Analysis of Clinical Trial Publication Data from the ISLR2 R Package

Sydney Murphy

2023-12-04

The Publication data in the ISLR2 R package contains information about the time to publication for the results of 244 clinical trials funded by the National Heart, Lung, and Blood Institute. Take some time to read more about this dataset in chapter 11.5.4 of ISL. You can also type ?Publication in R for more information after loading the dataset.

Load the Publication dataset in R

```
#install.packages("ISLR2")
library(ISLR2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Calculate for how many clinical trials the time associated with the event of interest (i.e., the time of publication) is observed and calculate for how many clinical trials the time of the event of interest is censored.

```
#??Publication
data("Publication")
head(Publication)

##   posres multi clinend     mech sampsize    budget impact      time status
## 1      0     0       1      R01    39876  8.016941 44.016 11.203285      1
## 2      0     0       1      R01    39876  8.016941 23.494 15.178645      1
## 3      0     0       1      R01     8171  7.612606  8.391 24.410678      1
## 4      0     0       1 Contract    24335 11.771928 15.402  2.595483      1
## 5      0     0       1 Contract    33357 76.517537 16.783  8.607803      1
## 6      0     0       1 Contract    10355  9.809938 16.783  8.607803      1
```

Produce and plot the Kaplan-Meier estimator for the time to publication of all the clinical trials in the dataset. Include 99% pointwise confidence bands in the plot.

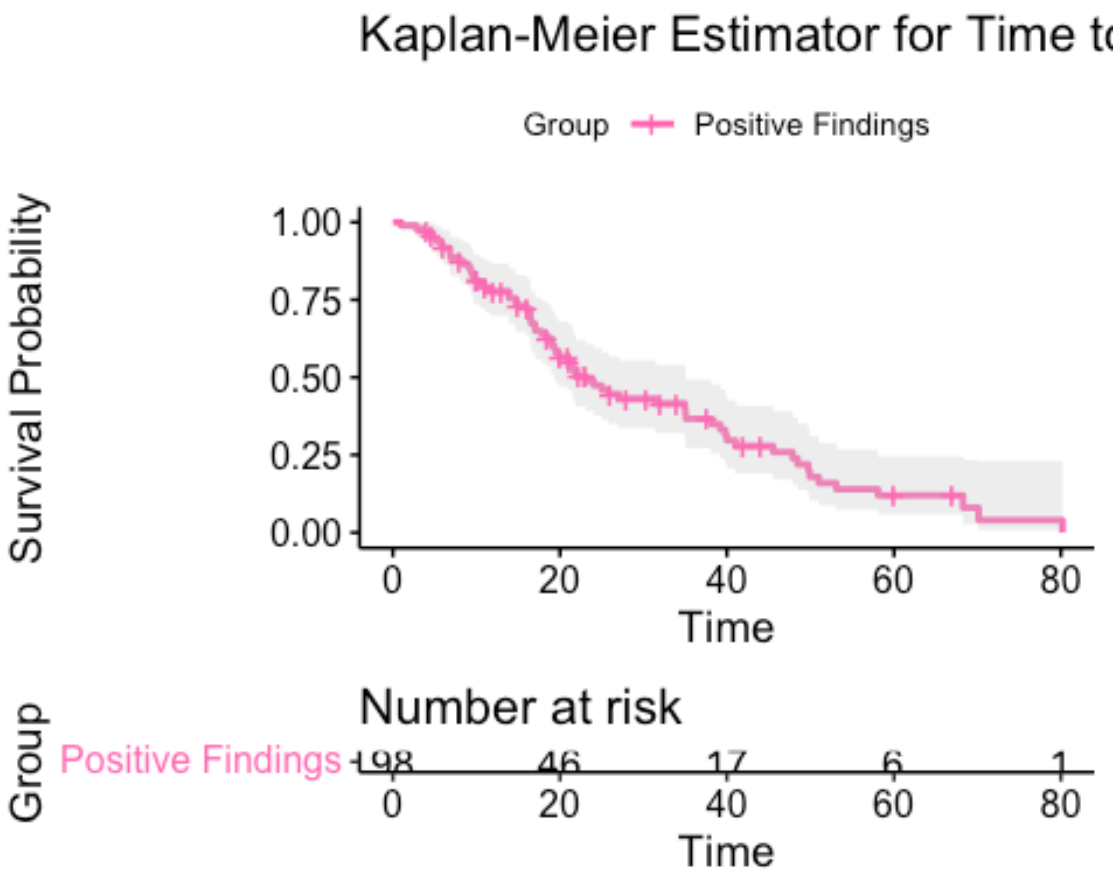
```r
num_observed <- sum(Publication$status == 1)
num_censored <- sum(Publication$status == 0)

