Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods

Lu, Min

UNIVERSITY OF MIAMI

ESTIMATING INDIVIDUAL TREATMENT EFFECT IN OBSERVATIONAL DATA USING RANDOM FOREST METHODS

By

Min Lu

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

August 2018

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

ESTIMATING INDIVIDUAL TREATMENT EFFECT IN OBSERVATIONAL DATA USING RANDOM FOREST METHODS

Min Lu

Approved:

Hemant Ishwaran, Ph.D. Professor of Biostatistics	J. Sunil Rao, Ph.D. Professor of Biostatistics
Daniel Feaster, Ph.D. Associate Professor of Biostatistics	Guillermo J. Prado, Ph.D. Dean of the Graduate School
Wei Sun, Ph.D. Assistant Professor of Management Science	

LU, MIN (Ph.D., Biostatistics)
Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods (August 2018)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Hemant Ishwaran.
No. of pages in text. (143)

Estimation of individual treatment effect in observational data is complicated due to the challenges of confounding and selection bias. A useful inferential framework to address this is the counterfactual model which takes the hypothetical stance of asking what if an individual had received both treatments. Making use of random forests (RF) within the counterfactual framework, I estimate individual treatment effects by directly modeling the response.

This thesis consists of five Chapters. Chapter 1 reviews the methodology in causal inference and provide mathematical notations. Major approaches reviewed include potential outcome approach, graphical approach and counterfactual approach. Chapter 2 discusses assumptions for counterfactual approach. P-values are useful in causal inference, but whenever it is used, caution must be taken. Section 2.3 and Section 2.4 propose machine learning methods as alternatives to p-values and checking proportional hazards assumption in survival analysis. These two sections are more general in content even beyond the scope of counterfactual approach. Chapter 3 describes six random forest methods for estimating individual treatment effects under counterfactual approach framework and discusses model consistency and convergence of random forest in Section 3.6. Chapter 4 demonstrates the performance of these methods in complex simulations and how the most appropriate method is used in a real dataset for continuous outcome. Chapter 5 addresses causal inference in survival analysis of ischemic cardiomyopathy. Treatment effect is viewed as a

dynamic causal procedure. New random forest methods are proposed in this chapter to assess individual therapy overlap. These methods possess the unique feature of being able to incorporate external expert knowledge either in a fully supervised way (i.e., we have a strong belief that knowledge is correct), or in a minimally-supervised fashion (i.e., knowledge is not considered gold-standard).

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to the people who have ever offered help to me in the past five years.

First and foremost, I am indebted to my amazing advisor, Dr. Hemant Ishwaran, who guided me through the transition from a student to an independent researcher. He has not only taught me versatile analytical techniques, but also has molded me into a scholar with comprehensive research skills. Throughout so many discussions we had together, I have benefited from his deep insights in statistics and his attitude towards work, which will have a lasting influence on my future career. I am honored to call him my advisor and mentor.

Next, I want to thank Dr. J. Sunil Rao, Dr. Xi Chen, Dr. Booil Jo, Dr. Daniel Feaster, and Dr. Wei Sun for serving on my thesis proposal and defense committees. Their suggestions and discussions have greatly improved my dissertation research. Dr. Rao and Dr. Feaster are very helpful for my job hunting and took their precious time writing recommendation letters for me. I would like to thank both Darlene and Dr. Rao for the enchanting parties in their home, which make the Fall semester especially cherishable.

I would like to thank Dr. Soyeon Ahn, Dr. Nicholas Myers and Dr. Cengiz Zopluoglu for their support in my first two years of study at UM. Their patience, encouragement, and insight have greatly enriched my PhD education and experience. A special thank you to Dr. Ahn whose encouragement is unconditional much like her support and confidence in me. I also thank all the instructors for the invaluable classes I have taken at UM. I thank Dr. Lily Wang and Dr. Eugene Blackstone for their recommendation letters. I have been very fortunate to be able to collaborate with Dr. Blackstone and Dr. Thomas W. Rice, and I want to thank them for bringing in exciting problems and trusting my knowledge in statistics.

The last five years could not be so delightful without the accompany of my friends and classmates, Qiuying Zhang, Shiyan Jiang, Guanhua Chen, Seniz Celimli, Marietta Suarez,

Amol Pande, Hongmei Liu, Jie Fan, Gang Xu, Huilin Yu, Alejandro Mantero, Mengying Li, Xing Wei, Feng Miao, Yifan Sha, Xiao Xiao, and Hang Zhang. I am sure our friendship will last well beyond graduation. I am also grateful to Marissa Kobayashi, Samuel Swift, Ji-Young Lee, Ashly Westrick and Vivek Singh, who were so supportive when I was serving as their teaching assistant. I am grateful to the staff at UM for making my time so rewarding and successful.

Finally, I thank my parents for all the support. Because of them, I have never felt alone or discouraged facing all the challenges in life. They taught me strength, patience, and perseverance.

Contents

List of Figures

List of Tables

Chapter 1

Causality

1.1 Introduction

Causality or causation, referring to observed associations on an informal basis, is one of oldest topics in philosophy. Aristotle once said in the *Posterior Analytics*, "We think we have knowledge of a thing only when we have grasped its cause". Today a formal theory of causal inference has been developed, with major contributions from Donald Rubin, James Robins, and Judea Pearl. Shpitser and Pearl (2008) suggest a hierarchy of queries in causal relationships: "**associative relationships**, derived from a joint distribution over the observable variables; **cause-effect relationships**, derived from distributions resulting from external interventions; and **counterfactuals**, derived from distributions that span multiple 'parallel worlds' and resulting from simultaneous, possibly conflicting observations and interventions." Examples would be:

> **Associative relationships**: "I took an aspirin after dinner, will I wake up with a headache?"
>
> **Cause-effect relationships**: "if I take an aspirin now, will I wake up with a headache?"
>
> **Counterfactuals** or "what-if" questions: "I took an aspirin, and my headache is gone; would I have had a headache had I not taken that aspirin?"

I will review approaches addressing the second query in Section 1.2 and 1.3, for instance, the widely used propensity score approaches in causal inference studies; however, this dissertation concentrates more in the counterfactual models.

An intervention's effectiveness is usually investigated in two settings. First, well-designed and implemented randomized controlled trials are considered the "gold standard". Fisher's book *The Design of Experiments* (1935) argued the challenges of confounding and emphasized randomization in experiments. The second way is to evaluate causality from observational studies. Even for a medical discipline steeped in a tradition of randomized trials, the evidence basis for only a few guidelines is based on randomized trials (Tricoci et al., 2009). In part this is due to continued development of treatments, in part to enormous expense of clinical trials, and in large part to the hundreds of treatments and their nuances involved in real-world, heterogeneous clinical practice. Thus, many therapeutic decisions are based on observational studies, which is the main theme of this dissertation.

However, comparative treatment effectiveness studies of observational data suffer from two major problems: only partial overlap of treatments and selection bias. Each treatment is to a degree bounded within constraints of indication and appropriateness. Thus, transplantation is constrained by variables such as age, a mitral valve procedure is constrained by presence of mitral valve regurgitation. However, these boundaries overlap widely, and the same patient may be treated differently by different physicians or different hospitals, often without explicit or evident reasons. Thus, a fundamental hurdle in observational studies evaluating comparative effectiveness of treatment options is to address the resulting selection bias or confounding. Naively evaluating differences in outcomes without doing so leads to biased results and flawed scientific conclusions.

Formally, let $\{(\mathbf{X}_1, T_1, Y_1), ..., (\mathbf{X}_n, T_n, Y_n)\}$ denote the data where $\mathbf{X}_i$ is the covariate vector for individual i, Y_i is the observed outcome, and T_i denotes the treatment group of i. For concreteness, let us say $T_i = 0$ represents the control group, and $T_i = 1$ the inter-

vention group. In the following sections of this chapter, I categorized approaches in causal inference into three schools and show how these apporaches use different mathematical languages that extend the limit of this notation. I would like to introduce how potential outcome approach and graphical approach deal with overall causal effect and how counterfactual approach addresses individual causal effect. However, these three schools are not mutually exclusive: some models in graphical approach can be used to estimate counterfactual individual causal effect.

1.2 The potential outcome approach

The potential outcome approach is often referred as Rubin causal model or the Neyman-Rubin causal model, contributed by Donald Rubin's work (1974) and Jerzy Neyman's work (1923). A key contribution of the potential outcome approach is addressing treatment assignment mechanism. Neyman's work uses randomized experiment to eliminate bias that could be potentially introduced by treatment assignment: he studies potential yield of v varieties of crops on m plots through an urn model (repeated-sampling). Rubin's work in obervational studies links the potential outcomes to the more general "missing data" mechanism. let $Y_i(0)$ and $Y_i(1)$ denote the potential outcome for i under treatments $T_i = 0$ and $T_i = 1$, respectively. The assignment mechanism can be written as

$$P(T|\mathbf{X}, Y(0), Y(1)).$$

Rubin (1978) defines the treatment assignment mechanism is "ignorable" when $P(T|\mathbf{X}, Y(0), Y(1)) = P(T|\mathbf{X}, Y)$ since probabilistic functions of recorded values are known, and Rosenbaum and Rubin (1983) define "strongly ignorable" or "unconfounded" as $P(T|\mathbf{X}, Y(0), Y(1)) = P(T|\mathbf{X})$ or $T \perp \{Y(0), Y(1)\}|\mathbf{X}$. For the latter definition,

assumption of strongly ignorable treatment assignment (SITA) is widely used. Under the assumption of SITA, we have

$$\begin{aligned}\tau(\mathbf{x}) =& E[Y(1)|T=1,\mathbf{X}=\mathbf{x}] - E[Y(0)|T=0,\mathbf{X}=\mathbf{x}] \\ =& E[Y|T=1,\mathbf{X}=\mathbf{x}] - E[Y|T=0,\mathbf{X}=\mathbf{x}].\end{aligned} \tag{1.1}$$

Thus, SITA ensures that $\tau(\mathbf{x})$ is estimable because it reduces estimating $\tau(\mathbf{x})$ to estimating conditional expectations of observable values. It should be emphasized that without SITA one cannot guarantee estimability of $\tau(\mathbf{x})$ because $E[Y(j)|\mathbf{X}=\mathbf{x}]$ is not estimable in general and $E[Y|T=j,\mathbf{X}=\mathbf{x}] = E[Y(j)|\mathbf{X}=\mathbf{x}]$ does not hold in general. SITA also provides a means for estimating the average treatment effect (ATE), a standard measure of performance in non-heterogeneous treatment settings. The ATE is defined as $\tau_0 = E[Y_i(1)] - E[Y_i(0)] = E[\tau(\mathbf{X})]$. By averaging over the distribution of $\mathbf{X}$ in (1.1),

$$\tau_0 = E\Big\{E[Y|T=1,\mathbf{X}=\mathbf{x}] - E[Y|T=0,\mathbf{X}=\mathbf{x}]\Big\} = E[Y|T=1] - E[Y|T=0]. \tag{1.2}$$

Thus SITA ensures that τ_0 is estimable.

Although direct estimation of (1.1) or (1.2) is possible by using mean treatment differences in cells with the same X as raw matching design, due to the curse of dimensionality this method will only work when X is low dimensional. Propensity score analysis proposed by Rosenbaum and Rubin (1983) is one means to overcome this problem. The propensity score is defined as the conditional probability of receiving the intervention given $\mathbf{X}=\mathbf{x}$, denoted here by $e(\mathbf{x}) = P\{T=1|\mathbf{X}=\mathbf{x}\}$. Under the assumption of SITA, the propensity score possesses the so-called balancing property. This means that T and X are conditionally independent given $e(\mathbf{X})$. Thus variables X are balanced between the two treatment groups after propensity score matching, thereby approximating a randomized clinical trial (Rubin, 2007). Importantly, the propensity score is the coarsest possible balancing score, thus

not only does it balance the data, but it does so by using the coarsest possible conditioning, thus helping to mitigate the curse of dimensionality. In order to use the propensity score for treatment effect estimation, Rosenbaum and Rubin (1983) further show that if the propensity score is bounded $0 < e(\mathbf{X}) < 1$ and SITA holds, then treatment assignment is conditionally independent of the potential outcomes given the propensity score; i.e., $T \perp \{Y(0), Y(1)\}|e(\mathbf{X})$. This result is the foundation for ATE estimators based on stratification or matching of the data on propensity scores, which contains three steps: firstly $e(\mathbf{X})$ is etimated through a statistical model like logit regression, and secondly, $\hat{e}(\mathbf{X}_i)$ is used for each data to get matched pairs or stratas; last, differences in Y_i within each pair or strata are averaged to get average treatment effect $\hat{\tau_0}$. However, this is not the only means for using the propensity score to estimate treatment effect. Others have directly used the SITA assumption to derive weighted estimators for the ATE. Analgous to (1.1), under SITA one has

$$E\Big[\frac{TY}{e(\mathbf{X})}|\mathbf{X}=\mathbf{x}\Big] = E[Y|T=1,\mathbf{X}=\mathbf{x}], E\Big[\frac{(1-T)Y}{1-e(\mathbf{X})}|\mathbf{X}=\mathbf{x}\Big] = E[Y|T=0,\mathbf{X}=\mathbf{x}],$$

which is the basis for ATE weighted propensity score estimator from a finite sample of size n:

$$\hat{\tau_0} = \frac{1}{n}\sum_{T_i=1}\frac{Y_i}{\hat{e}(\mathbf{X}_i)} - \frac{1}{n}\sum_{T_i=0}\frac{Y_i}{1-\hat{e}(\mathbf{X}_i)}. \tag{1.3}$$

Developed by Horvitz and Thompson (1952), Equation (1.3) is also refered as the difference of two Horvitz-Thompson estimators. If some estimated propensity score $\hat{e}(\mathbf{X}_i)$ is close to 0 (for a treated unit) or 1 (for a control unit), its inverse weight can become very large and unstable for estimating $\hat{\tau_0}$ in finite sample. One solution is to normalize the

weights within each treatment group:

$$\hat{\tau_0^*} = \Big(\sum_{T_i=1} \frac{1}{\hat{e}(\mathbf{X}_i)}\Big)^{-1} \sum_{T_i=1} \frac{Y_i}{\hat{e}(\mathbf{X}_i)} - \Big(\sum_{T_i=1} \frac{1}{1-\hat{e}(\mathbf{X}_i)}\Big)^{-1} \sum_{T_i=0} \frac{Y_i}{1-\hat{e}(\mathbf{X}_i))}.$$

See for example, Hirano et al. (2003) and Lunceford and Davidian (2004).

I only categorized those methods, which assume $T \perp \{Y(0), Y(1)\}|e(\mathbf{X})$ and utilize $e(\mathbf{X})$ for matching or weighting, as "potential outcome approach". Studies using raw oberved variables for matching belong to this category too (Alexander et al., 2002). In this section, the cause-effect relationships are simply reduced as average treatment effect τ_0 between two treatments, where all the other covariates $\mathbf{X}$ are treated as fixed: how these $\mathbf{X}$ may "cause" each other or "cause" the outcome is unknown. Section 1.2 addresses "cause of effect" in Section 1.1 through "balancing" on confoundness: Holland (1986) call this as "effects of causes". Section 1.3, graphical approach, is about directly dealing with "cause". Moreover, $\tau(\mathbf{x})$ here can be considered as counterfactual treatment effect in Section 1.1, but it is not estimated here; in Section 1.4, I will discuss more about $\tau(\mathbf{x})$. Notice that SITA assumption does not require correct form of $e(\mathbf{X})$ as propensity score approach does; therefore propensity score approach demands more restrict assumption. However, $0 < e(\mathbf{x}) < 1$ is till required for any causal inference on $\mathbf{x}$: more reason is in Section 1.4.

1.3 The graphical approach

This section is about Bayesian network (BN) model using directed acyclic graph (DAG). This graphical approach usually addresses more complex cause-effect relationships than potential outcome approach does: the goal here is to smell out a plot of all the variables. Based on Chapter 2 and 3 in the book of Koller and N. Friedman (2009) and Chapter 3 in Pearl (2009)'s book, I give some basic concept, assumption and example of this approach.

A graph is a data structure $\mathcal{K}$ consisting of a set of nodes, $\mathcal{X} = \{X_1, ..., X_n\}$, and a set of edges $\mathcal{E}$: $\mathcal{K} = (\mathcal{X}, \mathcal{E})$. A pair of nodes X_i, X_j can be connected by a directed edge $X_i \rightarrow X_j$ or an undirected edge X_i—X_j, or some edge, whether directed (in any direction) or undirected $X_i \rightleftharpoons X_j$. In $X_i \rightarrow X_j$, X_i is called *parent* of X_j, and X_j is called *child* of X_i. Let $Pa^{\mathcal{K}}_{X_i}$ denote the parents of X_i and $Ch^{\mathcal{K}}_{X_i}$ denote its children in $\mathcal{K}$. Let $NonDescendants_{X_i}$ be the variables in the graph that are not descendants of X_i. A cycle in $\mathcal{K}$ is a directed path $X_1 \rightarrow X_2, ..., \rightarrow X_k$ where $X_1 = X_k$. A graph is acyclic if it contains no cycles.

The use of Bayes theorem and the use of graphical models explain the choice of the name Bayesian network: A BN structure $\mathcal{G}$ is a directed acyclic graph defined above. $\mathcal{G}$ encodes the following set of conditional independence assumptions, called the *local independencies*, and denoted by $\mathbb{I}_l(\mathcal{G})$:

For each variable X_i: $(X_i \perp NonDescendants_{X_i} | Pa^{\mathcal{G}}_{X_i})$.

Since the local independencies state that each node X_i is conditionally independent of its nondescendants given its parents, a distribution P over the same space is called *factorizes* according to $\mathcal{G}$, when P can be expressed as a product
$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa^{\mathcal{G}}_{X_i})$. This equation is called the chain rule for Bayesian networks, denoted as $\mathbb{I}(P)$. The individual factors $P(X_i | Pa^{\mathcal{G}}_{X_i})$ are called conditional probability distributions (CPDs) or local probabilistic models.

The trail $T \rightleftharpoons X \rightleftharpoons Y$ is called active and X and Y are called blocked by T in either of the following four trails:

Causal trail $T \rightarrow X \rightarrow Y$: active if and only if X is not observed.
Evidential trail $T \leftarrow X \leftarrow Y$: active if and only if X is not observed.
Common cause $T \leftarrow X \rightarrow Y$: active if and only if Z is not observed.

Common effect $T \rightarrow X \leftarrow Y$ (v-structure): active if and only if either X or one of X's descendants is observed.

Let $\mathbb{I}(\mathcal{G})$ be $\{(T \perp Y|X)$: T and Y is *d-separated* by $X\}$, $\mathbb{I}(\mathcal{G})$ can be tested through $\mathbb{I}(P)$ since $\mathbb{I}(\mathcal{G}) \subseteqq \mathbb{I}(P)$.

Notations of causal inference in BN model are different: since it does not assume binary treatment variable by default as the potential outcome model does, probabilities of $Y(0)$ and $Y(1)$ are substituted by $P(y|do(t))$, $P(y|\hat{t})$ or $P_t(y)$, which is generally different from $P(y|T = t)$ given confounding variables $X \neq \varnothing$. Instead of working on the expectations of Y, Baysian Network model is a non-parametric method that works on the observed joint and conditional distributions, and estimates treatment effect through a manipulation on these distributions. Notation $do(\cdot)$ is used to manage extra information that even a full specification of a population density function does not permit us to predict beyond static conditions: for example, relationships would change from observational to controlled studies. In example illustrated in Figure 1.1, the total effect of fumigants T, on yields Y can be estimated consistently from the observed distribution of T, X_1, X_2, X_3 and Y through formula (1.4), where the quantities X_1, X_2, and X_3 denote the eelworm population before treatment, after treatment and at the end of the season, respectively. Last year's eelworm population X_0, is marked by a hollow circle because it is an unknown quantity, as is B, the populaiton of birds and other predators.

$$\begin{aligned} P(y|\hat{t}) = &\sum_{x_1}\sum_{x_2}\sum_{x_3} P(y|x_1, x_2, t)P(x_2|x_1, t) \\ &\times \sum_{t'} P(x_3|x_1, x_2, t')P(x_1, t') \end{aligned} \tag{1.4}$$

An intuitive question would be, can we get consistent estimate of $P(y|\hat{t})$ when B and X_0 is unknown? This question leads to the assumption in BN. Equavlent to SITA but in different demonstraton, two assumptions in BN are called back-door criteria and front-door

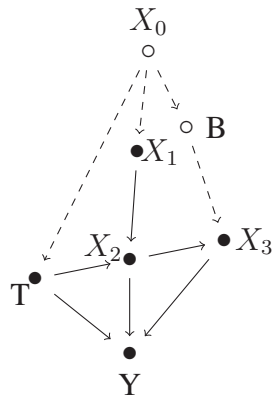

Figure 1.1: Causal diagram example from Pearl (2009)'s book

T=fumigants;
X=eelworm population: subscripts, 0, 1, 2, and 3, represent last year, before treatment, after treatment, and the end of the season, respectively;
B=the population of birds and other predators;
Y=yields;
Causal diagram represents the effect of fumigants (T) on yields (Y). X_0 and B are unknown in quantity denoting by hollow circle. Dashed arrows are used to connect unmeasured quantities and solid arrows are used to connect measured quantities

criteria, based on which Pearl (1995) derived causal calculus.

Back-Door A set of variables $\mathbf{X}$ satisfies the back-door criterion to an ordered pair of variables (T, Y) in a DAG if: (i) no node in $\mathbf{X}$ is a descendant of T and (ii) $\mathbf{X}$ blocks every path between T and Y that contains an arrow into T. For example, in Figure (1.4), $\{X_1, X_3\}$ meet the back-door criterion of T and Y.

Back-Door Adjustment If X satisfies the back-door criterion to (T, Y), then the causal effect of T on Y is identifiable and is given by the formula

$$P(y|\hat{t}) = \sum_x P(y|t, x)P(x) = \sum_x \frac{P(x, t, y)}{P(t|x)}, \tag{1.5}$$

which reflects back to the potential outcome approach in equation (1.3).

Front-Door A set of variables $\mathbf{X}$ satisfies the front-door criterion to an ordered pair of variables (T, Y) if: (i) $\mathbf{X}$ intercepts all directed paths from T to Y; (ii) there is no

back-door path from T to $\mathbf{X}$; and (iii) all back-door paths from $\mathbf{X}$ to Y are blocked by T.

Front-Door Adjustment If X satisfies the front-door criterion to (T, Y) and if $P(t, x) > 0$, then the causal effect of T on Y is identifiable and is given by the formula $P(y|\hat{t}) = \sum_x P(x|t) \sum_{t'} P(y|t', x)P(t')$.

In the front-door criterion, adjustment variable X can be the descendants of T. But the backdoor criterion was used twice: first computes the causal effect of T on X and then computes the causal effect of X on Y.

As Pearl (2009) states, the role of graphs is to provide convenient means of expressing substantive assumptions; to facilitate economical representation of joint probability functions; and to facilitate efficient inferences from observations. The potential outcome approach is not designed to rule out the possibility that outcome may cause the treatment: it is an unspoken assumption that the "cause" is treatment since usually treatment temporally precedes outcome. However, there is no such limit that the structure of the BN here has to be pre-determined by the researchers. Spirtes et al. (2000)'s book offers several Discovery Algorithms to identify the structure of networks, which is beyond the scope of this dissertation. Other kind of graphical models, such as undirected graphical models, Gaussian network models, and Markov Networks, can be found in Koller and N. Friedman (2009)'s book.

1.4 The counterfactual approach

Pearl (2009) defines *Counterfactual* in his book Chapter 7 as: Let X and Y be two subsets of variables. The counterfactual sentence "The value that Y would have obtained, had

X been $\mathbf{x}$" is interpreted as denoting the potential response $Y_{\mathbf{x}}(u)$[1]. $\tau(\mathbf{x})$ in Section 1.2 is still used here as counterfactual treatment effect, which can be rewritten as $\tau(\mathbf{x}) = E[Y_{\mathbf{x}}(1) - Y_{\mathbf{x}}(0)]$. If unobserved $Y_{\mathbf{x}}(1)$ or $Y_{\mathbf{x}}(0)$ can be "observed" through an imaginative counterfactual in term of methematical function

$$Y = g(T, \mathbf{X}, \epsilon_Y), \tag{1.6}$$

then $Y_{\mathbf{x}}(1) = g(1, \mathbf{x}, \epsilon_Y)$ and $Y_{\mathbf{x}}(0) = g(0, \mathbf{x}, \epsilon_Y)$. This function g reflects Laplace (1814)'s demon or Laplacian determinism in the history of science.

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.
>
> –Pierre Simon Laplace, A Philosophical Essay on Probabilities

It is very hard to talk about causality without being a Laplacian determinist. At the

[1] In BN, given model $\langle M, P(u)\rangle$, the conditional probability $P(B_A|e)$ of a counterfactual sentence "If it were A then B", given evidence e, can be evaluated using the following three steps.
1. **Abduction** - Update $P(u)$ by the evidence e to obtain $P(u|e)$.
2. **Action** - Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A.
3. **Prediction** - Use the modified model $\langle M_A, P(u)\rangle$ to compute the probability of B, the consequence of the counterfactual.
To compute, store and use $P(u|e)$, one can use the Twin Network method, which creates two networks, one to represent the actual world and one to represent the hypothetical word.

first glance, using equation (1.6) to solve causality seems rude to disbelievers of Laplacian determinism. In statistics language, this disbelif reflects the doubt that $Y_{\mathbf{x}}(1)$ and $Y_{\mathbf{x}}(0)$ are independant. On the other hand, potential-outcome approach, which also assumes independence of $Y_{\mathbf{x}}(1)$ and $Y_{\mathbf{x}}(0)$, still is Laplacian determinism even not that obvious. Another philosophical question would be predictability: Holland (1986) argued that only variables with "potential exposability" can be a cause. Here this means only $P(T|\mathbf{x}) > 0$ for both $T = 1$ and $T = 0$, T can be a cause on Y for $\mathbf{x}$, which requires an accurate estimate of $P(T|\mathbf{x})$ in equation (1.7),

$$P(T|\mathbf{x}) = e(\mathbf{x}). \tag{1.7}$$

Equation 1.6 and 1.7 are the key in this dissertation, which rise not only some philosophical thoughts, but also exciting transition from average treatment effect to individual treatment effect estimate(ITE).

Using equation (1.6) and (1.7) to get counterfactual treatment effect is also called functional approach in causality by Zhao (2016). Previous studies classified in this category include linear structure equation model (SEM), which simplifies all variables in function g and e as $S_i = \sum_{j \in Pa_{S_i}^{\mathcal{G}}} \alpha_{ij} S_j + \epsilon_i$ in a graph $\mathcal{G}$. Another model is the widely used "g-formula" algorithm for causal inference in the presence of time-varying covariates Robins et al. (1999), where we have similar $\tau_m(\mathbf{x}) = E[Y_{m,\mathbf{x}}(1)] - E[Y_{m,\mathbf{x}}(0)]$ under SITA, m being the index for time. $E[Y_{m,\mathbf{x}}(1)]$ and $E[Y_{m,\mathbf{x}}(0)]$ are obtained by substituting $\mathbf{x}$, $T = 0$ and $T = 1$ from a generalized linear model, whose coefficients are estimated through the generalized estimating equations (GEE) approach. Another example is Bayesian tree growing methods which have been successfully used to identify causal effects by directly modeling the response surface by Hill (2011). Although Hill estimates average treatment effect and conditional treatment effect, rather than counterfactual individual treatment effect, she uses the same manner to get $Y_{\mathbf{x}}(1) = g(1, \mathbf{x}, \epsilon_Y)$ and $Y_{\mathbf{x}}(0) = g(0, \mathbf{x}, \epsilon_Y)$, where g is a Bayesian

tree model. It is worth to mention that data imputation or augmentation method for causal inference (Dominici et al., 2006) is a special case in this category, which uses T to code Y as bivariate variable and make g as a distribution function to impute unobserved $Y(0)$ and $Y(1)$ for each $\mathbf{x}$. Section 3.4 applies this imputation method.

Although effectiveness of treatment in observational studies has traditionally been measured by the ATE, the practice of individualized medicine, coupled with the increasing complexity of modern studies, have focused recent efforts towards a more patient-centric view (Lamont et al.,2016). Accommodating complex individual characteristics in this new landscape has proven challenging, and for this reason there has been much interest in leveraging cutting-edge approaches addressing g and e, especially those from machine learning. Machine learning techniques such as random forests Breiman (2001b) (RF) provide a principled approach to explore a large number of predictors and identify replicable sets of predictive factors. In recent innovations these RF approaches have been used specifically to uncover subgroups with differential treatment responses (Su et al., 2009, 2011; Foster et al., 2011). Some of these, such as the virtual twins approach (Foster et al., 2011), build on the idea of counterfactuals. Virtual twins uses RF as a first step to create separate predictions of outcomes under both treatment and control conditions for each trial participant by estimating the counterfactual treatment outcome. In the second step, tree-based predictors are used to uncover variables that explain differences in the person-specific treatment and the characteristics associated with subgroups. In a different approach, Wager and Athey (2017) describe causal forests for ITE estimation. Others have sought to use RF as a first step in propensity score analysis in equation (1.7) as a means to nonparametrically estimate the propensity score. Lee et al. (2010) found that RF estimated propensity scores resulted in better balance and bias reduction than classical logistic regression estimation of propensity scores. More details about how to set up RF to get function g in equation (1.6) are in the Chapter 3. This dissertation focuses on estimating the counterfactual ITE using RF

methods. In Chapter 4, two sets of challenging simulations are used to assess performance of the various RF methods.

Chapter 2

Assumptions for counterfactual approach

This chapter discusses counterfactual approach using function g, $Y = g(T, \mathbf{X}, \epsilon_Y)$, to get individual counterfactuals $Y_{\mathbf{x}}(1) = g(1, \mathbf{x}, \epsilon_Y)$ and $Y_{\mathbf{x}}(0) = g(0, \mathbf{x}, \epsilon_Y)$, assuming (1): SITA: $P(T|\mathbf{X}, Y(0), Y(1)) = P(T|\mathbf{X})$ or $T \perp \{Y(0), Y(1)\}|\mathbf{X}$; (2): function g is consistent and (3): $P(T|\mathbf{x}) = e(\mathbf{x}) > 0$ for both $T = 1$ and $T = 0$. Apparently, SITA is unable to test. I focus on individual's treatment overlap from function e in assumption (3) in Section 2.1 and the specification of funciton g in Section 2.2. Since p-values are widely used in causal inference analysis, a machine learning alternative to p-values is proposed in Section 2.3. When parametric model is used for survival analysis, another common assumption is proportional hazards assumption, which is discussed in Section 2.4.

2.1 Complete overlap from function e

As discussed before in Section 1.4, causal inference can be done only for $\mathbf{x}$ when $P(T|\mathbf{x}) = e(\mathbf{x}) > 0$ for both $T = 1$ and $T = 0$. In other words, if $\mathbf{x}$ is not overlap/eligible for the treatment, there is no point to estimate treatment effect. This assumption is called "complete overlap assumption" in Chapter 5. Note that this assumption can be checked through other information and prior knowledge instead of modeling $e(\mathbf{X})$ from the data.

Consider example in Figure 2.1, treatment A and treatment B is compared for each

$\mathbf{X} = \mathbf{x}$, $\mathbf{X} = \{X_1, X_2, \ldots, X_p\}$. Marginal distributions $P(T = A, X_1 = x_1)$ and $P(T = B, X_1 = x_1)$ are ploted in red and blue respectively. Suppose $P(T = B, X_1 > C_{X_1,B}) = 0$, treatment effect $\tau(\mathbf{x}) \neq \varnothing$ only for $\{\mathbf{x} : x_1 \in [0, C_{X_1,B}]\}$. Region $\{\mathbf{x} : x_1 \in (C_{X_1,B}, \infty)\}$ is called "lack of overlap" region from complete overlap assumption, which will be deleted from the causality analysis. Even fuction e in $e(\mathbf{x}) = P(T|\mathbf{x})$ is consistent or correctly specified, the deleted region is usually defined in a fashion of α significant level, where the area of blue region in domain $(C_{X_1,\alpha_B}, \infty)$ is α. The question is, is it worthwhile to delete data within this α region?

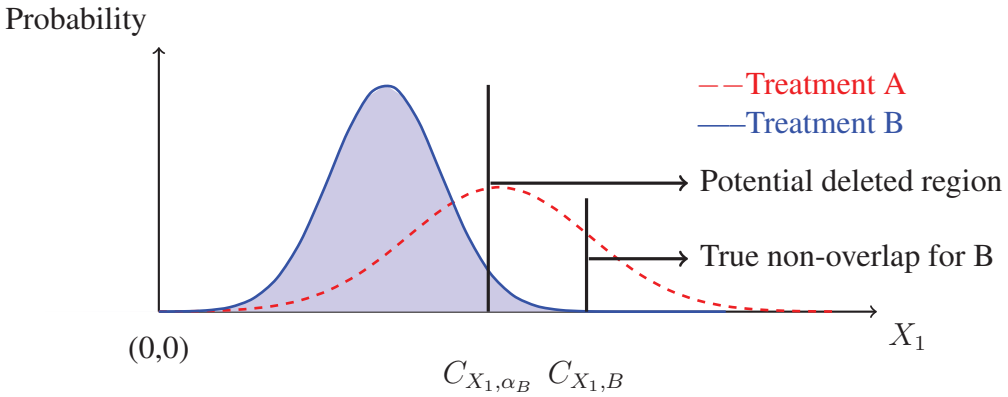

Figure 2.1: Marginal distribultion of overlap in X_1: red dash line represents $P(T = A, X_1 = x_1)$ and blue solid line represents $P(T = B, X_1 = x_1)$.

This α represents neither false positives nor false negatives of treatment effect: it only represent an error that claim treatment effect not exist ($\tau(\mathbf{x}) = \varnothing$) while the truth is that treatment effect does exist ($\tau(\mathbf{x}) \neq \varnothing$). If the goal is to give $\tau(\mathbf{x})$ for each $\mathbf{x}$, the quesion of whether $\tau(\mathbf{x}) = \varnothing$ does matter. If the goal is to detect the mechanism of $\tau(\mathbf{x})$, for example, further analysis $\tau(\mathbf{x})$ as a function of covariates $\mathbf{X}$ to detect moderators of treatments, one has to consider (1) is estimate $\hat{\tau}(\mathbf{x})$ still accurate when $P(T|\mathbf{x})$ is small; (2) is this X_1 related to Y at all. If the answer to (1) is Yes and the answer to (2) is No, there is no need to do this region deletion or data filtering, because X_1 is just a noise to outcome or $\tau(\mathbf{x})$ and filtering data according to a noise variable is not necessary. If the answer to (1)

is No and the answer to (2) is No, there is need to do region deletion or data filtering on those X other than X_1: deleted region=$\{\mathbf{x} : P(T|\mathbf{x}^{(Y)}) < \alpha\}$, $\mathbf{X}^{(Y)} = \{X_i : X_i \not\perp T \text{ and } X_i \not\perp Y\}$. Some machine learning approach, RF for example, is originated from nearist neighbor mechanism of data; for $\{\mathbf{x} : P(T|\mathbf{x}^{(Y)}) < \alpha\}$, there could be potential problem of unbalanceness of treatment assignment, making $\hat{\tau}(\mathbf{x})$ not as much as accurate in this α region.

The concept of $\mathbf{X}^{(Y)}$ reflects my recommandation of integrating a variable selection step in causal inference with large variable number p, and procedure in section 2.3 is useful to detect informative variables $\mathbf{X}^{(Y)}$. I am going to proof this point in Theorem 2.1.2 through equation (1.5), which is the foundation for both potential outcome approach and Baysian Network approach. Recall that potential outcome approach uses $e(\mathbf{x}) = P(t|\mathbf{x})$ to balance a set of covariates $\mathbf{X}$ between treatment groups, and Theorem 2.1.2 says balancing on those $\mathbf{X}^{(Y)^C} \in \{\mathbf{X}\}$ is not necessary.

Definition 2.1.1. $\mathbf{X}^{(Y)^C}$ and $\mathbf{X}^{(Y)}$:
$\mathbf{X}^{(Y)^C}$ is a subset of $\mathbf{X}$ which is independent from Y: $\mathbf{X}^{(Y)^C} = \{X_i : X_i \perp Y\}$, while $\mathbf{X}^{(Y)} = \{X_i : X_i \not\perp Y\}$; $\mathbf{X} = \{\mathbf{X}^{(Y)^C}, \mathbf{X}^{(Y)}\}$. When the conditional probability density function of Y is $g(t, \mathbf{x}, \sigma)$, we have $g(t, \mathbf{x}, \sigma) = g(t, \mathbf{x}^{(Y)}, \mathbf{x}^{(Y)^C}, \sigma) = g(t, \mathbf{x}^{(Y)}, \sigma)$; so $P(y|t, \mathbf{x}) = P\big(y|t, \mathbf{x}^{(Y)}\big)$.

