EVALUATION OF SUPERVISED LEARNING APPROACHES FOR ASSESSING HETEROGENEITY OF TREATMENT EFFECT IN CLINICAL TRIALS

By

Lisa Marie Levoir

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

May 9th, 2025

Nashville, TN

Approved:

Bryan S. Blette, Ph.D

Andrew J. Spieker, Ph.D

ACKNOWLEDGMENTS

I have been lucky to get to know and learn from many wonderful people during my graduate training. From mentors to friends, I have been surrounded by kind and supportive individuals. The process of enumerating the people who have encouraged me in one way or another these past few years reminds me just how darn lucky I am to be here and how much I have gained from this experience.

To my friends back home, for understanding the grad school experience and cheering me on the whole way. Specifically, Shilvi, Shreyas, Mahala, Marla, Alysha, and Leah. You all are awesome!

To my Vanderbilt community, I want to thank you or being my sounding board, teaching me and letting me teach you and for just being such warm friends. I want to acknowledge my cohort mates Alexis and Jeffrey specifically - their friendship and support made all the difference in my capacity to succeed in this program. I also want to thank Jamie, an incredible person who builds community everywhere she goes and who was always in my corner. Our weekly TV & takeout nights are actually when Eric and I felt at home in Nashville. I want to thank "older students" Max, Ruby, and Chiara for chatting with me and explaining so many concepts - you truly brightened my day so many times. I want to thank Elisa, Megan, Ashley, and Julia for your friendship and mentorship. Thank you for generously sharing your time and attention with me. I'll also take the chance to thank Caroline, Cara, Joanne, Marisa, Gabi, and Chih-Ting (Teresa). Thank you to Barre3 Nashville for being my community away from the academic realm and showing me the joy of movement again. I also want to thank my academic coach, Sam, who encouraged me to form systems to support my work and personal well being - recognizing that all sides of how we show up to in our lives are intertwined. As Eric says, "even Olympians have coaches" and I'm grateful that the Center for Student Wellbeing supports students with amazing coaches like Sam.

I've been lucky to work with truly kind mentors who were both incredible statisticians and patient enough to help me grow. Amber, thank you for taking the time to walk me through Probability concepts. Week after week, you helped me build a strong foundation. Andrew, your patience and skill as an instructor is endless and I will always be grateful for (and actually remember!) what I learned from you about communicating statistics. Jonathan, thank you for taking me on for the CLOVERS research project. I learned so much from you about diligence and attention to detail, and your kindness and humility shone through all of our interactions. Mario and Cindy, thank you both for your support and kindness. (I know I said kindness a lot, but it's true every time!) Tatsuki, thank you for advising me all along this journey. I'm glad I got to work with each of you and also get to know you more as people. Finally, I want to thank Chazlie. She was one of the first people I met when I was accepted to the program and has been an incredible advocate not only for me, but for all students over the years. Thank you for everything.

Rameela, thank you for taking a chance on me and hiring me as your research assistant. I have learned so much from you as my mentor and supervisor. Your consistent advocacy and backing has built up so many students' career development and collaboration skills, including my own.

Bryan, I'm so glad that we got to work together to create this thesis. Your balance of empathy and insights were exactly what I needed from a mentor. I'm very grateful to you and Andrew for shepherding this project from our initial brainstorming to the completed thesis it is now.

I want to thank my family, especially my parents and brother. You've been there for me since the literal beginning and I love how we laugh together. I'll also thank Eric's family for welcoming me and rooting for me in equal measure.

