AN ANALYSIS OF SURVIVAL DATA WHEN HAZARDS ARE NOT PROPORTIONAL: APPLICATION TO A CANCER TREATMENT STUDY

John Benjamin White

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the Department of Biostatistics,
Indiana University

December 2021

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Master's Thesis Committee

_____

Constantin T. Yiannoutsos, PhD Chair

_____

Giorgos Bakoyannis, PhD

_____

William F. Fadel, PhD

DEDICATION

To my ever-supportive family, and to Taylor Bone. Thank you for your love in my darkest times.

ACKNOWLEDGEMENT

I wish to thank Dr. Yiannoutsos for readily making the time to advise me in the writing of this thesis, and on the short notice he was provided. This project has provided me with a valuable opportunity to acquire new skills and experience in multiple areas of study.

John Benjamin White

AN ANALYSIS OF SURVIVAL DATA WHEN HAZARDS ARE NOT PROPORTIONAL: APPLICATION TO A CANCER TREATMENT STUDY

The crossing of Kaplan-Meier survival curves presents a challenge when conducting survival analysis studies, making it unclear whether any of the study groups involved present any significant difference in survival. An approach involving the determination of maximum vertical distance between the curves is considered here as a method to assess whether a survival advantage exists between different groups of patients. The method is illustrated on a dataset containing survival times of patients treated with two cancer treatment regimes, one involving treatment by chemotherapy alone, and the other by treatment with both chemotherapy and radiotherapy.

Constantin T. Yiannoutsos, PhD, Chair

Giorgos Bakoyannis, PhD

William F. Fadel, PhD

TABLE OF CONTENTS

LIST OF FIGURES

Background

The field of survival analysis focuses on the analysis of time-to-event data. Some examples of these would be time until remission of a disease, time until a test subject dies, or employee turnover at a company. Survival analysis deals with censoring, which is where time until an event is not observed for a subject prior to the end of an experiment. An example of that would be a test subject dropping out of an experiment before the remission of their disease. A common goal in survival analysis is viewing the differences in survival between two study groups and whether there is a significant difference in survival between them.

Important Concepts

Two fundamental quantities in survival analysis are the survival distribution and the hazard function. From these may be derived additional relations that will also be addressed.

Let T be the time until an event and let $F(t)$ be the cumulative distribution of T. We call $S(t)$ the survival distribution, defined as

$$S(t) = \mathrm{P}(T > t) = 1 - F(t)$$

The hazard function is the instantaneous risk of failure at some time $t + h$ for a member of the experiment that has yet to experience an event at t. Defining h as some small value, the hazard function is expressed as

$$\lambda(t) = \lim_{h \to 0} \frac{\mathrm{P}(\, t \leq T \leq t + h \mid T > t \,)}{h}$$

Finally, the cumulative hazard is the sum of instantaneous risks up until some time t, defined as

$$\Lambda(t) = \int_0^t \lambda(u) du$$

The cumulative hazard is related to the survival function via the expression

$$\mathrm{S}(t) = \exp[-\Lambda(t)]$$

Estimating the Survival Function

A common approach to estimation of the survival function is the Kaplan-Meier method. This divides the time prior to t into discrete intervals $[\tau_{j-1}, \tau_j)$.

To survive past time t, a person has to survive up to just prior to t, and then survive at t. Consequently, the probability of surviving past time t is given by the cumulative probability of surviving during all prior intervals up until the previous interval times the probability of surviving for the instant t. Therefore, given $j = 1, \ldots, J$, $\tau_0 = 0$, and $\tau_J = t$, the formula

$$\prod_{J=1}^{J} P(t_{j-1} \leq T < \tau_j) = \prod_{j=1}^{J} [1 - \lambda(\tau_j)]$$

provides what is called the "product limit estimator" of $S(t)$. It has been shown (Kaplan & Meier, 1958) that this estimate approaches $S(t)$ as $J \to \infty$.

Comparing Survival Curves

As previously mentioned, we are often interested in assessing differences in survival between two study groups. This might be for studying the efficacy of a treatment or factors that affect longevity of a product. We will examine two common approaches for analyzing the difference between two Kaplan-Meier curves before moving onwards to discuss the alternative presented in this thesis.