Theorem 2.1.2 (Subset $\mathbf{X}$ propensity score weighting)**.** *Suppose $\mathbf{X} = \{\mathbf{X}^{(Y)^C}, \mathbf{X}^{(Y)}\}$ defined before and assume $P(y|\hat{t})$ as equation (1.5) $P(y|\hat{t}) = \int_{\mathbf{X}} P(y|t, \mathbf{x})P(\mathbf{x})d\mathbf{x}$:*

$$P(y|\hat{t}) = \int_{\mathbf{X}^{(Y)}} P(y|t, \mathbf{x}^{(Y)})P(\mathbf{x}^{(Y)})d\mathbf{x}^{(Y)} = \int_{\mathbf{X}^{(Y)}} \frac{P(\mathbf{x}^{(Y)}, t, y)}{P(t|\mathbf{x}^{(Y)})}d\mathbf{x}^{(Y)}. \qquad (2.1)$$

For discrete variables:

$$P(y|\hat{t}) = \sum_{\mathbf{X}^{(Y)}} P(y|t, \mathbf{X}^{(Y)} = \mathbf{x}^{(Y)})P(\mathbf{X}^{(Y)} = \mathbf{x}^{(Y)}) = \sum_{\mathbf{X}^{(Y)}} \frac{P(\mathbf{X}^{(Y)} = x^{(Y)}, t, y)}{P(t|\mathbf{X}^{(Y)} = x^{(Y)})}.$$

Proof.

$$\begin{aligned}
P(y|\hat{t}) &= \int_{\mathbf{X}} P(y|t,\mathbf{x})P(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathbf{X}^{(Y)}} \int_{\mathbf{X}^{(Y)^C}} P(y|t,\mathbf{x})P(\mathbf{x})d\mathbf{x}^{(Y)^C}d\mathbf{x}^{(Y)} \\
&= \int_{\mathbf{X}^{(Y)}} \int_{\mathbf{X}^{(Y)^C}} P(y|t,\mathbf{x}^{(Y)})P(\mathbf{x})d\mathbf{x}^{(Y)^C}d\mathbf{x}^{(Y)} \\
&= \int_{\mathbf{X}^{(Y)}} P(y|t,\mathbf{x}^{(Y)}) \left[\int_{\mathbf{X}^{(Y)^C}} P(\mathbf{x})d\mathbf{x}^{(Y)^C} \right] d\mathbf{x}^{(Y)} \\
&= \int_{\mathbf{X}^{(Y)}} P(y|t,\mathbf{x}^{(Y)})P(\mathbf{x}^{(Y)})d\mathbf{x}^{(Y)}
\end{aligned}$$

Since

$$P(y|t,\mathbf{x}^{(Y)})P(\mathbf{x}^{(Y)}) = \frac{P(\mathbf{x}^{(Y)},t,y)}{P(t|\mathbf{x}^{(Y)})},$$

we also have:

$$P(y|\hat{t}) = \int_{\mathbf{X}^{(Y)}} \frac{P(\mathbf{x}^{(Y)},t,y)}{P(t|\mathbf{x}^{(Y)})}d\mathbf{x}^{(Y)}$$

□

2.2 Specification of function g

Using equation (1.6) and (1.7) to get counterfactual treatment effect first requires model consistency of function g: in other words for parametric function, correct specification. This section firstly addresses specification for parametric function g. More model specification and consistency in machine learning approach random forest are in Chapter 3.

A correct specified funciton g will control for confounding variables. Kish (1959) used the word "confounding" in the modern sense of the word, to mean "incomparability" of two or more groups (e.g., exposed and unexposed) in an observational study. Figure 2.2 gives an illustration of a simple relationship of confounding variable and treatment variable. If assuming all relationships in Figure 2.2 are linear, one can analysize treatment effect T,

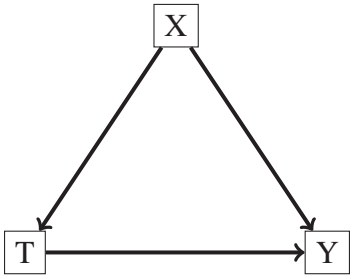

Figure 2.2: Causal diagram example as mediator analysis

T=Treatment variable; X=confounding variable; Y=Outcome variable;
X is confounding variable between T and Y; **also,** T **is mediator variable between** X **and** Y

as the coefficient relating the mediator T to the outcome Y adjusted for the independent variable X. Therefore, treatment effect β_2 can be simply given through estimating

$$Y = \beta_0 + \beta_1 X + \beta_2 T + \varepsilon. \tag{2.2}$$

as mediator analysis (MacKinnon, 2000). However, when there are many variables in complex moderator and non-linear relationships, it is dangerous to use regression functions similar to Equation (2.2) to get βs and interprete them as effect size with p-values, because these functions g may be mis-specified. In the next section, a machine learning alternative to p-values is introduced to provide information of model fitting and prediction power as well as performance under mis-specified model settings.

2.3 A Machine Learning Alternative to p-values

P-values are useful in causal inference. For example, if the outcome is survival and we assume confounding structure in Figure 2.2 (as well as proportional hazards assumption which will be discussed in the next section), we can fit a Cox regression as function g in similar form as Equation (2.2) and detect causal effect through p-values on β_2. The

question here is, can we get inference from function g based on other statistics, instead of p-values? This section presents a machine learning alternative to p-values in regression settings. Note that p-values are still used in this dissertation but not for function g; I only want to emphasize that when using p-values to detect causality, extra care must be used; therefore, I recommend this machine learning alternative.

This approach, whose origins can be traced to machine learning, is based on the leave-one-out bootstrap for prediction error. In machine learning this is called the out-of-bag (OOB) error. To obtain the OOB error for a model, one draws a bootstrap sample and fits the model to the in-sample data. The out-of-sample prediction error for the model is obtained by calculating the prediction error for the model using the out-of-sample data. Repeating and averaging yields the OOB error, which represents a robust cross-validated estimate of the accuracy of the underlying model. By a simple modification to the bootstrap data involving "noising up" a variable, the OOB method yields a variable importance (VIMP) index, which directly measures how much a specific variable contributes to the prediction precision of a model. VIMP provides a scientifically interpretable measure of the effect size of a variable, we call the *predictive effect size*, that holds whether the researcher's model is correct or not, unlike the p-value whose calculation is based on the assumed correctness of the model. We also discuss a marginal VIMP index, also easily calculated, which measures the marginal effect of a variable, or what we call the *discovery effect*. The OOB procedure can be applied to both parametric and nonparametric regression models and requires only that the researcher can repeatedly fit their model to bootstrap and modified bootstrap data. We illustrate this approach on a survival data set involving patients with systolic heart failure and to a simulated survival data set where the model is incorrectly specified to illustrate its robustness to model misspecification.

2.3.1 Background

The issue of p-values has taken center stage in the media with many scientists expressing grave concerns about their validity. “P values, the ’gold standard’ of statistical validity, are not as reliable as many scientists assume”, is the leading assertion of the highly accessed *Nature* article, “Scientific method: Statistical errors” (Nuzzo, 2014). Even more extreme is the recent action of the journal of Basic and Applied Social Psychology (BASP), which announced it would no longer publish papers containing p-values. In explaining their decision for this policy (Trafimow and Marks, 2015), the editors stated that hypothesis significance testing procedures are invalid, and that p-values have become a crutch for scientists dealing with weak data. These, and other highly visible discussions, so alarmed the American Statistical Association (ASA), that it recently issued a formal statement on p-values (Wasserstein and Lazar, 2016), the first time in its history it had ever issued a formal statement on matters of statistical practice.

A big part of the problem is that researchers want the p-value to be something that it was never designed for. At its heart, the p-value remains an awkward statistical concept wrapped in a stifled language that is odds with these needs. Consider the following language clarifying the p-value (some of these being taken from the ASA report):

1. A p-value is the probability of observing an equal or more extreme “event” than that calculated from the data under the assumption of a specific hypothesis assuming a pre-specified statistical model.
2. P-values only indicate how incompatible the data are with the pre-specified statistical model and null hypothesis.
3. P-values are dimensionless and cannot be interpreted in terms of a scientific effect size or the scientific importance of the result.
4. P-values do not provide a measure of evidence regarding the validity of the underly-

ing assumed model.

We see the terminology of statistical significance, null hypotheses, and model assumptions being used to explain the p-value. But researchers require a different type of language. Researchers want to make context specific assertions about their findings; they especially want a statistic that allows them to assert statements regarding scientific effect. Because the p-value cannot do this, and because the terminology is confusing and stifling, it is no wonder this leads to misuse and confusion.

Misinterpretation of the p-value is not the only issue. Another problem is verifying correctness of the model under which the p-value is calculated. If model assumptions do not hold, the p-value itself becomes statistically invalid. This is not an esoteric point. Commonly used models such as linear regression, logistic regression, and Cox proportional hazards can involve strong assumptions. Common practices such as fitting main effect models without interactions, assuming linearity of variables, and invoking distributional assumptions regarding the data, such as normality, can easily fail to hold. Moreover, the functional relationship between attributes and outcome implicit in some of these models, such as proportionality of hazards, may also fail to hold. Researchers rarely test for model correctness, and even when they do, they invariably do so by considering goodness of fit. But goodnesss of fit measures are notoriously unreliable for assessing the validity of a model (Breiman, 2001a).

This section focuses on the use of p-values in the context of regression models. All widely used statistical software provide p-value information when fitting regression models; typically p-values are given for the regression coefficients. These are provided in an ANOVA table with each row of the table displays the regression coefficient estimate, $\hat{\beta}$, for a specific coefficient, β, an estimate of its standard error, $\hat{\sigma}_{\beta}$, and then finally the p-value of

the coefficient, obtained typically by comparing a Z-statistic to a normal distribution:

$$Z_{\text{observed}} = \frac{\hat{\beta}}{\hat{\sigma}_{\beta}}, \quad \text{p-value} = P\{Z \geq |Z_{\text{observed}}|\}.$$

The p-value for the regression coefficient represents the statistical significance of the test of the null hypothesis H_0: $\beta = 0$. In other words, it provides a means of assessing whether a specific coefficient, in this case β, is zero. However, there is a subtle aspect to this where confusion can take place. When considering this p-value, it is important to keep in mind that its value is calculated not only under the null hypothesis of a zero coefficient value, but also assuming that *the model holds*. Thus, technically speaking, the null hypothesis is not just that the coefficient is zero, but is a collection of assorted assumptions, which should probably read something like:

$$H_0 \colon \Big\{\beta = 0, \text{ model holds, model assumptions hold (e.g. interactions not present)}\Big\}.$$

If any of these assumptions fail to hold, then the p-value is technically invalid.

Given these concerns with the p-value, we suggest a different approach using a quantity we call the variable importance (VIMP) index. Our VIMP index is based on variable importance, an idea that originates from machine learning. One of its earliest examples can be traced to Classification and Regression Trees (CART), where variable importance based on surrogate splitting was used to rank variables (see Chapter 5 of Breiman et al. (1984)). The idea was later refined for variable selection in random forest regression and classification models by using prediction error (Breiman, 2001b,a). Extensions to random survival forests were considered by Ishwaran et al. (2008). Our VIMP index uses the same idea as these latter approaches, but recasts it within the p-value context. Like those methods, it uses prediction error to assess the effect of a variable in a model. It replaces the statistical significance of a p-value with the predictive importance of a variable. Most importantly,

the VIMP index holds regardless of whether the model is true. This is because the index is calculated using test data and is not based on a presupposed model being true as the p-value does.

In statistics, effect size is a quantitative measure of the strength of a phenomenon, which includes as examples: Cohen's d (standard group mean difference); the correlation between two variables; and relative risk. In regression models, effect size is measured by the standardized $\hat{\beta}$ coefficient. Since VIMP is also a measure of the quantitative strength of a variable, we refer to its quantitative measure as *predictive effect size* to prevent readers from confusing it with the traditional effect size. With a simple modification to the VIMP procedure, we estimate another quantity we call marginal VIMP and refer to its quantitative measure as the *discovery effect size*. This refers to the discovery contribution of a variable, which will be explained in Section 4. An important aspect of both our procedures is that they can be carried out using the same models the researcher is interested in studying. Implementing them only requires the ability to resample the data, apply some modifications to the data, and calculate prediction error. Thus they can easily be incorporated with most existing statistical software procedures.

Section 2.3.2 outlines the VIMP index and provides a formal algorithmic formulation (see Algorithm 1). The VIMP index is based on out-of-bag (OOB) estimation, which relies on bootstrap sampling. These concepts are also discussed in Section 2.3.2. Section 2.3.3 illustrates the use of the VIMP index to a survival data set involving patients with systolic heart failure with cardiopulmonary stress testing. We show how to use this value to rank risk factors and assess their predictive effect sizes. In Section 2.3.4 we discuss the extension to marginal VIMP (Algorithm 2)) and show how this can be used to estimate discovery effect sizes in the systolic heart failure example. Section 2.3.5 studies how sample size (n) effects VIMP, comparing this to p-values to show robustness of VIMP to n, then in Section 2.3.6 we use a synthetically constructed data set where the model is incorrectly specified

to illustrate the robustness of VIMP in misspecified settings. We conclude the paper with a discussion in Section 2.3.7.

2.3.2 OOB prediction error and VIMP

OOB estimation is a bootstrap technique for estimating the prediction error of a model. While the phrase "out-of-bag" might be unfamiliar to readers, the technique has been known for quite some time in the literature, appearing under various names and seemingly different guises. In the statistical literature, the OOB estimator is refered to as the *leave-one-out bootstrap* due to its connection to leave-one-out cross-validation (Efron and Tibshirani, 1997). See also the earlier paper by Efron (1983) where a similar idea is discussed. It is also used in machine learning where it is refered to as OOB estimation (Breiman, 1998) due to its connections to the machine learning method, bagging (Breiman, 1996).

Calculating the OOB error begins with bootstrap sampling. A bootstrap sample is a sample of the data obtained by sampling with replacement. Sampling by replacement creates replicated values and on average one can expect a bootstrap sample to contain only 63.2% of the original data; this data is refered to as in-sample (inbag) data. The remaining 37% of the data, which is out-of-sample, and called the OOB data, represents test data used in the OOB calculation. Note that OOB data contains no replicated values. To calculate the OOB error, one begins by fitting the model to the inbag data. Then, taking the OOB data, and using it as test data, one calculates the prediction error for the model. This process is repeated B times, where B is some sufficiently large number, say $B = 1000$. The OOB error is obtained by averaging the B estimates of prediction error. Thus if Err_b is the OOB error for the bth sample, the OOB error rate is

$$\text{Err}_{\text{oob}} = \frac{1}{B}\sum_{b=1}^{B} \text{Err}_b.$$

See Figure 2.3 for an illustration of calculating OOB error.

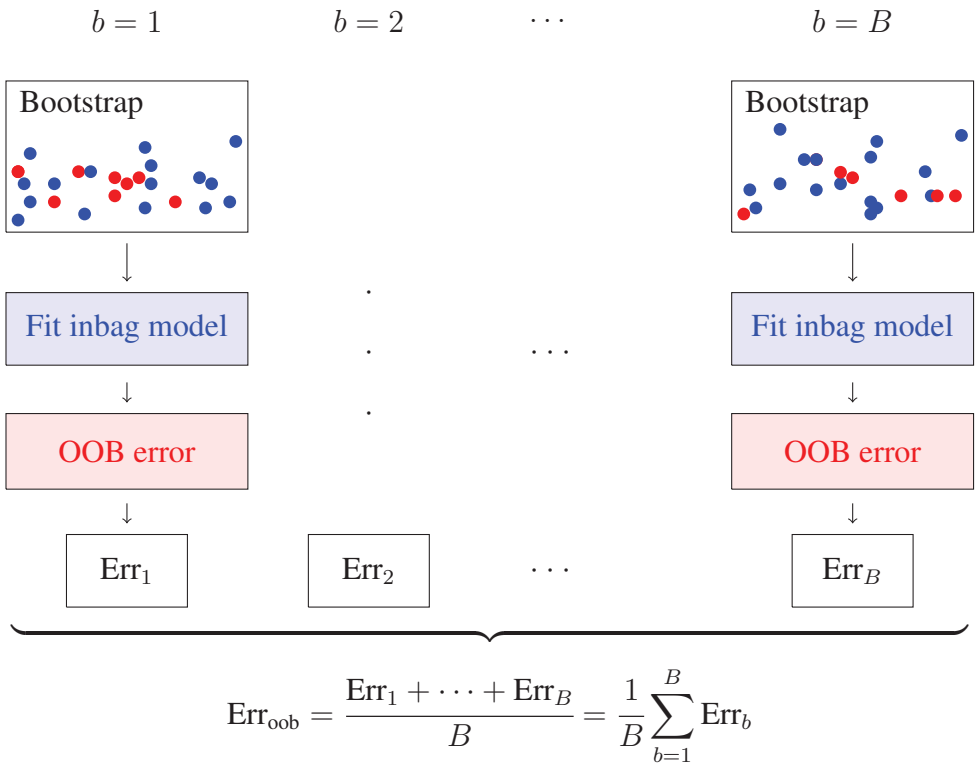

Figure 2.3: *Calculating the OOB prediction error for a model. Blue points depict inbag sampled values, red points depict OOB values. Model is fit using inbag data and then tested on OOB test data. Averaging the prediction error over the different bootstrap realizations yields the OOB prediction error.*

Calculating the VIMP index for a variable

The VIMP index for estimating the predictive effect size for a variable is obtained by a slight modification of the above procedure. For concreteness, call the variable of interest v and let Δ_v denote its VIMP index. To determine Δ_v we determine the VIMP index for v over each bootstrap sample b, call this $\Delta_{v,b}$, and average these values. Here is how $\Delta_{v,b}$ is calculated. For a given bootstrap sample b, take the OOB data for v and "noise it up". Noising the data is intended to destroy the association between v and the outcome and is a crucial step to determining variable importance. There are different ways to noise up

the data, but the simplest is to permute v's data (the OOB data for the other variables are unaltered). Use the new data (with the noised up v data) to calculate the prediction error for the model. Call this noised up prediction error $\text{Err}_{v,b}$. The vimp index for the bootstrap sample is

$$\Delta_{v,b} = \text{Err}_{v,b} - \text{Err}_b.$$

The prediction error for the noised up data will increase if v has a real effect in the model. Hence, comparing this prediction error to the original prediction error will yield a positive vimp index $\Delta_{v,b}$ if v is predictive. The vimp index for v is obtained by averaging these values over the bootstrap realizations:

$$\Delta_v = \frac{1}{B}\sum_{b=1}^{B} \Delta_{v,b} = \frac{1}{B}\sum_{b=1}^{B} \left[\text{Err}_{v,b} - \text{Err}_b\right].$$

For the reasons discussed above, it follows that a large positive value indicates a variable v that has a large test-validated effect size (predictive effect size). Algorithm 1 provides a formal statement of this procedure.

Algorithm 1 *VIMP index for a variable* v

for $b = 1, \ldots, B$ **do**
  Draw a bootstrap sample of the data.
  Fit the model to the bootstrap data.
  Calculate the prediction error, Err_b, using the OOB data.
  Noise up the OOB data for v.
  Calculate the prediction error, $\text{Err}_{v,b}$, using the noised up OOB data.
  Calculate the boostrap VIMP index $\Delta_{v,b} = \text{Err}_{v,b} - \text{Err}_b$
end for
Calculate the VIMP index by averaging: $\Delta_v = \sum_{b=1}^{B} \Delta_{v,b}/B$.
The OOB error for the model can also be obtained using $\text{Err}_{\text{oob}} = \sum_{b=1}^{B} \text{Err}_b/B$.

We make several remarks regarding the implementation of Algorithm 1.

1. As stated, the algorithm provides a VIMP index for a given variable v, but in practice

one applies the same procedure for all variables in the model. The same bootstrap samples are to be used when doing so. This is required because it ensures that the VIMP index for each variable is always compared to the same value Err_b.

2. Because all calculations are run independently of one another, Algorithm 1 can be implemented using parallel processing. This makes the algorithm extremely fast and scalable to big data settings. The most obvious way to parallelize the algorithm is on the bootstrap sample. Thus, on a specific computing machine on a cluster, a single bootstrap sample is drawn and Err_b determined. Steps 3-7 are then run for each variable in the model for the given bootstrap draw. Results from different computing machines on the computing cluster are then averaged as in Steps 9 and 10.
3. As mentioned earlier, noising up a variable is typically done by permuting its data. This approach is what is generaly used by nonparametric regression models. In the case of parametric and semiparametric regression models (such as Cox regression), in place of permutation noising up, the OOB data for the variable v is set to zero. This is equivalent to setting the regression coefficient estimate for v to zero which is the convenient way of implementing this procedure. Setting the coefficient to zero is a special feature of parametric models that provides a more direct and convenient way to noise up the data than permutation noising up used by nonparametric models.
4. As a side effect, the algorithm can also be used to return the OOB error rate for the model, Err_{oob} (see Step 10). This can be useful for assessing the effectiveness of the model and identifying poorly constructed models.
5. Algorithm 1 requires being able to calculate prediction error. The type of prediction error used will be context specific. For example in linear regression, prediction error can be measured using mean-squared-error, or standardized mean-squared errror. In classification problems, prediction error is typically defined by misclassification. In survival problems, a common measure of prediction performance is the Harrell's

concordance index. Thus unlike the p-value, the interpretation of the VIMP index will be context specific.

2.3.3 Risk factors for systolic heart failure

To illustrate VIMP, we consider a survival data set previously analyzed in Hsich et al. (2011). The data involves 2231 patients with systolic heart failure who underwent cardiopulmonary stress testing at the Cleveland Clinic. Of these 2231 patients, during a mean follow-up of 5 years, 742 died. In total, 39 variables were measured for each patient including baseline characteristics and exercise stress test results. Specific details regarding the cohort, exclusion criteria, and methods for collecting stress test data are discussed in Hsich et al. (2011).

We used Cox regression to fit the data using all cause mortality for the survival endpoint (as was used in the original analysis). Only linear variables were included in the model (i.e. no attempt was made to fit non-linear effects). Prediction error was assessed by the Harrell's concordance index as described in Ishwaran et al. (2008). For improved interpretation, prediction error was multiplied by 100. This is helpful because the resulting VIMP becomes expressible in terms of a percentage. For example, a VIMP index of 5% indicates a variable that improves by 5% the ability of the model to rank patients by their risk. We should emphasize once again that VIMP is cross-validated and provides a measure of predictive effect size.

Table 2.1 lists the results from the Cox regression analysis and from applying Algorithm 1 with $B = 1000$ replications. The first column lists patient variables. The second column with entry $\hat{\beta}$ lists the corresponding coefficient estimates obtained from the Cox regression of the original (non-bootstrapped) data. Column 3 is the p-value for the coefficient estimates of column 2. Column 4, $\hat{\beta}_{\text{inbag}}$, is the averaged coefficient estimates from the $B = 1000$ bootstrap Cox regression models. Notice that column 4 agrees closely with

Table 2.1: *Results from analysis of systolic heart failure data.*

	Cox Regression		VIMP			Marginal VIMP
Variable	$\hat{\beta}$	p-value	$\hat{\beta}_{\text{inbag}}$	Δ_β	Err_{step}	$\Delta_\beta^{\text{marg}}$
Peak VO_2	-0.06	0.002	-0.06	1.94	32.40	0.25
BUN	0.02	0.000	0.02	1.67	30.81	0.37
Exercise time	0.00	0.008	0.00	1.37	30.80	0.08
Male	0.47	0.000	0.47	0.52	30.01	0.37
beta-blocker	-0.23	0.006	-0.23	0.30	29.34	0.16
Digoxin	0.36	0.000	0.36	0.30	29.00	0.22
Serum sodium	-0.02	0.071	-0.02	0.20	28.93	0.07
Age	0.01	0.022	0.01	0.18	28.99	-0.03
Resting heart rate	0.01	0.058	0.01	0.14	28.93	0.04
Angiotensin receptor blocker	0.26	0.067	0.27	0.13	28.92	0.02
LVEF	-0.01	0.079	-0.01	0.11	28.86	0.03
Aspirin	-0.21	0.018	-0.21	0.11	28.83	0.03
Resting systolic blood pressure	0.00	0.158	0.00	0.07	28.83	0.00
Diabetes insulin treated	0.26	0.057	0.25	0.07	28.87	-0.02
Previous CABG	0.11	0.316	0.12	0.07	28.86	-0.02
Coronary artery disease	0.12	0.284	0.12	0.06	28.92	-0.04
Body mass index	0.00	0.800	0.00	0.00	28.96	-0.05
Potassium-sparing diuretics	-0.14	0.134	-0.14	-0.03	28.97	-0.01
Previous MI	0.29	0.012	0.30	-0.03	29.02	-0.01
Thiazide diuretics	0.04	0.707	0.04	-0.04	29.07	-0.05
Peak respiratory exchange ratio	0.12	0.701	0.12	-0.04	29.12	-0.05
Statin	-0.12	0.183	-0.13	-0.04	29.19	-0.07
Antiarrythmic	0.04	0.700	0.04	-0.04	29.25	-0.06
Diabetes noninsulin treated	0.01	0.930	0.00	-0.05	29.30	-0.06
Dihydropyridine	0.03	0.851	0.03	-0.05	29.35	-0.05
Serum glucose	0.00	0.486	0.00	-0.05	29.42	-0.07
Previous PCI	-0.06	0.557	-0.06	-0.05	29.48	-0.05
ICD	0.04	0.676	0.03	-0.05	29.55	-0.07
Anticoagulation	-0.01	0.933	-0.01	-0.06	29.61	-0.06
Pacemaker	-0.02	0.851	-0.01	-0.06	29.67	-0.06
Current smoker	0.03	0.807	0.03	-0.06	29.74	-0.06
Nitrates	-0.04	0.623	-0.04	-0.06	29.80	-0.06
Serum hemoglobin	0.00	0.923	0.01	-0.06	29.87	-0.07
Black	0.07	0.589	0.06	-0.07	29.95	-0.08
Nondihydropyridine	-0.30	0.510	-0.51	-0.07	30.03	-0.08
Loop diuretics	-0.07	0.541	-0.08	-0.07	30.09	-0.06
ACE inhibitor	0.10	0.371	0.11	-0.09	30.15	-0.06
Vasodilators	-0.08	0.606	-0.07	-0.09	30.25	-0.09
Creatinine clearance	0.00	0.624	0.00	-0.11	30.31	-0.06

column 2, which is to be expected if the number of iterations B is selected suitably large. Researchers can in fact use the closeness of column 2 to column 4 in their analyses as a way to assess if they have selected B appropriately large. Note that Table 2.1 has been sorted in terms of the VIMP index; these values are provided in column 5 under the entry Δ_β. It is interesting to observe that the magnitude of VIMP does not always match the corresponding p-value. For example, resting heart rate has a p-value of 6% which is very close to the widely used 5% cutoff value used as evidence of an important scientific effect. In contrast, however, its VIMP of 0.14% is relatively small compared with other variables. For example, the top variable identified by Algorithm 1 is peak VO_2 with a VIMP of 1.9%, which is over 13 times larger.

In addition to peak VO_2, Algorithm 1 identifies BUN and treadmill exercise time as two additional variables having large VIMP indices. Interestingly, the ranking of these three variables are identical to that in Hsich et al. (2011) obtained using a random survival machine learning analysis. Following these three variables is an assortment of variables with moderate VIMP: sex, use of beta-blockers, use of digoxin, serum sodium level, and age of patient. Then there are variables with small but non-zero VIMP, starting with patient resting heart rate, and terminating with presence of coronary artery disease. VIMP indices become zero or negative for the remaining variables.

These latter variables, with zero or negative VIMP indices, can be viewed as "noisy" variables which not only contribute no positive effect, but actually degrade model performance. This can be seen by considering column 6 of Table 2.1, labeled as Err_{step}. This column equals the OOB prediction error for each of the stepwise models ordered by VIMP. Table 2.2 lists the stepwise models that were considered:

Table 2.1 shows that Err_{step} decreases for models with positive VIMP variables, but rises once models begin to include noisy variables with zero or negative VIMP. Note that

Table 2.2: *Stepwise models used in calculating Err_{step}.*

Model Number	Stepwise Model
1	Model using the top variable only, {peak VO_2}
2	Model using top two variables, {peak VO_2, BUN}
3	Model using top three variables, {peak VO_2, BUN, exercise time}
⋮	⋮
39	Model using all 39 variables

because prediction error will be optimistic for models based on ranked variables, we reduce bias by calculating Err_{step} using the same bootstrap samples used by Algorithm 1. Thus, the value 30.31 in the last row of column Err_{step}, corresponding to fitting the entire model, coincides exactly with the OOB model prediction error obtained using Algorithm 1.

2.3.4 Marginal VIMP

Now we explain the meaning of the column entry $\Delta_\beta^{\text{marg}}$ in Table 2.1. To understand this, we return to the stepwise error, Err_{step}, listed in the table. Recall that Err_{step} measures the OOB prediction error for a specific stepwise model. Relative to its previous entry, it estimates the effect of a variable when it is added to the current model. To be concrete, consider Err_{step} for the third stepwise model:

Model containing the top three variables, {peak VO_2, BUN, exercise time}.

The value for Err_{step} is 30.80. Now since the value for Err_{step} for the second stepwise model (using the first two variables) is 30.81, we can conclude that the predictive effect size for adding exercise time is 0.01 (30.81 minus 30.80). This is much smaller than the VIMP index, Δ_β, for exercise time which equals 1.37. These values differ because the stepwise

error rate estimates the effect of *adding* treadmill exercise time to the model with Peak VO_2 and BUN. We call this the discovery effect size of the variable. The discovery effect measures a different predictive effect than the VIMP index. For exercise time, the VIMP index estimates the predictive effect size of exercise time in the *model using all variables* (including exercise time).

The stepwise error rate we constructed only estimates discovery effects for the specific stepwise models considered. It would be more helpful to calculate the discovery effect of a variable compared to the model containing all variables except that variable. This is what we call the marginal vimp, $\Delta_\beta^{\text{marg}}$. Table 2.3 summarizes these concepts.

Table 2.3: *Difference between VIMP and marginal VIMP.*

VIMP is calculated through noising up a variable.
Marginal VIMP is calculated through removing a variable.

Note that removing a variable from the model will change the coefficients of other variables, while noising up a variable will not change those.

The marginal VIMP is easily calculated by a simple modification to Algorithm 1. In place of noising up a variable v, a second model is fit to the bootstrap data, but with v removed. The OOB error for this model is compared to the OOB error for the full model containing all variables. Averaging these values over the bootstrap realizations yields Δ_v^{marg}. See Algorithm 2 for a formal description of this procedure.

Comparing the marginal VIMP of column 7 to the VIMP of column 5, we observe some very interesting differences. A first observation is that marginal VIMP is generally much smaller than VIMP. We can conclude that the discovery effect size is a conservative measure, as we would expect given the large number of variables in our model. Second,

Algorithm 2 *Marginal VIMP for a variable* v

1: **for** $b = 1, \ldots, B$ **do**
2: Draw a bootstrap sample of the data.
3: Fit the model to the bootstrap data and calculate its prediction error, Err_b, using the OOB data.
4: Fit a second model, but without variable v, and calculate its predicticition error, $\text{Err}^{\text{marg}}_{v,b}$ using the OOB data.
5: **end for**
6: Calculate the marginal VIMP by averaging: $\Delta^{\text{marg}}_v = \sum_{b=1}^{B} \left[\text{Err}^{\text{marg}}_{v,b} - \text{Err}_b\right] / B$.

as expected, the discovery effect of exercise time is substantially smaller than its VIMP. Third, there is a small collection of variables whose discovery effect is relatively large compared to their VIMP. The most interesting is sex, which has the largest discovery effect among all variables (being tied with BUN). The explanation for this is that adding sex to the model supplies new information not provided by other variables. Marginal VIMP is in some sense a statement about correlation. For example, the correlation of exercise time with peak VO_2 is 0.87, whereas the correlation of BUN with peak VO_2 is -0.40. Thus when peak VO_2 is included in the model, BUN is able to have a high discovery effect, while exercise time cannot. Differences between marginal VIMP and VIMP indices are conveniently summarized in Figure 2.4. For example, the right-hand plot displays the ranking of variables by the two methods. There is some overlap in the top variables (points in lower left hand side), but generally we see important differences.

2.3.5 Robustness of VIMP to the sample size

Here we demonstrate the robustness of VIMP to the sample size (n). We implemented the same procedures as before to the systolic heart failure data, but this time using only a fraction of the data. We used 10%, 25%, 50%, and 75% of the data. That is we subsampled the data without replacement, drawing a data set with sample size Fn, where F was the desired fraction of the data. This process was repeated 500 times independently. For each

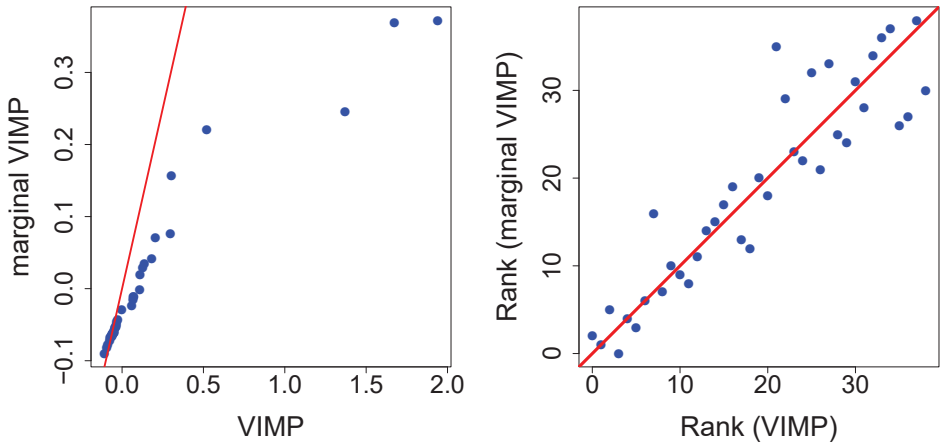

Figure 2.4: *Differences between marginal VIMP and the VIMP index for systolic heart failure data. Left-hand figure displays the two values plotted against each other. Right-hand figure compares the ranking of variables by the two methods.*

data set, we saved the p-values and VIMP indices for all variables. Figure 2.5 displays the logarithm of the p-values from the experiment (large negative values correspond to near zero p-values). Figure 2.6 displays the VIMP indices. What is most noticeable from Figure 2.6 is that VIMP indices are informative even in the extremely low sample size setting of 10%. For example, VIMP interquartile values (the lower and upper ends of the boxplot) are above zero for peak VO_2, BUN, and treadmill exercise time, showing that VIMP is able to consistently identify the top three variables even with limited data. In contrast, in Figure 2.5 for the low sample setting of 10%, no variable had a median log p-value below the threshold of $\log(0.05)$; showing that no variable met the 5% level of significance on average. Furthermore, even with 75% of the data, the upper end of the boxplot for exercise time is still above the threshold, showing its significance is questionable. These results demonstrate the sensitivity of p-values to sample size in contrast to the robustness of VIMP.

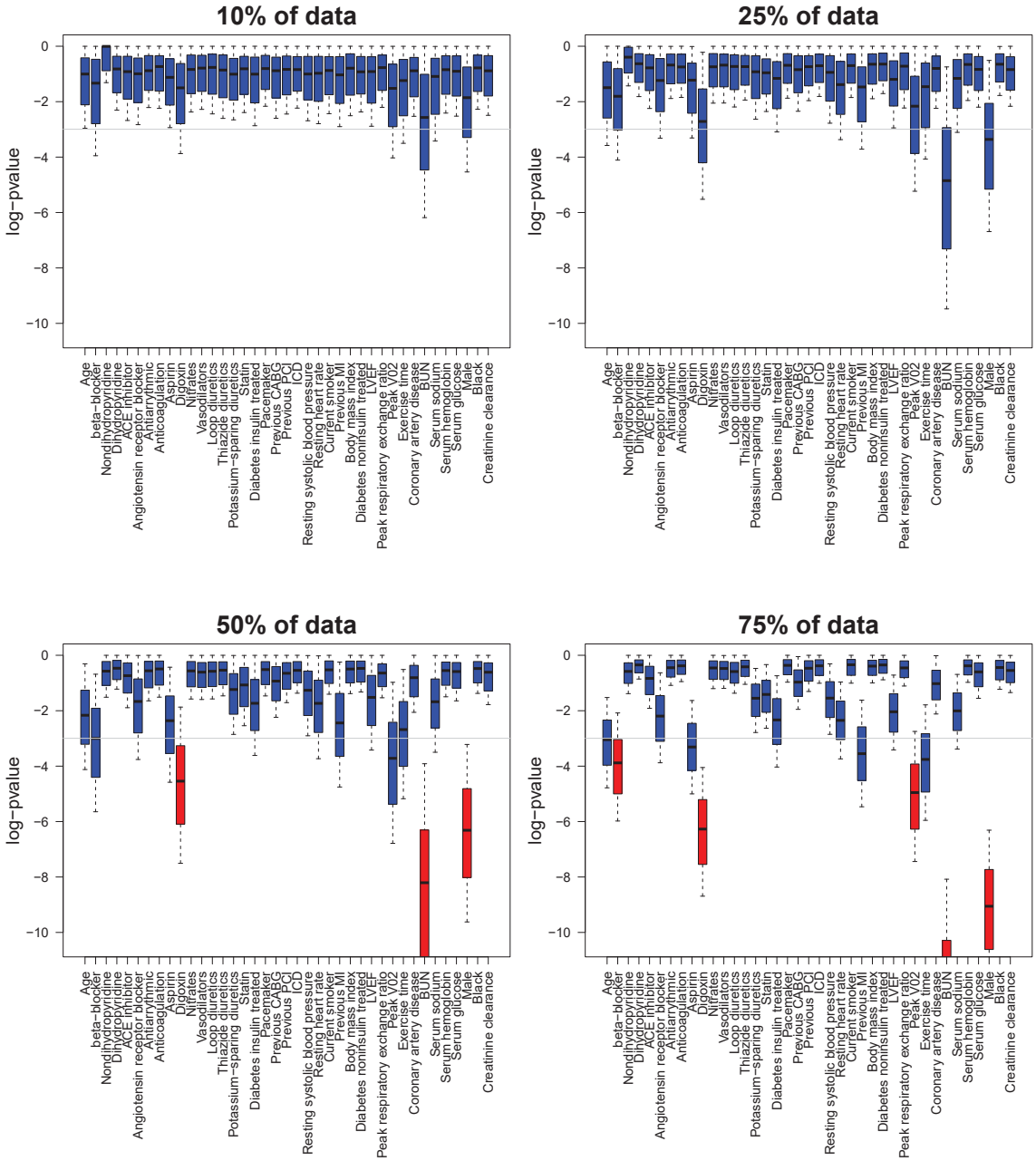

Figure 2.5: *Logarithm of p-value as a function of fraction of sample size for systolic heart failure data (large negative values correspond to near zero p-values). Values are calculated using 500 independently subsampled data sets. Horizontal line is* $\log(0.05)$*, the typical threshold used to identify a significant variable.*

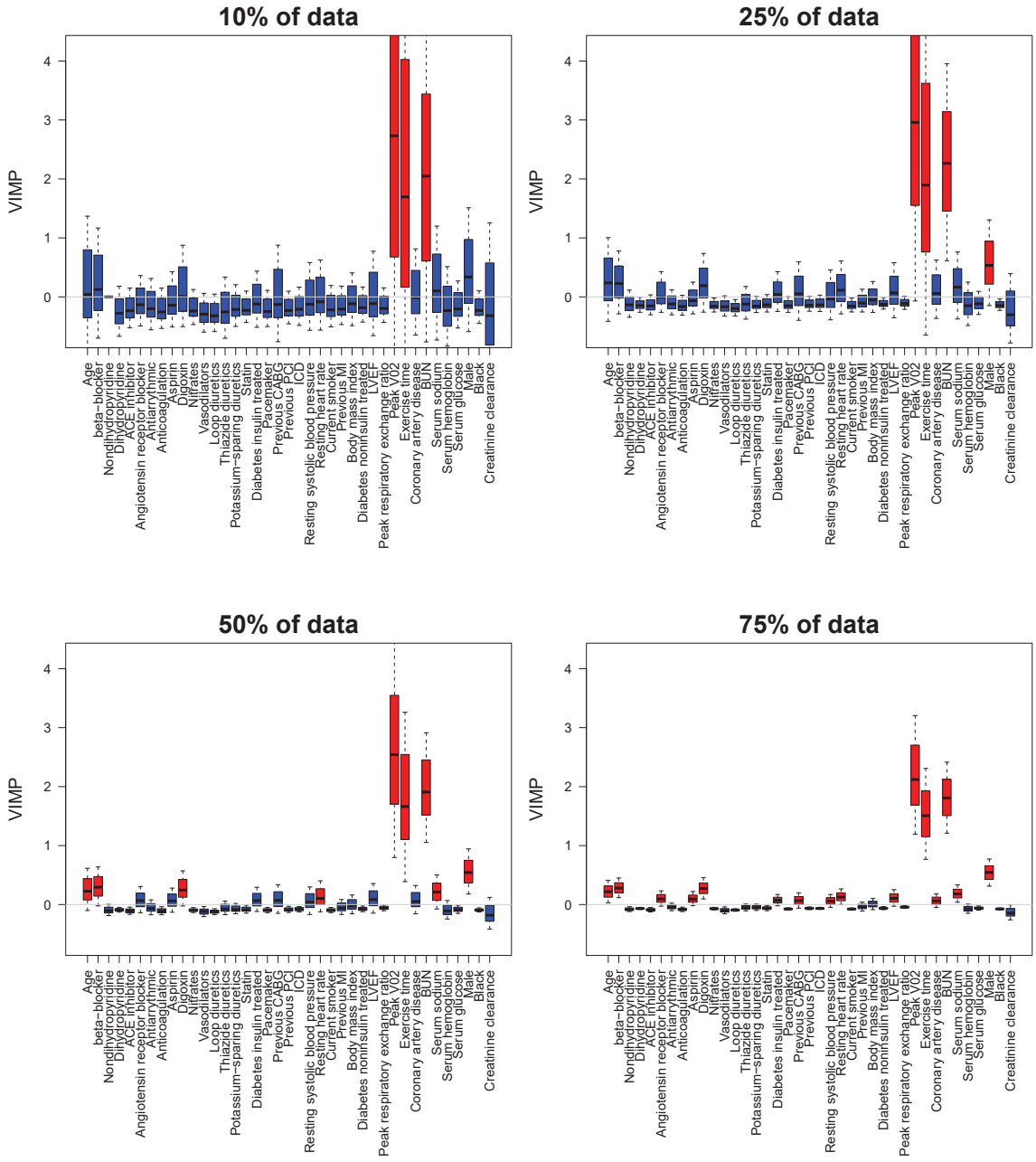

Figure 2.6: *Subsampled data is the same as Figure 2.5 but where VIMP is now reported.*

2.3.6 Misspecified model

For our second illustration we use a simulated survival data set. We use this to show the robustness of VIMP under model misspecification. For our simulation, we sampled $n = 1000$ values from a Cox regression model with five variables. The first two variables are "psa" and "tumor volume" and represent variables associated with the survival outcome. The remaining three variables are noise variables with no relationship to the outcome. These are called X_1, X_2, X_3. The variable psa has a linear main effect, but tumor volume has both a linear and non-linear term. The true regression coefficient for psa is 0.05 and the coefficient for the linear term in tumor volume is 0.01. A censoring rate of approximately 70% was used. The log of the hazard function used in our simulation is given in the left panel of Figure 2.7. Mathematically, our log-hazard function assumes the following function

$$\log(h(t)) = \alpha_0 + 0.05 \times \text{psa} + 0.01 \times \text{tumor volume} + \psi(\text{tumor volume})$$

where $\psi(x) = 0.04x^2 - 0.005x^3$ is a polyomial function with quadratic and cubic terms. The right panel of Figure 2.7 displays the log-hazard for the misspecified model that does not include the non-linear term for tumor volume.

