Effectiveness of Treatments in Prostate Cancer Patients

Spencer Yee

Table of Contents

Abstract

The aptitude to properly treat critical medical patients greatly increases with the ability to estimate survival time and probability in addition to identifying the most effective treatment options. It would be ignorant to explain the importance of survivability and the methods of this study play a huge role in the effective treatment and assessment of patients. This study examines a dataset of 502 prostate cancer patients. Kaplan-Meier estimates and Cox Proportional-Hazard models were fit to the treatment and survival portions respectively. Both modified and unmodified variables for patient treatment dosage, weight index, history of cardiovascular disease, electrocardiogram code, serum hemoglobin, size of primary tumor, and stage/histology grade index were found to be important determinants. The methods used will be useful in conducting survival analysis and finding viable treatment options.

Introduction

The methodology to properly treat patients and gauge the effectiveness of treatments takes a very meticulous approach. The importance of this is especially magnified when the diagnosis of patients is of critical scale, such as prostate cancer. There are numerous factors that go towards analyzing patterns in survival of these patients and the effectiveness of treatments thereafter. Such factors are not limited to; treatment options, patient age, and patient medical history. These variables can potentially be used to predict how long a patient will survive and to obtain predictive insight on the effectiveness (or lack of) of the treatment options. This is of utmost importance as the primary motive for this study is to obtain knowledge that can potentially help to improve the survival rate of cancer patients. There is no need to justify the significance of the ability to accurately treat these patients and we hope to enhance that crucial ability. This data can help to improve treatment methods and hopefully increase survival (if ever so slightly) for these patients. Even the smallest bit of change in chance of survival makes an immense impact. This study examines a prostate cancer patient dataset from a randomized trial comparing four treatments for state 3 and 4 prostate cancer, with almost equal numbers of patients on placebo and each of three doses of estrogen. The data comes from the book *The Choice of Treatment for Cancer Patients Based on Covariate Information: Application to Prostate Cancer* by D.P. Byar and S.B. Green. It contains the necessary features and content in order to conduct proper survival analysis, and produce both a Kaplan-Meier estimation (non-parametric) of survival as well as a semi-parametric Cox Proportional Hazards predictive model. Implemented treatment options, size of tumor, and blood pressure readings are logically likely to influence survival, however these and all other variables will be explored to determine if their effects are significant enough to be included in the model. The organization for the rest of the report is as follows. Section 2 will contain some data characteristics and exploration for choosing significant variables. Section 3 will delve into a discussion and comparison of the selected model and survival estimation. Finally, concluding remarks and recommendations will be found in section 4 along with a brief overview of the effectiveness of treatments.

Data Characteristics and Exploration

```
library(survival)
library(tidyverse)
library(rpart)
library(survminer)
prostate = read.csv("prostate.tsv", sep = "\t", header = TRUE)
any(is.na(prostate))
sum(is.na(prostate))
```

The data is part of the book *The Choice of Treatment for Cancer Patients Based on Covariate Information: Application to Prostate Cancer* by D.P. Byar and S.B. Green. As stated in the introduction, it is based on a randomized trial comparing four treatments for state 3 and 4 prostate cancer, with almost equal numbers of patients on placebo and each of three doses of estrogen. 19 observations containing missing values have been observed to have missing

values. As this study essentially concerns life vs. death, it is crucial that the data will not induce bias or reduce statistical power. Recursive partitioning will be used to fill in these missing values. The outcome of interest is whether a patient is alive or dead, however as we are not interested in cause of death for this study, we will create a new binary categorical variable called 'dead' where 1 = "dead" and 0 = "alive". Survival time is also a variable of interest and we will combine months of follow-up (dtime) with the new binary response as our target variable when constructing our hazards model. We have 16 other variables describing the patients, all of which are defined in the table below.

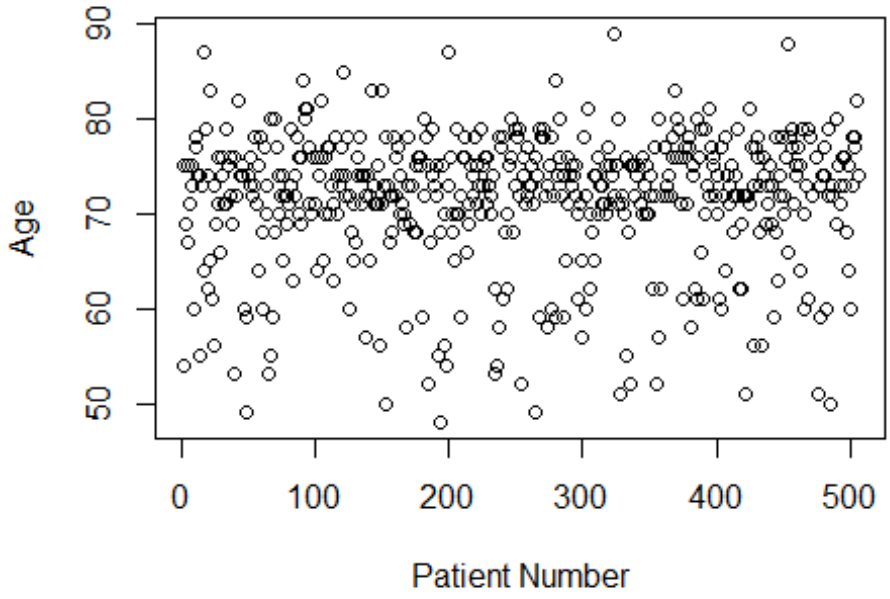
Item	Variable	Definition
1	patno	Patient number
2	stage	Stage
3	rx	Treatment
4	dtime	Months of follow-up
5	status	Follow-up status
6	age	Age in years
7	wt	Weight index = wt(kg) - ht(cm) + 200
8	pf	Performance rating
9	hx	History of cardiovascular disease
10	sbp	Systolic blood pressure /10
11	dbp	Diastolic blood pressure /10
12	ekg	Electrocardiogram code
13	hg	Serum Hemoglobin (gr / 100 ml)
14	sz	Size of primary tumor (cm squared)
15	sg	Combined index of stage and hist. grade
16	ap	Serum prostatic acid phosphatase
17	bm	Bone metastases
18	sdate	Date on study (as days since January 1, 1960)
19	dead	Dead/alive indicator (1 = dead, 0 = alive)