First, we will review the log-rank test. Define d_{0j} and d_{1j} as the number of failures in each group, n_{0j} and n_{1j} as the number at risk in each group (meaning subjects who have not experienced failure up to τ_j), d_j as the total number of failures in both groups, and n_j as the total at risk at τ_j, and k be the number of *distinct* failure times. To simplify the final expression, also define

$$v_{0j} = \frac{n_{0j} n_{1j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

The log-rank test statistic is given below.

$$L = \frac{\left[\sum_{j=1}^{k} d_{0j} - \frac{n_{0j} d_j}{n_j}\right]^2}{\sum_{j=1}^{k} v_{0j}} \rightarrow_d \chi^2(1)$$

The log-rank statistic is asymptotically distributed according to a chi-squared distribution with one degree of freedom.

An important note to make is the topic of proportional hazards. The log-rank test is most powerful when the hazards are proportional. Given the previous-mentioned link between the hazard function and the survival function, when hazards are proportional there is a widening difference in the distance between the curves as time moves forward and hazard is accumulated. There is a simple visual check to see whether the hazards are

proportional (and thus ascertain whether the log-rank statistic will have good power to detect differences in survival). A consequence of the proportionality of hazards is that, if the survival curves cross, the hazards are not proportional.

However, when hazards are not proportional, that is not evidence that the survival distributions are equal. It only indicates that the log-rank test will be much less powerful to detect any differences and comparisons become much more difficult to make.

by Tarone and Ware (1967). The Wilcoxon test statistic is defined as follows:

$$W = \frac{\left[\sum_{j=1}^{k} n_0 \left(d_{0j} - \frac{n_{0j} d_j}{n_j}\right)\right]^2}{\sum_{j=1}^{k} n_0^2 v_{0j}} \rightarrow_d \chi^2(1)$$

The Wilcoxon statistic is also asymptotically distributed according to a chi-squared distribution with one degree of freedom.

As shown by Tarone and Ware (1967), both the Log-Rank and the Wilcoxon statistic are part of a large class of tests which weigh failures differently. Whereas the Log-Rank test statistic weighs all failures the same, the Wilcoxon test statistic provides an alternative approach, similar to the log-rank test but putting more weight on the earlier rather than the later failure times. The intuition of the Wilcoxon test is that, given the larger sample size early in time, the early failures represent a great deal more information than the late failures.

Because of this weighting scheme, the Wilcoxon test may be more powerful than the Log-Rank test in situations where the proportional hazards assumption is violated. For this reason, we consider the use of the Wilcoxon test as an alternative to the Log-Rank test. In addition, in this thesis, we present another non-parametric approach to determine the more beneficial of the two treatments.

Motivating Example

Our motivating example is a cancer treatment trial described by Schein and colleagues (1982). The dataset is a record of two gastric carcinoma patient groups: one group was treated with both chemotherapy and radiotherapy, and the other was treated with chemotherapy alone. Estimating the survival in the two groups by the Kaplan-Meier approach, yields the survival curves displayed in the figure below.