Finally, I want to thank my spouse, Eric. His constant and unwavering support meant everything during the entire process - from encouraging me to persevere through the pre-reqs and application process before grad school to getting through the degree milestones. Life together has been such a gift. Thank you for being willing to move across the country with me (twice!) so I could learn and grow. Thank you for also walking Buddy so much when I was hitting the books - and thanks to Buddy! Buddy will often wait at the door for me when I'm out studying or will curl up to nap nearby while I work, which is proof to me that we don't deserve dogs but should always treasure our sweet companions.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS . . . . . **i**

LIST OF FIGURES . . . . . **iv**

1 Background **1**

1.1 Motivation 1
1.2 Causal Inference 1
1.2.1 Assumptions Sufficient to Identify the ATE 3
1.2.2 Plausibility of Causal Inference Assumptions in an RCT 3
1.3 Subgroup Analyses and the Realm of Precision Medicine 4
1.3.1 Characterizing HTE with Conditional Effects 5
1.4 Contributions of This Thesis 6

2 Methods **7**

2.1 Traditional HTE Methods 7
2.2 General Framework for CATE Estimation 8
2.3 Estimation of CATEs with Linear Regression 9
2.3.1 Confidence Intervals for the CATE with Linear Regression 10
2.4 Estimation of CATEs with Causal Forests 10
2.4.1 Random Forests 10
2.4.2 Causal Forests 13
2.4.3 Confidence Intervals for the CATE with Causal Forests 13

3 Simulations **14**

3.1 Data Generating Mechanisms 14
3.2 The Estimand 15
3.3 Methods for Comparison 15
3.3.1 Ordinary least-squares (OLS) Regression 15
3.3.2 Causal forest (CF) with Default Settings 15
3.3.3 Causal forest with Hyperparameter Optimization 16
3.4 Performance Metrics 17
3.4.1 Bias 17
3.4.2 Coverage 17
3.4.3 Standard Error 18

4 Results **19**

4.1 Results for One Scenario 19
4.1.1 Bias 19
4.1.2 95% Confidence Interval Coverage 20
4.1.3 Standard Error 21
4.2 Aggregated Results for the Linear DGM 22
4.2.1 Bias 22
4.2.2 95% Confidence Interval Coverage 23
4.2.3 Standard Error 25

4.3 Aggregated Results for the Nonlinear DGM 27
4.3.1 Bias 27
4.3.2 95% Confidence Interval Coverage 28
4.3.3 Standard Error 29

5 Discussion 30

5.1 Limitations 32

BIBLIOGRAPHY 33

LIST OF FIGURES

Figure		Page
1.1	Figure inspired by Hernan and Robins (2020) depicting the difference between association and causation	3
2.1	Example of a regression tree	11
4.1	Density plots for bias of the $\widehat{\text{CATE}}_i$ in one highlighted simulation scenario. A red vertical reference line is included for 0.	20
4.2	Density plots for confidence interval coverage proportions. Notice that some casual forest CIs for the CATE_i have very low coverage proportions.	21
4.3	Density plots for $\text{SE}(\widehat{\text{CATE}}_i)$.	22
4.4	Bias for the CATE under a linear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to 8 Bernoulli true effect modifiers (row wise facets) in each subplot. Labeled subplots show aggregate bias metrics for the additional combination of 0, 1, 4, or 8 Normally distributed true effect modifiers.	23
4.5	Coverage in the null setting for under a linear data generating mechanism. Situations did not include any true effect modifiers but did include 10 to 40 extraneous nuisance variables included in the model. Note the y axis range is from 0.9 to 1 in order to show detail.	24
4.6	95% confidence interval coverage under a linear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to 8 Bernoulli true effect modifiers (row wise facets) in each subplot. Labeled subplots show aggregate bias metrics for the additional combination of 0, 1, 4, or 8 Normally distributed true effect modifiers.	25
4.7	Standard error for the individual CATE estimator under a linear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to 8 Bernoulli effect modifiers (row wise facets) in each subplot. Labeled subplots show aggregate standard error metrics for the additional combination of 0, 1, 4, or 8 Normally distributed effect modifiers.	26
4.8	Bias for the individualized CATEs under a nonlinear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets). Labeled subplots show aggregate bias metrics for 1, 4, or 8 Normally distributed true effect modifiers. The y axis expands for the 4 and 8 effect modifiers subplots.	27
4.9	Coverage for under a nonlinear data generating mechanism with 1, 4, or 8 Normally distributed effect modifiers (top, middle, and bottom row) with 0 to 40 extraneous nuisance variables (columns).	28
4.10	Standard error for the CATE estimator under a nonlinear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to Bernoulli effect modifiers. Labeled subplots show aggregate standard error metrics for 1, 4, or 8 Normal effect modifiers.	29