cat("The number of observed clinical trials is:", num_observed, "\\n")

## The number of observed clinical trials is: 156 \n

cat("The number of censored clinical trials is:", num_censored)

## The number of censored clinical trials is: 88
```

Produce and plot the Kaplan-Meier estimator for the time to publication of the clinical trials for the two subgroups corresponding to the posres variable (i.e., for the group of clinical trials that resulted in positive findings and for the group of clinical trials that did not result in positive findings). Then, use the log-rank test to test the null hypothesis that the time to publication is not associated with whether or not the clinical trial resulted in a positive finding (posres). State in English the result of the log-rank test.

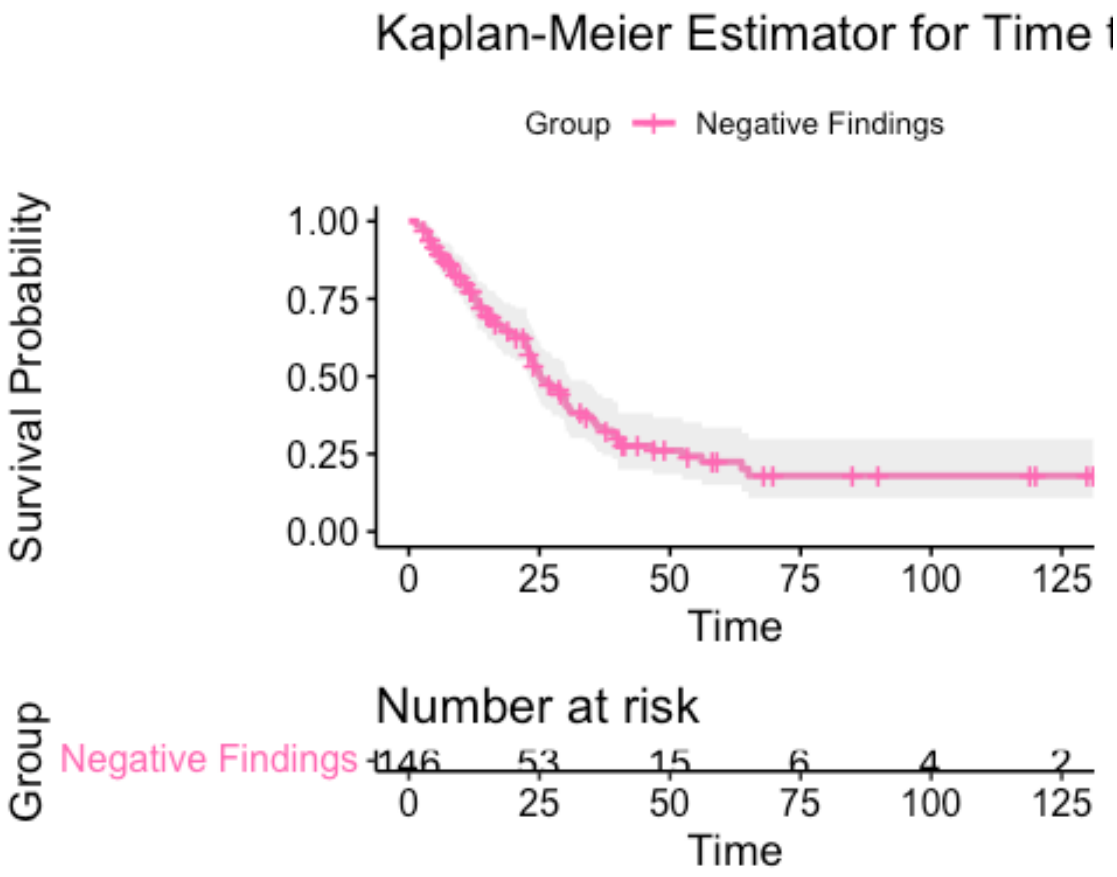
```r
#install.packages("survival")
#install.packages("survminer")
library(survival)
library(survminer)