Data Characteristics

It is apparent that the variable pertaining to patient number contains a unique value for each patient. We will use this variable to analyze the distribution of each numerical variable in order to assess data quality by plotting each against patient number. Following this process, we will briefly discuss the categorical variables.

A plot of patient age shows that a majority of patients are between 70 to 80 years of age while far less are outside of that range. As a result of this strange distribution, we will consider binning this variable into levels, each defined as a range of ages in order to recognize clearer and simpler patterns. There are no outliers. The age plot can be seen here:

```
plot(prostate$patno, prostate$age, xlab = "Patient Number", ylab = "Age")
```



The weight index variable (wt) is calculated by taking a patients height in centimeters, subtracting it from their weight in kilograms, and then adding a constant value of 200. The plot displays some outliers on the higher end. Thus, we will also consider binning or log transforming this variable to account for the unusually high values.

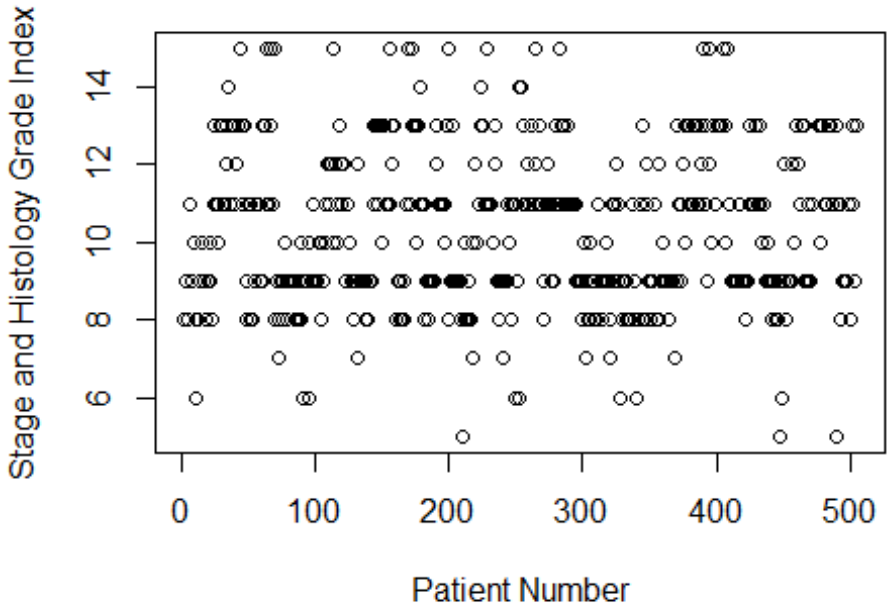
The two blood pressure variables, systolic (sbp) and diastolic (dbp), consist of a smaller range of whole number values. Shier minimum and maximum values are 8 and 30, and 4 and 18 respectively.

The variable defined as serum hemoglobin (hg) is a blood test that measures the level of free hemoglobin in the liquid part of blood. This is a real number value and has a range of 5.899 to 18.199. The plot against patient number determines outliers on the lower spectrum. We may transform this variable as a result.

The size of the primary tumor (sz) is measured in cm and also contains outliers. Binning or log transforming this variable will once again be considered.

The variable (sg), which is defined as the combined index of stage and histology grade, displays a strange plot. Converting this integer variable to categorical seems logical. The plot can be found below.

```
plot(prostate$patno, prostate$sg, xlab = "Patient Number", ylab = "Stage and
Histology Grade Index")
```



The variable concerning serum prostatic acid phosphatase (ap) contains a very strange distribution and obviously contains outliers. A vast majority of values appear closer to 0 and others reach as high as 999.88. We will very likely have to transform this variable. Defining this variable in context, it is simply the levels of enzyme produced by the prostate. Men with prostate cancer are associated with increased levels.

Our last integer variable refers to the date on study (days since Jan 1, 1960) and is denoted by (sdate). However this variable is not relevant and will be excluded from model consideration.