We first fit a Cox regression model to the data using only linear variables as one might typically do. Following this, Algorithms 1 and 2 were applied with $B = 1000$. The entire procedure was then repeated $M = 1000$ times. Each of these Monte Carlo runs consisted of simulating a new data set, fitting a Cox regression model to this simulated data, and running Algorithms 1 and 2. The results are summarized in Table 2.4. All reported values are averaged over the $M = 1000$ Monte Carlo experiments.

Table 2.4 shows that the p-value has no difficulty in identifying the strong effect of psa, which is correctly specified in the model. However, the p-value for tumor volume is

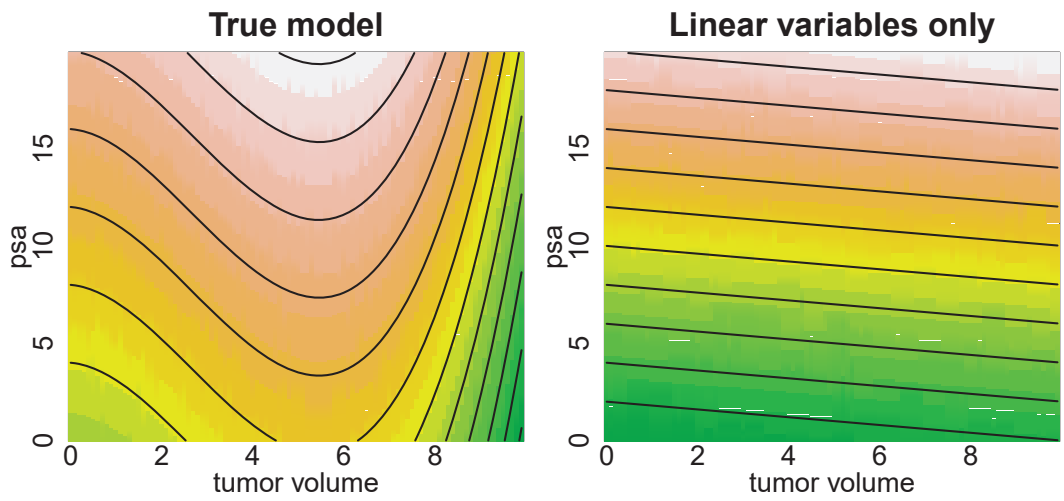

Figure 2.7: *Log-hazard function from Cox simulation example. Left figure displays the true log-hazard function which includes the non-linear term for tumor volume. Right figure displays the log-hazard function assuming linear variables only.*

0.267 which indicates a non-significant effect. The p-value tests whether this coefficient is zero, assuming the model is true, but the problem is that the fitted model is misspecified. The estimated Cox regression model inflates the coefficient for tumor volume in a negative direction (estimated value of -0.03, but true value is 0.01) in an attempt to compensate for the non-linear effect that was excluded from the model. This leads to the invalid p-value. In contrast, both the VIMP and marginal VIMP values for tumor volume are positive. Although these values are substantially smaller than the values for psa, VIMP is still able to identify an predictive effect size associated with tumor volume. Once again, this is possible because VIMP bases its estimation on test data and not a presumed model which can be incorrect. Also, notice that all three noise variables are correctly identified as uniformative. All have negative VIMP values.

Typically, a standard analysis would end after looking at the p-values in Table 2.4. However, a researcher with access to the entire table, might be suspicious of the nega-

Table 2.4: *Results from analysis of simulated Cox regression data set. The model is misspecified by failing to include the non-linear term for tumor volume.*

	$\hat{\beta}$	p-value	$\hat{\beta}_{\text{inbag}}$	Δ_β	$\Delta_\beta^{\text{marg}}$
psa	0.05	0.001	0.05	6.32	6.34
tumor volume	-0.03	0.267	-0.03	0.14	0.15
X_1	0.00	0.490	0.00	-0.25	-0.25
X_2	0.00	0.486	0.00	-0.25	-0.25
X_3	0.00	0.493	0.00	-0.27	-0.27

The overall OOB model error is 43%.

tive coefficient estimate for tumor volume, and they would be alerted by its small positive VIMP. This combined with the high OOB model error (equal to 43%) would alert them to consider more sophisticated modeling. This is easily done using standard statistical methods. Here we use B-splines (Eilers and Marx, 1996) to add non-linearity to tumor volume. This expands the design matrix for the Cox regression model to include additional columns for the the B-spline expansion of tumor volume. When noising up tumor volume all of these B-spline columns are noised up simultaneously (i.e. their coefficient estimates are set to zero). The extensions to Algorithms 1 and 2 are straightforward.

Table 2.5: *Results from Cox regression simulation using a B-spline to model non-linearity in tumor volume.*

	Δ_β	$\Delta_\beta^{\text{marg}}$
psa	4.20	4.23
tumor volume	2.27	2.31
X_1	-0.20	-0.20
X_2	-0.20	-0.20
X_3	-0.21	-0.21

The overall OOB model error is 40%.

The results from the B-spline analysis are displayed in Table 2.5. As before, the entire procedure was repeated $M = 1000$ times, with values averaged over the Monte Carlo runs.

Notice the large values of VIMP for tumor volume. The overall model performance has also improved to 40%. Overall, results have improved substantially.

2.3.7 Discussion

It seems questionable that the p-value can continue to meet the needs of scientists. It does not provide an interpretable scientific effect size that researchers desire and it is valid only if the underlying model holds, which can often be questionable given the restrictive assumptions often used with traditional modeling. In this section, I introduced VIMP as an alternative approach. VIMP provides an interpretable measure of effect size that is robust to model misspecification. It uses prediction error based on out-of-sample data and replaces statistical significance with predictive importance. The VIMP framework is feasible to all kinds of models including not only parametric models, such as those considered here, but also non-parametric models such as those used in machine learning approaches.

We discussed two types of VIMP measures: the VIMP index and the marginal VIMP. The scientific application will dictate which of these is more suitable. VIMP indices are appropriate in settings where variables for the model are already established and the goal is to identify the predictive effect size. For example, if several genetic markers are already identified as a genetic cause for coronary heart disease risk, VIMP can provide a rank for these and estimate the magnitude each marker plays in the prediction for the outcome. Marginal VIMP is appropriate when the goal is new scientific discovery. For instance, if a researcher is proposing to add a new genetic marker for evaluating coronary heart disease risk, marginal VIMP can yield a discovery effect size for how much the new proposed marker adds to previous risk models.

From a statistical perspective, VIMP idices are an OOB alternative to the regression coefficient p-value. However, what VIMP measures about a variable can be very flexible. It may be a linear effect, or quite easily a non-linear effect, such as modeled using B-splines.

An important feature is that degrees of freedom and other messy details required with p-values when dealing with complex modeling are never an issue with VIMP. Marginal VIMP is an OOB analog to the likelihood-ratio test. In statistics, likelihood-ratio tests compare the goodness-of-fit of two models, one of which (the null model with certain variables removed) is a special case of the other (the alternative model with all variables included). Marginal VIMP compares the prediction precision of these two scenarios.

Because both VIMP and marginal VIMP are measures of predictive importance, their values are standardized to the measure of prediction performance used. This makes it possible to compare values across different data sets. For example, a 0.05 VIMP value for two different variables from two different survival datasets is comparable—both imply a 5% contribution to the concordance index. Another feature which we touched upon briefly in our B-spline example is the ability to use VIMP to measure the effect of groups of variables. In our B-spline example, the cluster of variables used were the B-spline contributions to tumor volume, and were combined together to give an overall estimate of the effect of tumor volume. One could easily extend this to calculate cluster-VIMP as a better sense of the importance of a highly correlated group of variables.

2.4 Checking proportional hazards assumption using RF

Use T to denote the response variable, since the response is usually the time until an event. Define *survival function*, $S(t)$, as

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

where $F(t)$ is the cumulative distribution function for T. If the event is death, $S(t)$ is the probability that death occurs after time t, or the probability that the subject will survive at

least until time t. All subjects survive at least to time zero, so $S(0) = 1$. The accumulated risk up until time t is *cumulative hazard function*, denoted by $\Lambda(t)$. The hazard at time t, $\lambda(t)$, is related to the probability that the event will occur in a small interval around t, given that the event has not occurred before time t. $\lambda(t)$ is called the *hazard function*, or the *force of mortality*, or *instantaneous event (death, failure) rate*.

$$\lambda(t) = \lim_{u \to 0} \frac{P\{t < T \leq t + u | T > t\}}{u},$$

which becomes

$$\begin{aligned}
\lambda(t) &= \lim_{u \to 0} \frac{P\{t < T \leq t + u\}/P\{T > t\}}{u} \\
&= \lim_{u \to 0} \frac{[F(t+u) - F(t)]/u}{S(t)} \\
&= \frac{[F(t+u) - F(t)]}{S(t)} \\
&= \frac{\partial F(t)/\partial t}{S(t)} \\
&= \frac{f(t)}{S(t)} \\
&= -\frac{\partial \log S(t)}{\partial t},
\end{aligned}$$

where $f(t)$ is the probability density function of T evaluated at t. The integral of $\lambda(t)$ gives:

$$\Lambda(t) = \int_0^t \lambda(v) dv = -\log S(t) \quad \text{and} \quad S(t) = \exp\big[-\Lambda(t)\big].$$

To investigate ITE, one has to relate $S(t)$ with all the covariates $\mathbf{X}$. The form of the true population survival distribution function $S(t)$ is almost always unknown. For $S(t|\mathbf{x})$, $\lambda(t)$ is the key to explore mechanism of survival, and the most widely used survival regression

specification is to allow the hazard function $\lambda(t)$ to be multiplied by $exp(\mathbf{X}\beta)$:

$$\lambda(t|\mathbf{X}) = \lambda(t)exp(\mathbf{X}\beta). \tag{2.3}$$

$\lambda(t)$, sometimes called an *underlying hazard function* or a *hazard function for a standard subject* with $\mathbf{X}\beta = 0$, can be left completely unspecified without sacrificing the ability to estimate β by using Cox's semi-parametric proportional hazards (PH) model. Regression formulation (2.3) is called PH model, which generates:

$$S(t|\mathbf{X}) = S(t)^{exp(\mathbf{X}\beta)}. \tag{2.4}$$

To interprete β, equation (2.3) suggests $exp(\beta_i)$ as hazard ratio in unite change of in X_i:

$$\begin{aligned} \mathbf{X}^* : \mathbf{X} \text{ hazard ratio} &= \lambda(t)exp(\mathbf{X}^*\beta)/\lambda(t)exp(\mathbf{X}\beta) \\ &= exp(\mathbf{X}^*\beta)/exp(\mathbf{X}\beta) = exp\big[(\mathbf{X}^* - \mathbf{X})\beta_i\big] \\ &= exp(\beta_i), \end{aligned} \tag{2.5}$$

where $\mathbf{X}^*$ and $\mathbf{X}$ is the same except in the ith dimension $X_i^* - X_i = 1$.

One can also use treatment indicator variable as one of the covariates in X and interprete ATE to be $exp(\beta)$ as hazard ratio after matching treatment group and control group on propensity score or raw observed variables. Whenever Cox regression model is used, checking PH assumption is essential.

This section is about how to make sure Equation (2.3) is true. Note that PH assumption has to hold for each X_i. Let $h(t|\mathbf{x}) = \log\{-\log[S(t|\mathbf{x})]\}$, Equation (2.4) gives:

$$\begin{aligned} h(t|\mathbf{x}) &= \mathbf{x}\beta + \log\{-\log[S(t)]\} \\ &= \mathbf{x}\beta + \log[\Lambda(t)]. \end{aligned}$$

For $t_1 \neq t_2$,

$$h(t_1|\mathbf{x}) - h(t_2|\mathbf{x}) = \log\left[\Lambda(t_1)\right] - \log\left[\Lambda(t_2)\right] \tag{2.6}$$

holds for each dimension X_i. Define distance $d_{x_i}(t_1, t_2) = h(t_1|x_i) - h(t_2|x_i)$. From Equation (2.6),

$$d_{x_i}(t_1, t_2) \text{ is constant for any } x_i \in \mathbf{X}_i. \tag{2.7}$$

In other words, checking PH assumption for X_i is to check equal distance between log-log conditional survival curve $S(t|X_i)$.

I will use the partial plot from Random Survival Forest (Ishwaran et al., 2008; Ishwaran and Kogalur, 2017) to get conditional survival $S(t|X_i)$. It plots the marginal effect of an x variable on the class probability (in classification problem), response (in regression problem), mortality (in survival problem), or the expected years lost (in competing risk problem) from a random forest analysis after adjusting for other variables. The partial plot of the random forest model from Section 2.3 is demonstrated in Figure 2.8. The vertical axis on the partial plot is $-\log(-\log(S(t|X_i)))$ across different time. According to the defination of proportional hazards assumption, lines of different time should be paralell.

Formally, to check PH assumption in X_i, randomly choose M pairs of time points $\{(t_{11}, t_{12}), \ldots, (t_{m1}, t_{m2}), \ldots, (t_{M1}, t_{M2})\}$ and randomly choose N points $\{x_1 \ldots x_j \ldots x_N\}$ in X_i. $M \times N$ distances $d_{m,j}$ are calculated through $d_{m,j} = \log\{-\log\left[S(t_{m1}|x_j)\right]\} - \log\{-\log\left[S(t_{m2}|x_j)\right]\}$ for $x_j \in X_i$, $j = 1, \ldots, N$ and $m = 1, \ldots, M$. Define $e_{m,j} = d_{m,j} - d_{m,\cdot}$, where $d_{m,\cdot} = \frac{1}{N}\sum_{j=1}^{N} d_{m,j}$, and $e_{m,j}$ can be tested through Sign test: let W be the number of sign "+" for which $e_{m,j} > 0$; then W follows a binomial distribution

$$W \sim \mathcal{B}(M \times N, 0.5).$$

I chose three variables from the survival data in Section 2.3, made the partial plot in Figure 2.8 and used Sign test for checking PH assumption in Table 2.6.

Table 2.6: PH assumption checking for survial data in Section 2.3.

Variable	p-value	95% LL	95% UL	PH
Peak VO_2	0.39	0.36	0.41	UnSatisfied
beta-blocker	0.51	0.40	0.62	Satisfied
Potassium-sparing diuretics	0.51	0.40	0.62	Satisfied

Note: 95% LL=low limit for 95% confidence interval of p value; 95% UL=upper limit for 95% confidence interval of p value. If the 95% confidence interval of p value from Sign test covers 0.5, the PH assumption is satisfied.

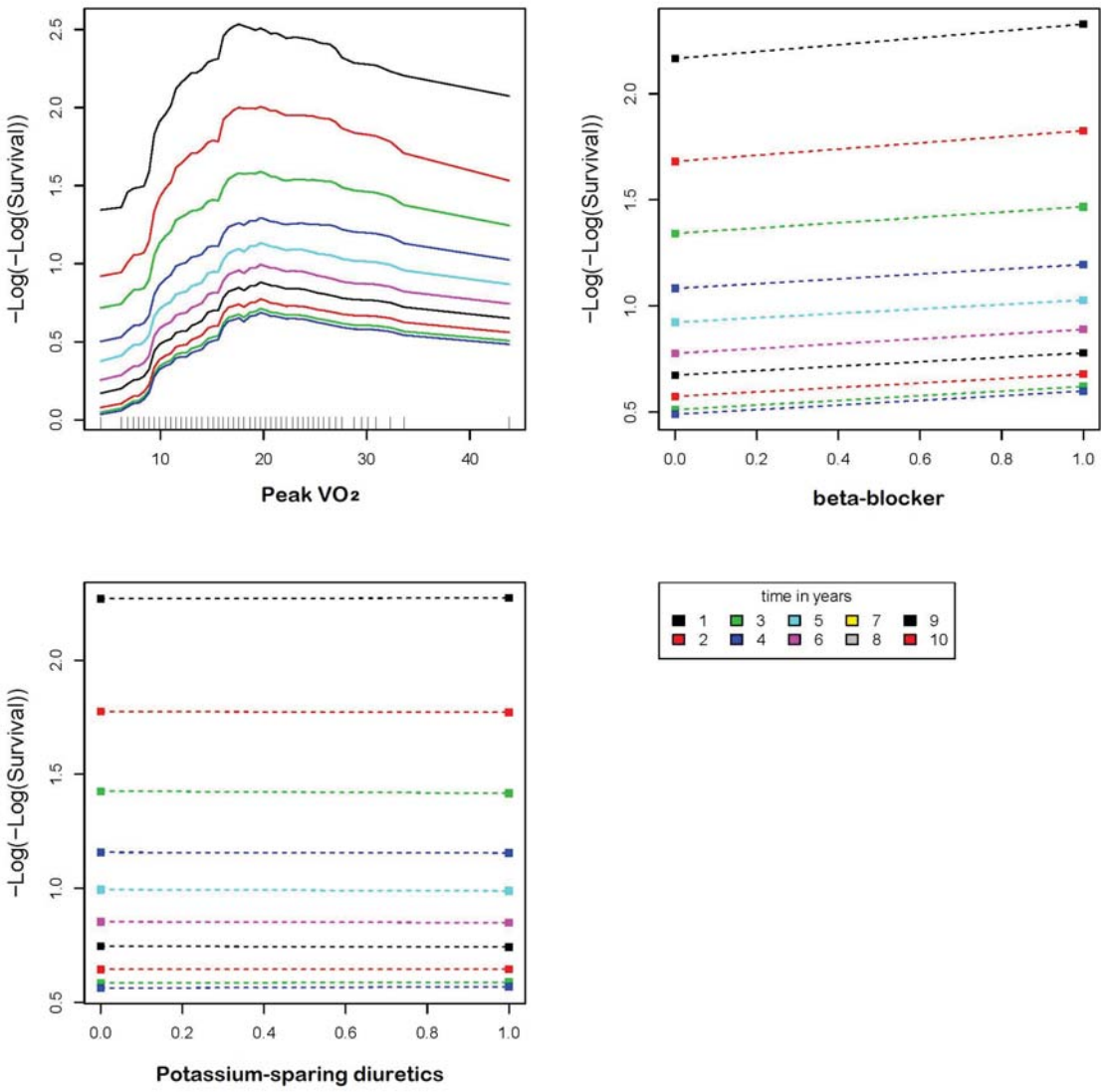

Figure 2.8: Partial plots for survial data in Section 2.3. Dash line displays categorical variable and solid line displays continuous variable.

Chapter 3

RF methods for estimating individual treatment effects

This chapter describes several RF methods used for estimating the ITE. Each method can be considered as a form of g in equation (1.6). Note that in equation (1.6), treatment variable T is considered as fixed; therefore, each of the following method provides a direct means for estimating the ITE without making use of the propensity score. As with models incorporating the propensity score, these methods will account for confounding as long as all confounding factors are observed and included in the feature set. Here I describe the proposed RF methods for estimating the ITE. The methods considered in this dissertation are as follows:

1. Virtual twins (VT).
2. Virtual twins interaction (VT-I).
3. Counterfactual RF (CF).
4. Counterfactual synthetic RF (synCF).
5. Bivariate RF (bivariate).
6. Honest RF (honest RF).

Virtual twins is the original method proposed by Foster et al. (2011) mentioned earlier. We also consider an extension of the method, called virtual twins interaction, which includes forced interactions in the design matrix for more adaptivity. Forcing treatment

interactions for adaptivity may have a limited ceiling, which is why we propose the counterfactual RF method. In this method we dispense with interactions and instead fit separate forests to each of the treatment groups. Counterfactual synthetic RF uses this same idea, but uses synthetic forests in place of Breiman forests, which is expected to further improve adaptivity. Thus, this method, and the previous RF methods, are all proposed enhancements to the original virtual twins method. All of these share the common feature that they provide a direct estimate for the ITE by estimating the regression surface of the outcome. This is in contrast to our other proposed procedure, bivariate RF, which takes a missing data approach to the problem. There has been much interest in the literature in viewing causual effect analysis as a missing data problem (Ghosh et al., 2015). Thus, we propose here a novel bivariate imputation approach using RF. In the following sections we provide more details about each of the above methods.

3.1 Virtual twins

Virtual Twins model uses a regular Breiman RF as function g in equation (1.6) and get $Y_{\mathbf{x}}(1) = g(1, \mathbf{x}, \epsilon_Y)$ and $Y_{\mathbf{x}}(0) = g(0, \mathbf{x}, \epsilon_Y)$. Figure 3.1 shows the framework of virtual twins RF approach. Foster et al. (2011) proposed a Virtual Twins (VT) approach for esti-

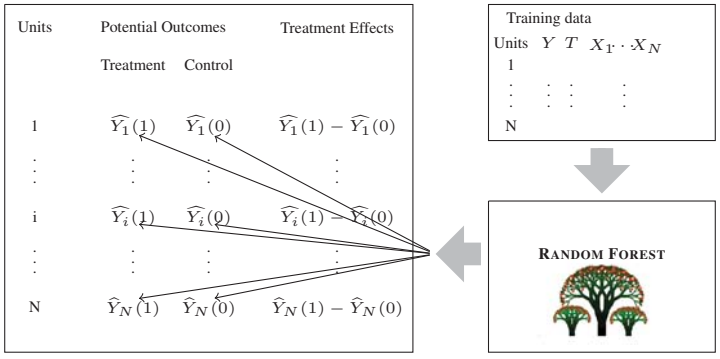

Figure 3.1: Illustration of virtual twins approach.

mating counterfactual outcomes. In this approach, RF is used to regress Y_i against $(\mathbf{X}_i, T_i)$. To obtain a counterfactual estimate for an individual i, one creates a VT data point, similar in all regards to the original data point $(\mathbf{X}_i, T_i)$ for i, but with the observed treatment T_i replaced with the counterfactual treatment $1 - T_i$. Given an individual i with $T_i = 1$, one obtains the RF predicted value $\hat{Y}_i(1)$ by running i's unaltered data down the forest. To obtain i's counterfactual estimate, one runs the altered $(\mathbf{X}_i, 1 - T_i) = (\mathbf{X}_i, 0)$ down the forest to obtain the counterfactual estimate $\hat{Y}_i(0)$. The counterfactual ITE estimate is defined as $\hat{Y}_i(1) - \hat{Y}_i(0)$. A similar argument is applied when $T_i = 0$. If $\hat{Y}_{VT}(\mathbf{x}, T)$ denotes the predicted value for $(\mathbf{x}, T)$ from the VT forest, the VT counterfactual estimate for $\tau(\mathbf{x})$ is

$$\hat{\tau}_{VT}(\mathbf{x}) = \hat{Y}_{VT}(\mathbf{x}, 1) - \hat{Y}_{VT}(\mathbf{x}, 0).$$

As noted by Foster et al. (2011), the VT approach can be improved by manually including treatment interactions in the design matrix. Thus, one runs a RF regression with Y_i regressed against $(\mathbf{X}_i, T_i, \mathbf{X}_i T_i)$. The inclusion of the pairwise interactions $\mathbf{X}_i T_i$ is not conceptually necessary for VT, but was observed to improve results. $\hat{\tau}_{VT-I}(\mathbf{x})$ denotes the ITE estimate under this modified VT interaction model. Figure 3.2 demonstrates the framework of virtual twins interaction approach.

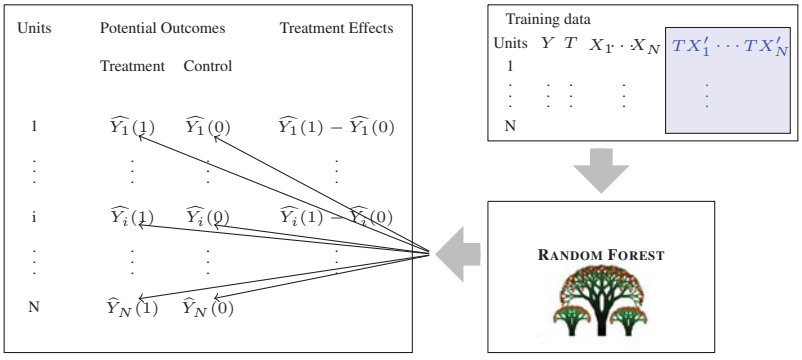

Figure 3.2: Illustration of virtual twins interaction approach.

When implementing the above procedures, out-of-bag (OOB) estimates is used whenever possible. Original experimentation with in-sample (in bag) estimates revealed they led to poorer performance due to increased variance. The advantage of OOB estimates is not made very clear in the RF literature and it is worth emphasizing this point here as readers may be unaware of this important distinction. OOB refers to out-of-sample (cross-validated) estimates and are helpful in reducing the variance of a RF estimator. Each tree in a forest is constructed from a bootstrap sample which uses approximately 63% of the data. The remaining 37% of the data is called OOB and are used to calculate an OOB predicted value for a case. The OOB predicted value is defined as the predicted value for a case using only those trees where the case is OOB (for example if 1000 trees are grown, approximately 370 are used). To illustrate how this applies to VT, suppose that case $\mathbf{x}$ is assigned treatment $T = 1$. Let $\hat{Y}^*_{VT}(\mathbf{x}, T)$ denote the OOB predicted value for VT $(\mathbf{x}, T)$. The OOB counterfactual estimate for $\tau(\mathbf{x})$ is

$$\hat{\tau}_{VT}(\mathbf{x}) = \hat{Y}^*_{VT}(\mathbf{x}, 1) - \hat{Y}_{VT}(\mathbf{x}, 0).$$

Note that $\hat{Y}_{VT}(\mathbf{x}, T)$ is not OOB. This is because $(\mathbf{x}, 0)$ is a new data point and technically speaking cannot have an OOB predicted value as the observation is not even in the training data. In a likewise fashion, if $\mathbf{x}$ were assigned treatment $T = 0$, the OOB estimate is

$$\hat{\tau}_{VT}(\mathbf{x}) = \hat{Y}_{VT}(\mathbf{x}, 1) - \hat{Y}^*_{VT}(\mathbf{x}, 0).$$

OOB counterfactual estimates for $\hat{\tau}_{VT-I}(\mathbf{x})$ are defined analogously.

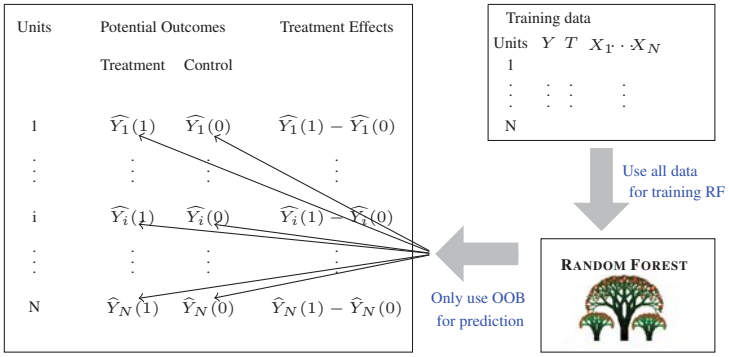

Figure 3.3: Illustration of Out-of-Bag estimates in virtual twins approach.

3.2 Counterfactual RF

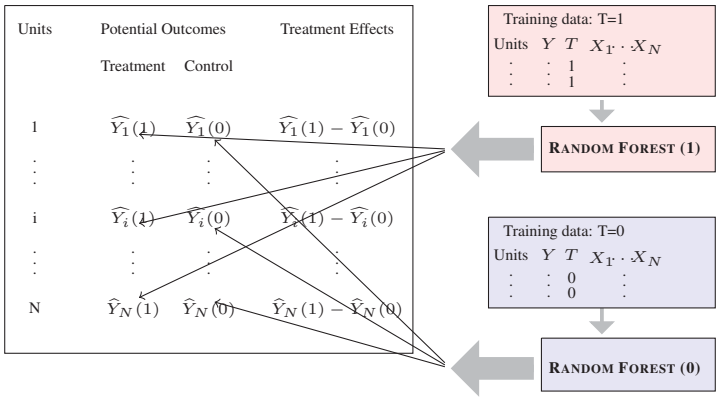

Figure 3.4: Illustration of Counterfactual RF approach.

So far g in equation (1.6) is in the same form between $(\mathbf{x}, y, T = 1)$ and $(\mathbf{x}, y, T = 0)$; another thought is setting up g as a step function as g_1 for $T = 1$ and g_0 for $T = 0$:

$$Y = g(T, \mathbf{X}, \epsilon_Y) = \begin{cases} g_0(\mathbf{X}, \epsilon_{Y0}) & T = 0 \\ g_1(\mathbf{X}, \epsilon_{Y1}) & T = 1 \end{cases}, \tag{3.1}$$

which is the idea of counterfactual RF. In an important extension to $\hat{\tau}_{VT-I}$, rather than fitting a single forest with forced treatment interactions, we instead fit a separate forest to each treatment group, shown in Figure 3.4. Doing so allows for much greater adaptivity to

differences between the two treatment groups. Forests CF_1 and CF_0 are fit separately to data $\{(\mathbf{X}_i, Y_i) : T_i = 1\}$ and $\{(\mathbf{X}_i, Y_i) : T_i = 1\}$, respectively. To obtain a counterfactual ITE estimate, each data point is run down its natural forest, as well as its counterfactual forest. If $\hat{Y}_{CF,j}(\mathbf{x}, T)$ denotes the predicted value for $(\mathbf{x}, T)$ from CF_j , for $j = 0, 1$, the counterfactual ITE estimate is

$$\hat{\tau}_{CF}(\mathbf{x}) = \hat{Y}_{CF,1}(\mathbf{x}, 1) - \hat{Y}_{CF,0}(\mathbf{x}, 0).$$

This modification to VT was mentioned briefly in the paper by Foster et al. (2011) although not implemented. A related idea was used by Dasgupta et al. (2014) to estimate conditional odds ratios by fitting separate RF to different exposure groups.

We note that just as with VT estimates, OOB values are utilized whenever possible to improve stability of estimated values. Thus, if $\mathbf{x}$ is assigned treatment $T = 1$, the OOB ITE estimate is

$$\hat{\tau}_{CF}(\mathbf{x}) = \hat{Y}^*_{CF,1}(\mathbf{x}, 1) - \hat{Y}_{CF,0}(\mathbf{x}, 0).$$

where $\hat{Y}^*_{CF,1}(\mathbf{x}, 1)$ is the OOB predicted value for $(\mathbf{x}, 1)$. Likewise, if $\mathbf{x}$ is assigned treatment $T = 0$, the OOB estimate is

$$\hat{\tau}_{CF}(\mathbf{x}) = \hat{Y}_{CF,1}(\mathbf{x}, 1) - \hat{Y}^*_{CF,0}(\mathbf{x}, 0).$$

3.3 Counterfactual synthetic RF

An substiution of Breiman RF as g_0 and g_1 in equation (1.6) is synthetic RF. In a modification to the above approach, Breiman RF regression used for the prediction $\hat{Y}_{CF,j}(\mathbf{x}, T)$ is replaced by with synthetic forest regression using synthetic forests (Ishwaran and Malley, 2014). The latter are a new type of forest designed to improve prediction performance

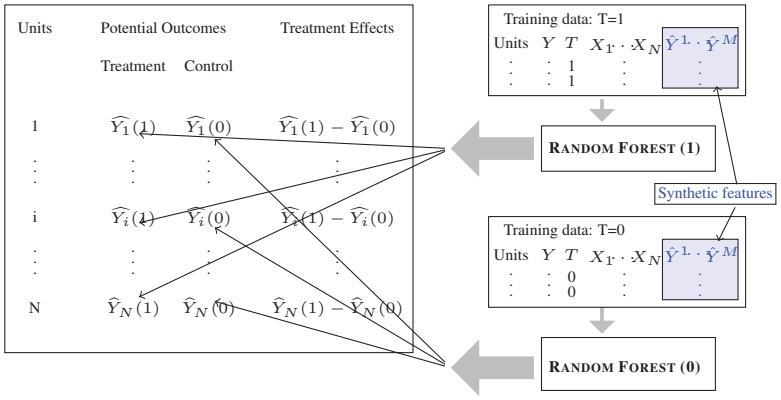

Figure 3.5: Illustration of Counterfactual synthetic RF approach.

of RF. Using a collection of Breiman forests (called base learners) grown under different tuning parameters, each generating a predicted value called a synthetic feature, a synthetic forest is defined as a secondary forest calculated using the new input synthetic features, along with all the original features. This process is shown in Figure 3.5. Typically, the base learners used by synthetic forests are Breiman forests grown under different nodesize and mtry parameters. The latter are tuning parameters used in building a Breiman forest. In RF, prior to splitting a tree node, a random subset of mtry variables are chosen from the original variables. Only these randomly selected variables are used for splitting the node. Splitting is applied recursively and the tree grown as deeply as possible while maintaining a sample size condition that each terminal node contains a minimum of nodesize cases. The two tuning parameters mtry and nodesize are fundamental to the performance of RF. Synthetic forests exploits this and uses RF base learners grown under different mtry and nodesize parameter values. To distinguish the proposed synthetic forest method from the counterfactual approach described above, we use the abbreviation synCF and denote its ITE estimate by $\hat{\tau}_{synCF}(\mathbf{x})$:

$$\hat{\tau}_{synCF}(\mathbf{x}) = \hat{Y}_{synCF,1}(\mathbf{x}, 1) - \hat{Y}^*_{synCF,0}(\mathbf{x}, 0).$$

where $\hat{Y}_{synCF,j}(\mathbf{x}, T)$ denotes the predicted value for $(\mathbf{x}, T)$ from the synthetic RF grown using data $\{(\mathbf{X}_i, Y_i) : T_i = j\}$ for $j = 0, 1$. As before, OOB estimation was used whenever possible. In particular, there are great efforts to ensure that bootstrap samples were held fixed throughtout in constructing synthetic features and the synthetic forest calculated from these features. This was done to ensure a coherent definition of being out-of-sample.