## Loading required package: ggplot2

## Loading required package: ggpubr

##
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':
##
##     myeloma

library(ggplot2)

data_positive <- subset(Publication, posres == 1)
data_negative <- subset(Publication, posres == 0)

survfit_positive <- survfit(Surv(time, status) ~ 1, data = data_positive)

survfit_negative <- survfit(Surv(time, status) ~ 1, data = data_negative)

ggsurvplot(list(positive = survfit_positive, negative = survfit_negative),
           data = Publication,
           palette = c("#FF69B4", "#8A2BE2"), # Hot pink and blue violet
colors
           conf.int = TRUE,
           pval = TRUE,
           risk.table = TRUE,
           title = "Kaplan-Meier Estimator for Time to Publication",
```

```
           xlab = "Time",
           ylab = "Survival Probability",
           legend.title = "Group",
           legend.labs = c("Positive Findings", "Negative Findings"))
```

```
## Warning in .pvalue(fit, data = data, method = method, pval = pval,
pval.coord = pval.coord, : There are no survival curves to be compared.
##  This is a null model.

## Warning in .pvalue(fit, data = data, method = method, pval = pval,
pval.coord = pval.coord, : There are no survival curves to be compared.
##  This is a null model.

## $positive
```




```
##
## $negative
```




```
##
## attr(,"class")
## [1] "list"            "ggsurvplot_list"
```

```
log_rank_test <- survdiff(Surv(time, status) ~ posres, data = Publication)

print(log_rank_test)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ posres, data = Publication)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## posres=0 146       87     92.6     0.341     0.844
## posres=1  98       69     63.4     0.498     0.844
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.4
```

Fit a Cox proportional hazards model to these data using the following predictors: • posres • multi • clinend • budget Also produce the model summary with the summary function.

```
cox_model <- coxph(Surv(time, status) ~ posres + multi + clinend + budget,
data = Publication)
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ posres + multi + clinend +
##     budget, data = Publication)
##
##   n= 244, number of events= 156
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## posres  0.533728  1.705278 0.178275 2.994  0.00275 **
## multi   0.633555  1.884298 0.227922 2.780  0.00544 **
## clinend 1.641604  5.163447 0.241385 6.801 1.04e-11 ***
## budget  0.002282  1.002285 0.001800 1.268  0.20477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## posres      1.705     0.5864    1.2024     2.418
## multi       1.884     0.5307    1.2054     2.945
## clinend     5.163     0.1937    3.2172     8.287
## budget      1.002     0.9977    0.9988     1.006
##
## Concordance= 0.731  (se = 0.021 )
## Likelihood ratio test= 81.39  on 4 df,   p=<2e-16
## Wald test            = 97.91  on 4 df,   p=<2e-16
## Score (logrank) test = 125.1  on 4 df,   p=<2e-16
```

```
#effect of each predictor on the hazard function corresponding to the time to publication
```

Do the global likelihood ratio, Wald, and score test suggest that the model is better than a model that does not use any predictor?

```
cat ("The global likelihood ratio, Wald, and score test yield extremely low p-values--much less than the conventional threshold of 0.05. This low p-value across all three tests indicates to me that the model including the predictors
posres, multi, clinend, and budget fits the data significantly better than a model without predictors. So, these predictors are important in explaining the
time to publication.")
```

```
## The global likelihood ratio, Wald, and score test yield extremely low
## p-values--much less than the conventional threshold of 0.05. This low p-
value
## across all three tests indicates to me that the model including the
predictors
## posres, multi, clinend, and budget fits the data significantly better than
a
## model without predictors. So, these predictors are important in explaining
the
## time to publication.
```

In English, interpret the estimated effect of each predictor on the hazard function corresponding to the time to publication.