The stage variable simply represents the present stage of prostate cancer for the given patient, denoted either stage 3 or 4. There is even spread across both values, which is expected. As this is merely an integer variable, we will convert it to categorical and call it "stage.cat" to retain the original variable.

```
prostate$stage.cat = as.factor(prostate$stage)
```

For treatment option (rx) we have four categories labeled 0.2 mg estrogen, 1.0 mg estrogen, 5.0 mg estrogen, placebo. These labels define themselves and even in spread. Obviously, this will be an important variable in gauging treatment effectiveness. We will also change this variable so that 'Placebo' is the reference level. This will allow us to easily compare the effects of different treatment options and draw better conclusions.

```
prostate$rx = fct_relevel(prostate$rx, "placebo")
```

The variable pertaining to history of cardiovascular disease (hx) is merely a binary indicator variable with "1" indicating history and "0" as no history. Like the stage variable, we will convert it to categorical

```
prostate$hx.cat = as.factor(prostate$hx)
```

Performance rating (pf) is determined by four levels named confined to bed, in bed < 50% daytime, in bed > 50% daytime, normal activity. We will modify this variable so that "normal activity" is the base level for ease of interpretation purposes.

```
prostate$pf = fct_relevel(prostate$pf, "normal activity")
```

Electrocardiogram code (ekg) contains eight levels named , benign, heart block or conduction def, heart strain, normal, old MI, recent MI, rhythmic disturb & electrolyte ch. As we can see, one category is blank which is odd. The table distribution for this variable also shows us that there are 8 patients that fall within this category. It should also be noted that for "recent MI", there is just a single patient. The level "Normal" will be changed to the basis level. Should this variable be found to be a significant predictor, we will consider possible transformations to address these strange values.

```
prostate$ekg = fct_relevel(prostate$ekg, "normal")
```

The last categorical variable referring to bone metastases (bm) is binary with "1" being true and "0" meaning false for the condition. This indicates whether the cancer cells have relocated to the bone.

```
prostate$bm.cat = as.factor(prostate$bm)
```

To create our response variable 'dead', we will transform the categorical status variable. This variable has 10 categories named alive, dead - cerebrovascular, dead - heart or vascular, dead - other ca, dead - other specific non-ca, dead - prostatic ca, dead - pulmonary embolus, dead - respiratory disease, dead - unknown cause, dead - unspecified non-ca. As we can see, this variable identifies the cause of death, which we are not interested in. We will turn status into a binary variable with 1 = 'dead' (indicating any of the nine dead levels regardless of cause), and 0 = 'alive'.

```
prostate$dead <- ifelse(prostate$status == "alive", 0, 1)
```

Addressing Missing Values

We will now briefly explain and perform recursive partitioning as an imputation method to fill in the 19 observations with missing values. For the combined stage and histology grade index variable (sg), it was observed that 11 observations had missing values. A new index variable was created called 'i.sg', which contains all of the original non-missing values in addition to new predicted values for the other 11 observations. This process defines surrogate variables (other independent variables) which are used to estimate the missing data values.

```
(rp <- rpart(sg ~ age + wt + hx + sbp + dbp + hg + sz + ap + bm,
             data = prostate))
```

```
idx <- which(is.na(prostate$sg))
prostate$i.sg <- prostate$sg
prostate$i.sg[idx] <- round(predict(rp, newdata = prostate[idx,]),0)

prostate[idx, c("ap", "sz", "bm", "sg", "i.sg")]

##             ap sz bm sg i.sg
## 2    0.6999512 42  0 NA   11
## 57   0.7999268 41  0 NA   11
## 123  0.1999817 17  0 NA    9
## 125  0.5000000 13  0 NA    9
## 169  4.5996094 27  1 NA   13
## 336  0.5999756  4  0 NA    9
## 418  0.8999023 22  0 NA    9
## 436  0.3999634 19  0 NA    9
## 472  1.0998535  5  0 NA   12
## 481  1.3999023 20  0 NA   12
## 502 22.1992188 33  1 NA   13

rm(idx, rp)

prostate$sg.cat = as.factor(prostate$i.sg)
```

We can see from this table that the missing data has been filled with predicted values. To judge the performance of these values, we can compare the means of the new imputed-values variable (i.sg) and the original variable. We can see below that the variable means are consistent, indicating that the imputed values do not significantly change the distribution of the old variable while increasing representativeness and potential statistical power.

```
rbind("original sg" = summary(prostate$sg),
      "imputed  sg" = c(summary(prostate$i.sg), 0))

##             Min. 1st Qu. Median     Mean 3rd Qu. Max. NA's
## original sg    5       9     10 10.30957   11.00   15   11
## imputed  sg    5       9     10 10.31673   11.75   15    0
```

This method will be replicated for size of tumor (sz), age, and weight index (wt) so that the dataset is free of missing values. These will be named 'i.sz', 'i.age' and 'i.wt'. It should be noted that these new variables will effectively render the old variables useless and thus will not be used in this study going forward (i.e 'i.age' becomes primary age variable).

```
(rp <- rpart(sz ~ age + wt + hx + sbp + dbp + hg + sg + ap + bm,
             data = prostate))

idx <- which(is.na(prostate$sz))
prostate$i.sz <- prostate$sz
prostate$i.sz[idx] <- round(predict(rp, newdata = prostate[idx,]),0)

prostate[idx, c("ap", "sg", "bm", "sg", "sz", "i.sz")]
```

```
rm(idx, rp)

rbind("original sz" = summary(prostate$sz),
      "imputed  sz" = c(summary(prostate$i.sz), 0))

rp <- rpart(age ~ wt + hx + sbp + dbp + hg + sg + ap + bm + sz,
            data = prostate)

idx <- which(is.na(prostate$age))
prostate$i.age <- prostate$age
prostate$i.age[idx] <- round(predict(rp, newdata = prostate[idx,]),0)

prostate[idx, c("hg", "sg", "sz", "hx", "ap", "dbp", "sbp", "age", "i.age")]
rm(rp, idx)

rbind("original age" = summary(prostate$age),
      "imputed  age" = c(summary(prostate$i.age), 0))

rp <- rpart(wt ~ age + hx + sbp + dbp + hg + sg + ap + bm + sz,
            data = prostate)

idx <- which(is.na(prostate$wt))
prostate$i.wt <- prostate$wt
prostate$i.wt[idx] <- round(predict(rp, newdata = prostate[idx,]),0)

prostate[idx, c("age", "dbp", "hg", "sz", "ap", "bm", "wt", "i.wt")]
rm(rp, idx)

rbind("original wt" = summary(prostate$wt),
      "imputed  wt" = c(summary(prostate$i.wt), 0))
```