3.4 Bivariate imputation method

Another modification for g in equation (1.6) is

$$\mathbf{Y} = g(\mathbf{X}, \epsilon_{\mathbf{Y}}),$$

where T is deleted and used for index of bivariate outcome $\mathbf{Y} = [Y(0), Y(1)]$, and g is served as a imputation function for unobserved $Y(0)$ or $Y(1)$. This is a new bivariate approach making use of bivariate RF counterfactuals. For each individual i, we assume the existence of bivariate outcomes under the two treatment groups. One of these is the observed Y_i under the assigned treatment T_i, the other is the unobserved Y_i under the counterfactual treatment $1 - Ti$. This latter value is assumed to be missing. To determine these missing outcomes we impute the data by using unsupervised RF imputation (Tang and Ishwaran, 2017; Ishwaran and Kogalur, 2017). Data used includes the bivariate Yi outcome (one of these being missing for each i) in addition to the covariate Xi. In unsupervised RF imputation, tree splitting is implemented without assuming an outcome value. As in supervised RF, mtry variables are selected at random. However, for each of these, a random subset of ytry variables are selected and defined as the multivariate pseudo-responses. A multivariate composite splitting rule of dimension ytry is then applied to each of the mtry multivariate regression problems and the node split on the variable leading to the best

split (Ishwaran and Kogalur, 2017). Using the imputed data obtained from unsupervised imputation, we define a bivariate Y_i for each i by using the observed Y_i and the imputed potential outcome Y_i associated with the counterfactual treatment. This yields the bivariate counterfactual estimate

$$\hat{\tau}_{bivariate}(\mathbf{x}) = \hat{Y}_{bivariate,1}(\mathbf{x}) - \hat{Y}_{bivariate,0}(\mathbf{x}).$$

Note that OOB values are not utilized in this approach.

3.5 Honest RF

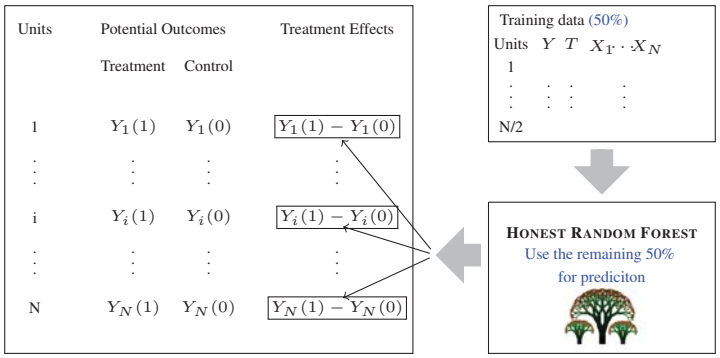

Figure 3.6: Illustration of honest RF approach.

Another substiution of Breiman RF is honest RF; however, instead of being g in equation (1.6), honest RF is more like a matching technique in potential outcome approach. Honest RF is not targeted in Y, but in treatment difference within a node: each node here is consider as a matched group, match based on raw observed variables. The honest causal procedure is described in Procedure 1 of Wager and Athey (2017). A RF is run by regressing Y_i on (X_i, T_i), but using only a randomly selected subset of 50% of the data. When fitting RF to this training data, a modified regression splitting rule.is used. Rather than

splitting tree nodes by maximizing the node variance, honest RF instead uses a splitting rule which maximizes the treatment difference within a node (see Procedure 1 and Remark 1 in Wager and Athey, 2017). Once the forest is grown, the terminal nodes of the training forest are repopulated by replacing the training Y with the Y values from the data that was held out. The purpose of this hold out data is to provide honest estimates and is akin to the role played by the OOB data used in our previous procedures. The difference between the hold out Y values under the two treatment groups is determined for each terminal node and averaged over the forest. This forest averaged value represents the honest forest ITE estimate. Figure 3.6 shows the framework of honest RF. We denote this estimate by $\hat{\tau}_{honestRF}(\mathbf{x})$.

3.6 Model consistency and convergence of RF

As discussed in Section 2.2, using RF as counterfactual approach requires model consistency. Although previous research already proved consistency of RF (Biau et al., 2008, 2016; Scornet et al., 2015; Ueda and Nakano, 1996; Wager and Guenther, 2015), discussions are addressed most from the aspect of trees, instead of forest. In this section, I will firstly discuss why ensemble works as increasing learner number and how sample size plays a role. Then I will re-discuss the consistency of RF as a kernel method. Another research question here I want to answer is that since counterfactual RF splits the data, is it doomed to lose efficiency? The answer is "No" and a simulation will be provided at the end of this section to show that splitting the data sometimes can enhance prediction performance.

3.6.1 Convergence properties in classifier ensemble

Consider data $(\mathbf{X}, Y)$ from a two class problem where $Y \in \{0, 1\}$ is the outcome and $\mathbf{X}$ denotes the p-dimensional feature. It is assumed that $(\mathbf{X}, Y)$ is sampled from some distri-

bution $\mathbb{P}_0$. Denote the learning data by $\mathscr{L} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ where $(\mathbf{X}_i, Y_i)$ are i.i.d. $\mathbb{P}_0$. Using the learning data $\mathscr{L}$, grow B base learners (for example trees) $T_1, \ldots, T_B$ with outputs $0 < T_b(\mathbf{X}) < 1$. Define $C_b(\mathbf{X}) = 1_{\{T_b(\mathbf{X})>1/2\}}$ to be the bth classifier based on the bth learner. The ensemble classifier is defined as

$$C_B^e(\mathbf{X}) = 1_{\left\{\sum_{b=1}^B C_b(\mathbf{X})>B/2\right\}}.$$

In other words, the ensemble classifies a case as 1 if the majority of the learners do, otherwise it classifies the case as 0.

The misclassification error for the ensemble can be made exponentially small with increasing B under certain assumptions. We begin by introducing two assumptions:

(A1) $T_1, T_2, \ldots, T_B$ are i.i.d. $\mathbb{P}$.

(A2) $\pi_1 = \mathbb{P}\{T_b(\mathbf{X}) > \frac{1}{2}|Y = 0\} < \frac{1}{2}$ and $\pi_2 = \mathbb{P}\{T_b(\mathbf{X}) \leq \frac{1}{2}|Y = 1\} < \frac{1}{2}$.

Assumption (A1) requires the base learners to be independently constructed using the same learning instructions. Assumption (A2) requires the conditional misclassification error for each class to be bounded above by 1/2. This is a very minimal assumption and basically requires only that each base learner is better than flipping a fair coin. However, as we will see shortly, this condition is not as trivial as it seems since the learner has to perform well in the whole sample space.

Remark. It is worth making a technical remark regarding the probability $\mathbb{P}$ used in the assumptions above. In (A2), this $\mathbb{P}$ involves the joint measure over $(\mathbf{X}, Y)$ (which has distribution $\mathbb{P}_0$) and the distribution for the learning data used to construct the learners. In (A1), $\mathbb{P}$ refers just to the learning data distribution. In general, when the integration involves quantities other than $\mathbb{P}_0$ we shall use the generic symbol $\mathbb{P}$.

We now prove the misclassification error converges to zero exponentially fast. Begin by noting that the misclassification error for C_B^e is

$$\mathbb{P}\{C_B^e(\mathbf{X}) \neq Y\} = \mathbb{P}\{C_B^e(\mathbf{X}) = 1, Y = 0\} + \mathbb{P}\{C_B^e(\mathbf{X}) = 0, Y = 1\}.$$

Consider the first term on the right-hand side. Define $Z_b(\mathbf{X}) = 1_{\{T_b(\mathbf{X})>1/2\}}$. Under assumption (A1),

$$Z_b(\mathbf{X})|(\mathbf{X} = \mathbf{x}, Y = 0) \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_1(\mathbf{x})), \text{ where } \pi_1(\mathbf{x}) = \mathbb{P}\Big\{T_b(\mathbf{x}) > \frac{1}{2}|Y = 0\Big\}.$$

It follows that

$$\begin{aligned}
&\mathbb{P}\{C_B^e(\mathbf{X}) = 1, Y = 0\} \\
&\quad = \mathbb{P}\left\{\sum_{b=1}^{B} C_b(\mathbf{X}) > B/2,\, Y = 0\right\} \\
&\quad = \mathbb{P}\left\{\sum_{b=1}^{B} 1_{\{T_b(\mathbf{X}) > 1/2\}} > B/2,\, Y = 0\right\} \\
&\quad = \mathbb{P}\left\{\sum_{b=1}^{B} Z_b(\mathbf{X}) > B/2 \,\middle|\, Y = 0\right\} \mathbb{P}_0\{Y = 0\} \\
&\quad = \mathbb{P}_{\mathbf{X}}\mathbb{P}\left\{\sum_{b=1}^{B} Z_b(\mathbf{X}) > B/2 \,\middle|\, \mathbf{X} = \mathbf{x}, Y = 0\right\} \mathbb{P}_0\{Y = 0\}.
\end{aligned}$$

The inner probability in the last line equals the tail probability of a sum of independent Bernoulli random variables with success probability $\pi_1(\mathbf{x})$. As there are B random variables, their average approaches $B\pi_1(\mathbf{x})$; thus the tail probability becomes exponentially small if $\pi_1(\mathbf{x}) < 1/2$. But this does not guarantee an exponential misclassification rate, because in order for the entire term to be exponentially small, $\pi_1(\mathbf{x})$ must be bounded above for each $\mathbf{x}$. Thus we need to strengthen Assumption (A2), as (A2) only requires the base learner to be on average better than flipping a coin. We therefore replace assumption (A2)

with the following stronger condition:

(A2*) For almost all $\mathbf{x}$ with respect to $\mathbb{P}_{\mathbf{X}}$ (the marginal distribution of $\mathbf{X}$ under $\mathbb{P}_0$):

$$\pi_1(\mathbf{x}) = \mathbb{P}\Big\{T_b(\mathbf{x}) > \frac{1}{2}|Y=0\Big\} < \frac{1}{2} \text{ and } \pi_2(\mathbf{x}) = \mathbb{P}\Big\{T_b(\mathbf{x}) \leq \frac{1}{2}|Y=1\Big\} < \frac{1}{2}.$$

To complete the proof we make use of the following well known result regarding the tail probability of a binomial random variable by Arratia and Gordon (1989).

Theorem 3.6.1. *Let Z be a binomial random variable with distribution Binomial(n,p). Then*

$$\mathbb{P}\{Z \geq k\} = F(k,n,p) \leq \exp\left(-nH(k/n,p)\right), \quad \textit{if } p < \frac{k}{n} < 1$$

where $H(a,p) = a\log(a/p) + (1-a)\log((1-a)/(1-p))$.

Apply Theorem 3.6.1 with $k = B/2, n = B$ and $p = \pi_1(\mathbf{x})$ to obtain

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{b=1}^{B} Z_b(\mathbf{X}) > B/2 \,\middle|\, \mathbf{X}=\mathbf{x}, Y=0\right\} \\
&= F(B/2, B, \pi_1(\mathbf{x})) \qquad \left(\text{by (A2}^*\text{) note } p = \pi_1(\mathbf{x}) < \frac{1}{2} = \frac{(B/2)}{B} = \frac{k}{n}\right) \\
&\leq \exp\left(-B\left[\frac{1}{2}\log\left(\frac{1}{2\pi_1(\mathbf{x})}\right) + \frac{1}{2}\log\left(\frac{1}{2(1-\pi_1(\mathbf{x}))}\right)\right]\right) \\
&\leq \exp\left(-\frac{B}{2}\log\left(\frac{1}{4\pi_1(\mathbf{x})(1-\pi_1(\mathbf{x}))}\right)\right) \\
&\leq \Big[4\pi_1(\mathbf{x})(1-\pi_1(\mathbf{x}))\Big]^{B/2}.
\end{aligned}$$

Consequently, it follows from our previous work that

$$\mathbb{P}\{C_B^e(\mathbf{X}) = 1, Y = 0\} \leq \mathbb{E}_{\mathbf{X}}\Big[4\pi_1(\mathbf{X})(1-\pi_1(\mathbf{X}))\Big]^{B/2}.$$

By invoking a similar argument using

$$(1 - Z_b(\mathbf{X}))|(\mathbf{X} = \mathbf{x}, Y = 1) \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_2(\mathbf{x})),$$

deduce under assumptions (A1) and (A2*) that

$$\mathbb{P}\{C_B^e(\mathbf{X}) \neq Y\} \leq \mathbb{E}_{\mathbf{X}}\Big[4\pi_1(\mathbf{X})(1 - \pi_1(\mathbf{X}))\Big]^{B/2} + \mathbb{E}_{\mathbf{X}}\Big[4\pi_2(\mathbf{X})(1 - \pi_2(\mathbf{X}))\Big]^{B/2}. \quad (3.2)$$

The exponential rate now follows under either smoothness conditions to $\pi_1(\mathbf{x})$ and $\pi_2(\mathbf{x})$ or by assuming a uniform bound. On the other hand, convergence to zero (without a rate) is already guaranteed without further assumptions. We state this formally in the following result.

Theorem 3.6.2. *Under assumptions (A1) and (A2*), $\mathbb{P}\{C_B^e(\mathbf{X}) \neq Y\} \to 0$ as $B \to \infty$.*

Proof. Let $0 < \gamma_B < 1$ and define sets $A_{j,B} = \{\mathbf{x} : \pi_j(\mathbf{x}) \leq (1 - \gamma_B)/2\}$ and $A_{j,B}^* = \{\mathbf{x} : (1 - \gamma_B)/2 \leq \pi_j(\mathbf{x}) < 1/2\}$ for $j = 1, 2$. Under assumption (A2*), $\pi_j(\mathbf{x}) < 1/2$ for almost all $\mathbf{x}$. Therefore we can decompose the integrals in (3.2) into integrals over $A_{j,B}$ and $A_{j,B}^*$. Setting $g_j(\mathbf{x}) = 4\pi_j(\mathbf{x})(1 - \pi_j(\mathbf{x}))$, we have

$$\mathbb{E}_{\mathbf{X}}\Big[4\pi_j(\mathbf{X})(1 - \pi_j(\mathbf{X}))\Big]^{B/2} = \int_{A_{j,B}} g_j(\mathbf{x})^{B/2}\, \mathbb{P}(d\mathbf{x}) + \int_{A_{j,B}^*} g_j(\mathbf{x})^{B/2}\, \mathbb{P}(d\mathbf{x}).$$

Let $g(\pi) = 4\pi(1 - \pi)$. Then $g(\pi) = 1 - (1 - 2\pi)^2$ which is a quadratic which attains its maximum over $[0, (1 - \gamma_B)/2]$ at $\pi = (1 - \gamma_B)/2$. Hence, for π over $[0, (1 - \gamma_B)/2]$

$$g(\pi)^{B/2} \leq (1 - \gamma_B^2)^{B/2} \asymp \exp(-B\gamma_B^2/2).$$

We want to choose γ_B such that the right hand side converges to zero while simultaneously

satisfying $\gamma_B \to 0$. In particular, setting

$$\gamma_B = \sqrt{\frac{2\log(B)}{B}}$$

we have $\gamma_B \to 0$ and

$$(1-\gamma_B^2)^{B/2} \asymp \exp(-\log B) = B^{-1} \to 0.$$

Hence, we have

$$\int_{A_{j,B}} g_j(\mathbf{x})^{B/2}\, \mathbb{P}(d\mathbf{x}) \leq (1-\gamma_B^2)^{B/2} \int_{A_{j,B}} \mathbb{P}(d\mathbf{x}) \leq (1-\gamma_B^2)^{B/2} \to 0.$$

Furthermore,

$$\int_{A^*_{j,B}} g_j(\mathbf{x})^{B/2}\, \mathbb{P}(d\mathbf{x}) \leq \int_{A^*_{j,B}} \mathbb{P}(d\mathbf{x})$$

which converges to zero by the Dominated Convergence Theorem because $A^*_{j,B} \to \varnothing$. □

It is clear assumption (A2*) cannot hold in general. For example, it cannot hold when the Bayes rule has nonzero misclassification error. Recall that the Bayes rule is the classifier C^o defined as

$$C^o(\mathbf{x}) = \begin{cases} 1 & \text{if } f(\mathbf{x}) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

where $f(\mathbf{x}) = \mathbb{P}_0\{Y = 1 | \mathbf{X} = \mathbf{x}\}$ is the true classification probability. The Bayes rule is the optimal rule under misclassification error. Let $C(\mathbf{X}) := C(\mathbf{X}, \mathscr{L})$ be an arbitrary classifier where $C : \mathscr{X} \to \{0, 1\}$. Optimality of the Bayes rule implies that

$$\mathbb{P}\{C^o(\mathbf{X}) \neq Y\} \leq \mathbb{P}\{C(\mathbf{X}) \neq Y\}.$$

This shows that if the Bayes misclassification error is nonzero, assumption (A2*) cannot hold, otherwise according to Theorem 3.6.2 we would have constructed a classifier with smaller misclassification error than the Bayes rule. In particular, because the Bayes rule has misclassification error

$$L^o = \mathbb{E}\Big[\min(f(\mathbf{X}), 1 - f(\mathbf{X}))\Big]$$

we must have

$$\mathbb{P}\{C_B^e(\mathbf{X}) \neq Y\} \geq \mathbb{E}\Big[\min(f(\mathbf{X}), 1 - f(\mathbf{X}))\Big].$$

We therefore replace assumption (A2*) with the following weaker condition: (A2$'$)

(A2$'$) For almost all $\mathbf{x}$ with respect to $\mathbb{P}_{\mathbf{X}}$ (the marginal distribution of $\mathbf{X}$ under $\mathbb{P}_0$):

$$\pi_1(\mathbf{x}) = \mathbb{P}\Big\{T_b(\mathbf{x}) > \frac{1}{2}|C^o(\mathbf{x}) = 0\Big\} < \frac{1}{2} \text{ and } \pi_2(\mathbf{x}) = \mathbb{P}\Big\{T_b(\mathbf{x}) \leq \frac{1}{2}|C^o(\mathbf{x}) = 1\Big\} < \frac{1}{2}.$$

Since formula (3.2) suggests learners have to have a uniform bound, which is, ideally, the Bayes risk.

Lemma 3.6.3. *Under assumptions (A1) and (A2$'$),* $\mathbb{P}\{C_B^e(\mathbf{X}) \neq C^o(\mathbf{X})\} \to 0$ *as* $B \to \infty$.

Proof. let $C(\mathbf{X})$ be $\mathbf{Y}$ in the proof of Theorem 3.6.2. □

(A2$'$) is not quite a strict assumption since previous studies showed the consistence of tree learner and we will further discuss the convergence rate of this consistence as increasing training sample size later. The following example shows the fast convergence of ensemble as learner number increase.

> Example: Consider a binary response variable of two classes, each of the class consists of a 2 dimensional Gaussian. The centers are equally spaced on a

circle around the origin with radius r. I used the default setting of the *"mlbench.2dnormals"* function in r package *"mlbench"* where $r = \sqrt{2}$. The right sub-figure in Figure 3.7 plotted the shape of the simulation data using 300 observations. 1000 fixed training data is simulated and fit into models of random forest based on learner number (RF model uses r package *"randomForestSRC"* with its default setting) 1 to 10, 20, 30 ,40, 50, 100, 200. 1000 separately. Through fixed 1000 test data. the averaged brier score and misclassification rate from 30 replications are plotted against the learner number in the left sub-figure. Bayes risk as misclassification rate of Bayesian classifier is also added as green dash line as a reference. In this simulation, both brier score (black solid line) and misclassification rate (red dash line) converges fast as learner number increases: after 50 leaners are added, both rates barely change. The misclassification rate will not reach the Bayes risk even leaner number goes to infinite because assumption (A2′) is not satisfied. Assumption (A2′) is about the sample size of training data, which is discussed in the next section. How RF gives different estimates to point A and point B in the right figure will be ploted in Figure 3.8.

3.6.2 Convergence of tree learners precision

This section will discuss $T_b(\mathbf{x})$ in a classification tree learner scenario. All content here is about a fixed given $\mathbf{x}$: performance all over $\mathbf{X}$ space is not discussed in this section. To prove the consistency of $T_b(\mathbf{x})$, I will first introduce the expectation of $T_b(\mathbf{x})$ as $P_B^e(\mathbf{x})$, prove $P_B^e(\mathbf{x}) \xrightarrow{p} P\{Y = 1|\mathbf{x} \in R(\mathbf{x})\}$ and then prove $R(\mathbf{x}) \xrightarrow{a.s.} \mathbf{x}$ in a pure random RF setting, where $R(\mathbf{x})$ is the expectation of terminal node for $\mathbf{x}$.

Recall that $T_1, \ldots, T_B$ are classification trees grown from $\mathscr{L} =$

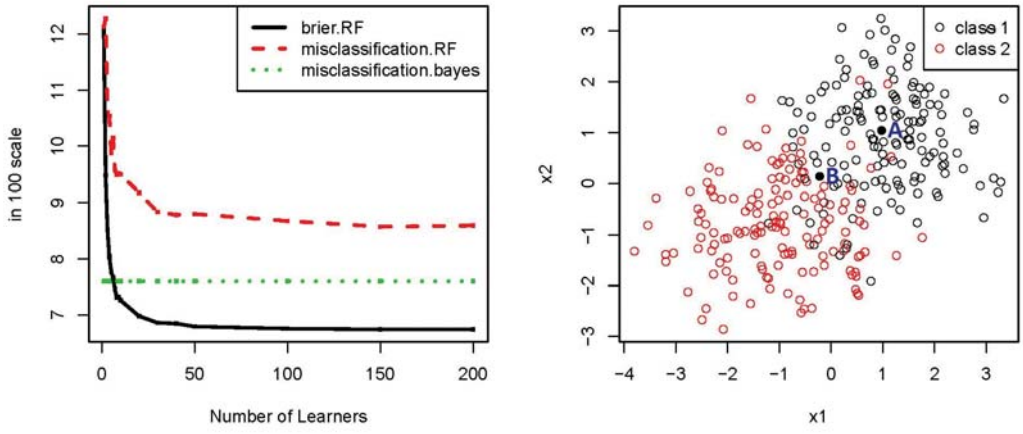

Figure 3.7: Convergence of ensemble through simulated 2-D Normal data

$\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ with outputs $0 < T_b(\mathbf{x}) < 1$ representing the predicted probability of $Y = 1$ for a given feature $\mathbf{x}$. Because of the recursive partitioning property of trees, each $\mathbf{x}$ must be a member of a unique terminal node of T_b which we shall denote by $R_b(\mathbf{x})$. The tree predicted probability of $Y = 1$ for $\mathbf{x}$ is

$$T_b(\mathbf{x}) = \frac{\sum_{i=1}^{n} 1_{\{\mathbf{X}_i \in R_b(\mathbf{X})\}} 1_{\{Y_i=1\}}}{\sum_{i=1}^{n} 1_{\{\mathbf{X}_i \in R_b(\mathbf{X})\}}}.$$

Define the expectation of $T_b(\mathbf{x})$ as $P_B^e(\mathbf{x}) = E(T_b(\mathbf{x}))$.

$$\begin{aligned} P_B^e(\mathbf{x}) =& \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^{n} 1_{\{\mathbf{X}_i \in R_b(\mathbf{X})\}} 1_{\{Y_i=1\}}}{N_b(\mathbf{x})} \\ =& \frac{1}{B} \sum_{b=1}^{B} \left(\frac{\sum_{\mathbf{X}_i \in R_b(\mathbf{x})} Y_i}{N_b(\mathbf{x})} \right), \end{aligned} \tag{3.3}$$

where $N_b(\mathbf{x}) = \sum_{i=1}^{n} 1_{\{\mathbf{X}_i \in R_b(\mathbf{X})\}}$, which is the node size for $R_b(\mathbf{x})$.

Here, the terminal node $R_b(\mathbf{x})$ is a hyperrectangle and also a random variable with p dimensions. Denote $R_b(\mathbf{x})$ follows a p dimension distribution with mean $R(\mathbf{x})$: $R(\mathbf{x}) = E_{b \in B}(R_b(\mathbf{x}))$ with Lebesgue measure denoted as $\mu(R(\mathbf{x}))$. In $R(\mathbf{x})$, suppose there are

q samples, which are denoted as $\{(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_q, \tilde{y}_q)\}$, $E(q) = \lfloor \frac{n}{k+1} \rfloor$ given sample size n and total splits number k, since on average there would be $\lfloor \frac{n}{k+1} \rfloor$ sample in one terminal node but variance of q can be very large. Formula (3.3) demonstrates that $P_B^e(\mathbf{x})$ is the mean of $\tilde{y}_1, \tilde{y}_2 ... \tilde{y}_q$: $P_B^e(\mathbf{x}) = \frac{1}{q} \sum_{i=1}^{q} \tilde{y}_i$. Note that these $\tilde{\mathbf{x}}_q$ and $\tilde{y}_q$ are not necessarily unique at all.

Theorem 3.6.4. *$P_B^e(\mathbf{x}) \xrightarrow{p} P_{R(\mathbf{x})}$ where $P_{R(\mathbf{x})} = P\{Y = 1 | \mathbf{x} \in R(\mathbf{x})\}$ as $k \to \infty$ and $\lfloor \frac{n}{k+1} \rfloor \to \infty$.*

Proof. When $\mu(R(\mathbf{x}))$ is very small as k is large, it is reasonable to assume $\tilde{y}_1, \tilde{y}_2 ... \tilde{y}_q$ are independent and identically distributed Bernoulli random variables

$$\tilde{y}_1, \tilde{y}_2 ... \tilde{y}_q \overset{\text{iid}}{\sim} \text{Bernoulli}(P_{R(\mathbf{x})}),$$

where $P_{R(\mathbf{x})} = P\{Y = 1 | \mathbf{x} \in R(\mathbf{x})\}$.

Let $\mu_{\tilde{y}} = E(\tilde{y})$, where $\tilde{y} = \sum_{i=1}^{q} \tilde{y}_i$, also $\mu_{\tilde{y}} = q P_{R(\mathbf{x})}$. Using Multiplicative Chernoff Bound by Chernoff (1952):

$$P\big(\tilde{y} > (1+\delta)\mu_{\tilde{y}}\big) < \Big[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\Big]^{\mu_{\tilde{y}}}$$

where δ denotes the relative error.

$$P\big(\frac{\tilde{y} - \mu_{\tilde{y}}}{q} > \frac{\delta}{q}\big) = P\big(P_B^e(\mathbf{x}) - P_{R(\mathbf{x})} > \epsilon\big) < \Big[\frac{e^{q\epsilon}}{(1+q\epsilon)^{1+q\epsilon}}\Big]^{q P_{R(\mathbf{x})}}$$

where $\epsilon = \frac{\delta}{q}$.

Further, according to Chernoff Bound by Chernoff (1952), denote S as the probability of an event, simultaneous occurrence on $\tilde{y}_1, \tilde{y}_2 ... \tilde{y}_q$ that more than $\frac{q}{2}$ of the events $\{\tilde{y}_i = 1\}$:

when $P_{R(\mathbf{x})} > \frac{1}{2}$, $S \gtrsim 1 - exp(-\frac{(P_{R(\mathbf{x})}-\frac{1}{2})^2}{2P_{R(\mathbf{x})}}q)$, in other words,

$$P\big[P_B^e(\mathbf{x}) > \frac{1}{2} | P_{R(\mathbf{x})} > \frac{1}{2}\big] = 1 - exp\big(-\frac{(P_{R(\mathbf{x})}-\frac{1}{2})^2}{2P_{R(\mathbf{x})}}q\big).$$

Here, we can see that $P\big(P_B^e(\mathbf{x}) - P_{R(\mathbf{x})}\big) > \epsilon$ and $P\big[P_B^e(\mathbf{x}) > \frac{1}{2} | P_{R(\mathbf{x})} > \frac{1}{2}\big]$ converges in different rate: the former is more related to the convergence rate of single learner and the later is more related to the convergence rate of ensemble. □

Consider a random tree classifier for simplicity as assumed by Breiman (2000): $\mathbf{X} \sim$ Uniform$[0,1]^d$. At each step of constructing the tree, a leaf is chosen uniformly at random; a split variable is then selected uniformly at random from the d candidates. Finally, a split k is chosen along the randomly chosen variable at a uniformly random location to create new hyperrectangle as a new node.

Theorem 3.6.5. $R(\mathbf{x}) \xrightarrow{a.s.} \mathbf{x}$ *as* $k \to \infty$ *in a random tree scenario.*

Proof. Suppose there is $\zeta \in R(\mathbf{x})$ and $\zeta \neq \mathbf{x}$, denote $d_m = [x_m, z_m]$ as both the interval and distance of $\mathbf{x}$ and ζ in the mth dimension. In a random tree classifier scenario, the $\lfloor k/d \rfloor$ splits, denoted as $k_m = \lfloor k/d \rfloor$ in each dimension m, are mapping to k_m split points, denoted as $S_{m1}, S_{mi}, ..., S_{mk_m}$which are all i.i.d Uniform[0,1] random variables. ζ and $\mathbf{x}$ are in the same terminal node $R(\mathbf{x})$ means in every dimension, there is no split point S_{mi} between $[x_m, z_m]$, with the probability $(1 - d_m)^{k_m}$. Therefore,

$$P(\zeta \in R(\mathbf{x})) = \prod_{m=1}^{p} \big[(1 - d_m)^{k_m}\big].$$

As $k \to \infty$, all $k_m \to \infty$:

$$P(\zeta \in R(\mathbf{x})) = \begin{cases} 0 & \text{if } d_m \neq 0 \\ 1 & \text{if } d_m = 0 \end{cases} \tag{3.4}$$

□

Lemma 3.6.6. $R(\mathbf{x}) \xrightarrow{a.s.} \mathbf{x}$ *as* $k_m \to \infty$ *for all* $m = 1, \ldots, d$ *dimensions in a general tree scenario.*

Proof. For non-random tree, as long as $k \to \infty$ and $k_m \to \infty$, $P\big(S_{m,i} \in [x_m, z_m]\big) = 1$ when $d_m \neq 0$ for $i \in \{1, .., k_m\}$, since there are infinite distinct split points in this mth dimension; Further, $P\big(\zeta \in R(\mathbf{x})\big) = \prod_{m=1}^{d} P\big(S_{mi} \notin [x_m, z_m]\big)$ making Equation (3.4) still holds. □

From Theorem 3.6.4 and Theorem 3.6.5, $P_B^e(\mathbf{x}) \xrightarrow{p} C^o(\mathbf{x})$; therefore (A2′) holds in a random classifier tree learner setting :

$$\pi_1(\mathbf{x}) = \mathbb{P}\Big\{T_b(\mathbf{x}) > \frac{1}{2} | C^o(\mathbf{x}) = 0\Big\} < \frac{1}{2} \text{ and } \pi_2(\mathbf{x}) = \mathbb{P}\Big\{T_b(\mathbf{x}) \leq \frac{1}{2} | C^o(\mathbf{x}) = 1\Big\} < \frac{1}{2},$$

as $k \to \infty$ and $q = \lfloor \frac{n}{k+1} \rfloor \to \infty$.

The aim of proving Theorem 3.6.5 is to give a insight of convergence rate of terminal nodes shrindage in probability. In order to observe the convergence behavior of $P_B^e(\mathbf{X})$, we call "$\frac{1}{B}\sum_{b=1}^{B}\left(\frac{\sum_{\mathbf{x}_i \in R_b(\mathbf{x})}}{N_b(\mathbf{x})}\right)$" as Kernal to get insight. Tree ensemble method is actually a weighting method on the training sample responses. Although Theorem 3.6.5 is based on random tree classifier scenario, it indicates information for regular tree with more complicated splitting rule and finite sample: $k_m = \lfloor k/d \rfloor$ for all dimensions in random tree secario; whereas usually learners tends to split more often on variables which are more informative to predict the outcome. If a variable, say $\tilde{m}$, is more important, then $k_{\tilde{m}} > k_m$, where $m = 1, \ldots, p$ and $\tilde{m} \neq m$, meaning the Kernal is sharper to the more important variable. Moreover, the split point is not usually uniformly randomly chosen: split points are more depended on the outcome value, meaning the Kernal is more flat to the Bayesian risk bound. Figure 3.8 (a) and (b) shows the kernal shape of two data points in the sim-

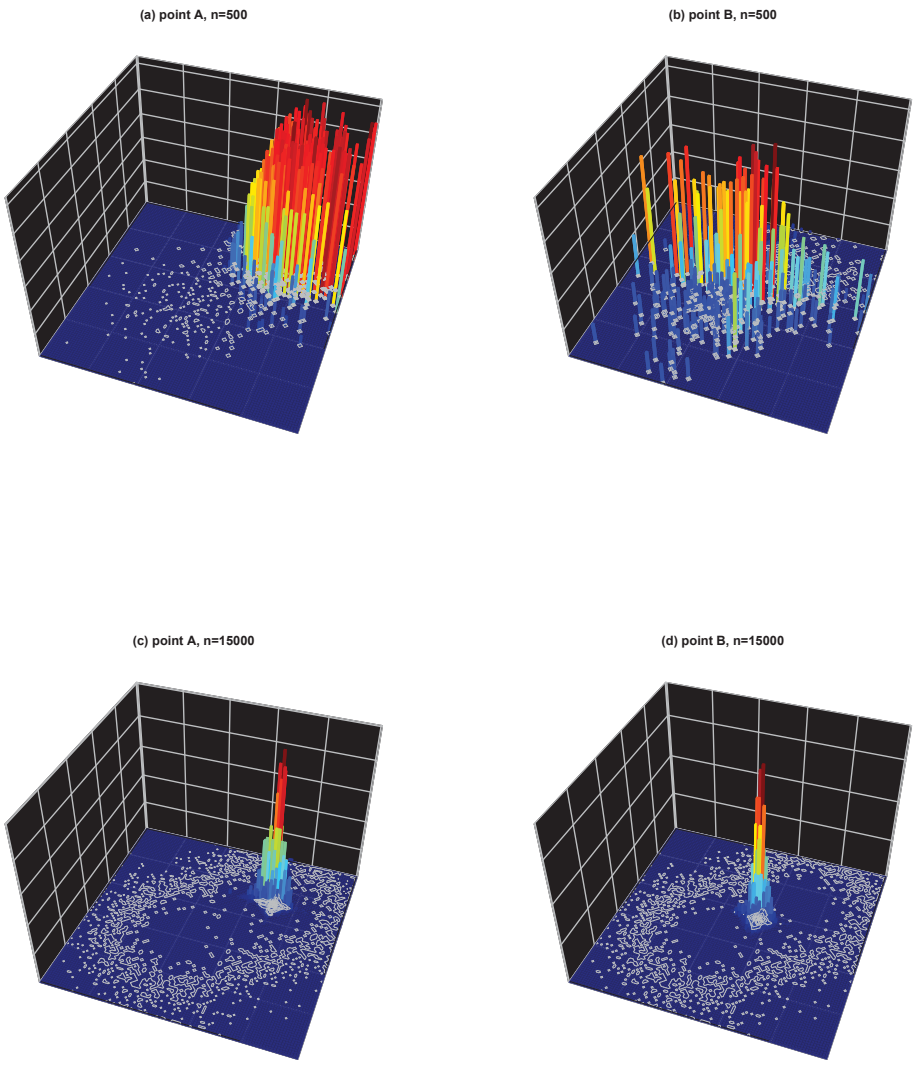

Figure 3.8: RF Kernel shape for two data points in simulation of Figure 3.7

Note. n=training data sample size; point A and point B is marked in the right sub-figure in Figure 3.7

ulation of Figure 3.7 with sample size $n = 500$, tree number $B = 1000$: for point **A** in the left sub-figure, the Kernal is more sharp than point **b** in (b) because the neighborhood data of the focal point equally predict the outcome, so the weights in Kernal only depend on the distance in $\mathbf{X}$; in contrast, for point **B** in the right sub-figure (b), the Kernal is more

flat along the Bayse boundary because the neighborhood data along the Bayse boundary is more predictive for the outcome Y, so the weights spread along the Bayse boundary. Figure 3.8 (c) and (d) shows the Kernal shape of the same two data points **A** and **B** with sample size $n = 15000$, tree number $B = 1000$. When sample size is large, k tends large; then $P_B^e(\mathbf{x}) \xrightarrow{p} C^o(\mathbf{x})$, and the Kernal is very sharp. For synthetic RF, the kernal will be more sharp because the synthetic variable is already a weighted average on the neighborhood data of the focal $\mathbf{x}$; therefore the synthetic RF has the "kernal of the kernal", which involves more broaden area in $\mathbf{X}$ but sharper on the tip area close to focal $\mathbf{x}$.

3.6.3 Convergence of tree learners correlation

Assumption (A1) in section 3.6.1 assumes independence of all learners, which is generally not true. Breiman (2000) firstly discussed the correlation of learners. In an infinite sample space, if all the learners are consistent, a "perfect" prediction is expected from them: every learner gives the same answer, so the correlation of learners would be 1.

Definition 3.6.7. Suppose $T_i(\mathbf{x})$ and $T_j(\mathbf{x})$ is the ith and jth learner 's forecast probability of $P(Y = 1|\mathbf{x})$, Denote

$$\rho_b(\mathbf{x}) = E_{i,j\in B}\left(\frac{cov(T_i(\mathbf{x}), T_j(\mathbf{x}))}{\sigma_{T_i(\mathbf{x})}\sigma_{T_j(\mathbf{x})}}\right)$$

as the learners' correlation, where

$$cov\big(T_i(\mathbf{x}), T_j(\mathbf{x})\big) = E[\big(T_i(\mathbf{x}) - P_B^e(\mathbf{x})\big)\big(T_j(\mathbf{x}) - P_B^e(\mathbf{x})\big)]$$

and

$$\sigma_{T_i(\mathbf{x})} = \sqrt{E\big[\big(T_i(\mathbf{x}) - P_B^e(\mathbf{x})\big)^2\big]},$$

similar for j.

As $k \to \infty$ and $q = \lfloor \frac{n}{k+1} \rfloor \to \infty$ in random tree RF, $\rho_b(\mathbf{x}) \to 1$, $\sigma_{T_i(\mathbf{x})} \to 0$ and $cov\big(T_i(\mathbf{x}), T_j(\mathbf{x})\big) \to 0$. From Theorem 3.6.4 and Theorem 3.6.6, $T_i(\mathbf{x})$ and $T_j(\mathbf{x}) \xrightarrow{p} P_B^e(\mathbf{x}) \xrightarrow{p} C^o(\mathbf{x})$ as $k \to \infty$ and $q = \lfloor \frac{n}{k+1} \rfloor \to \infty$.

Even $\rho_b(\mathbf{x}) \to 1$ makes assumption (A1) invalid, since $\sigma_{T_i(\mathbf{x})} \to 0$, all the learners already converge to Bayes classifier and generate the same predicted value in every $\mathbf{x}$, which is rarely true in real situation in a finite sample and non-pure random learner setting. However, $\rho_b(\mathbf{x})$ raises the issue that ensemble misclassification rate converges under correlated learner scenario.