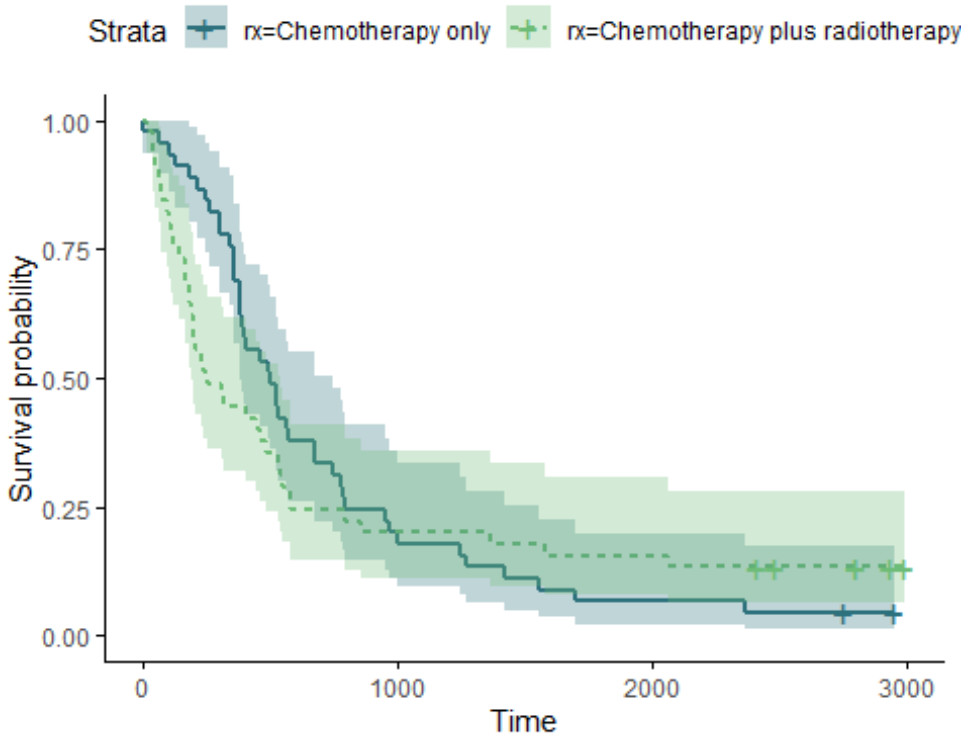


Figure 1: Crossing Survival Curves

Clearly, the survival curves cross, creating ambiguity in comparing the efficacy of the two treatments in terms of increasing patient survival. In fact, when applying the usual log-rank test to these data, we obtain a p value of 0.63, which is not statistically significant. This implies that, according to the log-rank test, no clear survival advantage

can be discerned with either of these treatments, despite some clear differences observed between the two survival curves. Analysis of the data in the motivating example by the Wilcoxon test produces a p-value of $0.007 < \alpha = 0.05$. The conclusion from this analysis is that, contrary to the log-rank test, the Wilcoxon test detects a statistically significant difference in the survival experience in the two treatment groups. We address an alternative analysis of these data in the following paragraphs.

Methodology Overview

As detailed by Bakoyannis (2020), the methodology is conducted according to the following steps: first, a Kaplan-Meier analysis is conducted on the data to produce estimates of the two survival curves. Following that analysis, the maximum vertical distance between the two curves is calculated. To obtain an estimate of the distribution of the maximum vertical difference, that same analysis is conducted many times on bootstrapped data. (For more details on resampling survival data, please refer to Efron (1981)). The resulting instances of the maximum differences based on these data, results in an estimate of the *empirical* distribution of this statistic. Finally, the p value is calculated from the results of these analyses, as the tail of the distribution to the right of the observed vertical distance in the motivating example.

Calculating Vertical Distance Between Curves

To calculate the maximum vertical distance, we begin by dividing the individual groups up into sets of ordered pairs of failure times and survival probabilities, (t_{1i}, s_{1i}) and (t_{2i}, s_{2i}), where the shorter of the two sets of pairs is extended to match the length of the other set by duplicating the last observation a sufficient number of times.

First, for group 1, we look at each t_{1i} and find $s_{2.i-1}$, and then subtract from s_{1i}, and then find the absolute value of the difference. This may be expressed as

$$D_{1i} = |s_{1i} - s_{2,i-1}|$$

We assess group 2 in a similar manner,

$$D_{2i} = |s_{2i} - s_{1,i-1}|$$

In the data set in the example we analyze, this may be visualized in figure 2.

