```
# Make some formatting changes
cmr <- cmr %>%
  # collapse categories for clinic (we will need this for analysis)
  mutate(Clinic = if_else(Clinic %in% c("ADMS", "CCST", "CODHDG", "FEECLS",
                                        "GMU", "HDGEN", "HDPROS", "LTDCAR"),
                          "Other",
                          Clinic),
         # tell the computer that 'No_surface' numbers represent *categories*
         No_Surfaces = as.character(No_Surfaces),
         # tell the computer that RCT is a yes/no outcome
         RCT = case_when(
           RCT == 0 ~  "No RCT",
           RCT == 1 ~ "RCT"),
         # make the IDs have readable names
         RecordID = paste0("Patient ", RecordID),
         # Represent the outcome (reintervention) in both numbers and words
         Event = Status,
         Status = case_when(
           Event == 0 ~ "Censored",
           Event == 1 ~ "Event"))
```

Descriptive analysis

Once I have verified that the data set is ready for analysis, I create a "Table 1" that summarizes each variable in the data set.

```
# Create Table 1
table1 <- tableby(~ .,
                  data = cmr %>%
                    dplyr::select(-c(RecordID, Event, No_Cases, Tooth_Number,
                                     Date_Repair, End_Date, Failure_Date)),
                  control = tableby.control(
                    test = FALSE
                  )
                  )

# Print out Table 1 in a readable format
summary(table1) %>%
  kable(digits = 2,
        format = "pipe",
        caption = "Table 1: Description of Data")
```

Table 1: Table 1: Description of Data

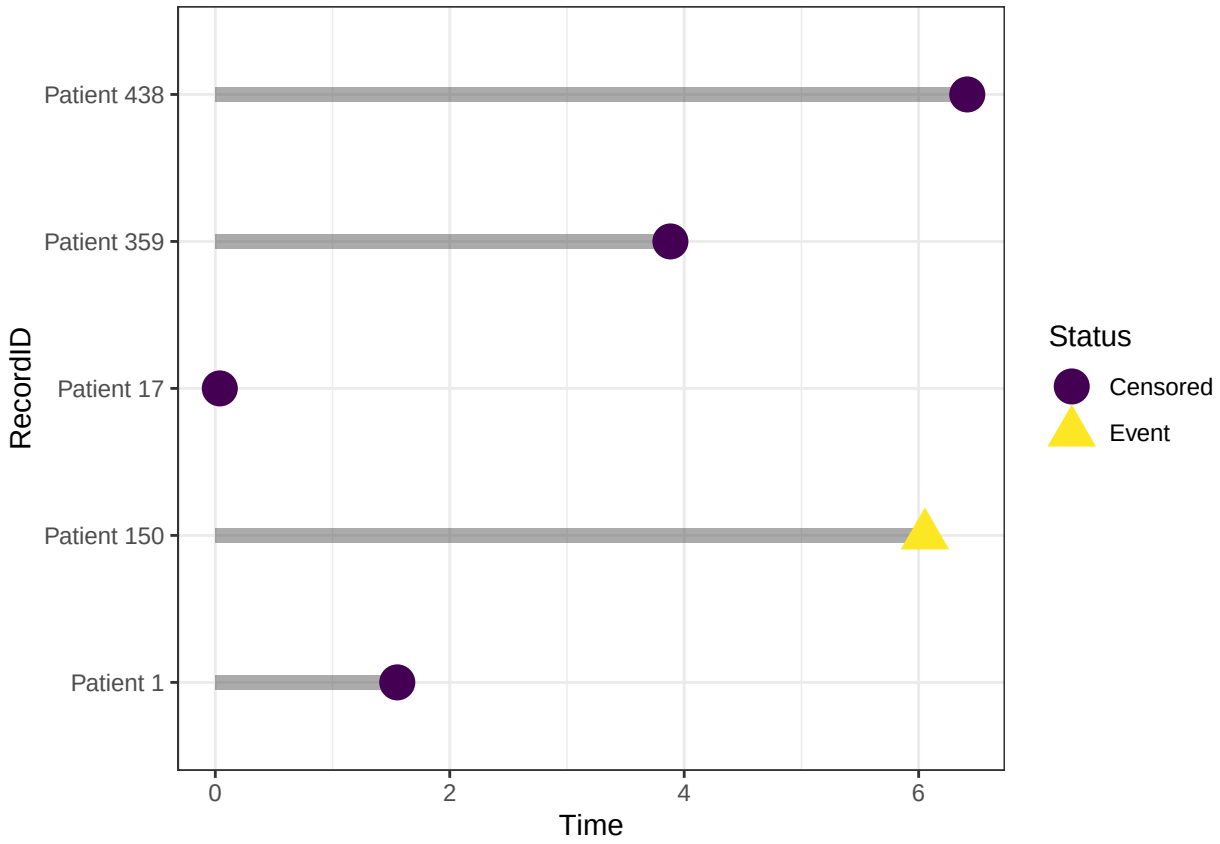
	Overall (N=1002)
Age	
N-Miss	445
Mean (SD)	74.530 (12.106)
Range	32.000 - 104.000
Gender	
N-Miss	446
F	294 (52.9%)
M	262 (47.1%)
CRA	
N-Miss	613
High Risk	159 (40.9%)
Not High Risk	230 (59.1%)
Tooth_Type	
N-Miss	4
A	190 (19.0%)
P	808 (81.0%)
Jaw	
Md	474 (47.3%)
Mx	528 (52.7%)
Repair_Material	
N-Miss	1
Amal	379 (37.9%)
GI	114 (11.4%)
RBC	92 (9.2%)
RMGI	416 (41.6%)
Surfaces	
B	403 (40.2%)
L	214 (21.4%)
Other	385 (38.4%)
No_Surfaces	
1	724 (72.3%)
2	278 (27.7%)
RCT	
N-Miss	6
No RCT	648 (65.1%)
RCT	348 (34.9%)
Crown_Type	
N-Miss	3
C	55 (5.5%)
Other	384 (38.4%)
PFM	560 (56.1%)
Clinic	
FAMD	440 (43.9%)
FDDAU	80 (8.0%)
FGP	181 (18.1%)
OPER	112 (11.2%)
Other	46 (4.6%)
PROS	65 (6.5%)
SPEC	78 (7.8%)
Provider_Type	