CHAPTER 1

Background

1.1 Motivation

Randomized controlled trials (RCTs) are the gold-standard study design used to generate high-quality evidence about the causal effect of a treatment or intervention (Piantadosi, 2017; Friedman et al., 2015). During the 20th century, clinical trials supported the discovery of effective drugs for the treatment of many diseases including cancer, infectious diseases, cardiovascular disease, and mental illnesses. Under straightforward and plausible assumptions, data from different arms of an RCT can be summarized and contrasted to consistently estimate an average treatment effect estimand, i.e., the difference in the average outcome of interest had everyone been assigned the intervention vs. had everyone been assigned the control.

Recent research has acknowledged the important role of treatment effect heterogeneity. The field of precision medicine has arisen in the 21st century to address differences among people and develop methods to address such heterogeneity (Collins and Varmus, 2015; Kosorok and Laber, 2019; Piantadosi, 2017). While the average treatment effect from RCTs can indicate which treatment may be superior *on average,* it does not answer the question of the practicing doctor: "What is the most likely outcome when this particular drug is given to a particular patient?" (Kent et al., 2018). Much research has moved toward targeting estimands which more closely reflect the patient-specific nature of clinical practice.

Further reflecting the important role of rigorously conducted clinical trials, evidence-based medicine has become the dominant paradigm for developing clinical recommendations and decision-making tools. Although the average treatment effect is identifiable under assumptions that are plausible in many randomized trials, average treatment effects *may not be ideal at the level used for individual decision making* because individual patients generally differ in at least one dimension from the average trial participant. In that way, many clinicians' initial concerns about evidence-based medicine reflect an incongruence between the overall average effect of treatment in a study population and deciding what is best for an individual based on their specific characteristics, needs, and desires (Kent et al., 2010). Given that clinicians are interested in determining the best treatment for a given patient (as opposed to inferring utility from the average trial result), there is growing interest in understanding how a treatment effect varies across patients - often termed Heterogeneity of Treatment Effect (HTE) (Kent et al., 2018; Varadhan et al., 2013; Willke et al., 2012).

1.2 Causal Inference

We will use tools from causal inference to define HTE and provide a brief review. Under the potential outcomes framework in causal inference, a treatment effect (TE) is a contrast between potential outcomes (hypothetical future

outcomes which would be observed under different possible treatment conditions) (Hernan and Robins, 2020). Stated more generally, causal effects of treatments are defined by contrasts of potential outcomes under treatment levels (Rubin, 2005). For simplicity, we focus this discussion on a two-arm trial. Let $W = 1$ indicate assignment to a treatment arm and $W = 0$ indicate assignment to a control arm. Let Y^1 be the potential outcome under treatment and Y^0 be the potential outcome under control. In this setting, only one of the two potential outcomes (Y_i^0 or Y_i^1) for a patient i can be observed in the real world.

The **fundamental problem of causal inference** is that both potential outcomes cannot be observed for the same individual, such that the treatment effect for any particular trial participant (the individual treatment effect, ITE) is unidentifiable. The observed potential outcome corresponds only to the received treatment. The other potential outcome is called the *counterfactual* because it is unobserved (latent) (Imbens and Rubin, 2015). For example, in a clinical trial for a new blood pressure medication, we can only observe the outcome (e.g., change in blood pressure) for the treatment that is actually received, not for other potential treatments.