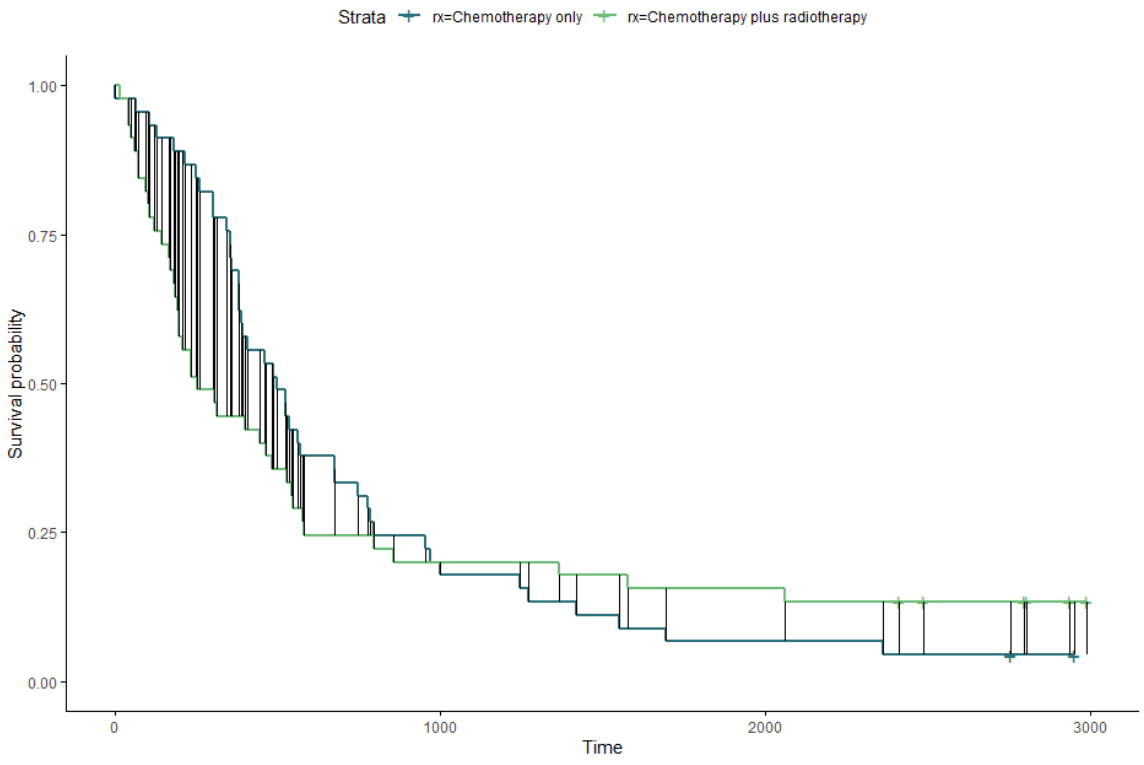


Figure 2: Vertical Distances at Failure Times in the Motivating Example

To find the maximum of all such vertical distances, we simply calculate

$$\widehat{\theta_n} = \max[\max(D_{1i}), \max(D_{2i})]$$

The algorithm defined above is implemented by the following steps:

1. Sample with replacement separately for each treatment group for as many observations are in each.
2. Generate the Kaplan-Meier estimates of the group survival distributions on the resampled data.
3. Calculate the maximum vertical distance for each of these results.
4. For each maximum vertical distance $\theta_n^{(b)}$, $b = 1, \dots, B$, determine whether $\theta_n^{(b)} - \hat{\theta}_n > \hat{\theta}_n$, where $\hat{\theta}_n$ is the observed vertical distance in the study data. If so, append 1 to d_b . Otherwise, append 0 to d_b.
5. Run the above steps for some B number of iterations.
6. Determine the value $a = \sum(d_b)$

The p value is determined by dividing a by the number of iterations B. The p value can thus be expressed mathematically as

$$\hat{p} = \sum_{b=1}^{B} \frac{I\{(\theta_n^b - \hat{\theta}_n) > \hat{\theta}_n\}}{B} = \sum_{b=1}^{B} \frac{d_b}{B}$$

where $I(\cdot)$ is the indicator function. The conclusion about the potential superiority of either group's survival distribution is determined by comparing this empirical p value against our chosen type-1 error level.

Analysis

The approach detailed above will now be implemented on our motivating example. I have developed statistical code in R (R Core Team, 2021), implementing this approach. The code is shown in the Appendix.

To begin, the program imports the study data and fits a Kaplan-Meier model. The two groups are then divided into separate data frames, each with two columns, one with failure time and the other with the corresponding survival probability.

Because the two groups will not necessarily be of the same length (as seen in this dataset), a function is defined that will find the shorter of the two and lengthen it by duplicating the final rows enough times to make the two data frames have an equal length.

Next, another function is defined for the purpose of finding the previous failure given some current failure, and an additional function is created to find the vertical distance between the survival probability at a certain failure time and the survival probability at a previous failure time. A final function is then created to find the maximum such distance. This is our "maximum vertical distance." The code then runs these functions to determine the maximum vertical distance in the dataset. As mentioned above, this value will be used in our final calculation to determine the p value.

Next, a value for B (the number of iterations to run the algorithm) is set, in this case 1000. Then, via a for loop, the dataset is bootstrapped and then the maximum vertical distance of that bootstrapped data is found and added to a list. As mentioned, this loop runs B times.

