**STA 723 HW07 Report** **Zach White**

# Executive Summary

Right heart Catheterization (RHC) is a procedure designed to measure how well the heart is pumping blood to the lungs. We analyze data on RHC procedures from five different medical centers in the United States for two purposes. First, we use a linear regression model with interactions to understand what factors contribute to an increased prognosis score after receiving RHC. Second, we use standardized mortality weighting to weight the propensity score to understand the treatment effect of RHC for the treated group (ATT). In general, we find that RHC is associated with lower survival rates, except in certain conditions, where we find that RHC is associated with higher prognosis scores.

# Introduction

Right Heart Catherization is a a procedure where a catheter is inserted into an artery (generally the femoral artery) and directed to the right atrium and ventricle of the heart to measure how well the heart is pumping blood. Although RHC can have good diagnostic properties due to its direct nature, there is debate whether or not the use of RHC leads to improved care due to its invasive nature. The goal of this analysis is two -fold. First, we seek to answer the question of whether or not RHC should be performed to assist treatment choices for new patients in order to maximize their individual prognosis scores. Second, we want to understand the average treatment effect of RHC for the treatment group.

In total, our data set contains 5735 subjects and were collected from five major medical centers in the United States. We have 55 total variables. These variables range from general information like age, sex, and race to health-related information like primary category disease, mean blood pressure, and heart-rate. Most of the covariates are shared across our two different goals. However, the response variables vary according to our overall goal. For our first problem, prediction and individual characteristics associated with an improvement, the response variable is the prognosis scores. However, for the second component where we are estimating average treatment effect (ATT), the response variable of interest is whether or not the individual died after 30-days.

The first question of interest can be seen as a double prediction. We need to build a model suited for prediction and then use that model for prediction where both when RHC is applied. We can then analyze the difference between applying RHC and not applying. Analyzing the covariates for the individuals where we predict an increase or decrease in the prognosis score, we can understand what factors can contribute to RHC positively impacting an individual.

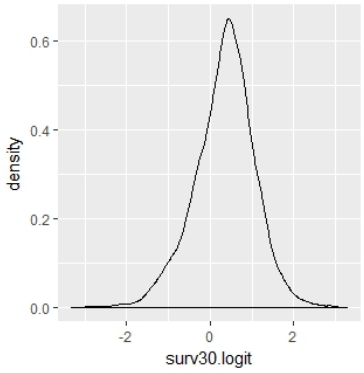
The goals for the second question are clearly different, and our analysis also has two parts. We first calculate propensity score, which is the probability that an individual is assigned the treatment of RHC. We then use that score to balance the treated and untreated patients and fit another model to estimate the average treatment effect on the treated. Since each of the components has different goals and challenges, we first analyze the first component and then analyze the second component. However, since there are shared covariates, we choose to share the exploratory data analysis for the most part.

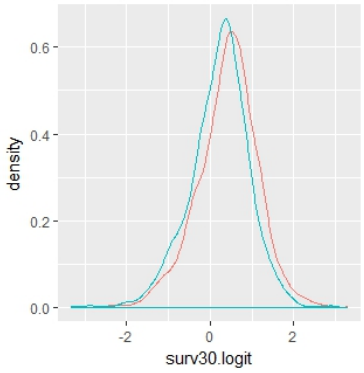
# Analysis

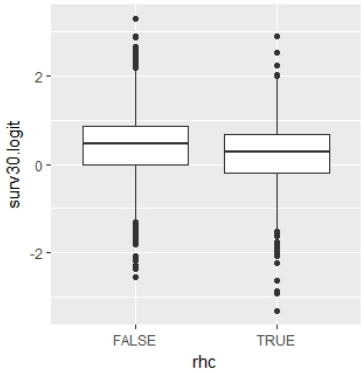
### Exploratory Data Analysis

For the first component, the response variable is a prognosis score, which is between zero and one. It can be interpreted as a probability of survival. For this reason, we use the logit transform, and we explore and model the $\log(\frac{\text{prognosis}_i}{1-\text{prognosis}_i})$. From now on, when we

discus prognosis score, we mean the logit transformed prognosis score unless explicitly state otherwise. After we perform this transform, there is one very clear outlier in the prognosis score. Although not ideal, we choose to remove this observation from our data set. Another thing of note in the first component is that we are dealing with 53 covariates, and thus, we don't include all of the exploratory plots, but rather discuss some overall features and points of note. The following plots show both marginal effects of the.