While the individual causal effects are not identifiable, the *average treatment effect in the population* can be identified under several assumptions (described below in Section 1.2.1). For this reason, the average treatment effect (ATE) is often the targeted estimand in clinical trials. The ATE is defined through "contrasts of the mean potential outcomes across the population" and exists when there is a mean difference between the treated and the control potential outcomes (Hernan and Robins, 2020):

$$\text{ATE} = E[Y^1 - Y^0] \stackrel{?}{=} E[Y|W = 1] - E[Y|W = 0] \tag{1.1}$$

The second equality is plausible in a RCT under the identifiability assumptions described in Section 1.2.1. In other literature, the ATE may also be called the Average Causal Effect (ACE). Even in settings where the *average* causal effect is equal to 0, there may still be *individual* causal effects (Hernan, 2004). There are two potential null settings where ATE = 0 overall: a) there is no causal effect for any individual (sometimes referred to as a sharp null) or b) there are individual causal effects but the mean potential outcomes do not differ under treatment and control. The null scenario will be detailed further in the simulation study results in Section 4.2.2.

There is an important distinction between measures of association versus measures of causation. The former compares $E[Y|W = w]$ across w while the latter compares $E[Y^w]$ across w, as described in Figure 1.1 (Hernan and Robins, 2020). While association does not imply causation *in general*, the design of a randomized controlled trial can make plausible a set of assumptions under which association and causation can align.

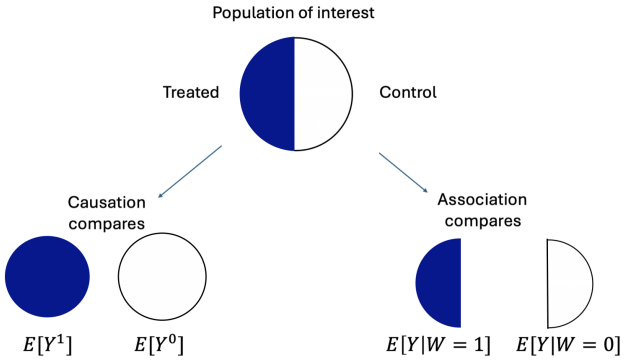

Figure 1.1: Figure inspired by Hernan and Robins (2020) depicting the difference between association and causation

1.2.1 Assumptions Sufficient to Identify the ATE

With observed data, causal effects may be identified and estimated under sets of conditions called identifiability assumptions (Hernan and Robins, 2020). The assumptions required to identify the ATE include:

Positivity: within strata defined by covariates $\mathbf{X}$, subjects have a probability > 0 of either having either treatment level. This can be violated if particular groups are ineligible for treatment.

Consistency: An individual with observed treatment $W = w$ will have outcome Y equal to Y^w.

No interference: A subject's potential outcome is not affected by other subjects' exposures.

Ignorability (unconfoundedness): Potential outcomes are independent from observed treatment. Independence between potential outcomes and treatment can be marginal or conditional on covariates.

Together, the assumptions of consistency and no interference are called the Stable Unit Treatment Value Assumption (SUTVA). As a part of SUTVA, we also assume there are not multiple versions of treatment (e.g., different treatment by location, or differently skilled interventionists) (Rubin, 2005). Potential outcomes are well-defined when we can assume SUTVA and positivity. To identify the ATE, we also need ignorability.

1.2.2 Plausibility of Causal Inference Assumptions in an RCT

Randomization is useful in generating convincing causal inferences (Hernan and Robins, 2020). In the setting of a RCT, the assumptions described in Section 1.2.1, which can be quite strong in observational settings, are often plausible:

- Positivity: Satisfied based on study design and eligibility criteria.
- Consistency: Satisfied when there is just one version of each treatment.