Finally, to calculate the p value, we take the sum of all Boolean values for the truth of the numerator's inequality and divide by the number of iterations. For the data assessed in this thesis, the calculation results in a p value of 0.007. Thus, with $p < \alpha = 0.05$, we can conclude that the null hypothesis is rejected and thus there is a significant difference in survival between the two groups.

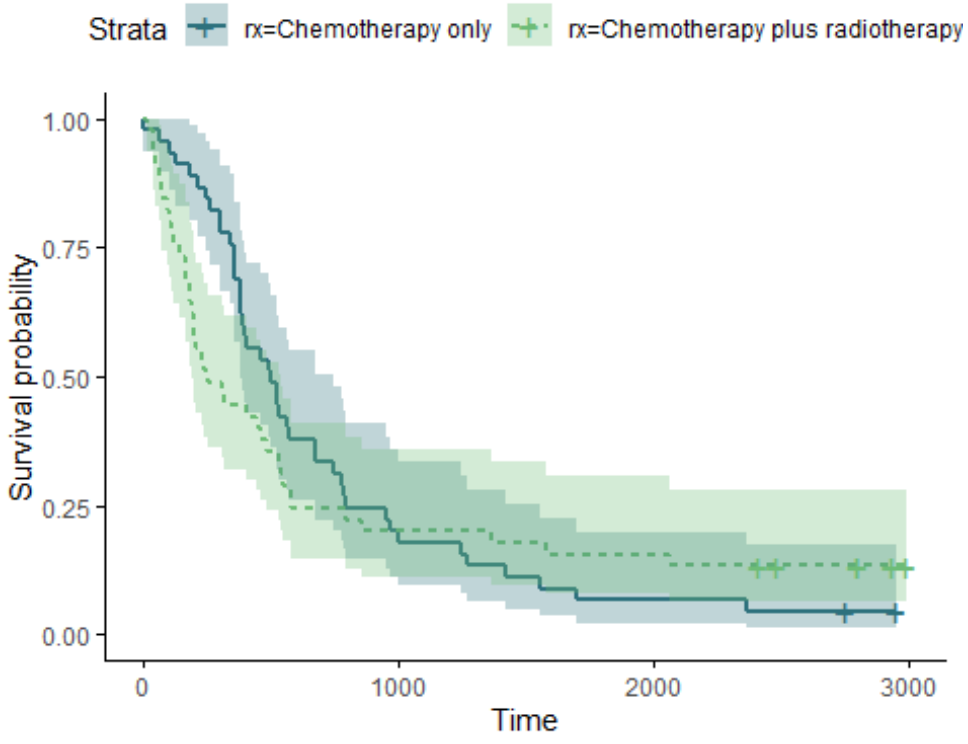
Conclusion

The crossing of survival curves creates ambiguity in considering the overall survival between two groups. In the case we assessed, it is unclear at an intuitive level whether there is a significant difference in the survival between the two treatment groups. Having examined an additional non-parametric approach, we have determined that there is indeed a significant difference between the two.

Appendices

Appendix I - Code

```
library(ggplot2)
library(ggfortify)
library(survival)
library(survminer)

## Loading required package: ggpubr

theme_set(theme_classic())
set.seed(45986)

study_data <- read.csv("stabkout.csv")

km_model <- survfit(Surv(survival, death) ~ rx, data = study_data)

ggsurvplot(km_model,
           conf.int = TRUE,
           risk.table.col = "strata",
           linetype = "strata",
           ggtheme = theme_classic(),
           palette = c("#2D717D", "#6DBC77"))
```



```r
model_attributes <- fortify(km_model)

chemo_only_fails <- model_attributes[which(model_attributes$strata=="Ch
emotherapy only"), ][c(1, 5)]
chemo_radio_fails <- model_attributes[which(model_attributes$strata=="C
hemotherapy plus radiotherapy"), ][c(1, 5)]
rownames(chemo_radio_fails) <- 1:nrow(chemo_radio_fails)

# Function to lengthen the shorter of the two data sets
# by duplicating the last entries.
make_same_length <- function(data1, data2) {

  len_1 <- length(data1[, 1])
  len_2 <- length(data2[, 1])

  length_dif <- max(len_1, len_2) - min(len_1, len_2)

  rownames(data1) <- 1:nrow(data1)
  rownames(data2) <- 1:nrow(data2)

  if (len_1 < len_2) {
    for (i in 1:length_dif) {
      data1[len_1 + i, ] <- data1[c("time", "surv")][len_1, ]
    }
  }

  else if (len_2 < len_1) {
    for (i in 1:length_dif) {
      data2[len_2 + i, ] <- data2[c("time", "surv")][len_2, ]
    }
  }

  colnames(data1) <- c("time_1", "surv_1")
  colnames(data2) <- c("time_2", "surv_2")
  return(data.frame(data1, data2))

}

# Given some value, find the previous failure in a
# set of failures
find_previous_failure <- function(data, failure) {

  other_fails <- c()
  other_survs <- c()
  failure <- failure[, "time"]

  # This portion is here to account for the situation where
  # the failure time provided is less than or equal to the first value
  # in the set of failures. I am not sure if this solution
  # solves the issue or not.
```