	Overall (N=1002)
faculty	356 (35.5%)
student	646 (64.5%)
Failure	
N-Miss	326
0	566 (83.7%)
1	1 (0.1%)
no follow up	109 (16.1%)
Last_exam	
N-Miss	435
Median	2019-01-18
Range	2013-02-19 - 2019-09-23
Failure_Reason	
N-Miss	675
Caries_or_Repair	145 (44.3%)
TE_or_Redo	182 (55.7%)
Status	
Censored	673 (67.2%)
Event	329 (32.8%)
Time	
Mean (SD)	2.448 (2.179)
Range	0.000 - 12.118

Timeline chart

To illustrate the stories of specific patients, I draw a timeline chart (some people call this type of graph a "swimlane diagram").

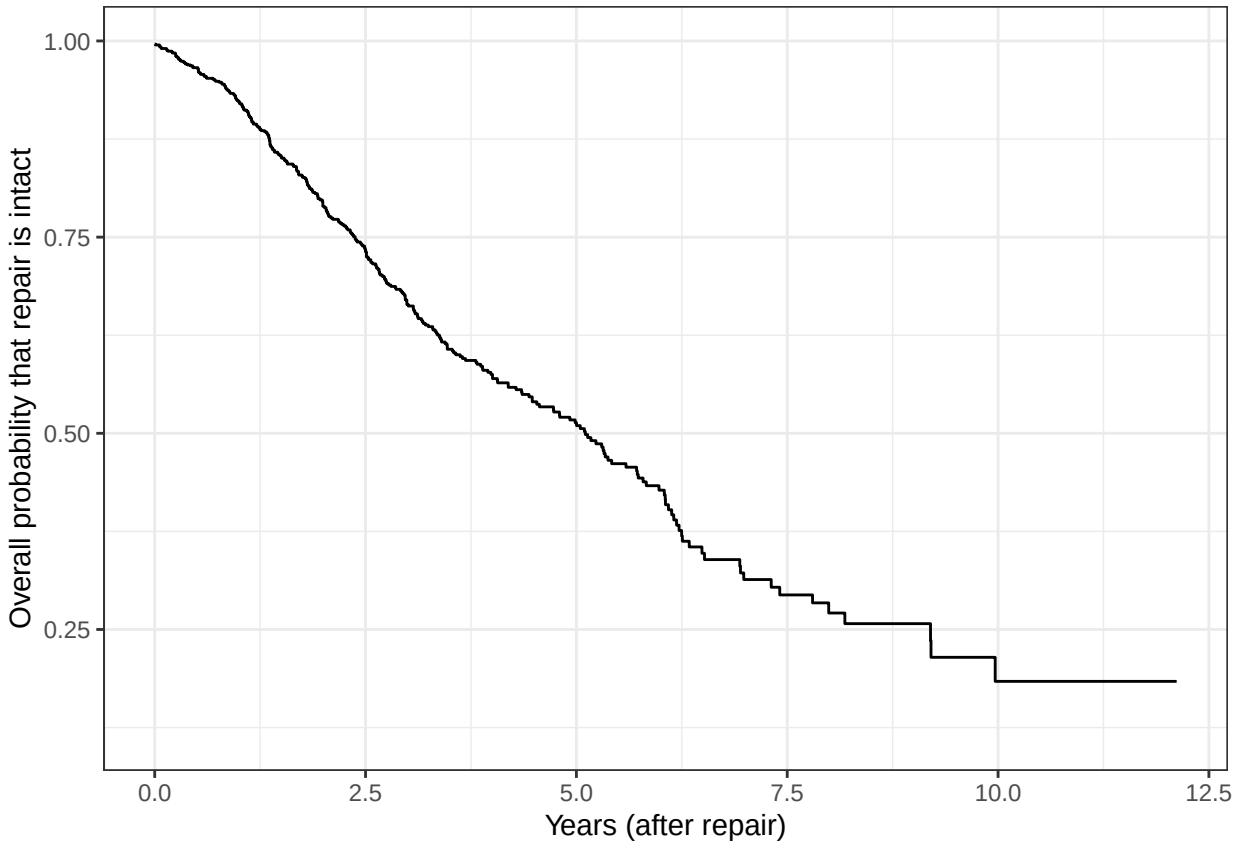
```
# Create a timeline chart
set.seed(52246) # this means the same patients are chosen each time I run this code
example_patients <- sample(1:nrow(cmr), size = 5)

timelines <- ggplot(cmr[example_patients, ], aes(x = RecordID, y = Time)) +
  geom_col(position = position_dodge(), width = 0.1, alpha = 0.5) +
  geom_point(data = cmr[example_patients, ],
             aes(RecordID, Time, color = Status, shape = Status),
             size = 6) +
  coord_flip() +
  scale_color_viridis(discrete = TRUE) +
  theme_bw()


(timelines)
```