The first plot above shows the marginal distribution for the prognosis score. This distribution should approximate a normal distribution, and it is clear it does. The second plot shows the distribution of the prognosis scores given the RHC status where blue represents an RHC status of true while red represents an RHC status of false. According to this plot, the patients to whom RHC was administered actually have a slightly lower prognosis score, which could indicate that RHC is detrimental to patients. The boxplot on the right confirms this. It appears that RHC is associated with a lower prognosis score. However, we wait to see if this is significant in our analyses.

For the second analysis, the figures shown above do not explore the goals for this analysis because the response of interest is death within 30 days, which is a binary response. Thus, these summaries are not applicable in this case. Analyzing the table shows results similar to the above plots. There seems to be a higher proportion of individuals that receive RHC die.

| | | Living | Death | |
|---|---|---|---|---|
| RHC | Control | 1088 | 2463 | 3551 |
| | Treatment | 830 | 1354 | 2175 |
| | Total | 1918 | 3817 | 5735 |

Overall, when we analyze the covariates' relationship with the different response variables, there aren't any glaring trends. The covariates are of different types: binary, categorical, numerical, and ordinal. There are some variables that are clearly associated with the prognosis score. The variable of probability of surviving two months is clearly very correlated with the prognosis score, which makes sense since the prognosis score is a probability of surviving one month.

**Model Choice: Component 1**

An important feature that we seek to understand and model is the heterogeneity of the study. We are interested in how individuals respond to the treatment RHC. We want to know what factors contribute to RHC being an effect diagnostic tool associated with higher prognosis scores. In order to do this, an essential feature of the model is the ability to have interactions between RHC and the other covariates. In this way, we can understand how RHC impacts an individuals prognosis score in the presence of other covariates. Even though these interactions are an important part of the study, we compare the mean squared error (MSE) for the different models we propose. We compare MSE for OLS with and without interactions, Random Forests, and Bayesian Additive Regression

Trees. Each of these techniques would allow us to achieve the goals of our first question, and we will choose the one with the lowest MSE.

The two OLS models that we will analyze are as follows:

$$\text{OLS: } \log(\frac{\text{prop}_i}{1-\text{prop}_i}) = \beta_0 + \beta 1 * \text{rhc}_i + \cdots + \beta_{53} * \text{wt0}_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

$$\text{OLS Int.: } \log(\frac{\text{prop}_i}{1-\text{prop}_i}) = \beta_0 + \beta 1 * \text{rhc}_i + \cdots + \beta_{53} * \text{wt0}_i + \beta_{1:2} * \text{rhc} \times \text{age} + \cdots + \beta_{1:53} * \text{rhc} \times \text{wt0} + \epsilon_i \epsilon_i \sim N(0, \sigma^2)$$

The first model includes all the covariates, and the second includes all covariates and the first order interaction terms that include rhc. Both of these models are quite simple, and for our goals, the latter will be better because of the interaction terms. We compare the MSE of these two models to that of Random Forests with different number of trees. Random Forests is a decision tree technique that can be used for regression regression. The formulation for BART is as follows:

$$Y = f(x_1, \ldots, x_p) + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$$
$$Y = h(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$
$$f(x) \approx h(x) = \sum_j^m g_j(x)$$

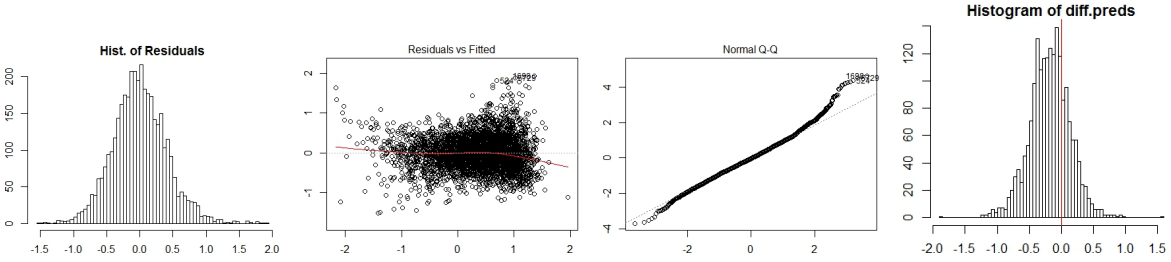
Under this construct, each $g_j(x)$ is a regression tree. BART is a way to naturally incorporate interaction effects, which ties into our goals well. This is a Bayesian approach, and thus we use Markov Chain Monte Carlo to approximate this.

| | OLS | OLS.Int | RF.100 | RF.1000 | BART.20 | BART.100 | BART.1000 |
|---|---|---|---|---|---|---|---|
| MSE | 0.194 | 0.173 | 0.415 | 0.404 | 0.219 | 0.203 | 0.193 |

It is clear that the model that minimizes MSE is the full model with e second-order interactions involving RHC included.