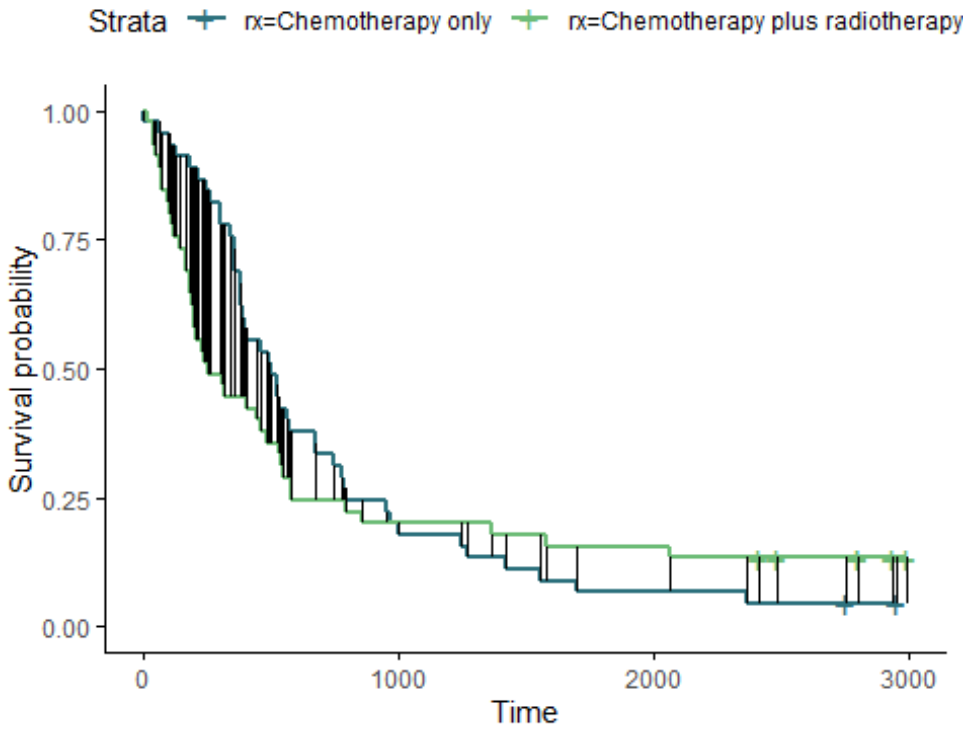
```
  if (failure <= data[, "time"][1]) {
    other_fails[1] <- 0
    other_survs[1] <- 1
  }

  else {
    for (i in 1:length(data[, "time"])) {
      if (data[, "time"][i] < failure) {
        other_fails[i] <- data[, "time"][i]
        other_survs[i] <- data[, "surv"][i]
      }
    }
  }

  out_data <- data.frame(other_fails, other_survs)[length(other_fails),
]
  colnames(out_data) <- c("time", "surv")
  return(out_data)

}

# Calculate the vertical distance at one of the failures
find_vert_dist <- function(obs, data1, data2) {
  return(abs(data1[obs, ]["surv"] - find_previous_failure(data2, data1[
obs, ]["time"])["surv"]))
}

# Find the maximum vertical distance
find_max_vert_dist <- function(data1, data2) {
  vert_dists_1 <- c()
  vert_dists_2 <- c()
  for (i in 1:length(data1[, 1])) {
    vert_dists_1[i] <- find_vert_dist(i, data1, data2)[, "surv"]
    vert_dists_2[i] <- find_vert_dist(i, data2, data1)[, "surv"]
  }
  return(max(max(vert_dists_1), max(vert_dists_2)))
}

# ----- ORIGINAL ESTIMATE ----- #
equalized_lengths <- make_same_length(chemo_only_fails, chemo_radio_fai
ls)

chemo_only_fails <- equalized_lengths[c("time_1", "surv_1")]
chemo_radio_fails <- equalized_lengths[c("time_2", "surv_2")]

colnames(chemo_only_fails) <- c("time", "surv")
colnames(chemo_radio_fails) <- c("time", "surv")
original_estimate <- find_max_vert_dist(chemo_only_fails, chemo_radio_f
ails)
# ----- ---------------- ----- #
```

```r
found_maxes <- c()

B <- 1000

for (i in 1:B) {
  index <- sample.int(length(study_data[,1]), replace=FALSE)
  boot_model <- survfit(Surv(study_data$survival, study_data$death) ~ s
tudy_data[index,]$rx)
  model_attributes_btstrp <- fortify(boot_model)

  model_attributes_btstrp_co <- model_attributes_btstrp[which(model_att
ributes_btstrp$strata=="Chemotherapy only"), ][c(1, 5)]
  model_attributes_btstrp_cr <- model_attributes_btstrp[which(model_att
ributes_btstrp$strata=="Chemotherapy plus radiotherapy"), ][c(1, 5)]
  rownames(model_attributes_btstrp_cr) <- 1:nrow(model_attributes_btstr
p_cr)

  equalized_lengths <- make_same_length(model_attributes_btstrp_co, mod
el_attributes_btstrp_cr)
  model_attributes_btstrp_co <- equalized_lengths[c("time_1", "surv_1")
]
  model_attributes_btstrp_cr <- equalized_lengths[c("time_2", "surv_2")