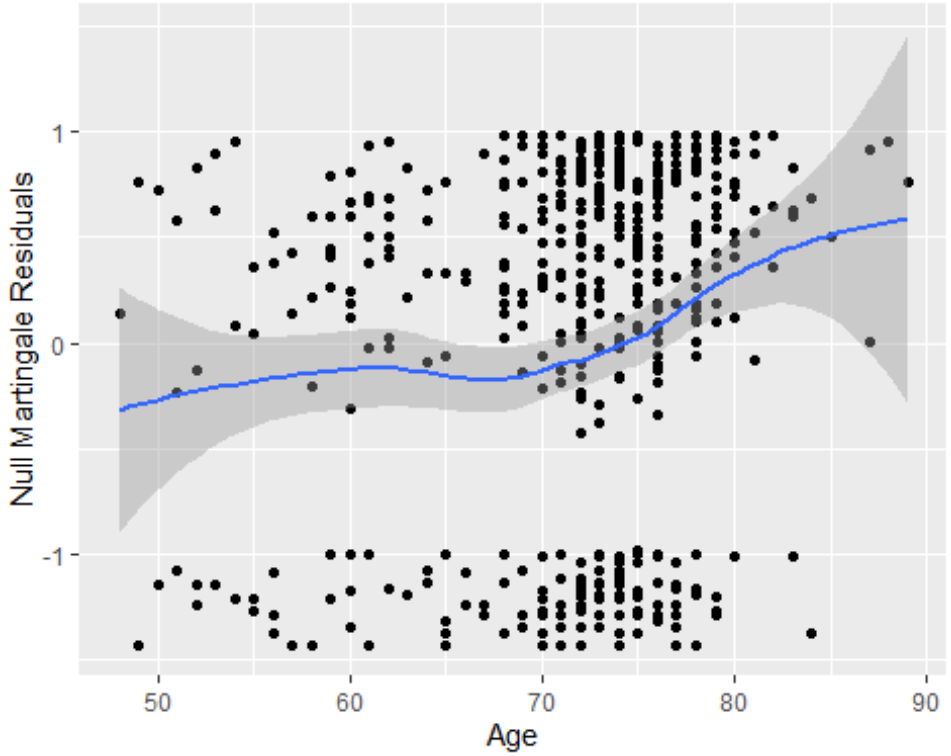
Data Exploration

Next, we will shift focus to exploration of potential predictors in order to determine whether any display association to the response variable. We will do this by first generating a null Cox Proportional Hazards model and calculating Martingale residuals from this null model (named 'null.mr'). Each quantitative variable will then be plotted against these residuals. For categorical variables, a boxplot against the residuals will be produced. An example of each plot can be found below for age (i.age) and performance (pf) variables respectively.

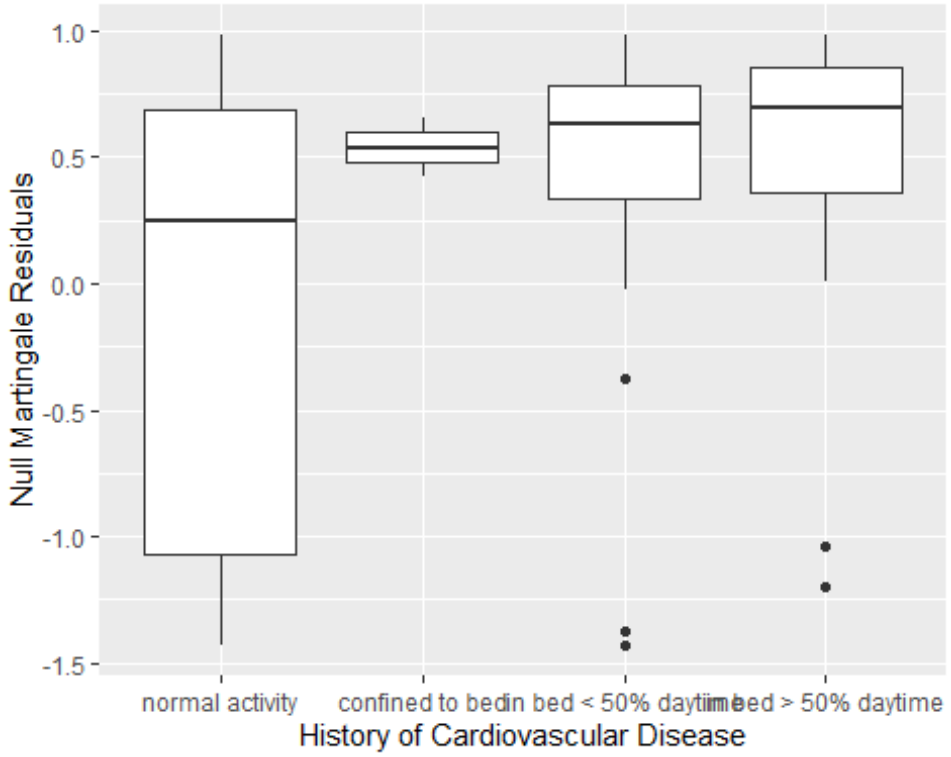
```
cph.null <- coxph(Surv(dtime, dead) ~ 1,
                  data = prostate)

prostate$null.mr <- residuals(cph.null, type = "martingale")

ggplot(prostate) +
  aes(x = i.age, y = null.mr) +
  geom_point() +
```

```
  geom_smooth() +
  labs(x = "Age", y = "Null Martingale Residuals")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggplot(prostate) +
  aes(x = factor(pf), y = null.mr) +
  geom_boxplot() +
  labs(x = "History of Cardiovascular Disease", y = "Null Martingale
Residuals")
```