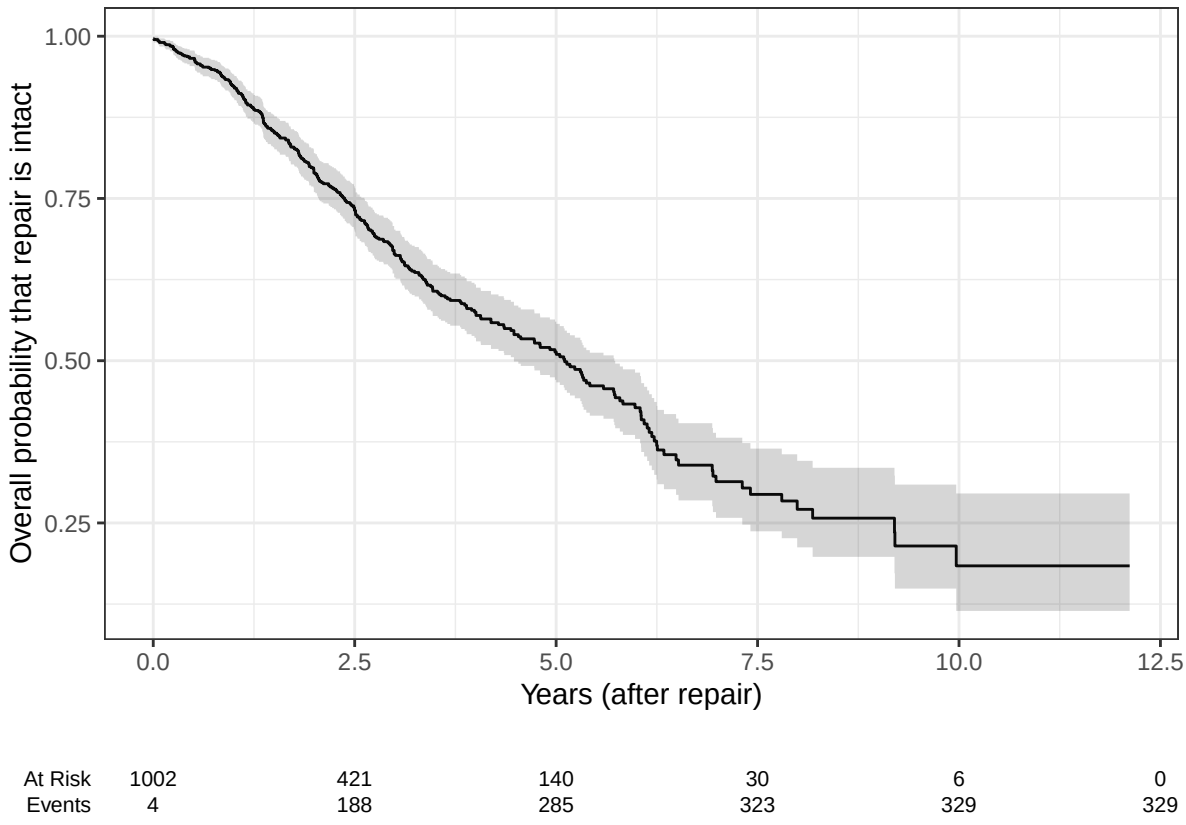
Kaplan-Meier plot

To get an idea of the overall trajectory of the time that the crown margin repairs are lasting, a create a Kaplan-Meier (KM) plot. As time goes on, there are fewer repairs upon which to draw estimates, so our estimates become more uncertain at later years. To visualize this uncertainty, we can add confidence intervals to the plots. A risk table provides details to supplement the general pattern illustrated in the KM plot.

```
# Kaplan Meier (KM) curve
km1 <- survfit2(Surv(time = Time, event = Event) ~ 1, data = cmr) %>%
  ggsurvfit() +
  labs(
    x = "Years (after repair)",
    y = "Overall probability that repair is intact"
  )
km1
```




```
# KM curve with confidence intervals and risk table
km2 <- survfit2(Surv(time = Time, event = Event) ~ 1, data = cmr) %>%
  ggsurvfit() +
  labs(
    x = "Years (after repair)",
    y = "Overall probability that repair is intact"
  ) +
  add_confidence_interval() +
  add_risktable()

km2
```



Analysis

Median survival

Remember: The median survival time is **not** just the median of all the survival time values. When we are talking about median survival, we have to account for the fact that some repairs are censored – we do not know everything about each repair in our study! The Kaplan-Meier method for calculating median survival takes censoring into account. We can do this calculation in `R` with the `survfit` function from the `survival` package.

```
# calculate median survival with the KM method
s1 <- survfit(Surv(Time, Event) ~ 1, data = cmr)

# create a table with the median survival time and its accompanying 95% confidence interval
s1_median <- summary(s1)$table

s1_median[c("median", "0.95LCL", "0.95UCL")] %>%
  t() %>%
 kable(digits = 3,
       col.names = c(
         "Median survival time",
         "95% CI (lower)",
         "95% CI (upper)"
       ))
```

Median survival time	95% CI (lower)	95% CI (upper)
5.106	4.476	5.722

Nth year survival

We are also often interested to estiamte the survival probability of a repair making it ___ number of years. Below, I use the same `survfit` function to estimate 1, 3, and 5 year survival with the KM method.