**Results: Component 1**

We proceed with this model, which yields the lowest MSE. However, we need to verify that this model is appropriate. Analyzing these residual plots, this model seems reasonable.

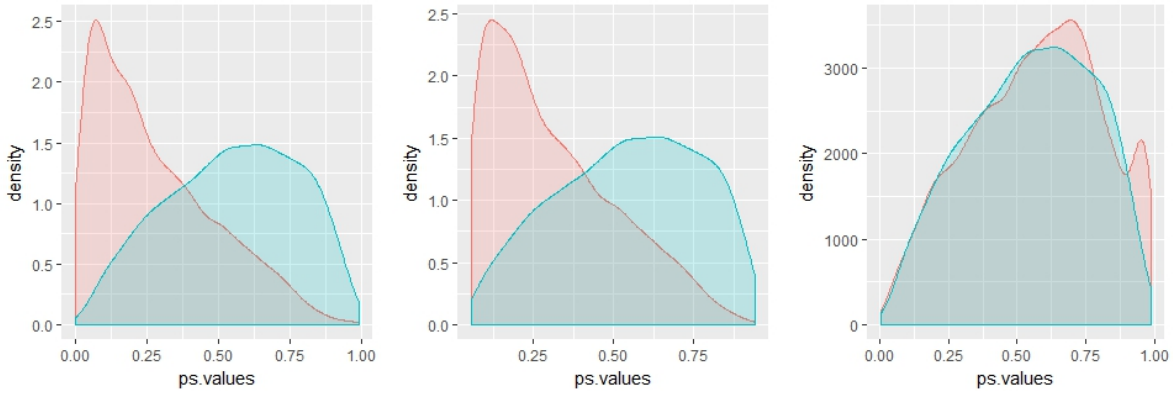



It seems linear, normal, and fairly homescedastic. However, the qq-plot, on the right, shows some deviation from the qq-line, which could indicate a lack of fit or a heavier tailed residual.

We perform double prediction in a sense. We set values for RHC to both true and false and predict values under both values. We then take the difference between treatment and control predicted values. The plot histogram on the right shows the distribution of these differences. It is clear that they are negative for the most part, which indicates that RHC can be associated with lower prognosis scores. When analyzing the values that are positive, we can see that many of the positive share some characteristics: male, immunosuppressant, transferred from another hospital, higher hematocrit count, and CHF.

**Model Choice: Component 2**

We now estimate average reatment effect on the treated (ATT), which is $E[Y_i(1) - Y_i(0)|W_i = 1]$. An important part of doing this i finding the propensity score, which is the probability of receiving the treatment: $e(X)Pr(W = 1|X)$ where X is the collection of covariate. We use logistic regression to estimate this propensity score, and we use this to balance the covariates between the treatment and control groups. We will use standardized mortality ratio weighting (SMRW) to help us balance the covariates. Under standardized mortality weights, if treatment is applied, the weight is 1, and if the individual is in the control group, the weight is $\frac{e(x)}{1-e(x)}$. When we estimate the propensity




scores, it is clearly unbalanced by the rhc status as shown in the left plot. We eliminate the bottom and top five quantiles of the propensity scores, which is shown in the middle. This seems more balanced, but when we add the weights, they are significantly more balanced. However, we need to perform a similar verification across all covariates. We can do this by performed t-tests across the deciles of the propensity scores. Under this weighting, the covariates do seem relatively balanced, and we proceed.

**Results: Component 2**

This weighting method does seem appropriate, and the propensity scores do seem to balance the covariates among the treatment and control groups. We then use logisic regression on death within 30 days of treatment with rhc and propensity scores with the SMR weights to estimate ATT. When we use this weights, the coefficient for the logistic regresion model is an estimate of ATT. When we transform the ATT into an appropriate scale, we find that it is -0.059 on average, which means that on the treated individuals, RHC lowered their probability of surviving nearly 6%, which is a significant amount.

## Conclusion

We find that in general, RHC is associated with lower prognosis scorse, but we do see more improvement with individuals with the following characteristics: male, immunosuppressant, transferred from another hospital, higher hematocrit count, and CHF. We also find the average treatment effect on the treated to be -0.059, which means that RHC decreases chances of survival on average. Thus, it is clear that RHC is associated with negative results. If we were to repeat this work, we would use the same multiple regression model with interactions, but we would work within the Bayesian framework to quantify the uncertainty on each individual more effectively. We would have bootstrapped for variance under both of these questions. Our weighting method handled the unbalanced covariates quite well. However, if it hadn't we could have included them in our final logistic regression model on death within 30 days, and the interpretation of the RHC coefficient still would have been ATT controlled for the unbalanced covariates.