For the quantitative variable plots, we are simply examining the horizontal blue line. A flatter, straight line would indicate that the plotted variable does not have a significant influence on our response variable (survival) while any curvature suggests it does. This age plot identifies with the latter, which supports an association and we will accept age as a significant predictor. As for the boxplots, we are looking at the difference in means between the sublevels of the variable. In this case the means do appear to be different, suggesting this variable is significant. We can also analyze each variable further by creating dummy models of determined significant variables to gauge the scale of effect on survival as performed below.

```
cph.null0 <- coxph(Surv(dtime, dead) ~ rx + pf,
                   data = prostate)
cph.null0

## Call:
## coxph(formula = Surv(dtime, dead) ~ rx + pf, data = prostate)
##
##                             coef exp(coef) se(coef)     z      p
## rx0.2 mg estrogen         0.0330    1.0335   0.1451  0.23 0.8203
## rx1.0 mg estrogen        -0.3389    0.7125   0.1579 -2.15 0.0318
## rx5.0 mg estrogen         0.0117    1.0118   0.1469  0.08 0.9364
## pfconfined to bed         0.7663    2.1519   0.7165  1.07 0.2848
## pfin bed < 50% daytime    0.5802    1.7864   0.1851  3.13 0.0017
## pfin bed > 50% daytime    0.6871    1.9880   0.3076  2.23 0.0255
##
```

```
## Likelihood ratio test=22.55  on 6 df, p=0.001
## n= 502, number of events= 354
```

As the coefficients of each level of performance are positive and rather large, we can effectively conclude that any form of bed-confinement negatively impacts survival time and capability. However for each variable we must always ensure we think more critically. Age and performance rating are logically expected to be good predictors as overall health naturally declines with age and increased time spent bed-ridden does not exactly provide assurance for survival (and survival time). We can make these types of assumptions for all variables and these plots may provide either supporting or nonsupporting evidence for such logical reasoning. This process was conducted for all variables.

In total, 10 variables were found to have a significant influence on survival. These variables were age, performance rating, stage, weight index, history of cardiovascular disease, electrocardiogram code, serum hemoglobin, size of tumor, stage/histology index, and bone metastases indicator. It should be mentioned that many transformations of numerous variables were performed and analyzed in the exploration of potential predictor variables, and two were found significant. One consisted of level grouping the electrocardiogram code variable, simplifying it to have only two levels; normal and abnormal. The other was binning the numerical stage/histology index variable into 3 categorical levels. It was determined that these two transformed variables along with the other eight variables could be useful predictors for survival.

```
prostate <- prostate %>%
  mutate(dbp.cat = case_when(
    dbp <= 5 ~ "<6",
    dbp <= 6 ~ "6",
    dbp <= 7 ~ "7",
    dbp <= 8 ~ "8",
    dbp <= 9 ~ "9",
    dbp <= 10 ~ "10",
    dbp <= Inf ~ ">10"))
prostate$dbp.cat <- factor(prostate$dbp.cat,
                    levels = c("<6","6","7", "8", "9", "10", ">10"))

prostate$b.ekg = prostate$ekg
prostate$b.ekg = fct_collapse(prostate$b.ekg,
                              abnormal = c("", "benign", "heart block or
conduction def",
                                           "heart strain", "old MI", "recent
MI",
                                           "rhythmic disturb & electrolyte
ch"),
                              normal = "normal"
)

prostate <- prostate %>%
  mutate(i.sg.cat = case_when(
    i.sg <= 10 ~ "(5,10]",
```

```
    i.sg <= 12 ~ "(10,12]",
    i.sg <= 15 ~ "(12,15]"))
prostate$dbp.cat <- factor(prostate$dbp.cat,
                           levels = c("<9","(9,10]","(10,12]", "(12,15]"))
prostate$i.sg.cat = fct_relevel(prostate$i.sg.cat, "(5,10]")
```

The next section will discuss the best potential proportional hazards model from these variables in addition to a non-parametric estimate of survival for treatment options.