]

  colnames(model_attributes_btstrp_co) <- c("time", "surv")
  colnames(model_attributes_btstrp_cr) <- c("time", "surv")

  data.frame(model_attributes_btstrp_co, model_attributes_btstrp_cr)

  found_maxes[i] <- find_max_vert_dist(model_attributes_btstrp_co, mode
l_attributes_btstrp_cr)
}

# Calculate the p-value
p_value <- sum(found_maxes > original_estimate)/B
print(p_value)
```

```
## [1] 0.007
```

```r
surv_plot_with_lines <- ggsurvplot(km_model,
                                   risk.table.col = "strata",
                                   ggtheme = theme_classic(),
                                   palette = c("#2D717D", "#6DBC77"))
surv_plot_with_lines <- surv_plot_with_lines$plot
```

```
for (i in 1:45) {
  surv_plot_with_lines <- surv_plot_with_lines + geom_segment(x = chemo
_only_fails[i, ]["time"][, ], y = find_previous_failure(chemo_radio_fai
ls, chemo_only_fails[i, ]["time"])["surv"][, ], xend = chemo_only_fails
[i, ]["time"][, ], yend = chemo_only_fails[i, ]["surv"][, ]) +
                                                    geom_segment(x = chemo
_radio_fails[i, ]["time"][, ], y = find_previous_failure(chemo_only_fai
ls, chemo_radio_fails[i, ]["time"])["surv"][, ], xend = chemo_radio_fai
ls[i, ]["time"][, ], yend = chemo_radio_fails[i, ]["surv"][, ])
}
surv_plot_with_lines
```



References

Kaplan, E., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*(282), 457-481. doi:10.2307/2281868

Tarone, R., & Ware, J. (1977). On Distribution-Free Tests for Equality of Survival Distributions. *Biometrika, 64*(1), 156-160. doi:10.2307/2335790

Schein, P. S. (1982). A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. *Cancer*, *49*(9), 1771–1777.

Bakoyannis, G. (2020). *Nonparametric tests for transition probabilities in nonhomogeneous Markov processes*. *Journal of Nonparametric Statistics*, **32**, 131–156

Efron, B. (1981) Censored data and the bootstrap. *Journal of the American Statistical Association*, **76**, 312–319.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Curriculum Vitae

John Benjamin White

Education

BS Mathematics, from Purdue University at IUPUI, December 2017

MS Biostatistics, from Indiana University at IUPUI, December 2021

Experience

Senior Analyst at Deloitte, May 2019-October 2020

Solution Consultant at Deloitte, October 2020-Present