```
cat ("The coefficient for posres is 0.533728 (or exponentiated = 1.705). This
     tells me that clinical trials with positive results are 1.705 times more
     likely to be published at any time compared to those without positive
     results (holding all variables constant). The p-value of posres is
0.00275,
     (less than 0.05) telling me that this effect is statistically
significant.")
```

```
## The coefficient for posres is 0.533728 (or exponentiated = 1.705). This
##      tells me that clinical trials with positive results are 1.705 times
more
##      likely to be published at any time compared to those without positive
##      results (holding all variables constant). The p-value of posres is
0.00275,
##      (less than 0.05) telling me that this effect is statistically
significant.
```

```
cat ("The coefficient for multi variable is 0.633555 (or 1.884). This tell me
     that multicenter trials have a hazard rate 1.884 times higher than
     single-center trials (probably published sooner). The p-value is
0.00544,
     showing a significant effect.")
```

```
## The coefficient for multi variable is 0.633555 (or 1.884). This tell me
##      that multicenter trials have a hazard rate 1.884 times higher than
##      single-center trials (probably published sooner). The p-value is
0.00544,
##      showing a significant effect.
```

```
cat ("The coefficient for cliniend variable is 1.641604 ( or 5.163). This
     tells me there is a strong association with trials with the specified
     clinical endpoint being more than five times likely to be published than
     others. The low p-value (1.04e-11) shows this predictor's
significance.")
```

```
## The coefficient for cliniend variable is 1.641604 ( or 5.163). This
##      tells me there is a strong association with trials with the specified
##      clinical endpoint being more than five times likely to be published
than
##      others. The low p-value (1.04e-11) shows this predictor's
significance.
```

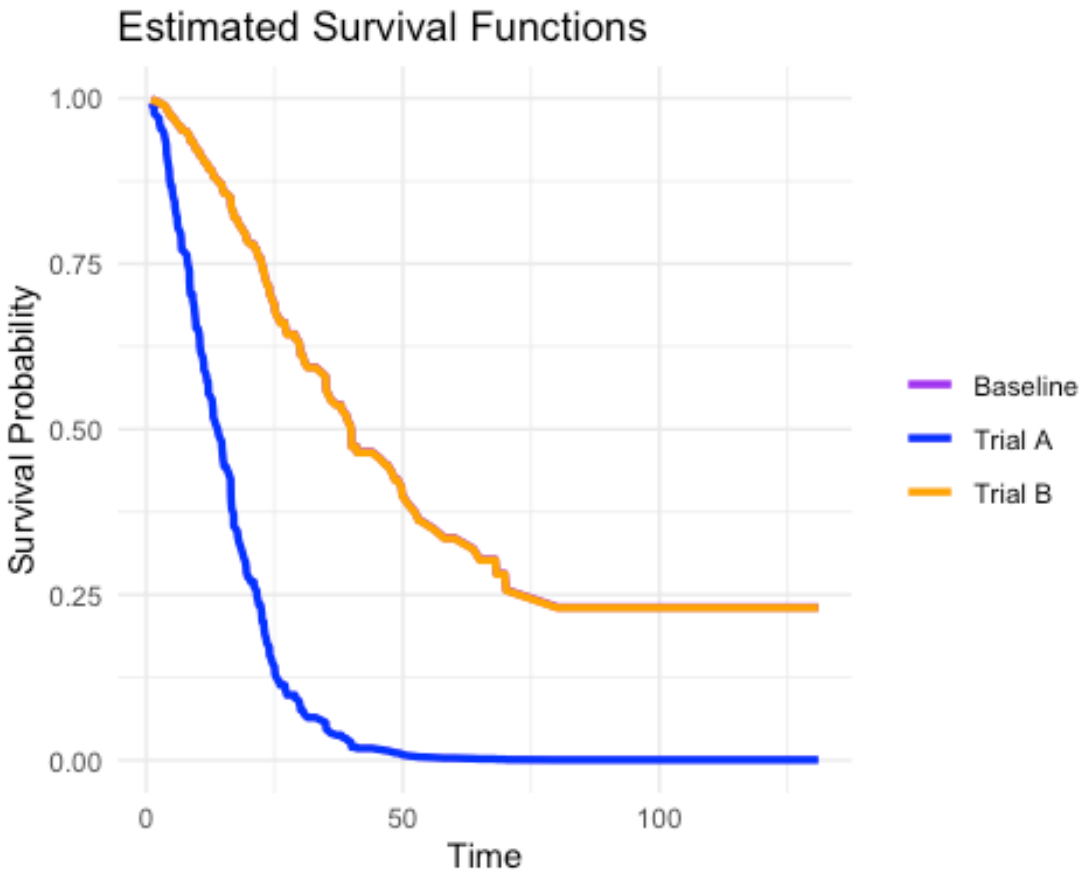
```
cat("The coefficient for budget is 0.002282 (or 1.002). This reflects a
    slight increase in the hazard rate with each unit increase in budget,
    telling me that as budget increases, likelihood of publication increases.
    But, the p-value is 0.20477, tells me that the effect of the budget on
the
    time to publication is not significant.")
```