How to lower the correlation of learners is to divide the sample. Assume a very large sample size $n = NB$. We partition $\mathscr{L}$ into blocks of size N denoted by $\mathscr{L}_b$, $b = 1, \ldots, B$ (i.e. $\mathscr{L} = \dot{\bigcup}_{b=1}^{B} \mathscr{L}_b$ and $|\mathscr{L}_b| = N$). Base learner T_b is constructed using $\mathscr{L}_b$. Its output is $T_b(\mathbf{X}) := T_b(\mathbf{X}, \mathscr{L}_b)$. By construction, $T_b(\mathbf{X}, \mathscr{L}_b) \overset{\text{iid}}{\sim} \mathbb{P}$, where $\mathbb{P}$ denotes the joint distribution for N learning data points: for example, the joint distribution for $(\mathbf{X}_1, Y_N), \ldots, (\mathbf{X}_n, Y_N)$. This satisfies assumption (A1). However, for assumption (A2*), we must ensure that the base learners are uniformly better than random guessing. Dividing the sample may make base learner less accurate. Therefore, doubts remain over whether counterfactual RF is a better technique as a data dividing model. Performance of counterfactual RF is compared with other RF method in the next chapter.

Example: Simulation in Figure 3.7 is used to show misclassification rate and Brier score convergences as sample size increasing. The left sub-figure in Figure 3.9 plotted how the misclassification rate converges as training sample size increases with a fixed learner number 50. Training data is of sample size 100, 400, 1000, 2000,3000,5000, 10000 as well as 20000. Through fixed 1000 test data, the averaged misclassification rate (red dash line) from 30 replications are plotted against the training sample size in the left sub-figure. Misclassification rate of Bayesian classifier is also added as a reference in color blue. Brier

score is displayed in right sub-figure. Results of single learner and random forest are compared in green and red dash line acccordingly. The solid black line is a model of ensemble random forests: I divided the training sample into subgroups of sample size=200; for each subgroup I grow a random forest (using default setting in r package *"randomForestSRC"* and ntree=50); the final result comes from the majority vote of these subgroup random forests result, called iid random forest result. From this simulation, both brier score and misclassification rate converges fast as training sample size increases. Random forest, as an ensemble method shows better performance than the single tree, as base learner, result. However, Random forest did not converge to the Bayes risk as sample size increasing partly due to the reason that assumption (A1) did not holds: learners are based on the same training sample so learners are correlated. One way to make the learners more independent is to divide the sample to grow separate learne as iid RF does. The result of iid random forest, as a model satisfied both assumption (A1) and (A2′), converges in to Bayes risk.

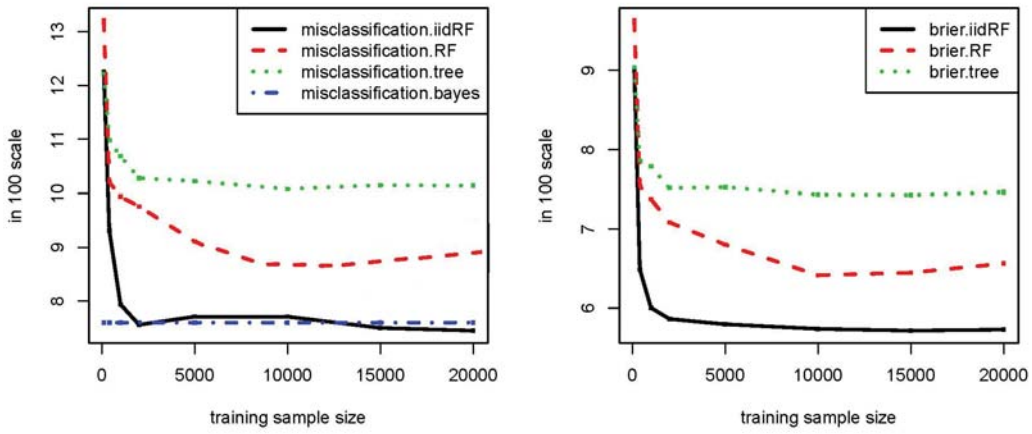

Figure 3.9: Performance of iid RF which splits the data into i.i.d groups and grows separate RFs

Chapter 4

Model application for continuous outcome

In this dissertation RF is used to model the response for direct estimation of the ITE. Our use of RF is motivated by its well known ability to provide accurate local prediction.

4.1 Simulation

Two sets of simulation models were used to assess performance of the different RF methods. All simulations assumed a continuous Y outcome and a binary exposure variable $T \in \{0, 1\}$ The first set of simulations were modified from those of Section 4.2 of Ghosh et al. (2015). We simulated 11 independent covariates $X_1, \ldots, X_{11}$ from a standard normal distribution, and nine independent covariates $X_{12}, \ldots, X_{20}$ from a Bernoulli(0.5). Three different models were used for the outcome Y, while a common simulation model was used for the exposure variable T. We refer to the three simulation models as G_1, G_2, G_3 (for Ghosh 1, 2, 3).

In G_1, G_2, G_3 the outcome was assumed to be $Y_i = f_{G_j}(\mathbf{X}_i, T_i) + \varepsilon_i$, where ε_i were

independent $N(0, \sigma^2)$ and

$$\begin{aligned} f_{G_1}(\mathbf{X}, T) =& 2.455 + .4T + .1X_1 - .154X_2 - .152X_{11} - .126X_{12} \\ f_{G_2}(\mathbf{X}, T) =& 2.455 + .4T + .1X_1 - .154X_2 - .152X_{11} - .126X_{12} - .3T_iX_{11} \\ f_{G_3}(\mathbf{X}, T) =& 2.455 + .4T + .1X_1 - .154(1-T)X_2 - .254TX_2^2 - .152X_{11} \\ & - .126X_{12} - .3T_iX_{11}. \end{aligned}$$

A logistic regression model was used to simulate T in which the linear predictor F (X) defined on the logit scale was

$$F(\mathbf{X}) = -2 + .028X_1 - .374X_2 - .03X_3 + .118X_4 - 0.394X_{11} + 0.875X_{12} + 0.9X_{13}.$$

Therefore in all three models, X_1, X_2, X_{11}, X_{12} were confounding variables, meaning that they were related to both the exposure and the outcome variable. In model 2, additionally X_{11} had a treatment interaction in the outcome model, while in model 3, both X_2 and X_{11} had a treatment interaction. Thus models G_2 and G_3 introduce a confounding heterogeneous treatment effect (CHTE). The top panel of Figure 4.1 depicts the relationship between the different variables.

The second set of simulations were modified from Setoguchi et al. (2008). In total 10 variables $\mathbf{W} = (W1, \ldots, W_{10})$ were simulated from a multivariate normal distribution with correlations between variables. Three exposure models for T were used, exposure models A, E and G. We refer to the resulting models as S_A, S_E, S_G (for Setoguchi A, E, G). In each of these, a logistic regression model was used where the linear predictor on the

logit scale was:

$$
\begin{aligned}
F_{S_A}(\mathbf{W}) =& .8W_1 - .25W_2 + .6W_3 - .4W_4 - .8W_5 - .5W_6 + .7W_7 \\
F_{S_E}(\mathbf{W}) =& .8W_1 - .25W_2 + .6W_3 - .4W_4 - .8W_5 - .5W_6 + .7W_7 - .25W_2^2 \\
& .4W_1W_3 - .175W_2W_4 - .2W_4W_5 - .4W_5W_6 \\
F_{S_G}(\mathbf{W}) =& F_{S_E}(\mathbf{W}).
\end{aligned}
$$

Three distinct models for Y were used based on a continuous normal measurement error model. The mean function $E(Y|\mathbf{W}, T)$ was respectively:

$$
\begin{aligned}
f_{S_A}(\mathbf{W}, T) =& 3.85 - .3W_1 + .4W_2 + .7W_3 + .2W_4 - .7W_8 + .2W_9 - .3W_{10} + .4T_i \\
f_{S_E}(\mathbf{W}, T) =& 3.85 - .3W_1 - .4W_2 + .7W_3 + .2W_4 - .7W_8 + .2W_9 - .3W_{10} + .4T_i \\
& + .2W_2^2 + .5W_1W_3 + .2W_1W_4 - .7TW_8 - .7TW_{10} \\
f_{S_G}(\mathbf{W}, T) =& 3.85 - .3W_1 - .4W_2 + .7W_3 + .2W_4 - .7W_8 + .2W_9 - .3W_{10} + .4T_i \\
& + .2W_2^2 + .5W_1W_3 + .2W_1W_4 - .7TW_1W_2.
\end{aligned}
$$

Note that simulations S_E and S_G use the same exposure model and employ nearly identical outcome models except that S_G has a treatment interaction with confounding variables W_1, W_2, and S_E has a treatment interaction with non-confounding variables W_8 and W_{10}. Thus S_G introduces CHTE while S_E only introduces a heterogeneous treatment effect (HTE). Simulation S_A has a fixed treatment effect and no heterogeneity is present. The bottom panel of Figure 3.1 displays the relationship between the different variables in S_A, S_E, S_G. Table 4.1 summarizes the exposure and outcome models for the Ghosh and Setoguchi models.

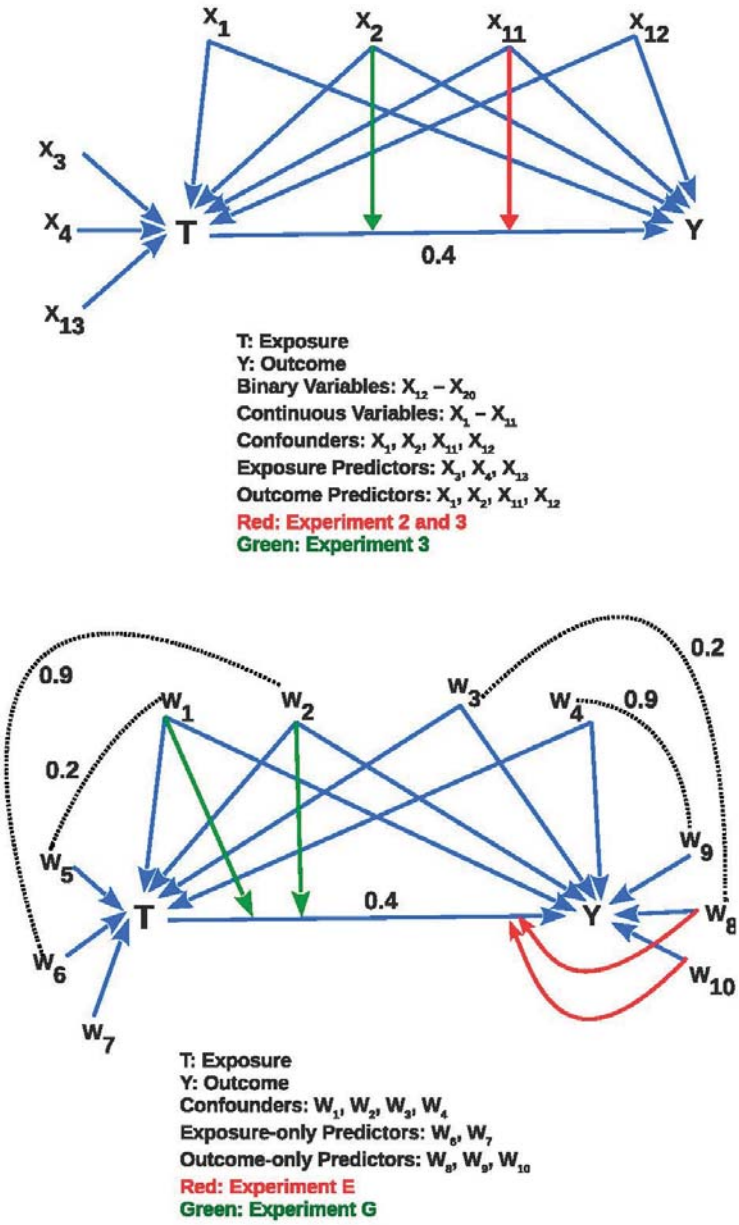

Figure 4.1: Top figure: simulation models from Ghosh et al. (2015). Bottom figure: simulation models from Setoguchi et al. (2008). Dashed lines indicate correlations between W variables.

Table 4.1: (a)Summary of exposure models used in Ghosh and Setoguchi simulations.

	Additive (main effects only)	Moderate non-additivity (two-way interaction terms involving confounders)
Linear (main effects only)	G_1, G_2, G_3, S_A	
Mild non-linearity (quadratic term)		S_E, S_G

Table 3.1: (b)Summary of outcome models used in Ghosh and Setoguchi simulations.

	Additive (main effects only)	Moderate non-additivity (two-way interaction terms involving confounders)
Linear (main effects only)	G_1, G_2, S_A	
Mild non-linearity (quadratic term)	G_3	S_E, S_G

4.1.1 Experimental settings and parameters

Simulations were run under two settings for the sample size, $n = 500$ and $n = 5000$. All sim- ulations used $\sigma^2 = 0.1$ for the variance of the normal error distribution used in the Y outcome models. The smaller sample size experiments $n = 500$ were repeated independently $B = 1000$ times, the larger $n = 5000$ experiments were repeated $B = 250$ times. All forests were based on 1000 trees with a nodesize of 3. One exception was for bivariate forests, where a nodesize of 1 was used for imputation using unsupervised forests (following the strategy recommended by Tang and Ishwaran, 2017). Another exception was

synthetic RF, where the RF base learners were constructed using all possible combinations of nodesize values 110, 20, 30, 50, 100 and mtry values of 1, 10 and 20 (for a total of 42 forest base learners). All RF computations were im- plemented using the *randomForestSRC* R-package (Ishwaran and Kogalur, 2017) (hereafter abbreviated as RF-SRC). The RF-SRC package implements all forms of RF data imputation, fits synthetic forests, multivariate forests, and utilizes openMP parallel processing, which allows for parallel processing on user desktops as well as large scale computing clusters; thus greatly reducing computational times.

4.1.2 Performance measures

Performance was assessed by bias and root mean squared error (RMSE). When calculating these measures we conditioned on the propensity score, $e(\mathbf{x})$. This was done to assess how well a procedure could recover treatment heterogeneity effects and to provide insight into its sensitivity to treatment assignment. A robust procedure should perform well not only in regions of the data where $e(\mathbf{x}) = 0.5$, and treatment assignment is balanced, but also in those regions where treatment assignment is unbalanced, $0 < e(\mathbf{x}) < .5$ and $1 > e(\mathbf{x}) > .5$. Assume the data is stratified into groups $G = \{\mathscr{G}_1, ..., \mathscr{G}_M\}$ based on quantiles $q_1, \ldots, q_M$ of $e(\mathbf{x})$. Given an estimator $\hat{\tau}$ of τ, the bias for group $\mathscr{G}_m$ was defined as

$$B(m) = \mathbb{E}\big[\hat{\tau}(\mathbf{X})|\mathbf{X} \in \mathscr{G}_m\big] - \mathbb{E}\big[\tau(\mathbf{X})|\mathbf{X} \in \mathscr{G}_m\big], m = 1, \ldots, M$$

Recall that our simulation experiments were replicated independently B times. Let $\mathscr{G}_{m,b}$ denote those $\mathbf{x}$ values that lie within the q_m quantile of the propensity score from realization b. Let $\hat{\tau}_b$ be the ITE estimator from realization b. The conditional bias was estimated by

$$\hat{B}(m) = \frac{1}{B}\sum_{b=1}^{B} \hat{\tau}_{m,b} - \frac{1}{B}\sum_{b=1}^{B} \tau_{m,b}$$

, where

$$\hat{\tau}_{m,b} = \frac{1}{\#\mathscr{G}_{m,b}} \sum_{\mathbf{x}_i \in \mathscr{G}_{m,b}} \hat{\tau}_b(\mathbf{x}_i), \tau_{m,b} = \frac{1}{\#\mathscr{G}_{m,b}} \sum_{\mathbf{x}_i \in \mathscr{G}_{m,b}} \tau_b(\mathbf{x}_i)$$

.Similarly, we define the conditional RMSE of $\hat{\tau}$ by

$$RMSE(m) = \sqrt{\mathbb{E}\big[\big(\hat{\tau}(\mathbf{X}) - \tau(\mathbf{X})\big)^2 | \mathbf{X} \in \mathscr{G}_m\big]}, m = 1, \ldots, M,$$

which we estimated using

$$\widehat{RMSE}(m) = \sqrt{\frac{1}{B}\sum_{b=1}^{B} \frac{1}{\#\mathscr{G}_{m,b}} \sum_{\mathbf{x}_i \in \mathscr{G}_{m,b}} \big[\hat{\tau}_b(\mathbf{x}_i) - \tau_b(\mathbf{x}_i)\big]^2}.$$

4.1.3 Results

Figure 4.2 displays the conditional bias and RMSE for each method for each of the six different simulation experiments. The left and right panels display small and larger sample sizes, $n = 500$ and $n = 5000$; the top and bottom panels display bias and RMSE, respectively. Each boxplot displays M values for the performance measure evaluated at each of the M stratified propensity score groups. We used a value of $M = 100$ throughout. One immediate observation from Figure 4.2 is that simulations S_A, S_E, S_G appear to be more difficult than G_1, G_2, G_3. This is likely due to the more complex nature of the outcome and exposure models used in these simulations (see Table 4.1). Also unlike the Ghosh simulations, the Setoguchi simulations used correlated features (see bottom panel of Figure 4.2). We summarize the performance of methods by separating results in terms of bias and RMSE below.

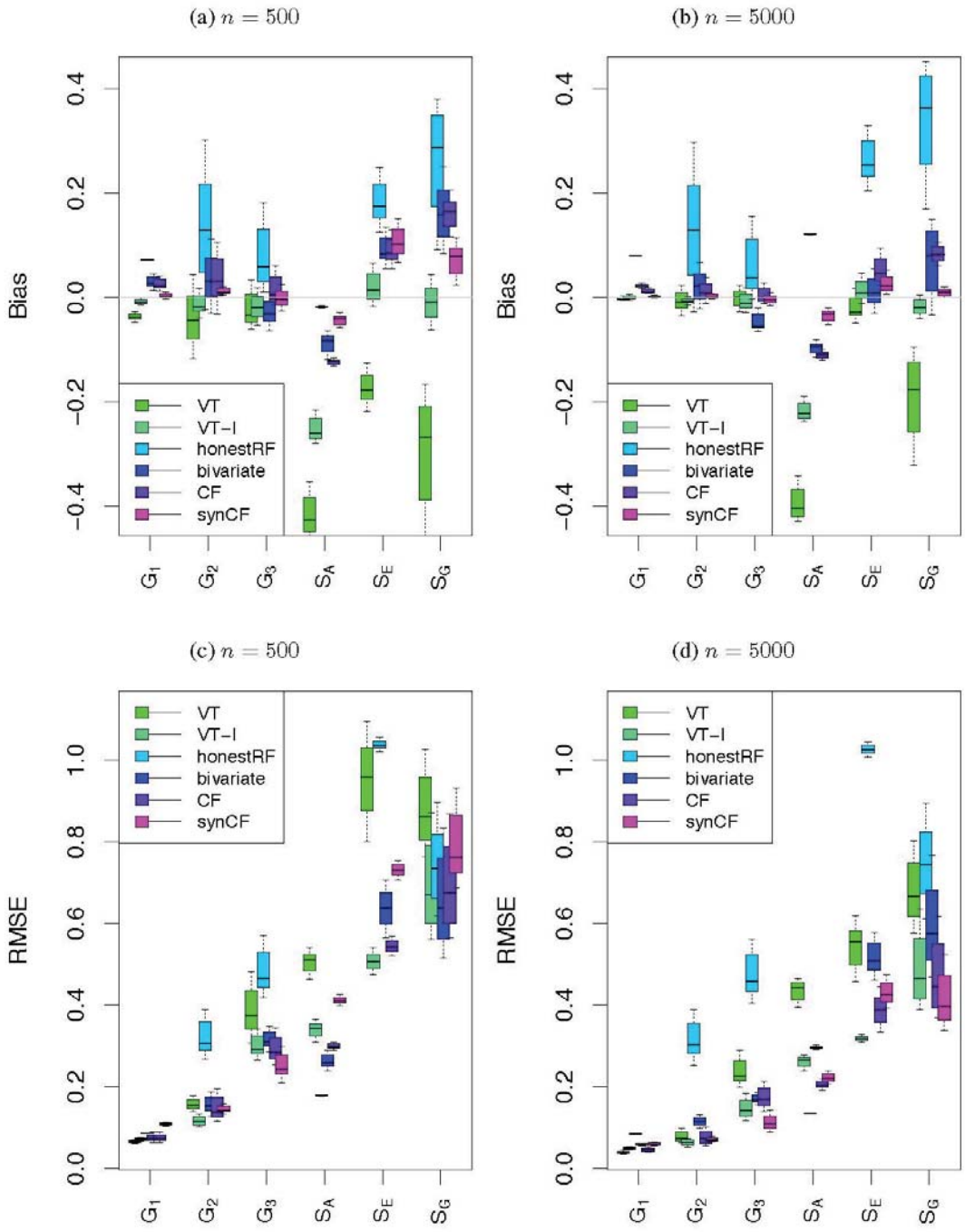

Figure 4.2: Conditional bias (top) and RMSE (bottom) from 6 simulation experiments for different sample sizes n. Boxplots display bias and RMSE values for each of the 100 percentiles of the propensity score.

Bias performance

Virtual twins interaction, $\hat{\tau}_{VT-I}$, and synthetic counterfactual forests, $\hat{\tau}_{synCF}$, are generally the best in terms of bias, with a slight edge going to $\hat{\tau}_{VT-I}$ in the smaller $n = 500$ sample setting (left top panel) for S and S_E and S_G. However, as n increases (right top panel), bias for $\hat{\tau}_{synCF}$ decreases significantly, especially for S_E and S_G, suggesting its performance is best in larger sample settings.Also, the variability of bias across the M propensity groups is smaller than $\hat{\tau}_{VT-I}$ (i.e. the boxplots for $\hat{\tau}_{synCF}$ are narrower than $\hat{\tau}_{VT-I}$). In general, $\hat{\tau}_{synCF}$ is among the methods with lowest bias in the larger sample simulations. It is also interesting to observe how bias of $\hat{\tau}_{VT-I}$ is improved over $\hat{\tau}_{VT}$ in the Setoguchi simulations. Augmenting the design matrix to include all pairwise treatment interactions improves adaptivity of VT forests in these more complex simulations.

RMSE performance

The results for RMSE performance generally mirror those for bias. As n increases (compare left bottom panel to right bottom panel), $\hat{\tau}_{synCF}$ improves dramatically and is generally among the best in terms of RMSE. For small n, both $\hat{\tau}_{VT}$ and $\hat{\tau}_{VT-I}$ outperform $\hat{\tau}_{synCF}$ in certain settings (such as S_E and S_G), but as n increases, $\hat{\tau}_{synCF}$ improves dramatically. Also, variability of RMSE for $\hat{\tau}_{synCF}$ over the M propensity scores is generally smaller than $\hat{\tau}_{VT-I}$ for large n. It is interesting to note that honest RF, $\hat{\tau}_{honestRF}$, does very well in simulations G_1 and S_A, but does poorly in all other simulations. One explanation for this is that G_1 and S_A are the only simulations with fixed treatment effects. Therefore, $\hat{\tau}_{honestRF}$ may not be suitable for complex HTE and CHTE settings. Finally, we note that the bivariate imputation method, $\hat{\tau}_{bivariate}$, generally performed well in all experiments, for example generally beating the standard implementation of VT, $\hat{\tau}_{VT}$, but was never the top performer in any setting.

Discussion

There is evidence that those methods with greatest adaptivity to potential confounding, when combined with out-of-sample estimation, do best. One particularly promising approach is a counterfactual approach in which separate forests are constructed using data from each treatment assignment. To estimate the ITE, each individuals predicted outcome is obtained from their treatment assigned forest. Next, the individuals treatment is replaced with the counterfactual treatment and used to obtain the counterfactual predicted outcome from the counterfactual forest; the two values are differenced to obtain the estimated ITE. This is an extension of the virtual twin approach, modified to allow for greater adaptation to potentially complex treatment responses across individuals. Furthermore, when combined with synthetic forests (Ishwaran and Malley, 2014), performance of the method is further enhanced due to reduced bias.

4.2 Project Aware: a counterfactual approach to understanding the role of drug use in sexual risk

Project Aware was a randomized clinical trial performed in nine sexually transmitted disease clinics in the United States. The primary aim was to test whether brief risk-reduction counseling performed at the time of an HIV test had any impact on subsequent incidence of sexually transmitted infections (STIs). The results showed no impact of risk-reduction counseling on STIs. Neither were there any substance use interactions of the impact of risk-reduction counseling; however, substance use was associated with higher levels of STIs at follow-up. Other research has shown that substance use is associated with higher rates of HIV testing, and Black women showing the highest rates of HIV testing in substance use treatment clinics (Hernández et al., 2016). Since substance use is associated with risky

sexual activity, detecting the dynamics of this relationship can contribute to preventive and educational efforts to control the spread of HIV. Our procedures for causal analysis of heterogeneity of effects in observational data should equalize the observed characteristics among substance use and non-substance use participants, thereby removing any impact of background imbalance in factors that may be related to relationship of substance use on sexual risk. Our procedure then allows an exploration of background factors that are truly related to this causal effect, conditional on all confounding factors being in the feature set.

To explore this issue of how substance use plays a role in sexual risk, we pursued an analysis in which the treatment (exposure) variable T was defined as drug use status of an individual ($0 =$ no substance use in the prior 6 months, $1 =$ any substance use in the prior 6 months leading to the study). For our outcome, we used number of unprotected sex acts within the last six months as reported by the individual. Although Project Aware was randomized on the primary outcome (risk-reduction counseling), analysis of secondary outcomes such as substance use should be treated as if from an observational study. Indeed, unbalancedness of the data for drug use can be gleaned from Table 4.3 which displays results from a logistic regression in which drug use status was used for the dependent variable ($n = 2813$, $p = 99$). The list of significant variables suggests the data is unbalanced and indicates that inferential methods should be considered carefully. Thus Table 4.4, which displays the results from a linear regression using number of unprotected sex acts as the dependent variable, should be interpreted with caution. Table 4.4 suggests there is no overall exposure effect of drug use, although several variables have significant drug-interactions.

However, in order to avoid drawing potentially flawed conclusions from an analysis like Table 4.4, we applied our counterfactual synthetic approach, $\hat{\tau}_{\text{synCF}}$. A synthetic forest was fit separately to each exposure group using number of unprotected sex acts as the dependent variable. This yielded estimated causal effects $\{\hat{\tau}_{\text{synCF}}(\mathbf{x}_i), i = 1, \ldots, n\}$ for $\tau(\mathbf{x})$ defined as the mean difference in number of unprotected sex acts for drug versus

Table 4.3: *Difference in variables by drug use illustrating unbalancedness of Aware data. Only significant variables (p-value < 0.05) from logistic regression analysis are displayed for clarity.*

	Estimate	Std. Error	Z	p-value
Race	-0.28	0.11	-2.50	0.01
Chlamydia	0.34	0.15	2.30	0.02
Site 2	-0.62	0.16	-3.94	0.00
Site 4	-0.53	0.16	-3.23	0.00
Site 6	0.44	0.18	2.43	0.01
Site 7	-0.65	0.15	-4.22	0.00
Site 8	0.95	0.21	4.51	0.00
HIV risk	0.17	0.03	5.14	0.00
CESD	0.02	0.01	3.13	0.00
Condom change 2	-0.24	0.12	-2.07	0.04
Marriage	0.08	0.03	2.76	0.01
In Jail ever	0.42	0.10	4.07	0.00
AA/NA last 6 months 1	0.69	0.23	3.04	0.00
Frequency of injection	0.18	0.07	2.49	0.01
Gender	-0.39	0.10	-4.00	0.00

non-drug users. The estimated causal effects were then used as dependent variables in a linear regression analysis. This is convenient because the estimated coefficients from the regression analysis can be interpreted in terms of subgroup causal differences (we elaborate on this point shortly). In order to derive valid standard errors and confidence regions for the estimated coefficients, the entire procedure was subsampled. That is, we drew a sample of size m without replacement. The subsampled data was then fit using synthetic forests as described above, and the resulting estimated causal effects used as the dependent variable in a linear regression. The procedure was repeated 1000 times independently. A subsampling size of $m = n/10$ was used. The confidence regions of the resulting coefficients are displayed in Figure 4.3. Table 4.5 displays the coefficients for significant values (p-values $< .05$). We note that bootstrapping could have have been used as another means to generate nonparametric p-values and confidence regions. However, we prefer subsampling because of its computational speed and general robustness (Politis et al., 1999).

Table 4.4: *Linear regression where dependent variable is number of unprotected sex acts from Aware data. Only variables with p-value < 0.10 from regression analysis are displayed for clarity.*

	Estimate	Std. Error	Z	p-value
Intercept	-5.06	22.17	-0.23	0.82
Drug	9.59	29.72	0.32	0.75
HCV2	8.14	4.09	1.99	0.05
Site 2	-14.00	6.82	-2.05	0.04
HIV risk	3.89	1.41	2.76	0.01
Condom change 3	-17.23	6.52	-2.64	0.01
Condom change 5	-21.98	6.65	-3.30	0.00
Visit opthamologist	-16.46	8.17	-2.01	0.04
Number visit opthamologist	9.13	3.93	2.33	0.02
Marriage	-2.13	1.17	-1.81	0.07
Smoke	53.13	16.26	3.27	0.00
Number cigarette per day	-14.63	4.76	-3.07	0.00
Drug x CESD	0.88	0.47	1.88	0.06
Drug x Condom change 2	-24.70	6.83	-3.62	0.00
Drug x Condom change 3	-25.82	8.75	-2.95	0.00
Drug x Condom change 4	-37.58	14.20	-2.65	0.01
Drug x Condom change 5	-28.78	9.33	-3.08	0.00
Drug x Visit dentist	-16.36	8.26	-1.98	0.05
Drug x Smoke	-34.71	20.53	-1.69	0.09
Drug x Number cigarette per day	9.81	5.95	1.65	0.10

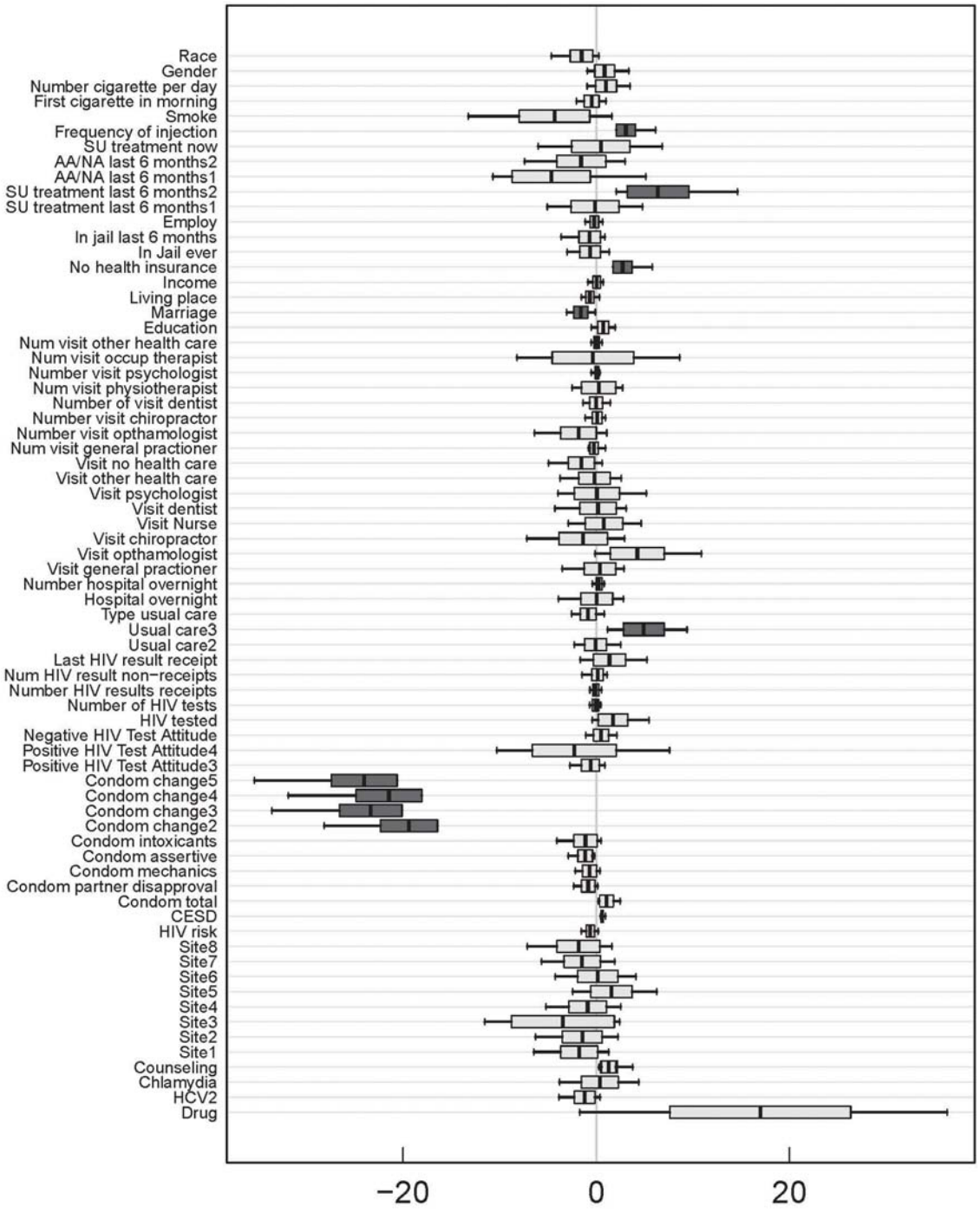

Figure 4.3: *Confidence intervals for all coefficients of linear model used in Table 3. Intervals determined using subsampling. Dark colored boxplots indicate variables with p-value* $< .05$.

To interpret the coefficients in Table 4.5, it is useful to write the true model for the outcome (number of unprotected sex acts) as $Y = f(\mathbf{X}, T) + \varepsilon$, where

$$f(\mathbf{X}, T) = \alpha_0 T + h(\mathbf{X}, T),$$

and h is some unknown function. Under the assumption of SITA, we have

$$\tau(\mathbf{x}) = f(\mathbf{x}, 1) - f(\mathbf{x}, 0) = \alpha_0 + h(\mathbf{x}, 1) - h(\mathbf{x}, 0).$$

Now since we assume a linear model $\alpha + \sum_{j=1}^{p} \beta_j x_j$ for the ITE, we have

$$\alpha_0 + h(\mathbf{x}, 1) - h(\mathbf{x}, 0) = \alpha + \sum_{j=1}^{p} \beta_j x_j.$$

From this we can infer that the intercept in Table 4.5 is an overall measure of the exposure effect of drug use, α_0 (this is why the intercept term is listed as drug use). Here the estimated coefficient is 17.0. The positive coefficient implies that on average drug users have significantly more unprotected sex acts than non-drug users (significance here is slightly larger than 5%).

The remaining coefficients in Table 4.5 describe how the effect of drug use on sexual risk is modulated by other factors. Under our linear model, we have

$$h(\mathbf{x}, 1) - h(\mathbf{x}, 0) = \sum_{j=1}^{p} \beta_j x_j.$$

Because $h(\mathbf{x}, 1) - h(\mathbf{x}, 0)$ represents how much a subgroup deviates from the overall causal effect, each coefficient in Table 4.5 quantifies the effect of a specific subgroup on drug use differences. Consider for example, the variable "Frequency of injection" which is a continuous variable representing frequency of injections in drug users. Because its esti-

Table 4.5: *Linear regression of Aware data with dependent variable equal to the estimated causal effects* $\{\hat{\tau}_{synCF}(\mathbf{x}_i), i = 1, \ldots, n\}$ *from counterfactual synthetic random forests. Causal effect is defined as the mean difference in unprotected sex acts for drug users versus non-drug users. Standard errors and significance of linear model coefficients were determined using subsampling. For clarity, only significant variables with p-value* < 0.05 *are displayed (the intercept is provided for reference but is not significant).*

	Estimate	Std. Error	Z
Intercept (drug use)	16.97	9.36	1.81
CESD	0.60	0.13	4.54
Condom change 2	-19.38	2.96	-6.56
Condom change 3	-23.33	3.23	-7.22
Condom change 4	-21.46	3.39	-6.32
Condom change 5	-24.02	3.41	-7.04
Usual care 3	4.91	2.11	2.33
Marriage	-1.61	0.73	-2.21
No health insurance	2.72	0.99	2.75
SU treatment last 6 months 2	6.38	3.20	2.00
Frequency of injection	3.59	1.77	2.02

mated coefficient is 3.6, this means the difference in unprotected sex acts between drug and and non-drug users, which is positive, becomes even wider for high frequency drug users. Another risky factor is "No health insurance", which is an indicator of lack of health insurance coverage. Because its estimated coefficient is 2.7, we can take this to mean that the increase in sexual risk for an individual without health insurance is more pronounced in drug users. As another example, consider the variable "Condom change" which is an ordinal categorical variable measuring an individual's stage of change with respect to condom use behavior. The baseline level is a "precontemplator", who is an individual who has not envisioned using condoms. The second level "contemplator" is an individual contemplating using condoms. Further increasing levels measure even more willingness to utilize condoms. All coefficients for Condom change in Table 4.5 are negative, and therefore if an individual is more willing to utilize safe condom practice (relative to the baseline condition), the difference in number of unprotected sex acts diminishes between drug and

non-drug users. Other variables that have a subgroup effect are Marriage (whether an individual is married), CESD (Center for Epidemiological Studies Depression Scale), and SU treatment last 6 months (substance abuse treatment in last 6 months). In all of these, the pattern is similar to before. With more risky behavior (with depression) the number of unprotected sex acts increases for non-drug users relative to drug users, but as risky behavior decreases (e.g. married), the effect of drug use diminishes.

Figure 4.4 displays a coplot of the RF estimated causal effects $\{\hat{\tau}_{\text{synCF}}(\mathbf{x}_i), i = 1, \ldots, n\}$ as a function of several variables. The coplot is another useful tool that can be used to explore causal relationships. We use it to uncover relationships that may be hidden in the linear regression analysis. The RF causal effects are plotted against CESD depression for individuals with and without health insurance. Conditioning is on the variables Condom change (vertical conditioning) and HIV risk (horizontal conditioning). HIV risk a self-rated variable and of potential importance and was included even though it was not significant in the linear regression analysis. For patients with potential to change condom use (rows 2 through 5), increased depression levels leads to an increased causal effect of drug use, which is slightly accentuated for high HIV risk (plots going from left to right). The effect of health insurance is however minimal. On the other hand, for individuals with low potential to change condom use (bottom row), the estimated exposure effect is generally high, regardless of depression, but is reduced if the individual has health insurance.