Model Selection & Interpretation

Section 2 determined that there is relevant association between survival and patient characteristics variables. It also described how some of these variables could be transformed to provide better predictors for a Cox Proportional-Hazards model. This model is intended to be useful in determining which variables are most influential to survival time. It can also help us to predict how long a patient will survive. This section summarizes the selection of the best model and will interpret the final model and its results in context. Additionally, a Kaplan-meier estimate will be produced which will help to describe the effectiveness of treatment options. Finally, this section will touch base on the utilization and benefits of this model in the professional world.

Kaplan-Meier Estimate for Treatment Options

Our focus will first shift towards the effectiveness of treatment options. We can produce a non-parametric Kaplan-Meier estimate in order to measure the fraction of patients living for a certain amount of time after different treatments. Our treatment options are placebo, 0.2mg estrogen, 1.0mg estrogen and 5.0mg estrogen. In order to gauge the effectiveness of treatments, we can first generate an overall survival estimate using the entirety of the data:

```
(kapm <- survfit(Surv(dtime, dead) ~ 1, data = prostate))

## Call: survfit(formula = Surv(dtime, dead) ~ 1, data = prostate)
##
##       n  events  median 0.95LCL 0.95UCL
##     502     354      34      30      39
```

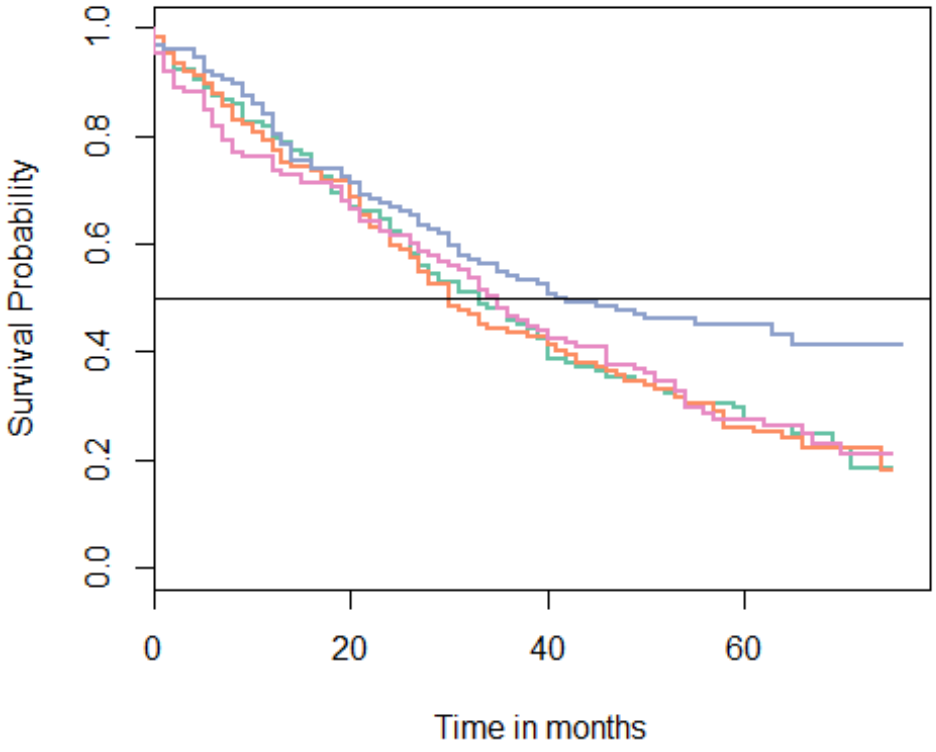
From this output we can see that the overall median survival time is 34 months. In other words, the survival probability of a patient is 50% when time is 34 months. Now, we can generate an estimate using treatment options (rx) to visualize the different impacts.

```
kapm.rx <- survfit(Surv(dtime, dead) ~ rx,
        data = prostate)
kapm.rx

## Call: survfit(formula = Surv(dtime, dead) ~ rx, data = prostate)
##
##                       n events median 0.95LCL 0.95UCL
## rx=placebo          127     95   33.0      27      40
## rx=0.2 mg estrogen 124     95   30.0      26      41
```

```
## rx=1.0 mg estrogen 126     71   41.5      31      NA
## rx=5.0 mg estrogen 125     93   35.0      28      46
```

From this, we can see that median survival time changes for each treatment dosage. This estimate suggests that both placebo and 0.2mg estrogen are rather ineffective as both are less than the overall time of 34 months. However, 1.0mg and 0.5mg are indeed effective. From these statistics, it is evident that the 1.0mg estrogen treatment is the most viable. We can visualize entirety of survival probability and time in the chart below:

```
par(mar = c(4,4,1,1))
plot(kapm.rx,
     xlab = "Time in months",
     ylab = "Survival Probability",
     col = c("#66c2a5", "#fc8d62", "#8da0cb", "#e78ac3"),
     lwd = 2)
abline(h = 0.5)
```



In this plot, the orange line represents 0.2mg, green is placebo, pink is 5.0mg, and blue is 1.0mg. This provides us with more insight in regards to treatment effectiveness. It involves the computing of probabilities of death occurrence at a certain point in time and multiplying these successive probabilities by any earlier computed probabilities to get the final estimate shown above. Despite somewhat similar effectiveness in the short-term, we can see that patients receiving 1.0mg estrogen appear to maintain a survival probability of approximately 0.4 (40%) after the 60th month, whereas the remaining three treatments dip well below 40% and approach 20% survival probability. This is enough evidence to

conclude 1.0mg estrogen as the most effective treatment. We will now shift focus towards building a Cox Proportional-Hazards model and will also include this variable in our model.