```
## The coefficient for budget is 0.002282 (or 1.002). This reflects a
##     slight increase in the hazard rate with each unit increase in budget,
##     telling me that as budget increases, likelihood of publication
increases.
##     But, the p-value is 0.20477, tells me that the effect of the budget on
the
##     time to publication is not significant.
```

Compute the estimated survival function Sb from a Cox proportional hazards model and plot it.

First, use the above commands to plot the estimated baseline survival function S0 for the time to publication (this is obtained by setting all predictors to 0 in newdata). Then, use the above commands to plot. the estimated survival function for the time to publication of a new clinical trial with predictor values – posres: 0 – multi: 0 – clinend: 1 – budget: 8.5 • the estimated survival function for the time to publication of a new clinical trial with predictor values – posres: 0 – multi: 0 – clinend: 0 – budget: 1.3 Based on the two estimated survival functions, which of these two clinical trials do you think is more likely to be published sooner? Explain.

```
baseline_data <- data.frame(posres = 0, multi = 0, clinend = 0, budget = 0)
trial_a_data <- data.frame(posres = 0, multi = 0, clinend = 1, budget = 8.5)
trial_b_data <- data.frame(posres = 0, multi = 0, clinend = 0, budget = 1.3)

#survival functions
baseline_survival <- survfit(cox_model, newdata = baseline_data)
trial_a_survival <- survfit(cox_model, newdata = trial_a_data)
trial_b_survival <- survfit(cox_model, newdata = trial_b_data)

#combined data frame for plotting
survival_data <- rbind(
  data.frame(time = baseline_survival$time, surv = baseline_survival$surv,
type = 'Baseline'),
  data.frame(time = trial_a_survival$time, surv = trial_a_survival$surv, type
= 'Trial A'),
  data.frame(time = trial_b_survival$time, surv = trial_b_survival$surv, type
= 'Trial B')
)

#plot
g <- ggplot(survival_data, aes(x = time, y = surv, color = type)) +
  geom_line(linewidth = 1.2) +
  scale_color_manual(values = c('Baseline' = 'purple', 'Trial A' = 'blue',
'Trial B' = 'orange')) +
  labs(title = "Estimated Survival Functions", x = "Time", y = "Survival
Probability") +
  theme_minimal() +
  theme(legend.title = element_blank())
print(g)
```




```
cat ("Trial A (blue) is more likely to be published sooner than Trial B.
Trial A's survival curve drops more quickly than Trial B, showing me that it
has a higher rate of publication over time OR higher hazard function.")
```

```
## Trial A (blue) is more likely to be published sooner than Trial B.
## Trial A's survival curve drops more quickly than Trial B, showing me that
it
## has a higher rate of publication over time OR higher hazard function.
```

```
cat ("Trial B doesn't have a clinical endpoint and has a smaller budget,
     which shows a more gradual decrease in its survival function. This
     tells me that its chances of being published quickly are lower than
Trial A.")
```

```
## Trial B doesn't have a clinical endpoint and has a smaller budget,
##      which shows a more gradual decrease in its survival function. This
##      tells me that its chances of being published quickly are lower than
Trial A.
```