4.3 Discussion

In observational data with complex heterogeneity of treatment effect, individual estimates of treatment effect can be obtained in a principled way by directly modeling the response outcome. However, successful estimation mandates highly adaptive and accurate regression methodology and for this we relied on RF, a machine learning method with well known

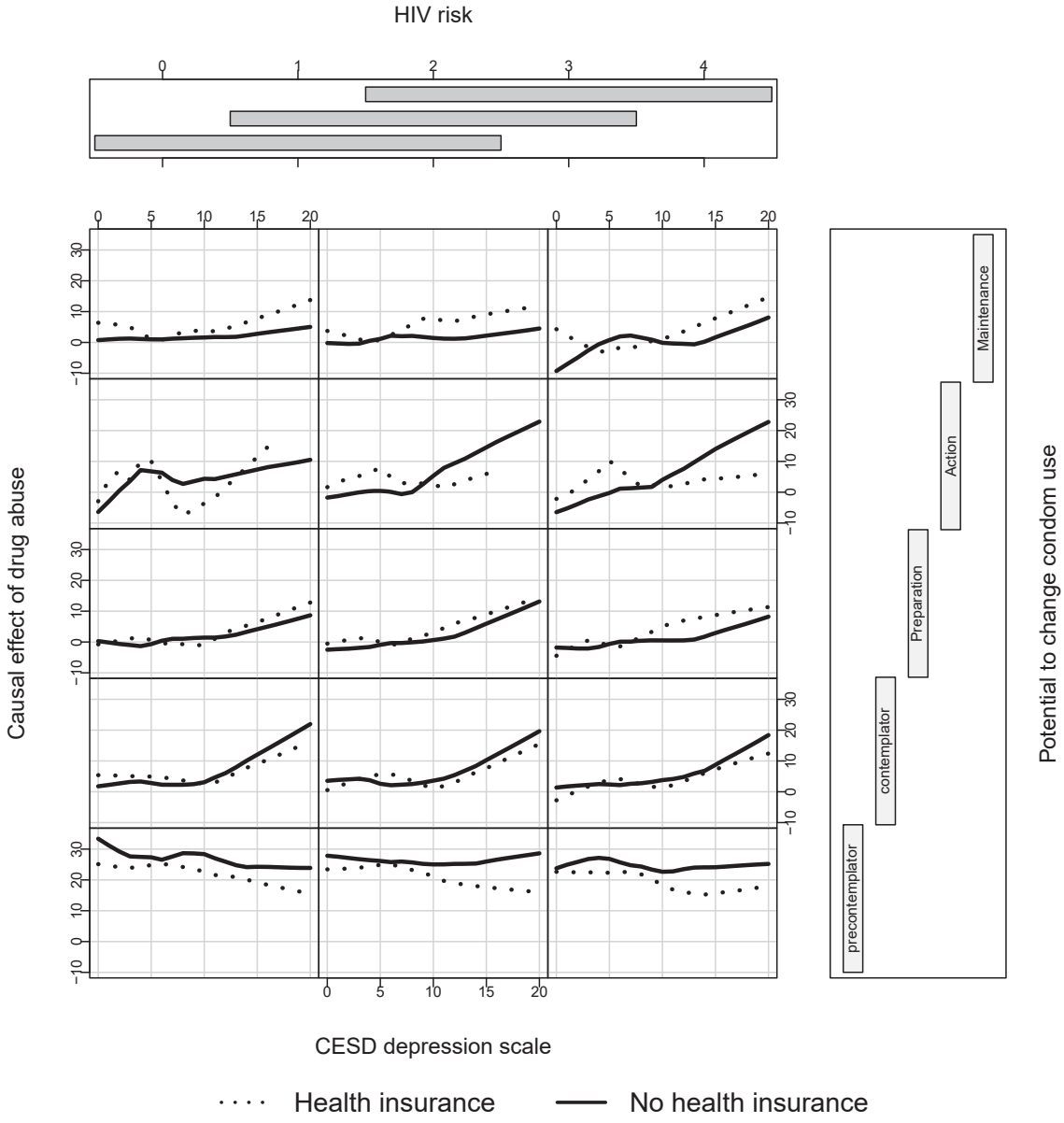

Figure 4.4: *RF estimated causal effect of drug use plotted against CESD depression for individuals with and without health insurance. Values are conditioned on Condom change (vertical conditional axis) and HIV risk (horizontal conditional axis).*

properties for accurate estimation in complex nonparametric regression settings. However, care must be used when applying RF for casual inference. We encourage the use of out-of-bag estimation, a simple but underappreciated out-of-sample technique for improving accuracy. We also recommend that when selecting a RF approach, that it should have some means for encouraging adaptivity to confounding, i.e. that it can accurately model potentially separate regression surfaces for each of the treatment groups. One example of this is the extension to VT, which expands the design matrix to include all pairwise interactions of variables with the treatment, a method we call $\hat{\tau}_{\text{VT-I}}$, and described in the paper by Foster et al. (2011). We found that this simple extension, when coupled with out-of-bagging, significantly improved performance of VT. Another promising method was counterfactual synthetic forests $\hat{\tau}_{\text{synCF}}$, which generally had the best performance among all methods, and was superior in the larger sample size simulations, outperforming even the highly adaptive BART method. The larger sample size requirement is not so surprising as having to grow separate forests causes some loss of efficiency; this being however mitigated by its superior bias properties which take hold with increasing n.

In looking back, we can now see that the success of counterfactual synthetic RF can be attributed to three separate effects: (a) fitting separate forests to each treatment group, which improves adaptivity to confounding; (b) replacing Breiman forests with synthetic forests, which reduces bias; and (c) utilizing OOB estimation, which improves accuracy. Computationally, counterfactual synthetic RF are easily implemented with available software and have the added attraction that they reduce parameter tuning. The latter is a consequence and advantage of using synthetic forests. A synthetic forest is constructed using RF base learners, each of these being constructed under different nodesize and mtry tuning parameters. Correctly specifying mtry and nodesize is important for good performance in Breiman forests. The optimal value will depend on whether the setting is large n, large p, or large p and large n. With synthetic forests this problem is alleviated by building RF base

learners under different tuning parameter values.

Importantly, and underlying all of this, is the potential outcomes model, a powerful hypothetical approach to causation. The challenge is being able to properly fit the potential outcomes model and for this, as discussed above, we relied on the sophisticated machinery of RF. We emphasize that the direct approach of the potential outcomes model is well suited for personalized inference via the ITE. Estimated ITE values from RF can be readily analyzed using standard regression models to yield direct inferential statements for not only overall treatment effect, but also interactions, thus facilitating inference beyond the traditional ATE population-centric viewpoint. Using the Aware data we showed how counterfactual ITE estimates from counterfactual synthetic forests could be explored to understand causal relations. This revealed interesting connections between risky behavior, drug use, and sexual risk. The analysis corrects for any observed differences by the exposure variable, so to the extent that we have observed the important confounding variables, this result can tentatively be considered causal, though caution should be used due to this assumption. Clearly, this type of analysis, which controls for observed confounding gives additional and important insights above simple observed drug usage differences. We also note that although we used linear regression for interpretation in this analysis, it is possible to utilize other methods as well. The counterfactual synthetic forest procedure provides a pipleline that can be connected with many types of analyses, such as the conditional plots that were also used in the Aware data analysis.

Chapter 5

Personalized Treatment in ischemic cardiomyopathy

Estimation of multiple treatment effects in observational survival data is complicated due to confounding, heterogeneity, and selection bias. A key challenge is assessing overlap and, possibly, estimation of effects strictly within overlapping populations that are eligible for the corresponding treatments. Unfortunately, treatments do not always have clearly defined evidence based eligibility criteria. Therefore, we propose new random forest methods to address individual therapy overlap. These methods possess the unique feature of being able to incorporate external expert knowledge either in a fully supervised way (i.e., we have a strong belief that knowledge is correct) using multilabel analyses, or in a minimally supervised fashion (i.e., knowledge is not considered gold-standard) using multiclass analyses. We directly estimate individual treatment effect (ITE) and average treatment effect (ATE) through comparison of survival under counterfactual treatment assignments using an extension to random survival forests we call virtual twin random survival forests interaction. Treatment effect is viewed as a dynamic causal procedure to making treatment decisions. Motivation for our methodology arose from the problem of current treatment management for ischemic cardiomyopathy. Using a large observational survival data set, four well established therapies are compared: coronary artery bypass grafting (CABG), CABG combined with surgical ventricular reconstruction (SVR), CABG combined with mitral valve anuloplasty (MVA), and listing for heart transplantation (LCTx).

5.1 Background

It is common in medical settings for multiple treatment options to be available for a patient. However determining the therapy most applicable for a patient that maximizes patient outcome, such as survival, is often difficult. One problem is that therapy will not always have clearly defined evidence based eligibility criteria. Thus, the same patient may be treated differently by different physicians, or at different hospitals, seemingly without explicit or evident reasons. A second problem is that even when treatment eligibility can be decided, the efficacy of eligible treatments may be controversial.

Consider ischemic cardiomyopathy, a cardiovascular condition in which the arteries that supply blood to the heart become narrowed. Over time, the lack of blood causes the heart to become enlarged and dilated and ultimately reduces its ability to pump blood to the body. Treatments for ischemic cardiomyopathy include coronary artery bypass grafting (CABG), a surgical treatment developed in the late 1960's for patients with severe obstructive coronary artery disease. Other treatments include surgical ventricular reconstruction (SVR) performed together with CABG; and mitral valve anuloplasty (MVA) performed together with CABG. Patients may also be listed for heart transplant (LCTx) in cases where the above surgical interventions may not be appropriate. For convenience, Table 1 lists the four treatments and their acronyms.

Determining appropriate therapy for ischemic cardiomyopathy is difficult because eligibility for therapy depends on many factors. For heart transplant listing (LCTx), presence of severe clinical symptoms is required, for example as measured by NYHA functional class, and patients must generally be less than 70 years of age, although this restriction may additionally depend on the hospital and the state. For MVA, severe mitral valve regurgitation grade is generally required, but for CABG, patients typically have minimal mitral valve regurgitation. Further complicating matters is that even when eligibility of therapy can be

Table 5.1: Abbreviations and terminology used throughout the paper

Abbreviation	Definition
CABG	Coronary artery bypass grafting alone
MVA	Coronary artery bypass grafting with mitral valve anuloplasty
SVR	Coronary artery bypass grafting with surgical ventricular reconstruction
LCTx	Listing for cardiac transplantation
ATE	Average treatment effect
ATT	Average treatment effect on the treated
ITE	Individual treatment effect
ITR	Individualized treatment rule
RSF-VT-I	Random survival forests virtual twins interactions

agreed upon, efficacy of therapy can be highly controversial. For example, the Surgical Treatment for Ischemic Heart Failure (STICH) trial, a large study funded by the National Heart, Lung, and Blood Institute, showed SVR reduced left ventricular volume compared with CABG, but did not significantly reduce the rate of death or hospitalization due to cardiac related causes (Jones et al., 2009). Another example is ischemic mitral regurgitation, an affliction affecting thousands of Americans after a heart attack, which is caused by papillary muscle displacement, leaflet tethering, and ventricular dilatation. There is evidence that MVA performs better than CABG in diminishing postoperative mitral regurgitation and improving early symptoms (Mihaljevic et al., 2007). However, whether, or to which subgroup of patients, MVA improves long-term functional status and survival for patients with chronic mitral regurgitation, remains debatable.

5.1.1 Patients

In this paper, we consider the problem of estimating individual treatment effects and determining optimal treatment for ischemic cardiomyopathy. We base our analysis on data from 1468 patients who were treated for ischemic cardiomyopathy at Cleveland Clinic

from 1997 to 2007. Ischemic cardiomyopathy was defined as having severe left ventricular systolic dysfunction with a measured or estimated ejection fraction of less than 30%. Of the 1468 patients in the study, 386 underwent CABG, 360 SVR, 212 MVA, and 510 LCTx. The primary outcome was all-cause mortality, including in-hospital mortality after surgical procedures and interim deaths while awaiting transplantation. Mean duration of follow-up was 3.8 years. See Yoon et al. (2010) for a detailed description of the data.

5.1.2 Approach

The observational nature of this study, which makes heterogeneity and confounding quite likely, and the lack of overlap in treatment, required us to develop innovative causal methodology for our analysis. Building on previous work of random forests for causal inference (Wager and Athey, 2017; Lu et al., 2018), we propose a novel extension for estimating the individual treatment effect (ITE) in observational survival data. The extension to the survival setting from previous work (Lu et al., 2018), which looked at regression, involved dealing with nuances unique to survival. The definition of ITE, for example, as well as other quantities for estimating treatment effectiveness, are more complicated as these values depend not only on patient pre-treatment variables, but also survival time. As well, censoring had to be accommodated. Definitions and a framework for this extension are given in Section 5.2. Another innovation involved dealing with patient treatment overlap. A unique feature of this particular data was that "expert" knowledge was available for determining overlap. Section 4 describes new random forest methods for incorporating this information ranging from being fully supervised, to partially supervised, for determining overlap of treatment.

Section 5.5 describes our new ITE approach which is an extension of the virtual twins random forests method proposed by Foster et al. (2011) Although virtual twins was originally described in the context of randomized studies involving continuous outcomes, the

idea rests on a counter-factual framework which can be extended to ITE analysis for observational data (Lu et al., 2018). In order to extend the method to survival settings, we make use of random survival forests (Ishwaran et al., 2008), and introduce an extension which we call random survival forests virtual twins interaction. This strategy is different from outcome weighted learning approaches (Zhao et al., 2012, 2014, 2017; Zhang et al., 2012; Bai et al., 2017; Zhu et al., 2017) and inverse probability of censoring weighting (IPCW) methods (Robins et al., 2008; Goldberg and Kosorok, 2012). This is because our approach directly models the target outcome. This is often referred to as an outcome-regression approach (Robins, 1986, 2004; Murphy, 2003; Moodie et al., 2007; Qian and Murphy, 2011; Hill, 2011; Lu et al., 2018). It is known that the performance of outcome-regression methods depends critically on the predictive performance of the estimated model. Parametric models are likely to perform poorly in observational data settings due to complex interactions, non-linear effects, and departures from model assumptions that are likely to be at play. In contrast, random survival forests, which forms the basis of our approach, is a nonparametric and highly robust procedure yielding accurate estimation of survival curves. Lu et al. (2018) demonstrated that random forests, when cast in a virtual twins interaction framework, is highly accurate for outcome-regression modeling.

5.1.3 Contributions and outline

There are several unique and innovative aspects to our work which we highlight below.

1. In Sections 2 and 5, we show that not only does the ITE provide us with insight into personalized treatment, but it also directly yields population measures of treatment effectiveness, such as the average treatment effect (ATE). In contrast, ATE estimates are otherwise obtained indirectly using techniques such as matching or propensity score (PS) analyses. For the latter, matching on the PS, stratification on the PS, or

weighting with the PS are utilized (Austin, 2011). See Parast and Griffin (2017) for advanced propensity score approaches for survival outcomes using landmark estimation.

2. Hill and Su (2013) and McCaffrey et al. (2013) provide guidance for checking overlap in observational data. A unique feature of our study was the availability of expert knowledge defining treatment eligibility. We use this to assess overlap using two strategies (Section 4). One strategy, which is fully supervised, uses expert knowledge in a novel multilabel analysis. The other strategy, partially supervised, estimates overlap using treatment assignment.

3. Often lack of overlap results in sample size reduction, however we utilize all data points in our outcome-regression strategy when estimating the conditional survival function, thereby mitigating loss of statistical efficiency. Our approach does not use treatment assignment probability to weight patients, but rather restricts consideration of ITE to covariates where overlap holds.

4. We obtain direct estimates of potential outcome survival curves. This allows us to dynamically view treatment effect as a function of time and yields important insight in our analysis by visualizing treatment effectiveness over time over multiple treatments simultaneously.

5. Section 3 draws a direct link between the ITE and the individualized treatment rule (ITR) (Qian and Murphy, 2011). Importantly, our analysis is able to interpret these rules in meaningful clinical terms and to identify functional relationships between variables for multi-dimensional visualization of how patient information impacts gain of life under optimal therapy.

5.2 Treatment effect for observational survival data

In this section, we describe the formal framework and assumptions used for our causal inference of observational survival data. Let $\{(\mathbf{X}_1, Z_1, T_1, \delta_1), \ldots, (\mathbf{X}_n, Z_n, T_n, \delta_n)\}$ denote the data, assumed to be independently distributed from a common distribution $\mathbb{P}$, where $\mathbf{X}_i$ denotes the covariate vector for individual i, (T_i, δ_i) is the observed survival outcome, and Z_i denotes i's assigned treatment group. We asssume Z_i is coded as an integer value from $\{1, \ldots, M\}$, where $M > 1$ is the total number of available treatments. The individual's survival data is as (T_i, δ_i), where $T_i = \min(T_i^o, C_i^o)$ is the observed survival time and $\delta_i = \mathbf{1}_{\{T_i^o \leq C_i^o\}}$ is the observed censoring variable. Here T_i^o denotes the true event time and C_i^o the censoring time. We say i is right-censored at time T_i if $\delta_i = 0$; otherwise the individual is said to have an event at T_i.

5.2.1 Unconfoundedness

Without additional assumptions, it is generally not possible to estimate treatment effects in observational studies. A standard assumption is unconfoundedness (Rosenbaum and Rubin, 1983), which assumes all relevant covariates are available to the analyst and in particular that there are no unmeasured covariates associated with both treatment and potential outcomes. Unconfoundedness is typically formulated as a conditional independence of treatment and potential outcomes, conditional on pre-treatment variables.

In our framework, the potential survival outcome is the potential event time $T^o(j)$ and potential censoring time $C^o(j)$ under a given treatment $Z = j$. The potential survival outcomes are related to the actual survival outcome (T^o, C^o) via the consistency assumption:

$$(T^o, C^o) = \sum_{j=1}^{M} \mathbf{1}_{\{Z=j\}}(T^o(j), C^o(j)). \tag{5.1}$$

We will assume a type of weak unconfoundedness, a less stringent assumption of unconfoundedness (Imbens, 2000). Weak unconfoundedness asserts conditional independence of *each* potential outcome, and not the stronger assumption of *joint-multivariate* conditional independence of the potential outcomes (Rosenbaum and Rubin, 1983). The following is a slightly different formulation than used by Imbens (2000).

Definition 5.2.1. Weak unconfoundedness holds for treatment j at $\mathbf{X} = \mathbf{x}$ if

$$\mathbb{P}\{T^o(j) > t_1, C^o(j) > t_2 \,|\mathbf{X} = \mathbf{x}, Z = j\} = \mathbb{P}\{T^o(j) > t_1, C^o(j) > t_2 \,|\mathbf{X} = \mathbf{x}\}$$

for any real valued t_1, t_2.

5.2.2 Treatment overlap

Another key concept in our development is treatment overlap.

Definition 5.2.2. Let $p_j(\mathbf{x}) = \mathbb{P}\{Z = j|\mathbf{X} = \mathbf{x}\}$. Complete overlap for all treatments is said to hold for $\mathbf{x}$ if $p_j(\mathbf{x}) > 0$ for $j = 1, \ldots, M$. Overlap between treatments j and k is said to hold for $\mathbf{x}$ if $p_j(\mathbf{x}) > 0$ and $p_k(\mathbf{x}) > 0$. Finally, if $p_j(\mathbf{x}) > 0$, we say $\mathbf{x}$ satisfies overlap for treatment j.

Treatment overlap is necessary to ensure covariate distributions between treatment groups have common support, thereby ensuring treatment effectiveness can be estimated. The function $p_j(\mathbf{x})$ is referred to as the propensity score in the casual literature. The propensity score has many uses including dealing with the common support problem. One strategy used to estimate population measures of treatment effectiveness is to construct estimators using only those data that fall within a suitable range of propensity score values (Dehejia and Wahba, 1999; Heckman et al., 1997, 1998; Morgan and Harding, 2006). The resulting estimators are then interpreted within a narrower treatment effect perspective: the common

support treatment effect (Heckman et al., 1997, 1998). In the analysis of individualized treatment, as considered here, the assumption of overlap is required for counter-factual assignment to be plausible. By consistency (5.1), if overlap does not exist for treatment j, then the potential outcome under treatment j is not plausible for $\mathbf{x}$. Therefore, ITE analysis for $\mathbf{x}$ must exclude j in treatment comparisons. We will come back to this point shortly.

We can conceptualize overlap as a binary function that identifies whether treatment j can be administered to $\mathbf{x}$. We formally define the overlap function as follows

$$o_j(\mathbf{x}) = \mathbf{1}_{\{p_j(\mathbf{x})>0\}}.$$

In practice, the propensity score is unknown and therefore must be estimated by some estimator $\hat{\mathbf{p}}_j(\mathbf{x})$. Because this provides only an approximation, the overlap function should be estimated using

$$\hat{\mathrm{o}}_j(\mathbf{x}; C) = \mathbf{1}_{\{\hat{\mathbf{p}}_j(\mathbf{x})>C\}}, \tag{5.2}$$

where $0 < C < 1$ is an appropriately selected cutoff value.

Assessing overlap can be difficult (Hill and Su, 2013) and therefore it is useful to consider methods other than the propensity score to achieve this goal. A unique feature of our study was the availability of an expert database identifying patient eligibility for treatment. This presented us with the opportunity to assess overlap using expert knowledge. Let $E_j \in \{0, 1\}$ denote eligibility for treatment j. Define $e_j(\mathbf{x}) = \Phi\{E_j = 1|\mathbf{X} = \mathbf{x}\}$ as the eligibility probability defined by expert knowledge. Lack of overlap can be assessed by the magnitude of $e_j(\mathbf{x})$: large values identify when overlap for j is highly likely. If these probabilities are high across all treatments, it may be possible for $\sum_{j=1}^{M} e_j(\mathbf{x}) > 1$, thus reflecting a high overlap in all treatment groups. This motivates the following empirical estimator for overlap:

$$\hat{\mathrm{o}}_j(\mathbf{x}; C) = \mathbf{1}_{\{\hat{\varepsilon}_j(\mathbf{x})>C\}}, \tag{5.3}$$

where $0 < C < 1$ is an appropriately selected cutoff and $\hat{\varepsilon}_j(\mathbf{x})$ is an estimator for $e_j(\mathbf{x})$. Section 4 discusses a novel multilabel procedure for obtaining estimates of $e_j(\mathbf{x})$. We also describe a calibration method for determining threshold values C for (5.3) and also for (5.2).

5.2.3 Ignorable treatment assignment

Now we return to the issue of how overlap affects ITE inference. In the causal literature, it is common to combine the assumption of unconfoundedness and overlap into a single assumption. Rosenbaum and Rubin (1983) define strongly ignorable treatment assignment (SITA) as the combined requirements of strong unconfoundedness and treatment assignment overlap; the latter often being referred to as the positivity condition. However, positivity is a strong assumption as complete treatment overlap may not always exist. Even in Imbens (2000), weak unconfoundedness is implicitly assumed to hold for all $\mathbf{x}$, thus implying the existence of all potential outcomes regardless of $\mathbf{x}$. Neither of these assumptions are always realistic in multiple treatment settings ($M > 2$). This is because a specific treatment, or set of treatments, may be implausible for certain $\mathbf{x}$. This is certainly true for ischemic cardiomyopathy where treatment option depends strongly on clinical make up of a patient (for example, see Figure 5.1 which displays the eligibility status for our patients determined by experts).

The important point is that even though certain treatments may be precluded for a given $\mathbf{x}$, ITE analysis is still possible as long as we consider treatment comparisons among treatments satisfying overlap. To accommodate this scenario, we introduce a more flexible definition of ignorable treatment assignment.

Definition 5.2.3. Weak ignorable treatment assignment (WITA) holds if weak unconfoundedness holds for treatment $j = 1, \ldots, M$ for all $\mathbf{x}$ satisfying overlap, $\{\mathbf{x} : o_j(\mathbf{x}) = 1\}$.

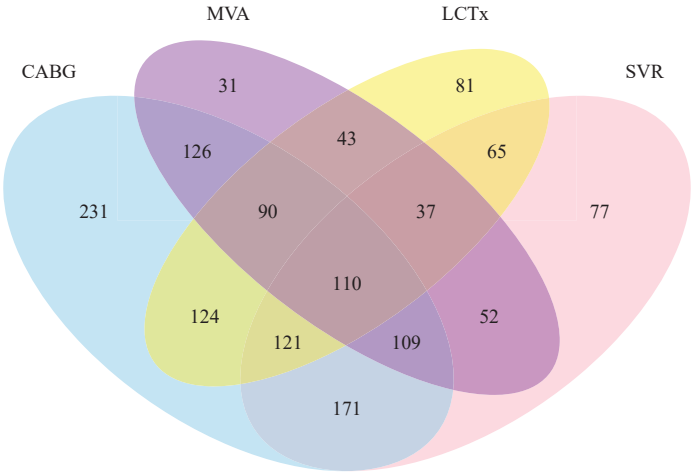

Figure 5.1: *Number of patients eligible for treatment determined by expert knowledge (total sample size,* $n = 1468$*). The many non-overlapping sets provides strong evidence of lack of overlap.*

WITA lifts the requirement that weak unconfoundedness must hold for all $\mathbf{x}$. Thus, it can be seen to be a slightly weaker condition than weak unconfoundedness of Imbens (2000). This slightly more flexible definition is better suited for ITE inference and fits naturally within our two-step approach to ITE estimation. In the first step, we estimate the survival function using random survival forests, using all the data for all treatments simultaneously without restrictions to the data regarding overlap of covariate distributions, such as treatment assignment probability or eligibility probability. This allows us to use the full sample size so as not to decrease efficiency of our survival function estimator. Overlap, assumed by WITA, is then taken into account in downstream analyses (second step) of the resulting estimated survival function. Thus, in assessing the ITE (Sections 2,4, 2.5) we only consider treatment comparisons for $\mathbf{x}$ where overlap holds. The same is true for population measures such as the ATE where we restrict its calculation to patients satisfying overlap (see Section 2.6).

5.2.4 Individual treatment effect (ITE)

We now define various useful quantities for assessing treatment effectiveness and show how these can be estimated under our assumptions. We begin by providing a definition for ITE in survival settings. Causal inference in survival settings typically focuses on a specific time point, however we emphasize the dynamic aspect of treatment effectiveness in the following definition of ITE which is a function of both $\mathbf{x}$ and t.

Definition 5.2.4. The individual treatment effect (ITE) at time t for covariate $\mathbf{x}$ for treatment j over treatment k is

$$\tau_{j,k}(t,\mathbf{x}) = \psi\Big(S_j\big(t|\mathbf{x}\big), S_k\big(t|\mathbf{x}\big)\Big),$$

where $\psi(\cdot,\cdot)$ is a known function specified by the analyst (examples are given in Section 2.5), and $S_l\big(t|\mathbf{x}\big) = \mathbb{P}\{T^o(l) > t|\mathbf{X}=\mathbf{x}\}$ is the survival function for the potential outcome $T^o(l)$ conditioned on $\mathbf{X}=\mathbf{x}$ for $l = j,k$.

Combined with model identification and consistency, WITA ensures $S_j(t|\mathbf{x})$ is estimable from the observed data. For all $\mathbf{x}$ satisfying overlap for treatment j,

$$\begin{aligned}
&S_j\big(t|\mathbf{x}\big)\\
&\quad= \mathbb{P}\{T^o(j) > t|\mathbf{X}=\mathbf{x}\}\\
&\quad= \mathbb{P}\{T^o(j) > t|\mathbf{X}=\mathbf{x}, Z=j\} \quad \text{(weak unconfoundedness; } t_1 = t, t_2 = -\infty)\\
&\quad= \mathbb{P}\{T^o > t|\mathbf{X}=\mathbf{x}, Z=j\} \qquad \text{(consistency (5.1))}\\
&\quad= S\big(t|\mathbf{x}, Z=j\big),
\end{aligned}$$

where $S(t|\mathbf{x}, Z=j)$ is the survival function for T^o conditioned on $\mathbf{X}=\mathbf{x}$ and $Z=j$. Model identifiability is required in order for the survival function $S(t|\mathbf{x}, Z=j)$ to be es-

timable. We will rely on the standard assumption that time and censoring are conditionally independent,

$$T_i^o \perp C_i^o | (\mathbf{X}_i, Z_i). \tag{5.4}$$

Under this assumption, $S(t|\mathbf{x}, Z = j)$ is identifiable and estimable from the observed data, and by the above, equals the potential outcomes survival function. Thus, the unobservable potential outcomes survival function $S_j(t|\mathbf{x})$ can be estimated by the survival function $S(t|\mathbf{x}, Z = j)$ from the observable data under the assumption of consistency, conditional independence of T^o and C^o, and the assumption of WITA.

Remark. We have assumed conditional independence on the actual survival outcomes, which is a standard assumption, and therefore familiar to most readers. However, under our assumptions, this implies conditional independence of the potential survival outcomes, which is a condition we would naturally anticipate. For all $\mathbf{x}$ satisfying overlap for treatment j,

$$\begin{aligned}
&\mathbb{P}\{T^o(j) > t_1, C^o(j) > t_2 \,|\mathbf{X} = \mathbf{x}\} \\
&= \mathbb{P}\{T^o(j) > t_1, C^o(j) > t_2 \,|\mathbf{X} = \mathbf{x}, Z = j\} && \text{(weak unconfoundedness)} \\
&= \mathbb{P}\{T^o > t_1, C^o > t_2 \,|\mathbf{X} = \mathbf{x}, Z = j\} && \text{(consistency (5.1))} \\
&= \mathbb{P}\{T^o > t_1 \,|\mathbf{X} = \mathbf{x}, Z = j\}\mathbb{P}\{C^o > t_2 \,|\mathbf{X} = \mathbf{x}, Z = j\} && \text{(independence (5.4))} \\
&= \mathbb{P}\{T^o(j) > t_1 \,|\mathbf{X} = \mathbf{x}\}\mathbb{P}\{C^o(j) > t_2 \,|\mathbf{X} = \mathbf{x}\},
\end{aligned}$$

where the last line follows by a combined application of consistency and weak unconfoundedness.

5.2.5 Expressing the ITE in terms of the estimable survival function

Hereafter we will assume that consistency (5.1), independence (5.4), and WITA holds. By the above, this implies the ITE for any two treatments j and k is estimable and can be written as

$$\tau_{j,k}(t, \mathbf{x}) = \psi\Big(S\big(t|\mathbf{x}, Z = j\big), S\big(t|\mathbf{x}, Z = k\big)\Big) \tag{5.5}$$

for all $\mathbf{x}$ satisfying overlap $o_j(\mathbf{x}) = o_k(\mathbf{x}) = 1$. Given an estimator $\hat{S}(t|\mathbf{x}, Z)$ for the survival function (which we will estimate using random survival forests), we estimate the ITE by substituting $\hat{S}$ into ψ. Examples of $\psi(\cdot, \cdot)$ that can be used to define the ITE, include

$$\tau^{(1)}_{j,k}(t, \mathbf{x}) = S\big(t|\mathbf{x}, Z = j\big) - S\big(t|\mathbf{x}, Z = k\big), \tag{5.6}$$

where $\psi(a, b) = a - b$, so that $\tau_{j,k}(t, \mathbf{x})$ is the difference of two survival curves. Another way to measure ITE is through survival curve domination,

$$\tau^{(2)}_{j,k}(t, \mathbf{x}) = \mathbf{1}_{\{S(t|\mathbf{x}, Z = j) > S(t|\mathbf{x}, Z = k)\}},$$

which corresponds to $\psi(a, b) = \mathbf{1}_{\{a>b\}}$.

Andersen (2013) defines the expected number of years alive before time t_0 as the survival function integrated from $[0, t_0]$ (also commonly referred to as the restricted mean survival time, RMST (Irwin, 1949; Andersen et al., 2004; Royston and Parmar, 2011; Kim et al., 2017). Typically, t_0 is chosen to equal the maximum observed follow-up time. In a similar manner, integrating over $t \in [0, t_0]$, we define the ITE before time t_0 as

$$\tau_{j,k}([0, t_0], \mathbf{x}) = \int_0^{t_0} \tau_{j,k}(t, \mathbf{x})\, dt. \tag{5.7}$$

For example, if $\tau_{j,k}(t, \mathbf{x}) = \tau^{(1)}_{j,k}(t, \mathbf{x})$, this can be interpreted as the difference in the RMST for treatment j over k, thus assessing gain (or loss) in restricted lifetime in treatment j over k.

5.2.6 Average treatment effect (ATE)

From the ITE we can directly calculate many useful quantities for assessing treatment effectiveness, such as the ATE.

Definition 5.2.5. The average treatment effect (ATE) at time t for treatment j over treatment k is

$$\tau_{j,k}(t) = \mathbb{E}\big[\tau_{j,k}(t, \mathbf{X}) \big| o_j(\mathbf{X}) = 1, o_k(\mathbf{X}) = 1\big]. \tag{5.8}$$

For example if $\tau_{j,k}(t, \mathbf{x}) = \tau^{(1)}_{j,k}(t, \mathbf{x})$, the ATE at time t, denoted as $\tau^{(1)}_{j,k}(t)$, equals the conditional population average difference in survival curves at time t for treatment j compared to k. If the ITE is $\tau^{(2)}_{j,k}(t, \mathbf{x})$, then the ATE is $\tau^{(2)}_{j,k}(t)$, which equals the conditional population average domination of one survival curve over the other. *Lack of overlap* for treatments j and k for a given $\mathbf{x}$ implies the existence of an individual with covariate $\mathbf{x}$ with zero assignment probability for at least one of the treatments. Notice that the conditioning in (5.8) excludes such cases, which only permits individuals satisfying overlap.

In a similar fashion to (5.7), we define the ATE before time t_0 as

$$\tau_{j,k}([0, t_0]) = \int_0^{t_0} \tau_{j,k}(t)\, dt. \tag{5.9}$$

For example if $\tau_{j,k}(t, \mathbf{x}) = \tau^{(1)}_{j,k}(t, \mathbf{x})$, this equals the average difference in RMST before t_0, which can be interpreted as the average gain (or loss) in restricted lifetime.

5.3 Individualized treatment rules

In this section, we draw a direct connection between the ITE and the Individualized Treatment Rule (ITR) (Qian and Murphy, 2011) to connect our work to the literature on optimal treatment decision making. As in Qian and Murphy (2011) let R be some continuous real valued quantity representing the target response, where larger values are better. An ITR is a decision rule $d : \mathscr{X} \to \mathscr{Z}$ which maps an individual's feature $\mathbf{X} \in \mathscr{X}$ to the action $Z \in \mathscr{Z} = \{1, \ldots, M\}$ of possible treatments. Denote the distribution of $(\mathbf{X}, Z, R)$ by Φ and let Φ^d denote the distribution of $(\mathbf{X}, Z, R)$ constrained to $Z = d(\mathbf{X})$. The expectation of the target response, R, with respect to the distribution $\mathbb{P}^d$ is called the expected reward for d, denoted by $V(d)$. Assuming that $p(Z|\mathbf{X}) > 0$ almost everywhere, it can be show that (Qian and Murphy, 2011)

$$V(d) = \mathbb{E}\left[\frac{\mathbf{1}_{\{Z=d(\mathbf{X})\}}}{p(Z|\mathbf{X})}R\right] = \mathbb{E}\Big[Q_0(\mathbf{X}, d(\mathbf{X}))\Big], \tag{5.10}$$

where $Q_0(\mathbf{X}, Z) = \mathbb{E}(R|\mathbf{X}, Z)$ is called the quality of treatment Z for $\mathbf{X}$. The optimal treatment d^{opt} is the ITR in the space of decision rules $\mathscr{D}$ with maximum reward

$$d^{\text{opt}} = \underset{d\in\mathscr{D}}{\text{argmax}}\big\{V(d)\big\}.$$

To see how this is related to the ITE, define R to be the RMST; other choices are of course possible:

$$R = \int_0^{t_0} S\big(t|\mathbf{X}, Z\big)\, dt.$$

The difference in quality of treatment under treatment j and k for $\mathbf{X} = \mathbf{x}$ satisfying overlap

$o_j(\mathbf{x}) = o_k(\mathbf{x}) = 1$ is

$$\begin{aligned} & Q_0(\mathbf{x}, Z = j) - Q_0(\mathbf{x}, Z = k) \\ &= \int_0^{t_0} S\big(t|\mathbf{x}, Z = j\big)\, dt - \int_0^{t_0} S\big(t|\mathbf{x}, Z = k\big)\, dt \\ &= \int_0^{t_0} \tau_{j,k}^{(1)}(t, \mathbf{x})\, dt := \tau_{j,k}^{(1)}([0, t_0], \mathbf{x}), \end{aligned}$$

which is the ITE before t_0 under $\tau_{j,k}^{(1)}$ and equals the expected gain (or loss) in restricted years for $\mathbf{x}$.

By the above, treatment j is prefered to treatment k for $\mathbf{X} = \mathbf{x}$ if and only if $Q_0(\mathbf{x}, Z = j) > Q_0(\mathbf{x}, Z = k)$. It is clear because the optimal decision rule is

$$d^{\text{opt}}(\mathbf{x}) = \underset{\{l: o_l(\mathbf{x})=1\}}{\text{argmax}} \left\{ \int_0^{t_0} S\big(t|\mathbf{x}, Z = l\big)\, dt \right\},$$

that d^{opt} is uniquely determined by the ITE.

Thus our strategy for estimating d^{opt} is the same as our strategy for estimating the ITE. We use random survival forests to directly estimate the survival function $S\big(t|\mathbf{x}, Z = l\big)$, which provides not only an estimate for the ITE, but as we have now just shown, also an estimate for the optimal treatment rule. This is different than strategies that have been used up to this point. Even in Qian and Murphy (2011), which is an outcome-regression approach like ours, we find differences. Putting aside that they consider clinical trial data and focus on regression, an important distinction is that they estimate the conditional mean for the response, which is used to estimate the optimal decision rule. This is unlike our approach where we estimate the conditional survival function (i.e. viewed from a regression perspective, we are estimating the conditional distribution function rather than the conditional mean). Another strategy used for estimating d^{opt} is Outcome Weighted Learning (OWL, also refered as O-learning) (Zhao et al., 2012, 2014). This is different than

our outcome-regression modeling. O-learning makes use of the first identity in (5.10): $\mathbb{E}[R\mathbf{1}_{\{Z=d(\mathbf{X})\}}/p(Z|\mathbf{X})]$ and notes that maximizing this value is equivalent to minimizing $\mathbb{E}[R\mathbf{1}_{\{Z\neq d(\mathbf{X})\}}/p(Z|\mathbf{X})]$. This can be viewed as a weighted classification problem with binary outcomes $\mathbf{1}_{\{Z\neq d(\mathbf{X})\}}$ and weights $R/p(Z|\mathbf{X})$.