NB: to make the following tables have a readable format, I wrote my own `R` function `nth_yr_surv()` - the code for this function is the `data\R.R` file.

```
# one year survival ---------------------------------------------------
one_yr_surv <- summary(survfit(Surv(Time, Event) ~ 1, data = cmr), times = 1)

nth_year_survival(one_yr_surv) %>% kable(digits = 3)
```

Time	Number at risk	Number of events	Probability of survival	Standard Error	95% CI (lower)	95% CI (upper)
1	690	64	0.923	0.009	0.904	0.941

```
# 1 and 3 year survival -----------------------------------------------
three_yr_surv <- summary(survfit(Surv(Time, Event) ~ 1, data = cmr), times = c(1, 3))

nth_year_survival(three_yr_surv) %>% kable(digits = 3)
```

Time	Number at risk	Number of events	Probability of survival	Standard Error	95% CI (lower)	95% CI (upper)
1	690	64	0.923	0.009	0.904	0.941
3	342	161	0.664	0.019	0.628	0.702

```
# 1, 3, and 5 year survival -------------------------------------------
five_yr_surv <- summary(survfit(Surv(Time, Event) ~ 1, data = cmr), times = c(1, 3, 5))

nth_year_survival(five_yr_surv) %>% kable(digits = 3)
```

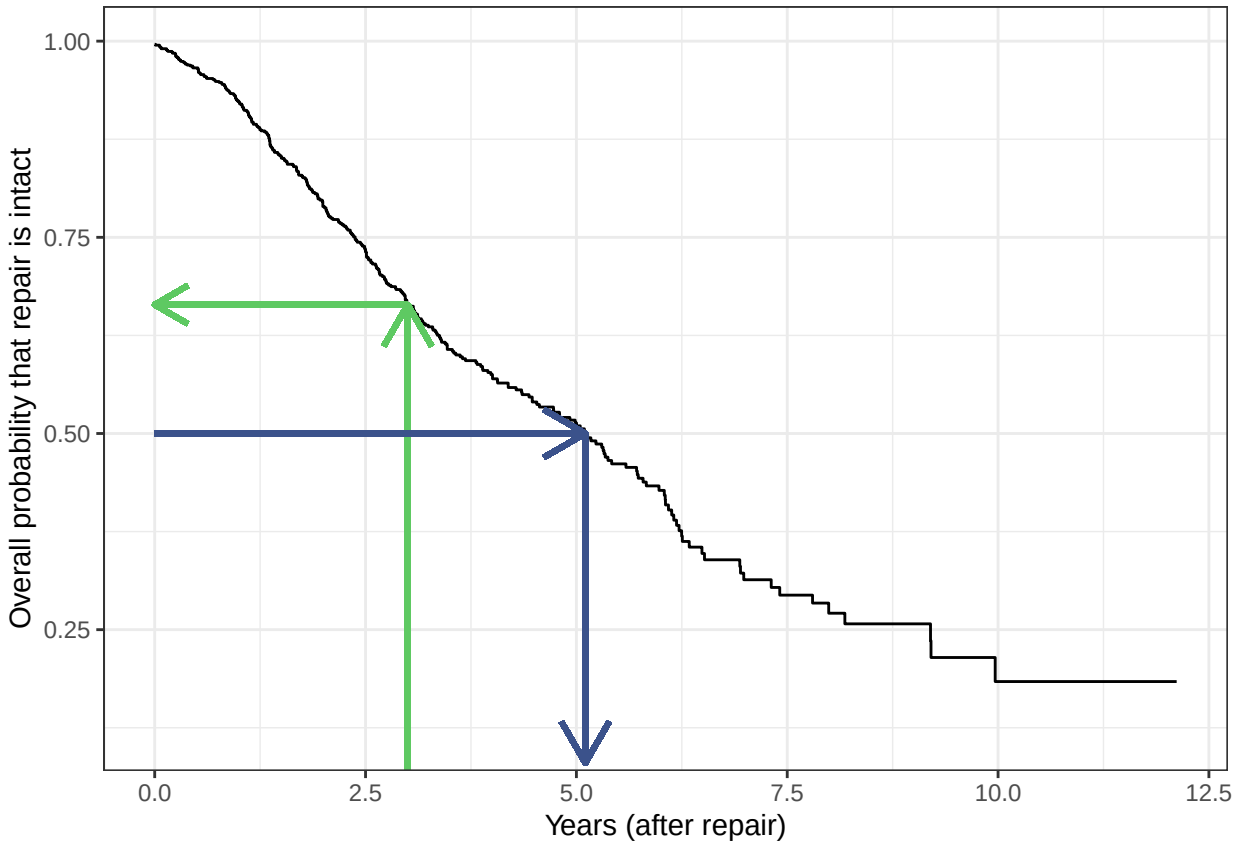
Time	Number at risk	Number of events	Probability of survival	Standard Error	95% CI (lower)	95% CI (upper)
1	690	64	0.923	0.009	0.904	0.941
3	342	161	0.664	0.019	0.628	0.702
5	140	60	0.513	0.023	0.471	0.560