Cox Proportional-Hazards Model for Survival

From the significant variables found in section 2, a Cox Proportional-Hazards model was fitted. Out of these 11 predictors; treatment (rx), weight (i.wt), history of cardiovascular disease (hx.cat), electrocardiogram code (b.ekg), serum hemoglobin (hg), tumor size (i.sz), and stage/histology index (i.sg.cat) were included in the final model for a total of 7 systematic components. These were ultimately determined to be the most significant and influential for survival. The remaining 4 variables of age, performance rating, stage and bone metastases were omitted. See Table *A1* for the model's full summary statistics. A table of exponentiated coefficients for the model (also known as hazard ratios) can be found below:

```
cph.mod1 <- coxph(Surv(dtime, dead) ~ rx + i.wt + hx.cat + b.ekg + hg +
                    i.sz + i.sg.cat, data = prostate)

exp(coef(cph.mod1))
```

```
## rx0.2 mg estrogen rx1.0 mg estrogen rx5.0 mg estrogen             i.wt
##         0.9863036         0.6642947         0.9950472        0.9885932
##           hx.cat1     b.ekgabnormal                hg             i.sz
##         1.6829616         1.4525919         0.9276110        1.0151469
##   i.sg.cat(10,12]   i.sg.cat(12,15]
##         1.4360956         1.6686595
```

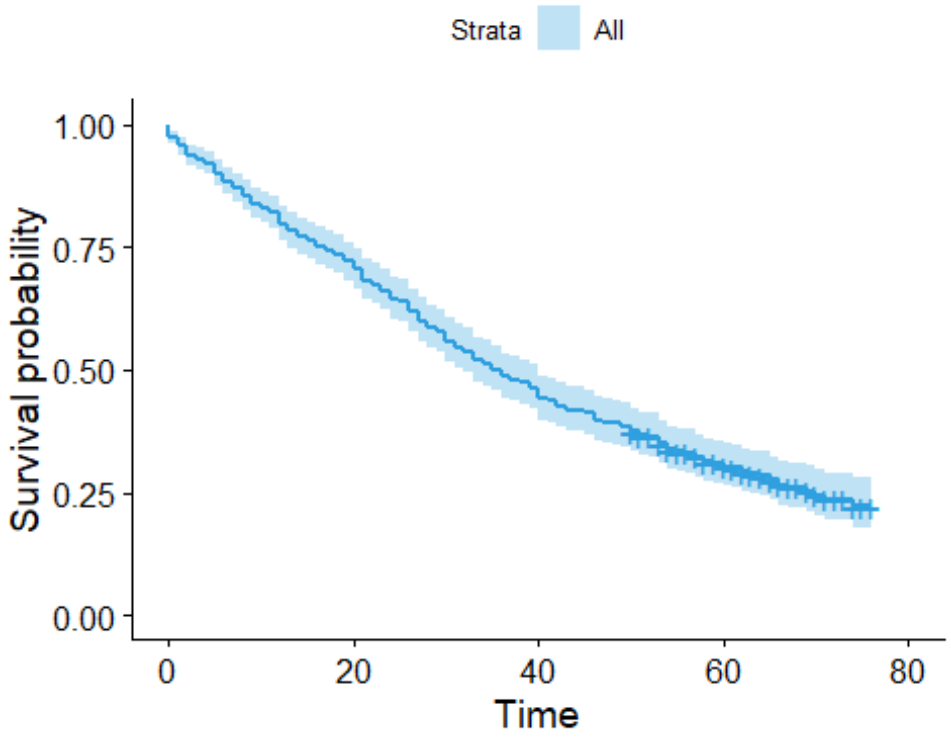
These ratios give the effect size for these covariates. For interpretation purposes, let us consider the coefficient for history of cardiovascular disease (hx.cat) holding all other variables constant. As this is a binary categorical variable, this coefficient indicates that a '1' denoting history of cardiovascular disease would increase the hazard by a factor of 1.68 (or 68%) and we can conclude that having such history for disease is strongly associated with increased risk of survival. In appendix table *A1* we can also see that the variable's p-value is low at 2.48e-6 which also indicates a strong relationship between the two. The interpretation for all included variables can be viewed in similar fashion. It is important to remember that positive coefficients increase hazard and negative coefficients reduce it. Variables with negative coefficients would hence be more desirable for survival, a formula which is supported by the 1.0mg estrogen treatment option with negative coefficient 0.66. Remember, we found this treatment option to be most effective in our Kaplan-Meier estimate results and these coefficients can provide us with more statistical insight of the exact effect they provide for survival.

From Table *A1* we can see that the p-values for all included components is low and significant for at least the 0.05 significance level (save for treatment options). The reasoning for this can be explained by the individual confidence intervals found in the same table. Because the intervals for 0.02mg and 5.0mg contain a value of 1, it indicates that these treatment options make a smaller contribution to the difference in the hazard ratio after adjusting for other variables. Additionally, the hazard ratios for the two treatments

are 0.98 and 0.99 respectively, which concludes that these treatment options only reduce daily hazard of death by a factor of 2% and 1% which are not significant contributions. 1.0mg estrogen is clearly the most statistically significant and effective treatment option.