5.4 Assessing overlap using expert knowledge

As described in the introduction, a unique aspect of our study was the availability of expert knowledge for assessing overlap. Guidelines used by experts for determining treatment eligibility are provided in Table 5.2. Criteria used were based solely on clinical make up, thereby allowing for objective treatment determination. Of our 1468 total patients, expert knowledge judged 1082 (74%) patients to be eligible for CABG; 742 (50%) eligible for SVR; 598 (41%) eligible MVA, and 671 (46%) eligible for LCTx. All patients were found eligible for at least one treatment; 110 were eligible for all $M = 4$ treatments. Figure 5.1 presented earlier displays the number of eligible patients for the $\sum_{j=1}^{4}\binom{4}{j} = 15$ possible eligibility subsets.

Table 5.2: Expert knowledge used for determining treatment eligibility

Treatment	Expert Knowledge Eligibility Criteria
CABG	(a) Ischemic symptoms (angina); viable myocardium with diseased but by-passable coronary arteries. If (a) was not available, eligibility was determined using: (b) ACC/AHA guidelines for CABG based on angina and coronary artery disease.
SVR*	Anterior wall akinesia/dyskinesia; left ventricular end-diastolic diameter>6 cm.
MVA	3+/4+ mitral regurgitation (MR) present.
LCTx*	Age<70 years; NYHA functional class III/IV; creatinine level<1.7 mg$\cdot dL^{-1}$.

*Treatments where expert knowledge is considered less accurate for determining eligibility.

The guidelines listed in Table 5.2 should be considered cautiously as they only represent the current state of clinical knowledge of ischemic cardiomyopathy. There are scenarios where no "gold standard" criteria or universal rules exist for defining treatment eligibility. This is why, not surprisingly, we found instances in our data where expert decision differed from actual treatment assignment. Take LCTx for instance. Each case is painstakingly discussed at great length and multiple times as the actual decision is made to list a given patient; there are a multitude of objective, subjective, geographic, and idiosyncratic regulations that govern transplantation beyond simple clinical criteria such as age. SVR is another example. SVR is a complex non-standardized procedure that has both proponents (particularly in Europe and on the West Coast of the U.S.) and skeptics as to its efficacy. So indications for SVR procedure are controversial, even following the randomized trial of SVR vs. CABG alone (Jones et al., 2009). Among the four treatments, only eligibility of CABG and MVA can reasonably be considered as gold standard.

Let $\mathbf{E}_{n\times M} = \{E_{ij}\}$ denote the eligibility data from our $n = 1468$ patients for the $M = 4$ treatments. Here $E_{ij} \in \{0, 1\}$ denotes individual i's eligibility indicator for the jth treatment. As discussed above, there may be patients for which expert eligibility status may be suspect. Therefore, we adopt two different strategies for assessing overlap. In the first strategy, which makes minimal use of expert knowledge, we estimate the propensity score using a multiclass analysis (two different procedures are considered). Using the estimator $\hat{\mathbf{p}}_j(\mathbf{x})$ we determine the threshold C for the overlap function $\hat{\mathrm{o}}_j(\mathbf{x}; C) = \mathbf{1}_{\{\hat{\mathbf{p}}_j(\mathbf{x})>C\}}$ by making use of expert knowledge. Thus, strategy 1 uses $\mathbf{E}_{n\times M}$ in a minimally supervised way. In strategy 2, we adopt a multilabel approach and directly model $\mathbf{E}_{n\times M}$, thus making full use of expert knowledge. We estimate $e_j(\mathbf{x}_i) = \Phi\{E_j = 1|\mathbf{X} = \mathbf{x}_i\}$, the probability that a patient with feature $\mathbf{X} = \mathbf{x}_i$ is eligible for treatment j. We then use expert knowledge to determine the threshold C for the overlap $\hat{\mathrm{o}}_j(\mathbf{x}; C) = \mathbf{1}_{\{\hat{\varepsilon}_j(\mathbf{x})>C\}}$. The details are given below.

5.4.1 Random forest classification approach

Our first approach adopts strategy 1 and estimates the propensity score $p_j(\mathbf{x}) = \mathbb{P}\{Z = j|\mathbf{X} = \mathbf{x}\}$ using a multiclass random forest analysis. Specifically, a random forest comprised of classification trees (RF-C) is constructed using treatment received, for the outcome and patient covariates for the features. The propensity score is estimated using the predicted probabilities from RF-C.

5.4.2 Random forest distance approach

Our second approach also follows strategy 1 and applies RF-C. However, in place of random forest predicted probabilities to estimate $p_j(\mathbf{x}_i)$, we use a novel distance based measure. The general idea is to determine the likelihood patient i is assigned to treatment j by using a new random forest distance to measure distance of i to treatment j patients. We call this RF-D.

Typically, distance between individuals in random forests is estimated using proximity (Breiman, 2001b). The proximity $p_{i,i'}$ between individuals i and i' is defined as the forest average number of times i and i' share a terminal node. However, proximity does not define data distance accurately because it is too conservative. For example, two individuals i and i' may have terminal nodes that are side by side in a tree. Thus they differ only in their last executed tree-node split. The definition of proximity does not take this into account and by definition i and i' are considered out of proximity with a maximal distance of one. Because random forest trees are typically very deep, this means that the i and i' in our example, which are likely very close to one another, are assigned a maximal distance on the basis of a deep split that is likely to be weak and of little consequence. In contrast, if i and i' diverge quickly at a split occurring near the root node, which generally indicates a strong difference between i and i', they will also receive a maximal distance of one. Proximity

does not distinguish between these two settings.

Therefore, we propose a new distance measure for forests. We define the distance between two points i and i' as the ensemble fractional distance across the forest to the closest common ancestor of i and i'. More precisely, consider a single tree. Let d_i^A be the count of the edges from i to the closest common ancestor of i and i'. Similarly, let $d_{i'}^A$ count the edges from i' to the closest (i, i') common ancestor. Define $D_{i,i'}^A = d_i^A + d_{i'}^A$. Let d_i^R and $d_{i'}^R$ be the count of the edges from i and i' to the root node and define $D_{i,i'}^R = d_i^R + d_{i'}^R$. The distance is defined as

$$d_{i,i'} = \frac{D_{i,i'}^A}{D_{i,i'}^R}.$$

Thus, $d_{i,i'} = 0$ if i and i' are co-terminal. If the closest common ancestor of i and i' is the root node, then $d_{i,i'} = 1$. In addition, notice that $d_{i,i'} \leq 1 - p_{i,i'}$ and that distance is symmetric (diagonal elements are $d_{i,i} = 0$).

Figure 5.2 provides an illustration of our new distance measure. Terminal nodes for i and i' are highlighted in red. The common ancestor node is denoted by N_A and the root node by N_R. The distance is the ratio of the number of edges connecting the red nodes to the ancestor, N_A, to the number of edges connecting the red nodes to the root node, N_R. Thus $d_{i,i'} = (2 + 1)/(4 + 3) = 3/7$. This should be compared to the distance under proximity, which is $1 - p_{i,i'} = 1$.

The above describes distance for a single tree. The forest distance is defined as the forest averaged distance, which we denote by $\overline{d}_{i,i'}$. We define the probability of assigning i to treatment j by the closeness of i to treatment j patients,

$$\hat{\mathbf{p}}_j(\mathbf{x}_i) = \frac{\sum_{i':Z_{i'}=j}(1 - \overline{d}_{i,i'})}{\sum_{i'}(1 - \overline{d}_{i,i'})}.$$

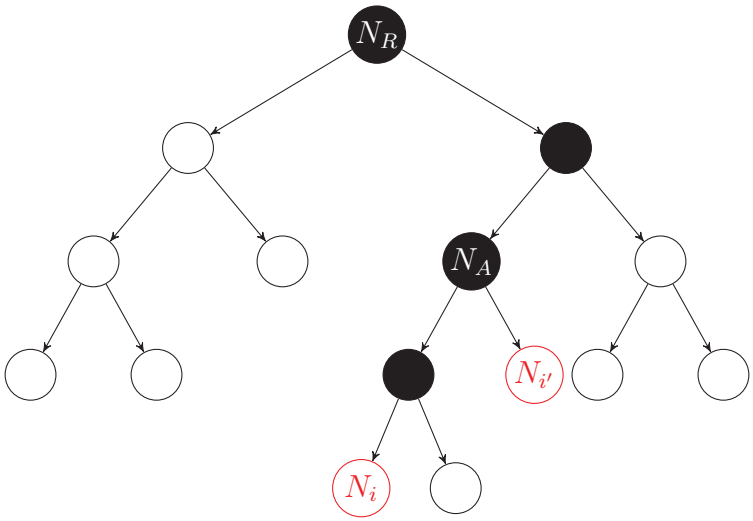

Figure 5.2: *Example illustrating random forest distance between i and i'.*

5.4.3 Multivariate random forest multilabel approach

The previous two approaches are examples of strategy 1. Our third approach adopts strategy 2 and utilizes expert knowledge by directly modeling $\mathbf{E}_{n\times M}$. Specifically, we estimate $e_j(\mathbf{x}_i) = \Phi\{E_j = 1|\mathbf{X} = \mathbf{x}_i\}$ using $\{E_{i,j}\}$ as multivariate outcomes in an M-dimensional multivariate classification analysis with $\mathbf{x}_i$ for features. Observe that the outcomes in this approach an be viewed as multilabels and the results as a multilabel analysis. For example, patient i could be eligible for two treatments CABG and SVR (the multilabel $\{$CABG, SVR$\}$ coded as $(1, 0, 1, 0)$), or patient j could be eligible for CABG, SVR, and LCTx (multilabel $\{$CABG, SVR, LCTx$\}$ coded as $(1, 0, 1, 1)$). We refer to this analysis as MRF. We note that in our implementation we use multivariate random forests as in Ishwaran and Kogalur (2017) which differs from Segal and Xiao's (2011) definition.

5.4.4 Determining the cutoff for the overlap function and validation

When determining overlap, the value for the cutoff C will likely be subjectively chosen; for example, by using a preset value such as $C = 0.05$ or $C = 0.10$. However, because expert

knowledge was available in this study, we use this data to provide an objective means for determining C.

Let $\hat{\text{o}}_j(\mathbf{x}_i; C)$ denote a procedures estimated overlap function. We define the misclassification error (ME) for the cutoff value C to be

$$\text{ME} = \frac{1}{n}\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{M}\mathbf{1}_{\{E_{ij}\neq\hat{\text{o}}_j(\mathbf{x}_i;C)\}}. \tag{5.11}$$

The cutoff value for a procedure is chosen by finding that $0 < C < 1$ which minimizes (5.11). Note that to avoid over-training, we use out-of-bag (OOB) predicted values for each of our procedures. In general, we use OOB estimated values from our forests whenever possible as OOB estimates are known to be generally more reliable and more accurate than inbag (in-sample) values (Breiman, 1998).

As we have remarked, RF-C and RF-D only minimally use eligibility data for determining the cutoff C. This is unlike MRF which uses eligibility data directly in its modeling and for determining C. This semi-supervised utilization of expert knowledge is a useful feature that can protect one from possible problems with expert knowledge databases. As discussed, here some of the treatment decisions, such as LCTx and SVR, are highly controversial and expert eligibility can disagree with actual treatment assignment.

On the other hand, CABG and MVA are treatments where expert knowledge is considered to be highly accurate. Therefore, it is interesting to develop a separate cutoff value using only data from these two treatments. Let $M' = \{j_1, j_2\}$ denote the subset of treatment groups corresponding to CABG and MVA. The cutoff using the CABG and MVA data is defined as follows:

$$C^* = \underset{0<c<1}{\text{argmin}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\sum_{j'\in M'}\mathbf{1}_{\{E_{ij'}\neq\hat{\text{o}}_{j'}(\mathbf{x}_i;c)\}}\right\}. \tag{5.12}$$

Table 5.3: *Cutoff values for estimating treatment eligibility*

Method	Cutoff Value	Misclassification Error	
		CABG MVA	All four treatments
RF-C	0.08	0.26	0.32
RF-D	0.12	0.18	0.35
MRF	0.61	0.04	0.13

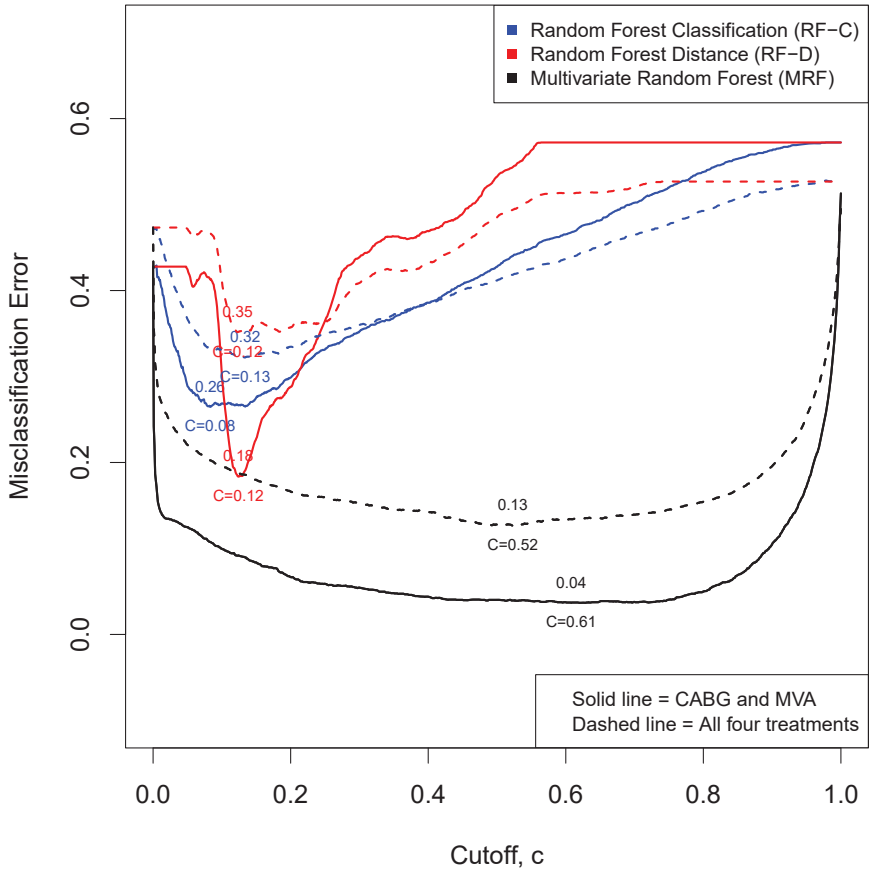

Figure 5.3: *Misclassification error as a function of the cutoff value c. The minimum point for each line is displayed above the line and its corresponding cutoff parameter $\hat{c}$ is marked below using $C = \hat{c}$.*

The results from optimizing (5.11) and (5.12) for the three methods are displayed in Table 5.3 and Figure 5.3. The dashed lines in Figure 5.3 correspond to the cutoff (5.11) based on all four treatments; the solid lines correspond to (5.12) restricted to CABG and

MVA treatments. All dashed lines are generally above their corresponding solid lines, which means that all of the three methods are able to estimate treatment eligibility for CABG and MVA more accurately than over all four treatments. For RF-C and RF-D, both the solid and dash lines reach their optimal values at relatively small cutoff values, thus indicating they can robustly estimate treatment eligibility even when expert knowledge may not be known, or even when expert knowledge is inaccurate. The new distance method, RF-D, substantially outperforms RF-C for the restricted treatment optimization (5.12), and while RF-C is slightly better than RF-D over all four treatments, we generally prefer RF-D. Also, it is not surprising to find that MRF, which was trained on the expert data, is clearly the best performer. This is especially true for the CABG/MVA treatment group (solid black line).

5.4.5 Robustness

All random forests calculations were based on default tuning parameters using the R-package `randomForestSRC` (Ishwaran and Kogalur, 2017). In particular, default nodesize parameter values of 1 and 3 were used for RF-C/RF-D and MRF calculations, respectively. To assess robustness of cutoff values to tuning values, we recalculated the cutoff C^* of (5.12) under different nodesize values. The results are displayed in Figure 5.4 (a). While all methods were generally found robust to nodesize, RF-D was found to be especially robust.

We also assessed robustness of estimated overlap indicator functions when the number of treatment options was varied. For example, if we only consider CABG and SVR, how well does each procedure perform in estimating the eligibility? That is, rather than using information from all 4 treatments, what happens if we only use CABG and SVR information for the forests defined in Sections 5.1, 5.2, and 5.3? Using the resulting C^*, what is the concordance between the estimated eligibility and the expert knowledge in this case?

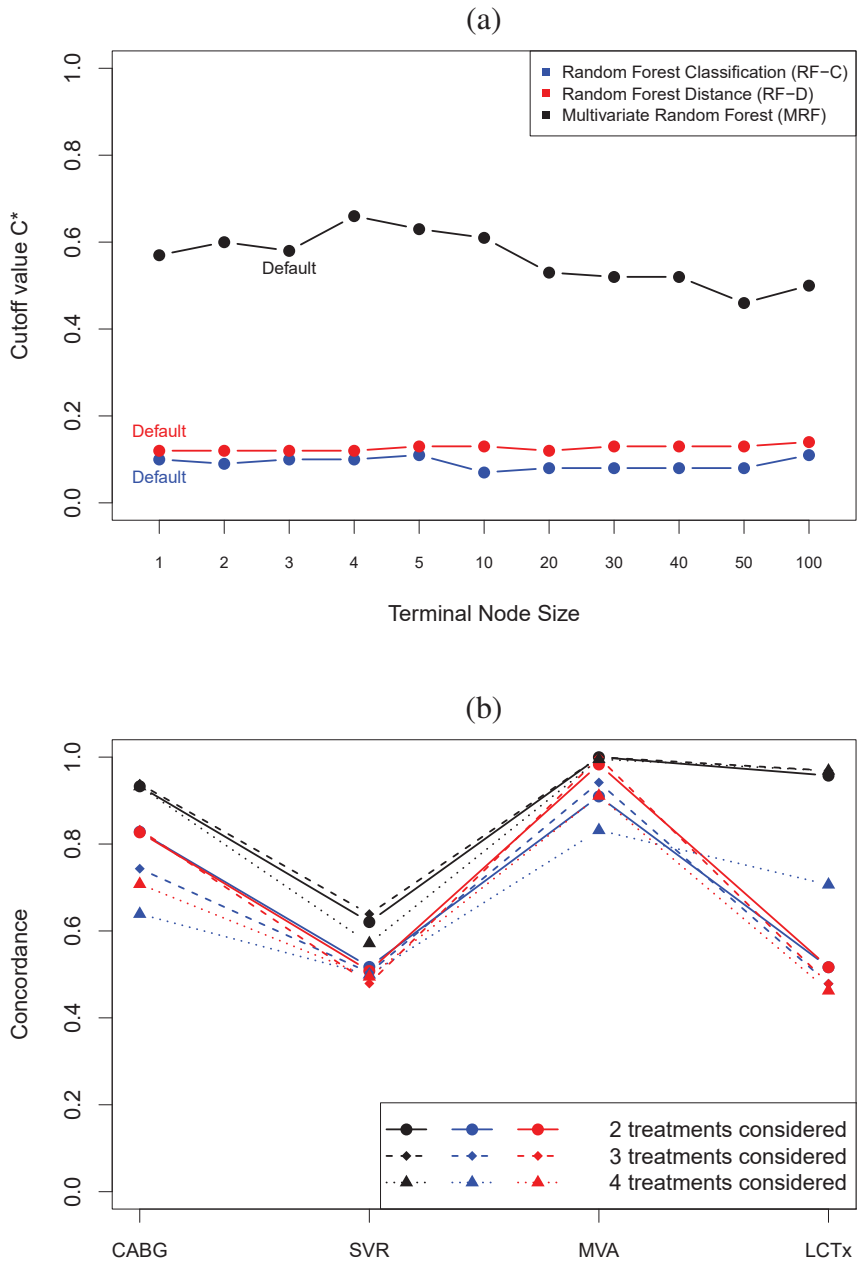

Figure 5.4: (a) Cutoff value C^* as function of random forest terminal node size; (b) OOB concordance between estimated overlap indicators and expert knowledge under different number of treatments. Subpanel (a) demonstrates general robustness to nodesize. Subpanel (b) shows that concordance for a given treatment is generally robust to the number of treatments for MRF and RF-D but less so for RF-C. Definition for line types are given in the legend; colors used are the same as the legend in panel (a).

Figure 5.4 (b) displays the concordance for each treatment under all possible 2 treatment, 3 treatment, and 4 treatment scenarios. Ideally, we would like the concordance for a treatment to remain the same regardless whether 2, 3, or 4 treatments were used. The results show that MRF and RF-D are highly robust to the number of treatments, but RF-C less so. In particular, RF-D has a higher concordance than RF-C for CABG and MVA. Because this expert knowledge is considered accurate, this provides strong evidence of superiority of RF-D to RF-C.

5.5 Counterfactual analysis using random survival forests

As we have discussed, our strategy is to estimate the survival function $S(t|\mathbf{x}, Z)$ using random survival forests (Ishwaran et al., 2008). Specifically we use what we call random survival forests virtual twins interactions (RSF-VT-I), named after the virtual twins approach of Foster et al. (2011). Virtual twins was originally introduced in the context of randomized studies, however Lu et al. (2018) showed that virtual twins could be applied to observational data and that treatment effect estimators for continuous outcomes from virtual twins could be improved by adding all possible interactions between the treatment variable Z and covariates $\mathbf{X}$ to the design matrix. The extension of this method to the survival setting is what we call RSF-VT-I.

To obtain RSF-VT-I counter-factual estimates, we run random survival forests, but where independent variables are taken to be the original features as well as all $(Z, \mathbf{X})$ interactions. To obtain a counterfactual estimate of $\tau_{j,k}(t, \mathbf{x}_i)$, we create a virtual twin data point, similar in all regards to i's original data $(\mathbf{x}_i, Z_i)$, but with the observed treatment replaced with a counterfactual treatment. Specifically, suppose that $Z_i = j$. We calculate the OOB estimated survival value $\hat{S}^*(t|\mathbf{x}_i, Z_i = j)$ based on i's original (unaltered) data. We then obtain i's counterfactual estimate defined as $\hat{S}(t|\mathbf{x}_i, Z = k)$ by using i's original

$\mathbf{x}_i$ feature but with i's treatment altered to equal $Z = k$. The counterfactual ITE estimate is defined as

$$\hat{\tau}^*_{j,k}(t, \mathbf{x}_i) = \psi\big(\hat{S}^*(t|\mathbf{x}_i, Z = Z_i = j), \hat{S}(t|\mathbf{x}_i, Z = k)\big).$$

In a similar fashion if $Z_i = k$, the counterfactual estimate is

$$\hat{\tau}^*_{j,k}(t, \mathbf{x}_i) = \psi\big(\hat{S}(t|\mathbf{x}_i, Z = j), \hat{S}^*(t|\mathbf{x}_i, Z = Z_i = k)\big).$$

Useful measures of treatment effectiveness derived from the ITE are the ATE at time t (5.8) and the ATE before time t_0 (5.9). These values are estimated by using ITE OOB estimates as follows:

$$\hat{\tau}^*_{j,k}([0, t_0]) = \int_0^{t_0} \hat{\tau}^*_{j,k}(t)\, dt \tag{5.13}$$

where

$$\hat{\tau}^*_{j,k}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\hat{\mathrm{o}}^*_j(\mathbf{x}_i;C)=1, \hat{\mathrm{o}}^*_k(\mathbf{x}_i;C)=1\}} \hat{\tau}^*_{j,k}(t, \mathbf{x}_i)}{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\hat{\mathrm{o}}^*_j(\mathbf{x}_i;C)=1, \hat{\mathrm{o}}^*_k(\mathbf{x}_i;C)=1\}}}$$

and $\hat{\mathrm{o}}^*_l(\mathbf{x}_i; C)$ are OOB estimates of the overlap function.

Another important concept for treatment decision support is treatment effect on the treated.

Definition 5.5.1. The average treatment effect on the treated (ATT) at time t for the treated j, for treatment j over treatment k, is

$$\tau_{\textcircled{j}k}(t) = \mathbb{E}\Big[\tau_{j,k}(t, \mathbf{X})\Big| Z = j,\, o_j(\mathbf{X}) = 1, o_k(\mathbf{X}) = 1\Big].$$

Likewise, the ATT for the treated k, for treatment j over k, is

$$\tau_{j\textcircled{k}}(t) = \mathbb{E}\Big[\tau_{j,k}(t, \mathbf{X})\Big| Z = k,\, o_j(\mathbf{X}) = 1, o_k(\mathbf{X}) = 1\Big].$$

Furthermore, define the ATT before time t_0 as

$$\tau_{\textcircled{j}k}([0,t_0]) = \int_0^{t_0} \tau_{\textcircled{j}k}(t)\,dt, \quad \tau_{j\textcircled{k}}([0,t_0]) = \int_0^{t_0} \tau_{j\textcircled{k}}(t)\,dt.$$

We estimate these quantities using OOB values. For example, to estimate $\tau_{\textcircled{j}k}(t)$, we use

$$\hat{\tau}^*_{\textcircled{j}k}(t) = \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Z_i=j\}}\mathbf{1}_{\{\hat{\mathrm{o}}^*_j(\mathbf{x}_i;C)=1,\hat{\mathrm{o}}^*_k(\mathbf{x}_i;C)=1\}}\hat{\tau}^*_{j,k}(t,\mathbf{x}_i)}{\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Z_i=j\}}\mathbf{1}_{\{\hat{\mathrm{o}}^*_j(\mathbf{x}_i;C)=1,\hat{\mathrm{o}}^*_k(\mathbf{x}_i;C)=1\}}}. \tag{5.14}$$

5.5.1 Results

Figure 5.5 displays ATE and ATT estimates (5.13) and (5.14) from the RSF-VT-I analysis. All forest estimates were calculated using the R-package `randomForestSRC` (Ishwaran and Kogalur, 2017) which provides a general implementation of Breiman random forests (Breiman, 2001b). Complete overlap was defined using the cutoff criteria C obtained by optimizing (5.12) restricted to the CABG/MVA treatment groups. We could have optimized (5.11), but as we have discussed, expert knowledge could only be considered as gold standard for the CABG and MVA treatment groups. Furthermore, Figure 5.3 shows that the optimized C for RF-C and RF-D are nearly the same under (5.11) and (5.12), and for MRF, performance is robust to C.

ATE and ATT estimates in Figure 5.5 are displayed for eligibility defined by each of the methods, RF-C, RF-D, and MRF. All values are based on the ITE, $\tau^{(1)}_{j,k}(t,\mathbf{x})$, defined as the difference between two survival curves (5.6). Thick lines display estimates using MRF eligibility, thin dashed lines with circles are based on RF-C, and thick dashed lines with triangles are derived from RF-D eligibility. Generally, RF-D and MRF values agree, while RF-C differs substantially for certain treatment comparisons. This confirms our analysis from the previous section which suggested RF-D to be superior to RF-C. Also, because RF-D generally matches MRF, we can trust that MRF is not biased by inaccurate eligibility

data.

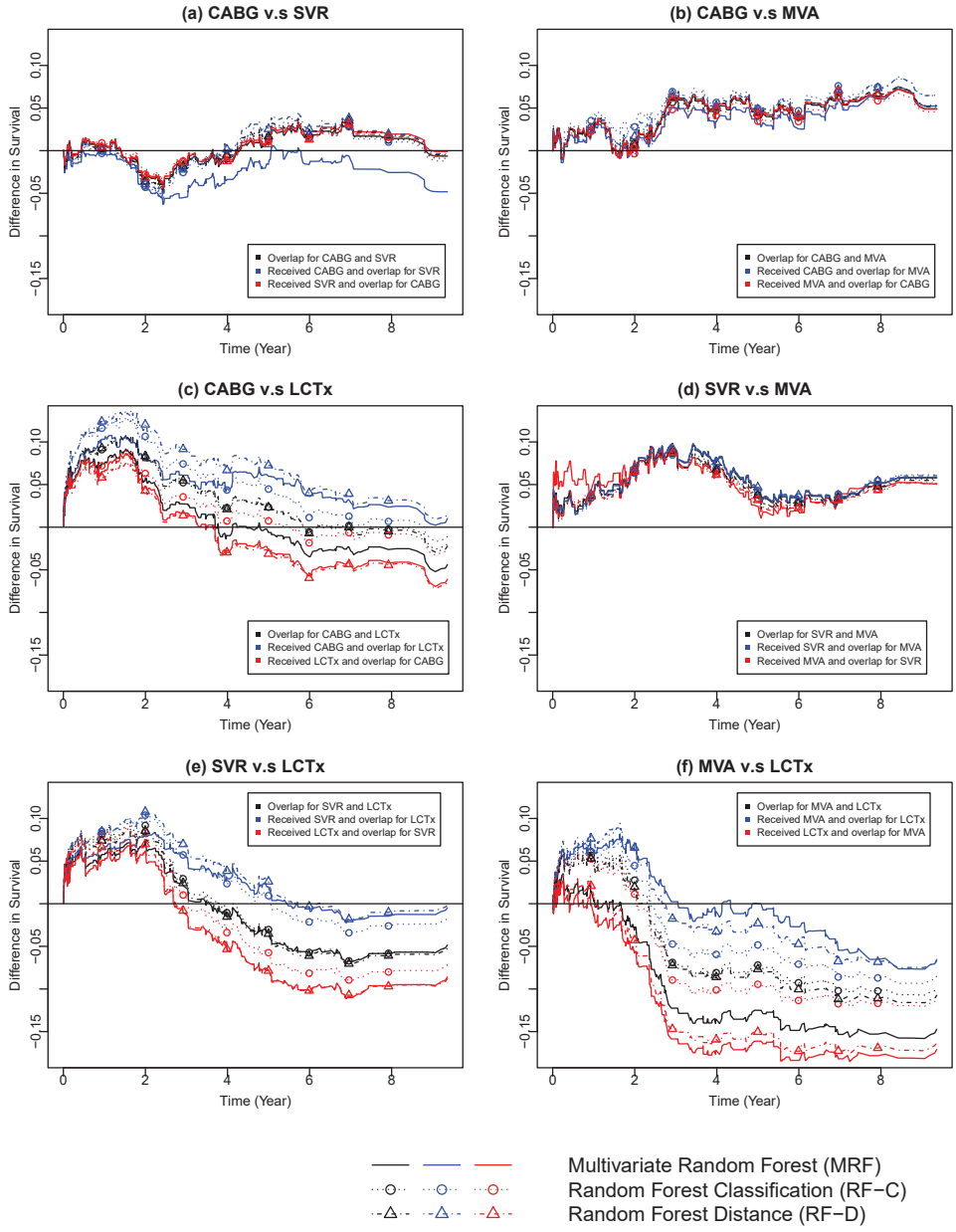

Figure 5.5: *ATE (5.13) and ATT (5.14) estimated values where overlap was determined using the three methods RF-C, RF-D, and MRF. Each subfigure title indicates the pairwise comparison for treatment j versus k. Black lines are ATE values $\hat{\tau}^*_{j,k}(t)$; blue and red lines are ATT values, where blue is $\hat{\tau}^*_{\bigcirc k}(t)$, where j is the treated group, while red is $\hat{\tau}^*_{j\bigcirc}(t)$, where k is the treated group.*

Interpreting the ATE and ATT

To understand Figure 5.5, and what it is saying about treatment effectiveness, it is helpful to focus on one of the subfigures. Take panel (a) for instance, which displays the treatment effect for CABG compared to SVR. Our chosen measure of treatment effectiveness is $\tau_{j,k}^{(1)}(t, \mathbf{x})$, which means the values display the difference between the survival function for CABG compared to SVR. Values above zero signify when CABG is the preferred treatment. The black line is the ATE and represents the average treatment effect for patients eligible for both treatments, the blue line is the ATT for patients who received CABG and were eligible for SVR, while the red line is the ATT for patients who received SVR and were eligible for CABG. What we see is that regardless of the treatment assignment, the ATT generally tracks the horizontal zero line, with low variability. This is in contrast to other subfigures, where the magnitude of treatment effect is generally in the range of 5-10%, or higher. Treatment effect for CABG and SVR in subfigure (a) has the lowest magnitude, which reflects the controversy of SVR (Jones et al., 2009).

In general, the dynamics of treatment effect over the various subpanels is interesting. For ease of description, let us focus on interpreting values based on MRF eligibility (as discussed above, we can be confident in this eligibility classification). We observe that all ATT and ATE lines pass zero at some time point except for subfigure (d). When the various lines pass zero, the two survival curves under the two treatments cross each other, which generally implies that one treatment is beneficial in the short term, while the other treatment is beneficial in the long term. Take subfigure (f) for instance. The blue line, which represents patients who received MVA, crosses zero at around the third year, meaning that MVA is only beneficial in the first three years. This suggests, that if possible, LCTx should always be recommended over MVA. However, it may be impossible to give LCTx to all these patients due to the limited number of transplant donors. This seems to be the case, because the current recommendation of LCTx (red line) has already selected many suitable

patients, who greatly benefit from LCTx (red line is below zero for almost all t).

Average number of months alive

The areas under the black, blue, and red lines of Figure 5.5 equal the ATE and ATT before t_0 (the maximum observed follow-up time), and thus represent the difference in number of years alive before t_0 (here $t_0 = 9.36$ years). We describe these values below, but first, because the notation for the ATE and ATT before t_0 will be awkward to work with, we introduce the following simplified notation:

$$\begin{aligned}
&\text{ATE}^o_{jk} = \tau_{j,k}([0, t_0]) && \text{ATE before } t_0 \text{ (black line)}\\
&\text{ATT}^o_{jk} = \tau_{\textcircled{j}k}([0, t_0]) && \text{ATT before } t_0 \text{ where } j \text{ is the treated (blue line)}\\
&\text{ATT}^o_{kj} = \tau_{j\textcircled{k}}([0, t_0]) && \text{ATT before } t_0 \text{ where } k \text{ is the treated (red line).}
\end{aligned}$$

Table 5.4 summarizes these values (for convenience, values have been converted to patient months). Values for ATE^o_{jk} are given for eligibility determined using MRF, RF-C, and RF-D. For simplicity, ATT^o_{jk} values are provided only for MRF. For ATE^o_{jk}, we find values calculated using MRF and RF-D are closer to those than RF-C. Once again, this reflects a lack of accuracy of RF-C. To interpret Table 5.4, consider the first row for ATT^o_{jk} and ATT^o_{kj}, corresponding to CABG versus SVR (panel (a) of Figure 5.5). For patients who received CABG (ATT^o_{jk}), on average they lose 2.67 months (standard error, 3.74) than if they had received SVR; such a loss is not statistically significant. For patients receiving SVR (ATT^o_{kj}), they gain 0.7 months of life (standard error, 0.93). This is also not statistically significant.

Generally speaking, when the blue line is above zero and the red line is below zero in Figure 5.5, and when ATT^o_{jk} is positive and ATT^o_{kj} is negative in Table 5.4, the current treatment decision is supported by evidence, because on average, patients received the

Table 5.4: Difference in number of months alive before maximum follow-up time, $t_0 = 9.36$ years.

Treatment j vs. k	ATE^o_{jk}			ATT^o_{jk}		ATT^o_{kj}	
	MRF	RF-C	RF-D	Mean	SE	Mean	SE
(a) CABG vs. SVR	0.31	0.29	0.60	-2.67	3.74	0.70	0.93
(b) CABG vs. MVA	4.88	5.06	5.21	4.20	2.89	5.02	1.55
(c) CABG vs. LCTx	0.85	3.67	3.50	5.85	2.26	-0.74	1.11
(d) SVR vs. MVA	5.95	5.49	5.47	5.97	1.41	5.70	5.61
(e) SVR vs. LCTx	-1.40	-0.55	-1.08	2.57	1.52	-4.81	1.53
(f) MVA vs. LCTx	-11.80	-6.08	-6.81	-0.84	2.62	-14.97	1.36

treatment most beneficial to them. For example, this is what happens in subpanels (c) and (e).

Treatment effect heterogeneity

Standard errors in Table 5.4 were calculated using subsampling. We used a 25% subsampling rate with 1000 replications. Subsampling can also be used to calculate confidence regions for ITE. Confidence intervals for ITE at the fifth year are provided in Figure 5.9. Figure 5.5 only gives us general treatment information without details of treatment effect heterogeneity. Treatment effect heterogeneity can be tested through regression models. Previous studies typically regress the outcome on the covariates using experimental data. Crump et al. (2008) use regression within each treatment group and test whether treatment effect conditional on covariates is identical for all subpopulations; that is, they test the null hypothesis that there is no heterogeneity in average treatment effects among subpopulations defined by covariates. Imai and Ratkovic (2013) detect treatment effect heterogeneity through detecting interaction terms between treatment variable and pre-treatment covariates. We regressed the ITE on the pre-treatment covariates and used subsampling to derive confidence intervals for each coefficient. The results are shown in Figures S2 and S3 of Supplementary Materials. Significant coefficients suggest evidence of heterogeneity. All 6

treatment comparisons show evidence of heterogeneity.

Subgroup analysis

Subgroup analysis is needed when there is treatment effect heterogeneity. The goal of subgroup analysis is to find features related to treatment effect to better guide patient treatment decision making. We use patients who received either treatment j or k and who were eligible for both treatments, and fit a bump hunting model (Friedman and Fisher, 1999; Duong, 2015). To improve efficiency of the algorithm, we only used variables found important by using random forest variable selection. Variables were identified using variable importance using a random forest regression model in which the estimated ITE was used for the outcome and all pre-treatment covariates as independent variables. Results from the subgroup analysis are provided in Table 5.5.

Future researchers can use these criteria to conduct randomized controlled trial as further validation of subgroup treatment effect. We did not find subgroups with a positive treatment effect of MVA. Since Table 5.4 shows a large negative treatment effect of MVA, we conclude that fewer patients should be assigned to MVA.