Plot 3 year survival probability and median survival time

To show the difference between estimating ___ year survival and estimating the median survival time, I can plot these both on the same plot.

```
# start with KM curve
km1 +
```

```
  # visualize 3 year survival (green)
  geom_segment(x = 3, xend = 3,
               y = 0, yend = three_yr_surv$surv[2],
               size = 1,
               arrow = arrow(),
               colour = "#5ec962") +
  geom_segment(x = 3, xend = 0,
               y = three_yr_surv$surv[2], yend = three_yr_surv$surv[2],
               size = 1,
               arrow = arrow(length = unit(0.2, "inches")),
               colour = "#5ec962") +
  # visualize median survival time (blue)
geom_segment(x = s1_median["median"], xend = s1_median["median"],
             # the first y-val must be slightly > 0 for the arrow to print correctly
             y = 0.08, yend = 0.5,
             size = 1,
             arrow = arrow(ends = "first"),
             colour = "#3b528b") +
  geom_segment(x = 0, xend = s1_median["median"],
               y = 0.5, yend = 0.5,
               size = 1,
               arrow = arrow(ends = "last"),
               colour = "#3b528b")
```



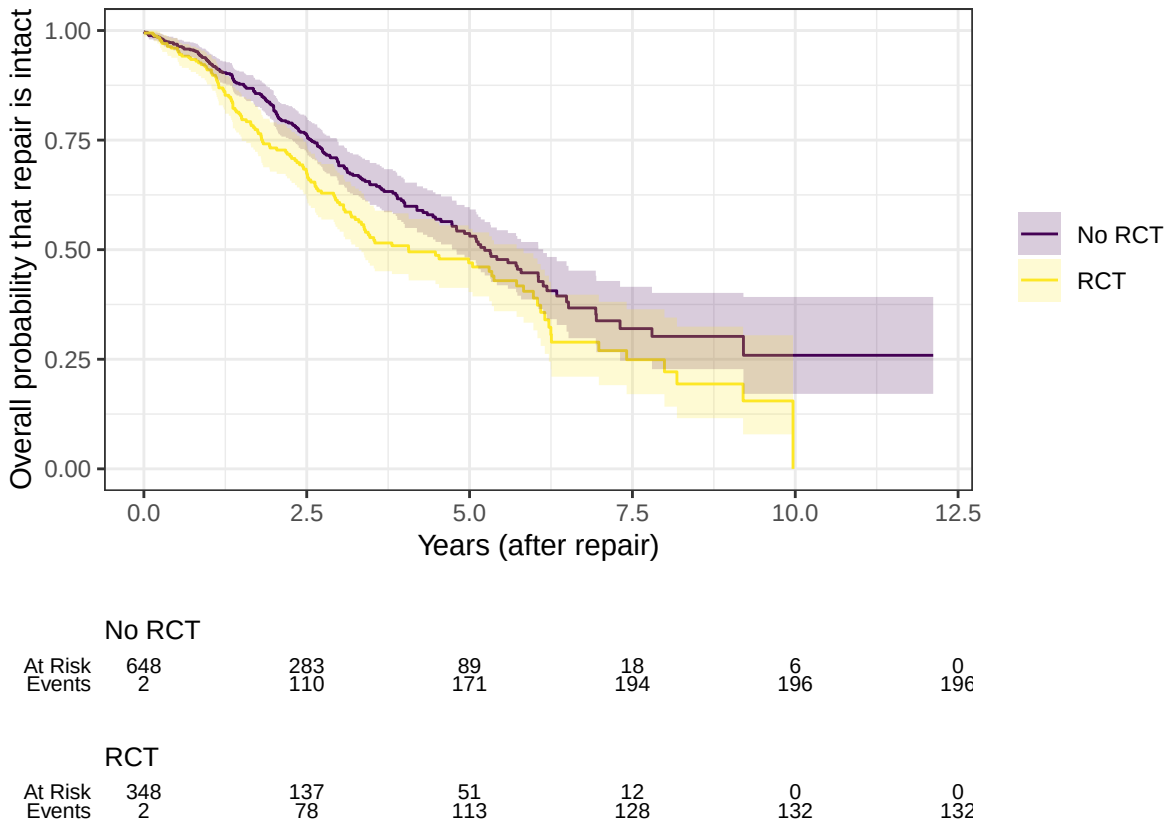
Up to this point, we have been studying a Kaplan-Meier plot that describes the entire data set (all crown margin repairs). In practice, the objective is often to compare two subgroups from within the data set – for instance, suppose we are interested in comparing how well crown margin repairs lasted between the root