Let us now consider the overall significance of our entire model by briefly looking at the three alternative tests for overall model significance. In Table *A1* these are the likelihood ratio test, wald test and score (logrank) test. The methods are equivalent, with values of 94.19, 91.21 and 92.55 respectively which is desirable and indicative of good model strength. This justifies the decision to remove age, performance rating, stage and bone metastases and in turn reduces model complexity. Below, we can obtain a visual of survival time and probability for this model, using the mean values of all covariates.

```
ggsurvplot(survfit(cph.mod1), data = prostate, color = "#2E9FDF")
```



Summary and Concluding Remarks

It is inevitable that all patients diagnosed with prostate cancer will be different and possess unique medical traits. However it is certainly possible to establish significant determinants of survival and to distinguish a single treatment which generally maximizes survival. The best-fitted Cox Proportional-Hazards model concluded that survival can be estimated by patient treatment dosage, weight index, history of cardiovascular disease, the grouped normal vs. non-normal electrocardiogram code, serum hemoglobin, size of primary tumor, and also the stage/histology grade index after being split into a specific range of values.This was motivated by evidence that these transformed variables better fit the variation in

response variable data and that fewer explanatory variables would provide more accurate predictions. One might logically argue that other variables could be included such as age, however these variables were found to insignificant contribution to survival estimation when adjusting for other determinants.

The non-parametric Kaplan-Meier estimate and Cox Proportional-Hazards model are both very useful methods in analyzing survival. The purpose of this study is not to depreciate one over the other, but rather to exhibit the functionality of each. A Kaplan-Meier estimate would be most appropriate in measuring the scale of a single intervention. In this study, the intervention of interest is treatment and we were able to conclude 1.0mg estrogen as the most effective treatment for increasing survival probability. Meanwhile, the hazards model provides us with the ability to account for multiple other factors, as is frequently the case when dealing with medical patients. There are oftentimes several known covariates which can account for patient prognosis, many of which are provided by this dataset. Thus, it is always of utmost importance to adjust for the impacts that these variables bring. These two methods are meant to co-exist to increase the capacity to which we can draw conclusions for survival estimation.

This study was based on a randomized trial of 502 prostate cancer patients and contained sufficient variables relating to patient characteristics that may prove to influence survival. This allowed us to build complex models in order to determine the best for estimating survival, survival time and best treatment options.

The analysis of this data is based on a randomized trial of prostate cancer patients comparing treatment options. There will always be strange circumstances, however the methods explored in this report should be directly applicable to a majority of prostate cancer patients. This in turn will greatly increase the ability to effectively treat patients and to assess survival rates at given points in time.

References

D.P. Byar and S.B. Green, **The Choice of Treatment for Cancer Patients Based on Covariate Information: Application to Prostate Cancer**, Bulletin Cancer, Paris, 67:477-488, 1980.

Appendix

Table **A1**

```
cph.mod1 <- coxph(Surv(dtime, dead) ~ rx + i.wt + hx.cat + b.ekg + hg +
                    i.sz + i.sg.cat, data = prostate)
summary(cph.mod1)

## Call:
## coxph(formula = Surv(dtime, dead) ~ rx + i.wt + hx.cat + b.ekg +
##     hg + i.sz + i.sg.cat, data = prostate)
```

```
##
##   n= 502, number of events= 354
##
##                         coef exp(coef)  se(coef)      z Pr(>|z|)
## rx0.2 mg estrogen -0.013791  0.986304  0.146318 -0.094 0.924907
## rx1.0 mg estrogen -0.409029  0.664295  0.158779 -2.576 0.009992 **
## rx5.0 mg estrogen -0.004965  0.995047  0.149592 -0.033 0.973522
## i.wt              -0.011472  0.988593  0.004355 -2.634 0.008436 **
## hx.cat1            0.520555  1.682962  0.110528  4.710 2.48e-06 ***
## b.ekgabnormal      0.373350  1.452592  0.118063  3.162 0.001565 **
## hg                -0.075143  0.927611  0.029926 -2.511 0.012041 *
## i.sz               0.015033  1.015147  0.004496  3.344 0.000826 ***
## i.sg.cat(10,12]    0.361928  1.436096  0.129168  2.802 0.005079 **
## i.sg.cat(12,15]    0.512021  1.668659  0.147958  3.461 0.000539 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## rx0.2 mg estrogen    0.9863     1.0139    0.7404    1.3139
## rx1.0 mg estrogen    0.6643     1.5054    0.4866    0.9068
## rx5.0 mg estrogen    0.9950     1.0050    0.7422    1.3341
## i.wt                 0.9886     1.0115    0.9802    0.9971
## hx.cat1              1.6830     0.5942    1.3552    2.0900
## b.ekgabnormal        1.4526     0.6884    1.1525    1.8308
## hg                   0.9276     1.0780    0.8748    0.9836
## i.sz                 1.0151     0.9851    1.0062    1.0241
## i.sg.cat(10,12]      1.4361     0.6963    1.1149    1.8498
## i.sg.cat(12,15]      1.6687     0.5993    1.2486    2.2300
##
## Concordance= 0.648  (se = 0.017 )
## Rsquare= 0.171   (max possible= 1 )
## Likelihood ratio test= 94.19  on 10 df,   p=8e-16
## Wald test            = 91.21  on 10 df,   p=3e-15
## Score (logrank) test = 92.55  on 10 df,   p=2e-15
```