Optimal individual treatment decisions

The previous analysis focuses on average effect. Here we now consider patients on a case-by-case basis to determine if they received optimal treatment. For each patient i, with covariate $\mathbf{x}_i$, we calculated RMST before t_0,

$$\hat{R}_{i,j} = \int_0^{t_0} \hat{S}\big(t|\mathbf{x}_i, Z = j\big)\, dt,$$

where $\hat{S}$ was the estimated survival function defined as the forest OOB estimate when $Z = Z_i$ was actual treatment assignment, and the forest predicted estimate when $Z \neq Z_i$ was a

Table 5.5: Subgroup detection using bump hunting after variable selection. CATE^o_{jk} equals the conditional ATE before t_0, conditioned on subgroup criteria.

Treatment j vs. k	Subgroup	CATE^o_{jk}/ATE^o_{jk}	Size/Total	% in j	% in k
CABG vs. SVR	BSA>2.23	-4.08/0.31	44/246	28.57	16.51
CABG vs. SVR	Regurgitation Grade>0	-7.26/0.31	31/246	10.71	12.84
CABG vs. LCTx	Blood Urea Nitrogen<30 Creatinine<1.8 BMI>27.04 GFR>44.75	5.31/0.85	125/406	59.18	21.75
SVR vs. LCTx	Blood Urea Nitrogen<25 LDL<133.31 BSA>1.83 BMI>27.77 55.29<GFR<120.80	7.66/-1.40	60/292	30.37	12.10

BSA=body surface aera (m^2); BMI=body mass index; GFR=glomerular filtration rate; LDL=low-density lipoprotein cholesterol.

counterfactual assignment. The optimal treatment $\hat{d}^{\text{opt}}(\mathbf{x}_i)$ was defined as the treatment j with maximal $\hat{R}_{i,j}$ over those treatments for which i satisfied overlap (defined using MRF).

Figure 5.6 displays the numbers for patients who correctly received optimal therapy and those who would have benefited from an alternate optimal therapy for which they satisfied overlap. Years gained under alternate optimized therapy are displayed as "gain" (specifically, gain equals $\hat{R}_{i,j_i} - \hat{R}_{i,Z_i}$ where j_i is the optimal therapy for i). For example, for patients receiving CABG (top left plot), only 21% would have benefited from another therapy, thus showing CABG is generally being assigned correctly. An example where treatment assignment is generally poor is MVA (bottom left plot), where 64% of patients receive suboptimal treatment. For a large fraction of these ($n = 95$), the optimal treatment is CABG, with an average gain of more than 1 year of restricted life. The superiority of CABG over MVA agrees with our previous ATE and ATT analysis, Figure 5.5(b) and Table 5.4 line (b), although the magnitude of effect of 1 restricted year gained is much higher than suggested by the ATE/ATT analysis; thus demonstrating the importance of evaluating

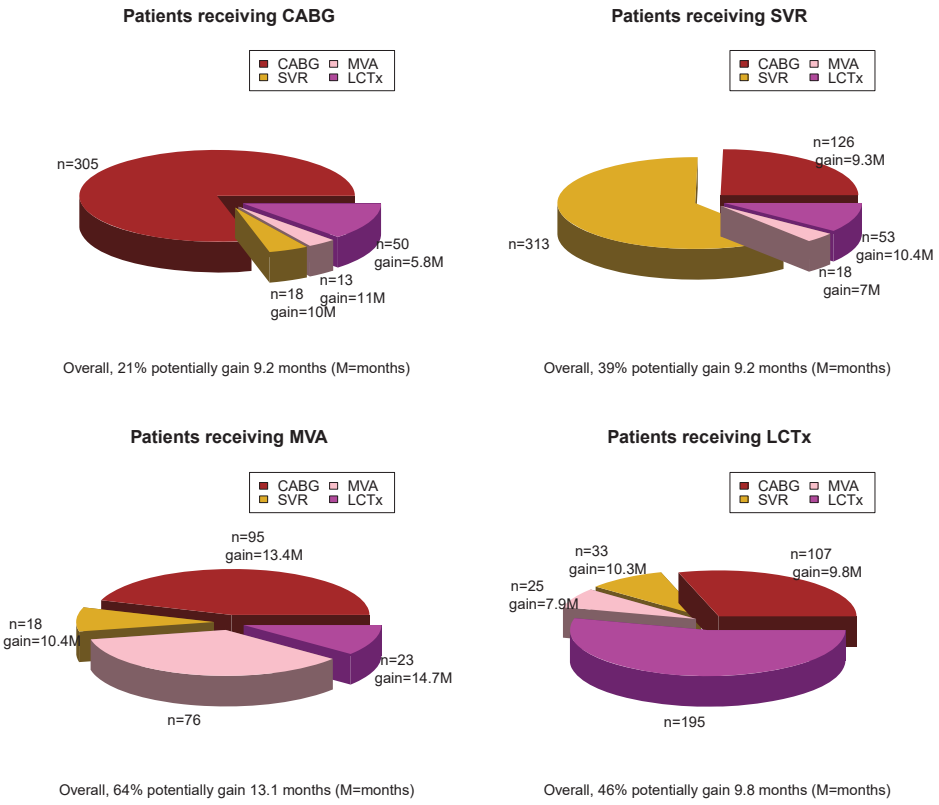

Figure 5.6: *Identifying patients who received optimal treatment and those who did not. Optimal therapy is defined as treatment maximizing restricted mean survival time (RMST). Pie charts display gain in months for alternative optimized therapies and their respective sample sizes. If optimized treatment is the assigned treatment, gain is defined as zero.*

gain under individual optimized overlap therapy. Another example of a problematic therapy is LCTx (bottom right plot), where 46% of patients received a non-optimal treatment. A sizeable fraction of these ($n = 107$) would have been better off with CABG, with an average restricted life gain of almost 10 months.

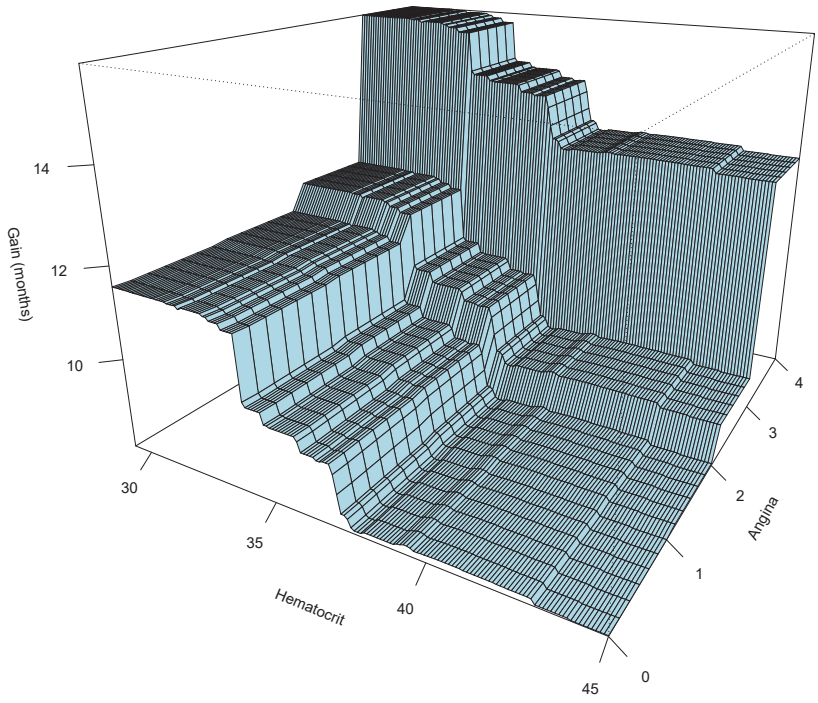

Figure 5.7: *Gain in months for patients who received SVR but where optimal therapy was CABG. Gain is plotted against hematocrit level and angina pectoris grade.*

It is also interesting to consider patients who received SVR. Although a reasonably large fraction receive the correct therapy (61%), there is a sizeable population ($n = 126$) that would have benefited from CABG. This is interesting as SVR versus CABG is controversial. Looking at those patients that would have benefited by CABG, we find that their improved survival is strongly related to preoperative grade of angina pectoris (Canadian Cardiovascular Society grading scale) and preoperative hematocrit. In particular, there are sizeable gains for CABG when hematocrit levels are low and when there is strong evidence of angina (Figure 5.7). We believe these findings are important and useful to treatment

management for ischemic cardiomyopathy.

Treatment decisions

Figure 5.8 summarizes our methodology and conclusions for treatment decision making. When ATT lines (blue and red lines) merge in Figure 5.5, this indicates that current recommendations for therapy do not alter treatment effect. This occurs for subfigures (a), (b), and (d). For (a) CABG vs. SVR, because Figures 5.9 and 5.10 identify heterogeneity, a subgroup analysis was conducted. This showed SVR benefits patients with larger body surface area and with higher aortic valve regurgitation grade. For these patients, treatment decision should be modified. Regardless of whether ATT blue and red lines in Figure 5.5 merge, ideally we would want the blue line to be on top of the red line and above zero. Otherwise, current treatment decision is not supported by evidence. For example, subfigure (a) suggests that current treatment decision for SVR is slightly worse than random assignment. For (b) CABG vs. MVA, (d) SVR vs. MVA, and (f) MVA vs. LCTx, Table 5.4 and Figure 5.9 demonstrate that most patients do not benefit from MVA. We conclude that current treatment decision should be modified: in general, fewer patients should be assigned to MVA. When blue and red ATT lines do not merge in Figure 5.5, with the blue line above, and when ATT_j is positive and ATT_k is negative in Table 5.4, patients are benefiting from their treatment and the current treatment decision is supported by evidence. For (c) CABG vs. LCTx, and (e) SVR vs. LCTx, current treatment decision is supported by evidence. However, Figures 5.9, 5.10, and 5.11 indicate presence of heterogeneity for these groups. Table 5.5 shows that the Conditional Average Treatment Effect (CATE) for certain subgroups differ from their ATE values of Table 5.4, thus showing that current treatment decision can be improved.

5.6 Concluding remarks

Estimation of treatment overlap and individual treatment effects play an essential role in causal inference and therapy management, such as in the case of ischemic cardiomyopathy. A contribution of this paper is to offer estimation methods for assessing treatment overlap under the scenario that some treatments may have either gold standard expert knowledge, or controversial knowledge for judging eligibility. We described a novel fully supervised multilabel procedure and a novel distance based multiclass semi-supervised procedure as new tools to complement and improve upon the conventional multiclass approach of assessing overlap using estimated treatment assignment. Subject matter knowledge available in observational data studies should be viewed as a powerful tool that can be incorporated into causal analyses and our novel strategies for systematically utilizing such information may prove highly useful. Another contribution is our direct outcome-regression approach to estimating the individual treatment effect (ITE) and the individualized treatment rule (ITR). Our approach leverages powerful machine learning methods such as random forests and random survival forests to provide a direct approach to this challenging issue. Another unique aspect of our work is the ability to view treatment effectiveness as a dynamic process. In the case of survival data this provides an important tool for assessing and understanding treatment differences, especially when multiple treatments are at play. Applying our methodology to a large observational survival data set, we studied the current standards for treatment management of ischemic cardiomyopathy, and found standard treatment management was not always optimal, and in some cases, even suboptimal. We believe our findings of non-optimal therapy management are not unique to ischemic cardiomyopathy and that this problem may be more common than generally appreciated. Our methodology is general and easily applied with available public software and can be instrumental in studying this issue.

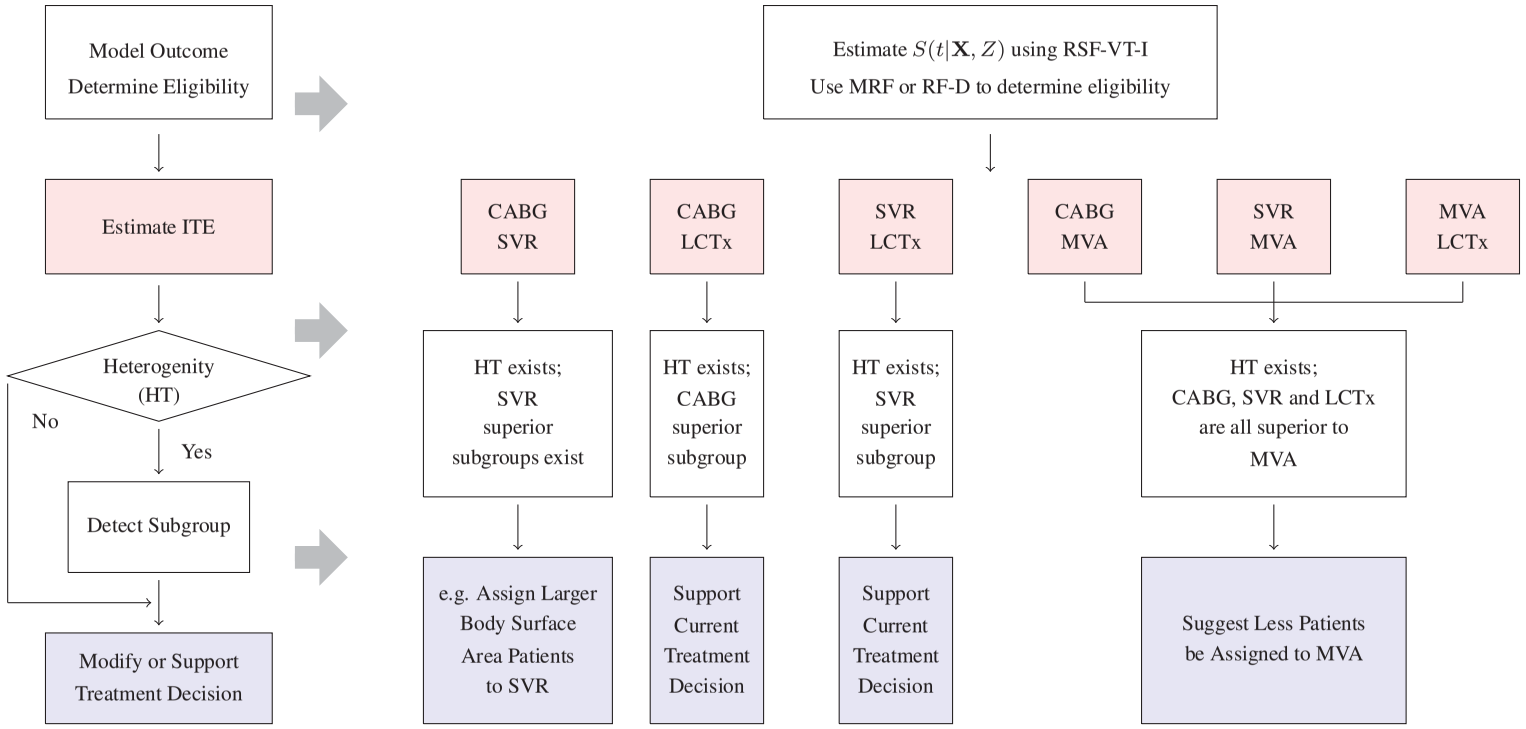

Figure 5.8: Paradigm for Individual Causal Inference and Treatment Decision Making for Ischemic Cardiomyopathy.

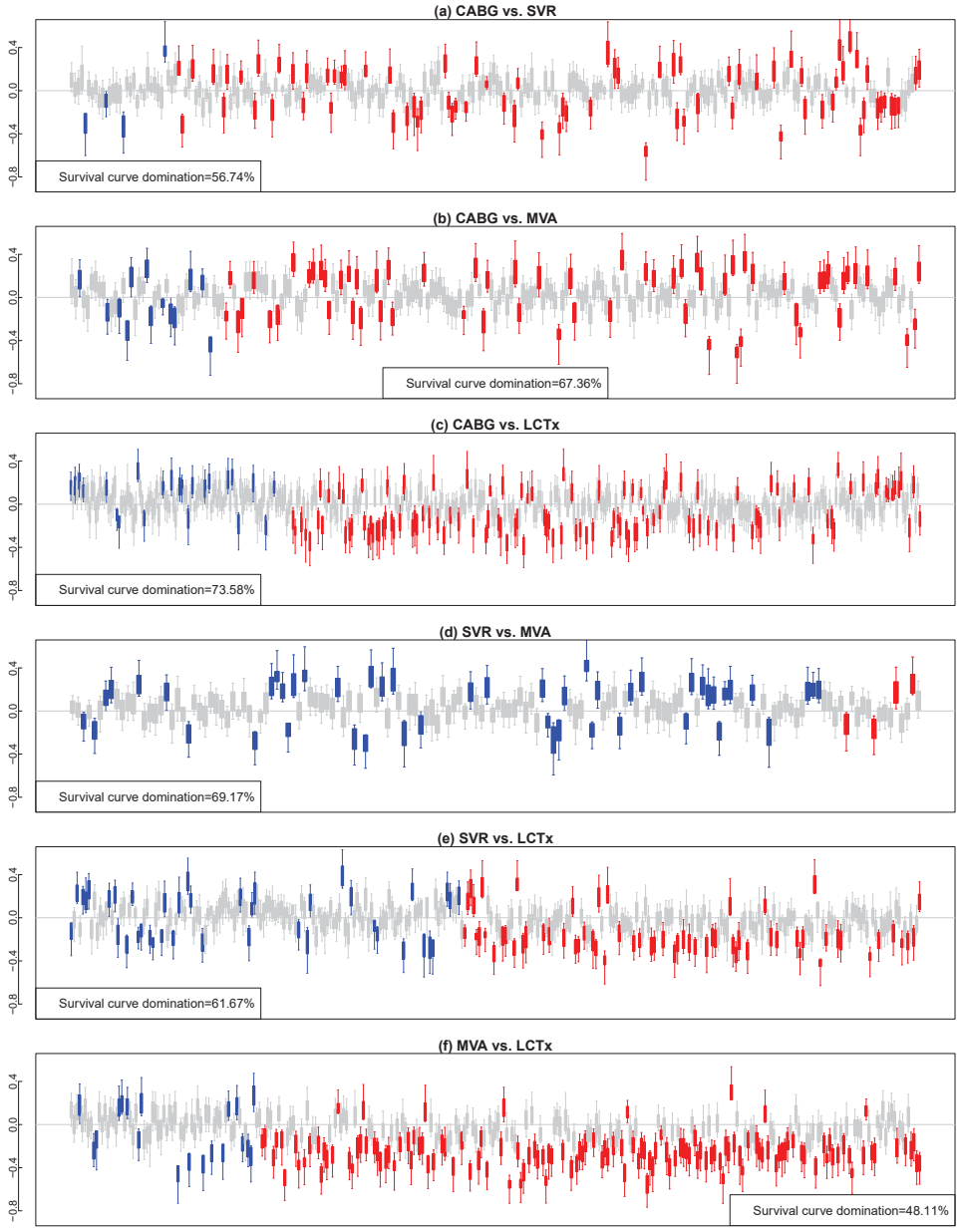

Figure 5.9: Confidence intervals for individual treatment effects (5.5) at $t = 5$ years. Each subfigure indicates a pairwise comparison for treatment j versus k. Red and blue indicate patients with significant treatment effect (p-value $< .05$), where blue are from treatment j group and red are from treatment group k. Thus, blue and red boxes correspond to some of the patients from blue and red lines in Figure 5.5. Survival curve domination is defined as $\tau^{(2)}_{j,k}(t)$.

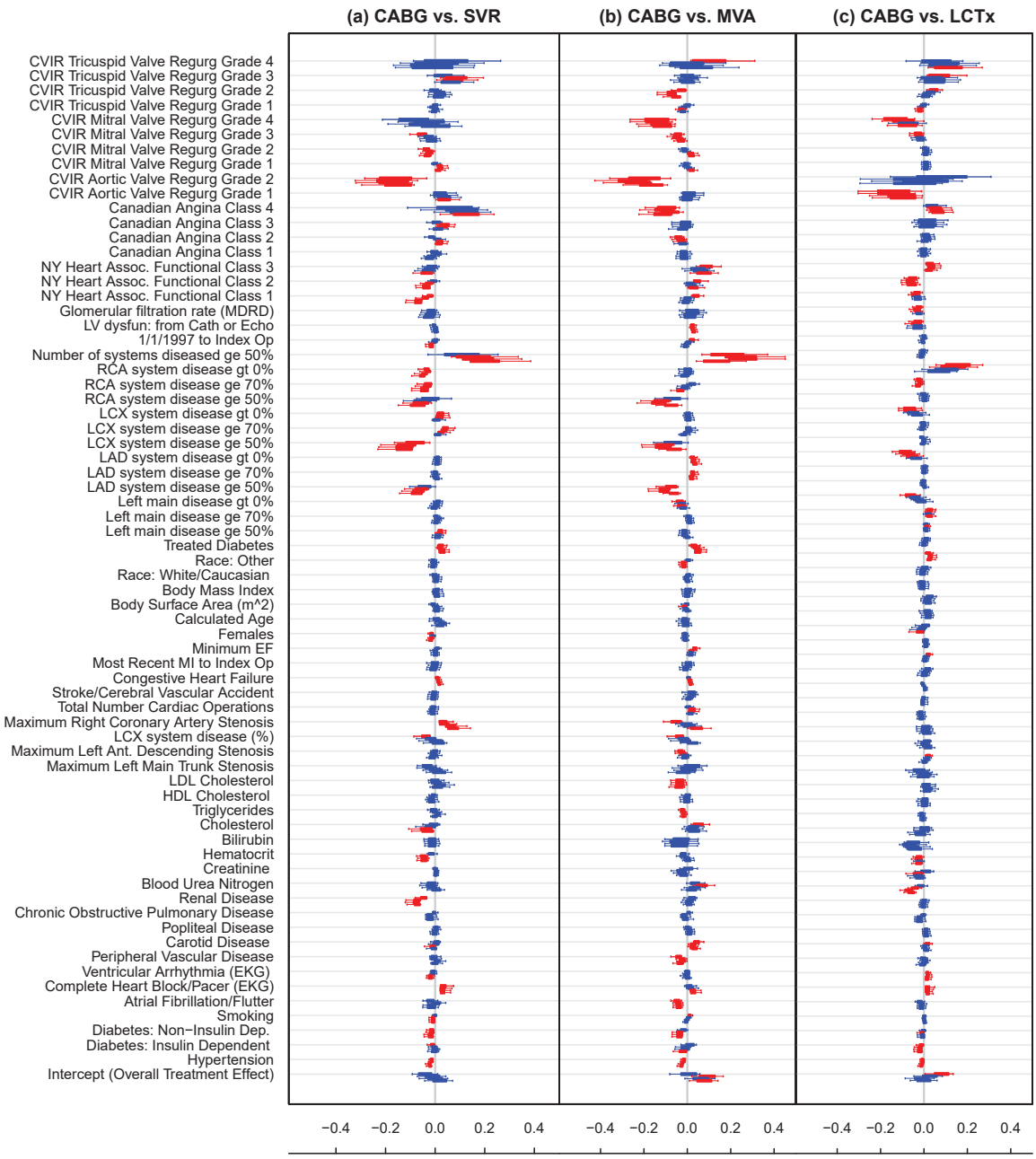

Figure 5.10: Confidence intervals for coefficients from linear regression of estimated individual treatment effect for pairwise comparison of treatment j versus k. Regression included patients receiving either treatment j or k and who were eligible for both treatments. For each variable, there are 4 boxplots corresponding to coefficients for that variable for $t = 2, 4, 6, 8$ (years).

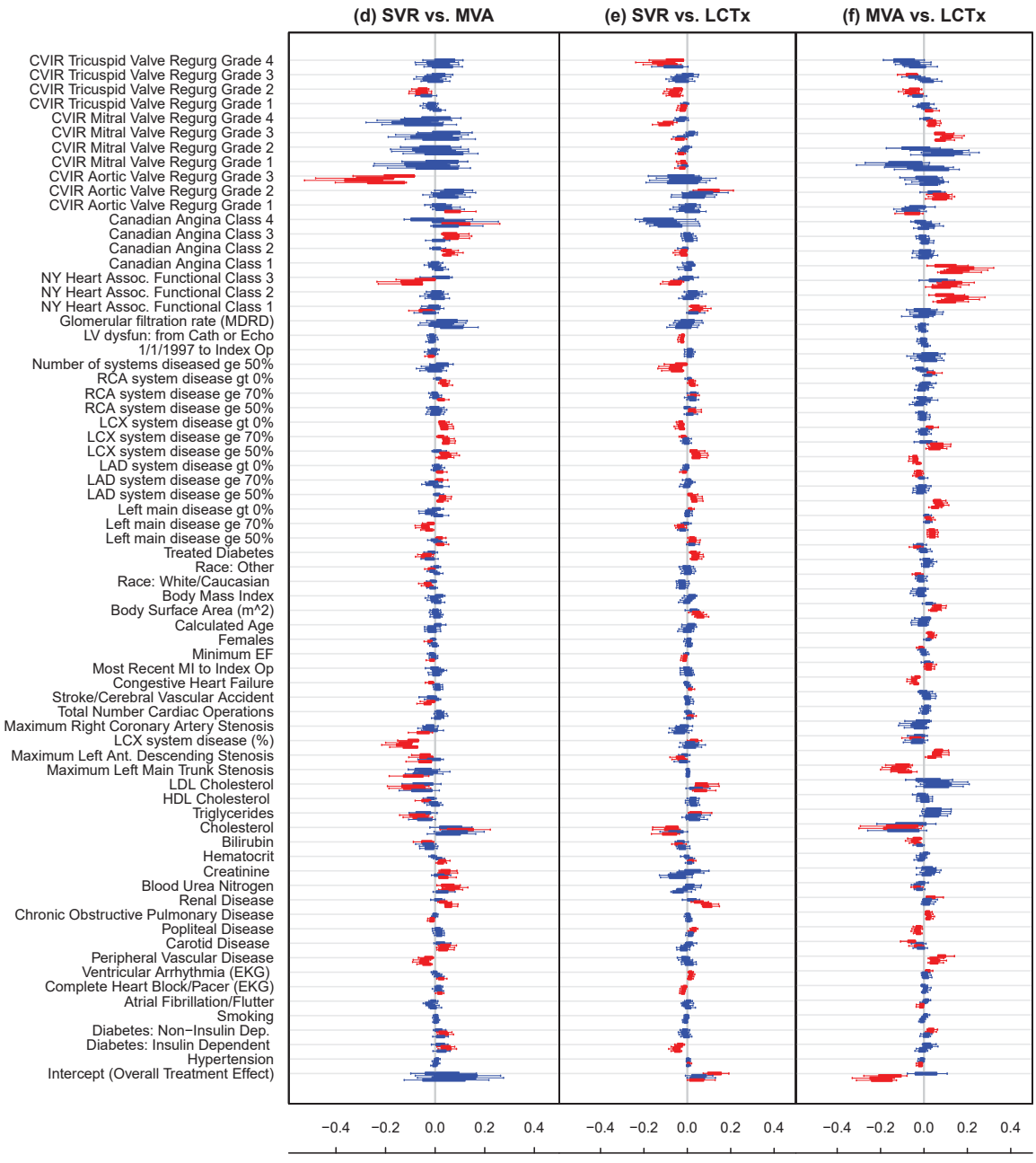

Figure 5.11: Linear regression results continued from Figure 5.10.

Reference

Alexander, G. E., Chen, K., Pietrini, P., Rapoport, S. I., and Reiman, E. M. (2002). Longitudinal PET evaluation of cerebral metabolic decline in dementia: a potential outcome measure in Alzheimer's disease treatment studies. *American Journal of Psychiatry* **159** 738–745.

Andersen, P. K. (2013). Decomposition of number of life years lost according to causes of death. *Stat. Med.* **32** 5278–5285.

Andersen, P.K., Hansen, M.G., and Klein, J.P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Anal.* **10** 335–350.

Arratia, R. and Gordon, L. (1989). Tutorial on large deviations for the binomial distribution. *Bulletin of Mathematical Biology* **51** 125–131.

Austin, P. C. (2011). A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research* **46** 119–151.

Bai, X., Tsiatis, A.A., Lu, W., and Song, R. (2017). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Anal.* **23** 585–604.

Bhattacharyya, N. (2003). A matched survival analysis for squamous cell carcinoma of the head and neck in the elderly. *The Laryngoscope* **113** 368–372.

Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research* **9** 2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST* **25** 197–227.

Breiman L., Friedman J.H., Olshen R.A., and Stone C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, California, 1984.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

Breiman, L. (1998). Out-of-bag estimation. Technical report, Statistics Dept., University of California at Berkeley, CA.

Breiman, L. (2000) Some infinite theory for predictor ensembles. *Technical Report 577, Statistics Department*, UC Berkeley. http://www.stat.berkeley.edu/breiman.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* **16** 199–231.

Breiman, L. (2001) Random forests. *Machine Learning* **45** 5–32.

Chernoff, H. (1952). Asymptotic efficiency for tests based on the sum of observations. *Ann. Math. Stat.* **23** 493-507.

Chipman, H. and McCulloch, R. (2016). *BayesTree: Bayesian Additive Regression Trees*. R package version 0.3-1.4.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298.

Crump, R.K., Hotz, V.J., Imbens, G.W. and Mitnik, O.A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* **90** 389–405.

Crump, R.K., Hotz, V.J., Imbens, G.W. and Mitnik, O.A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199.

Dasgupta, A., Szymczak, S., Moore, J. H., Bailey-Wilson, J. E., and Malley, J. D. (2014). Risk estimation using probability machines. *BioData mining* **7** 1–17.

Dehejia, R.H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J Amer. Stat. Assoc.*, **94** 1053–1062.

Duong, T. (2015). Patient Rule Induction Method (PRIM) for bump hunting in high-dimensional data. R package version 1.0.16.

Dominici, F., Zeger, S. L., Parmigiani, G., Katz, J., and Christian, P. (2006). Estimating percentilespecific treatment effects in counterfactual models: a casestudy of micronutrient supplementation, birth weight and infant mortality., *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **55** 261–280.

Efron B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Stat. Assoc.* **78** 316–33.

Efron B. and Tibshirani R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Amer. Stat. Assoc.* **92** 548–560.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11** 89–102.

Feller, A., Grindal, T., Miratrix, L. and Page, L.C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *Ann. Appl. Stat.* **10** 1245–1285.

Fisher, R.A. *The Design of Experiments*, Macmillan. ISBN 0-02-844690-9. 1935.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880.

Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing* **9** 123–143.

Frumento, P., Mealli, F., Pacini, B. and Rubin, D.B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Amer. Statist. Assoc.* **107** 450–466.

Ghosh, D., Zhu, Y., and Coffman, D. L. (2015). Penalized regression procedures for variable selection in the potential outcomes framework. *Stat. Med.* **34** 1645–1658.

Goldberg, Y. and Kosorok, M.R. (2012). Q-learning with censored data. *Ann. Statist.* **40** 529–560.

Heckman, J.J., Ichimura, H., and Todd, P.E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review Econ. Studies* **64** 605–654.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, **66** 1017–1098.

Hernández, D., Feaster, D. J., Gooden, L., Douaihy, A., Mandler, R., Erickson, S. J., Kyle, T., Haynes, L., Schwartz, R., Das, M., et al. (2016). Self-reported HIV and HCV screening rates and serostatus among substance abuse treatment patients. *AIDS and Behavior* **20** 204–214.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20** 217–240.

Hill, J. and Su, Y. S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.* **7** 1386–1420.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *J Amer. Stat. Assoc.* **81** 945–960.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series* **1** 1–50.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *J Amer. Stat. Assoc.* **47** 663-685.

Hsich E., Gorodeski E.Z., Blackstone E.H., Ishwaran H., and Lauer M.S. (2011). Identifying important risk factors for survival in systolic heart failure patients using random survival forests. *Circ. Cardiovasc. Qual. Outcomes* **4** 39–45.

Imbens, G.W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7** 443–470.

Irwin, J.O. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *The Journal of Hygiene* **47** 188–189.

Ishwaran, H. and Kogalur, U. (2017). *Random Forests for Survival, Regression and Classification (RF-SRC).* R package version 2.5.0.

Ishwaran, H. and Lu, M. (2017). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat. Med. (in press).*

Ishwaran H., Kogalur U.B., Blackstone E.H., and Lauer M.S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860.

Ishwaran, H. and Malley, J. D. (2014). Synthetic learning machines. *BioData Mining***7** 28–40.

Jones, R.H., Velazquez, E.J., Michler, R.E., et al. (2009). Coronary bypass surgery with or without surgical ventricular reconstruction. *N Engl J Med.* **360** 1705–17.

Kern, H. L., Stuart, E. A., Hill, J. L. and Green, D. P. (2013). Assessing methods for generalizing experimental impact estimates to target samples. Technical report, Univ. South Carolina, Columbia, SC.

Kim, D.H., Uno, H., and Wei, L.J. (2017). Restricted mean survival time as a measure to interpret clinical trial results. *JAMA Cardiology* **2** 1179–1180.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review* **24** 328–338.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.

Lamont, A. E., Lyons, M., Jaki, T. F., Stuart, E., Feaster, D., Ishwaran, H., Tharmaratnam, K., and Van Horn, M. L. (2016). Identification of predicted individual treatment effects (PITE) in randomized clinical trials. *Statistical Methods in Medical Research* **0** 1–19.

Laplace, P.-S. (1814). *A Philosophical Essay on Probabilities.* English translation by Truscott, Frederick Wilson and Emory, Frederick Lincoln from the original French 6th ed. Dover Publications (New York, 1951) p.4

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Stat. Med.* **29** 337–346.

Lipkovich, I., Dmitrienko, A., Denne, J. and Enas, G. (2011). Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat. Med.* **30** 2601–2621.

Liu, K. and Meng, X.L. (2016). There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Application* **3** 79–111.

Lu M., Sadiq S., Feaster D.J. and Ishwaran H. (2018). Estimating individual treatment effect in observational data using random forest methods *Journal of Computational and Graphical Statistics,* **27** 209–219.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* **23** 2937–2960.

MacKinnon, DP. *Contrasts in multiple mediator models.* In: Rose, JS.; Chassin, L.; Presson, CC.; Sherman, SJ., editors. Multivariate Applications in Substance Use Research: New Methods for New Questions. Mahwah, NJ: Erlbaum; 2000. p. 141-60.

McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., et al. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Medicine* **32** 3388–3414.

Metsch, L. R., Feaster, D. J., Gooden, L., Schackman, B. R., Matheson, T., Das, M., Golden, M. R., Huffaker, S., Haynes, L. F., Tross, S., et al. (2013). Effect of risk-reduction counseling with rapid HIV testing on risk of acquiring sexually transmitted infections: the AWARE randomized clinical trial. *JAMA* **310** 1701–1710.

Mihaljevic, T., Lam, B. K., Rajeswaran, J., Takagaki, M. et al. (2007). Impact of mitral valve annuloplasty combined with revascularization in patients with functional ischemic mitral regurgitation. *J. Amer. College of Cardiology* **49** 2191–2201.

Moodie, E.E., Richardson, T.S., and Stephens, D.A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63** 447–455.

Morgan, S.L. and Harding, D.J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Soc. Methods Research*, **35** 3–60.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B* **65** 331–355.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (Translated and edited by Dabrowska, DM and Speed, TP from the Polish original, Statistical Science (1990), 5, 465–472). *Annals of Agricultural Sciences* **10** 1–51.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature* **506** 150–152.

Parast, L. and Griffin, B.A. (2017). Landmark estimation of survival and treatment effects in observational studies. *Lifetime Data Anal.* **232** 161–182.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–688.

Pearl, J. (2009). *Causality.* Cambridge University Press, 2009.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling: Springer Series in Statistics*. Springer, Berlin.

Qian, M. and Murphy, S.A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Math. Modelling* **7** 1393–1512.

Robins, J. (2004). Optimal structural nested models for optimal sequential decisions. *Proceedings of the Second Seattle Symposium in Biostatistics*. Springer, New York.

Robins, J. M., Greenland, S., and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *J Amer. Stat. Assoc.*, **94** 687–700.

Robins, J., Orellana, L., and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* **27** 4678–4721.

Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41-55.

Royston, P. and Parmar, M.K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat. Med.* **30** 2409-2421.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization, *The Annals of Statistics* **1** 34–58.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* **26** 20–36.

Scornet, E., Biau, G. and Vert, J.P. (2015). Consistency of random forests *The Annals of Statistics* **43** 1716–1741.

Segal, M. and Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1** 80–87.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety* **17** 546-555.

Shpitser, I., and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* **9** 1941–1979.

Spirtes, P., C. N. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search.* MIT Press.

Su, X., Meneses, K., McNees, P., and Johnson, W. O. (2011). Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60** 457–474.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* **10** 141–158.

Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **10** 363–377.

Trafimow, D. and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology* **37** 1–2.

Tricoci, P., Allen, J. M., Kramer, J. M., Califf, R. M., and Smith, S. C. (2009). Scientific evidence underlying the ACC/AHA clinical practice guidelines. *Journal of the American Medical Association* **301** 831–841.

Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. In *Neural Networks, IEEE International Conference on* (Vol. 1, pp. 90-95). IEEE.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc. (in press)*

Wager, S. and Guenther, W. (2015). Uniform convergence of random forests via adaptive concentration arXiv preprint, arXiv:1503.06388.

Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician* **70** 129–133.

Yoon, D. Y., Smedira, N. G., Nowicki, E. R. et al. (2010). Decision support in surgical management of ischemic cardiomyopathy. *J. Thoracic and Cardiovascular Surgery* **139** 283-293.

Zhang, J.L., Rubin, D.B. and Mealli, F. (2009). Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking. *J. Amer. Statist. Assoc.* **107** 80-92.

Zhang, J.L., Rubin, D.B. and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *J. Amer. Statist. Assoc.* **104** 166-176.

Zhang, B., Tsiatis, A.A., Davidian, M., Zhang, M., and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103-114.

Zhao, Q. *Topics in causal and high dimensional inference*, Diss. Stanford University, 2016.

Zhao, Y., Zeng, D., Rush, A.J., and Kosorok, M.R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107**(499) 1106–1118.

Zhao, Y.Q., Zeng, D., Laber, E.B., Song, R., Yuan, M., and Kosorok, M.R., (2014). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* **102** 151-168.

Zhao, Y.Q., Zeng, D., Laber, E.B., Song, R., Yuan, M., and Kosorok, M.R. (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electron. J. Stat.* **11** 3927-3953.

Zhu, R., Zhao, Y.Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics* **73** 391-400.