canal treated (RCT) and non-RCT groups. The plot below draws two Kaplan-Meier survival curves – one for each of these subgroups. We notice that across time, the curve representing the RCT teeth is consistently below the curve representing the non-RCT teeth. This indicates that the curve for the RCT teeth is *dropping (decreasing) faster*, illustrating that the crown margin repairs done on RCT treated teeth do not last as long as the repairs done on non-RCT teeth.

In addition to the curves in this graph, we also see the confidence intervals at each time point illustrated by the tinted area around each curve. The yellow and purple tinted areas overlap with each other quite a bit, which symbolizes that the difference between crown margin repairs done on RCT teeth and non-RCT teeth is subtle – the repairs last only slightly longer on the non-RCT teeth.

```
km3 <- survfit2(Surv(time = Time, event = Event) ~ RCT, data = cmr) %>%
  ggsurvfit() +
  labs(
    x = "Years (after repair)",
    y = "Overall probability that repair is intact"
  ) +
  add_confidence_interval() +
  add_risktable() +
  scale_color_viridis(discrete = TRUE) +
  scale_fill_viridis(discrete = TRUE) +
  theme_bw()
```

```
km3
```

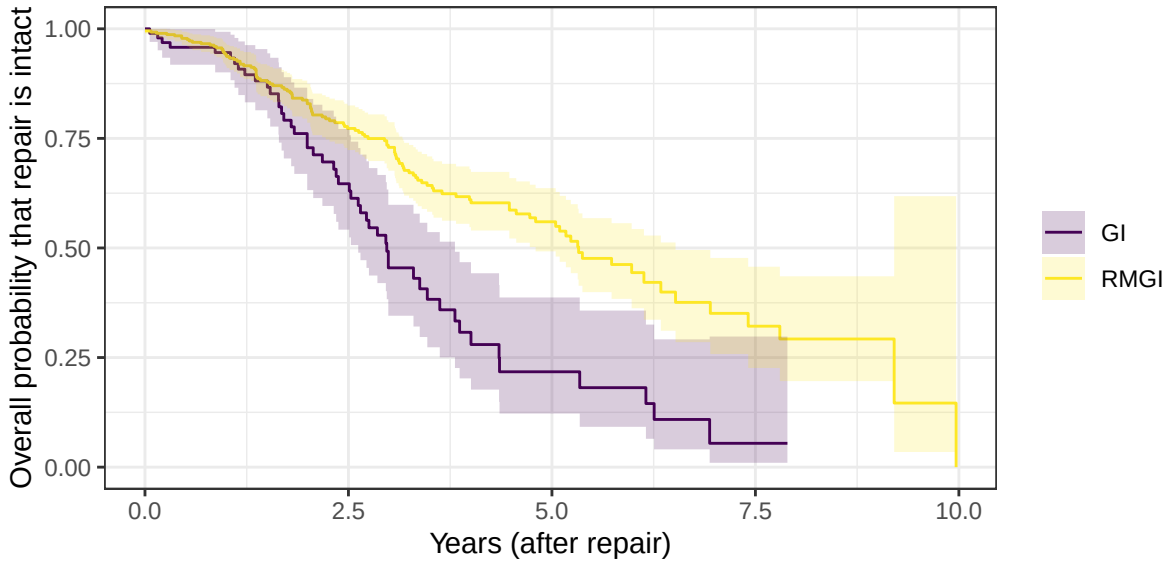


As a second example of comparing Kaplan-Meier plots between groups, let us suppose that we are working in a materials science context, where we are interested in comparing crown margin repairs that were done with glass ionomer (GI) to repairs done resin-modified glass ionomer (RMGI). We see in the next figure that the survival curve representing the GI group is much lower than the curve for the RMGI group for all times

after two years. We also see that the space between the two curves increases over time - the two curves are diverging. The confidence intervals do not overlap much at all after 2.5 years. These survival curves indicate that the crown margin repairs done with RMGI lasted notably longer than the repairs done with GI. There is evidence in this data set that the modification to GI makes a positive impact on the expected lifespan of crown margin repairs.

```
km4 <- survfit2(Surv(time = Time, event = Event) ~ Repair_Material,
                # NB: we are considering only those repairs that were done with GI and Amalgam
                data = cmr %>% filter(Repair_Material %in% c("RMGI", "GI"))) %>%
  ggsurvfit() +
  labs(
    x = "Years (after repair)",
    y = "Overall probability that repair is intact"
  ) +
  add_confidence_interval() +
  add_risktable() +
  scale_color_viridis(discrete = TRUE) +
  scale_fill_viridis(discrete = TRUE) +
  theme_bw()
```

```
km4
```



Our final survival analysis tool for this tutorial is a Cox proportional hazards model. This model examines each of the variables (i.e. the independent variables) in relationship the time-to-event outcome. A table summarizes the results of this Cox model using hazard ratios (HR), 95% confidence intervals (CI), and p-values.

```
coxph(Surv(time = Time, event = Event) ~ No_Cases + Age + Gender + CRA + Tooth_Type + Jaw +
        Repair_Material + Surfaces + No_Surfaces + RCT + Crown_Type + Clinic +
```

```
      Provider_Type,
    data = cmr) %>%
tbl_regression(exp = TRUE)
```

Characteristic	HR	95% CI	p-value
No_Cases	1.15	1.03, 1.28	0.012
Age	1.01	0.99, 1.02	0.5
Gender			
F			
M	1.20	0.80, 1.80	0.4
CRA			
High Risk			
Not High Risk	0.80	0.52, 1.22	0.3
Tooth_Type			
A			
P	1.84	0.97, 3.52	0.064
Jaw			
Md			
Mx	0.59	0.36, 0.95	0.030
Repair_Material			
Amal			
GI	2.51	1.24, 5.09	0.011
RBC	1.42	0.56, 3.61	0.5
RMGI	1.29	0.80, 2.07	0.3
Surfaces			
B			
L	1.27	0.72, 2.25	0.4
Other	1.65	0.96, 2.86	0.071
No_Surfaces			
1			
2	1.41	0.89, 2.22	0.14
RCT			
No RCT			
RCT	1.56	1.02, 2.38	0.038
Crown_Type			
C			
Other	1.11	0.39, 3.15	0.8
PFM	1.03	0.37, 2.84	>0.9
Clinic			
FAMD			
FDDAU	1.87	1.00, 3.50	0.050
FGP	2.33	0.84, 6.45	0.11
OPER	0.99	0.54, 1.82	>0.9
Other	1.36	0.54, 3.46	0.5
PROS	1.90	0.58, 6.16	0.3
SPEC	1.39	0.53, 3.66	0.5
Provider_Type			
faculty			
student	1.99	0.93, 4.28	0.076

References