- Interference: Could occur in trials of infectious disease treatments, educational programs, or other settings where one person's outcome is influenced by interactions with others (Hudgens and Halloran, 2008). This should be evaluated in the context of each clinical trial; we will focus on settings where this is assumed to be satisfied.
- Marginal ignorability (unconfoundedness): Satisfied in a randomized experiment (treated and untreated groups are exchangeable)

Under these assumptions, the ATE can be identified as $E[Y|W=1]-E[Y|W=0]$ and thus estimated using a simple contrast of means. In essence, RCTs typically put more burden on the design and implementation phase of a study, but yield data from which it is often trivial to draw causal inference compared to observational studies, leading to the equivalence in Equation 1.1.

1.3 Subgroup Analyses and the Realm of Precision Medicine

While straightforward to identify and estimate, the average treatment effect does not closely mirror the personalized nature of clinical practice. We have established that the ATE is insufficient to learn about heterogeneity in treatment effect across the population. So, how can we evaluate treatment effects across the patient population? Conventional approaches include one variable at a time "subgroup" analyses which often do not detect differences within strata and falsely enforce a "consistency of treatment effect" across subgroups (Kent et al., 2010). Furthermore, these methods have other shortcomings - we generally would need to conduct too many subgroup analyses across the multitude of covariates collected in modern clinical trials, rendering them underpowered and vulnerable to detecting spurious associations (Kent et al., 2010; Brookes et al., 2004). Statistical best practices include documenting prespecified analysis plans. While these plans can help avoid the detection of spurious associations and an increased Type I error rate associated with multiple testing, the trade off with prespecifying all analyses could mean missing "strong but unexpected" treatment effect heterogeneity (Wager and Athey, 2018).

Even in the setting where there is a large and clinically meaningful treatment effect difference across strata, conventional "one variable at a time" comparisons are poorly calibrated to detect differences between risk strata because conducting inference on one variable at a time does not reflect the fact that patients may have multiple characteristics that influence risk simultaneously (Kent et al., 2010). For example, in the blood pressure medicine trial mentioned earlier, patients will differ in their set of covariates but some variables will frequently co-occur (ie., high biomarkers for both blood sugar and cholesterol) such that it would be hard to find a set of patients with *just* one factor and not the other. Another related limitation is that this approach only considers two-way interactions, ignoring the potential for interaction between two or more covariates with the treatment. In other words, higher-order interactions between covariates will be missed by conventional "one variable at a time" methods. That being said, there is a "degrees-of-freedom" tradeoff the practitioner should consider when deciding which hypothesis tests to conduct.

1.3.1 Characterizing HTE with Conditional Effects

Similar assumptions to those used to identify the ATE can be used to identify conditional average treatment effects (assumptions like positivity must be broadened to hold across potential effect modifiers as well). The conditional average treatment effect (CATE) is defined by:

$$\text{CATE}(\mathbf{x}) = E[Y^1 - Y^0 | \mathbf{X} = \mathbf{x}] \tag{1.2}$$

where the $\mathbf{x}$ of interest will vary depending on the scientific question (in practice the full set of baseline covariates and the set of potential effect modifiers can be distinct; we use $\mathbf{X}$ generally moving forward to consider these sets to be identical for simplicity). Trivially, in the setting where there are no effect modifiers, the $\text{CATE}(\mathbf{x}) = \text{ATE}$ for all $\mathbf{x}$. Otherwise HTE is present. In words, the CATE is the average treatment effect *conditional* on belonging to a subgroup defined by $\mathbf{x}$. "Conditional" refers to the idea that the treatment effect may vary across different subgroups in a population depending on their covariate values. In the context of this work, the covariates, $\mathbf{x}$ are measured at baseline. Traditionally, one might condition on a set of single baseline covariates in a series of subgroup analyses, as described in Section 1.3. However, estimation of "individualized" treatment effects is becoming more popular (i.e., estimation of a CATE_i corresponding to each unique $\mathbf{x_i}$ observed in a trial) to account for potentially complex HTE which cannot be easily captured by subgroup analyses. Note that such "individualized" effects are CATE parameters and **not** individual-level treatment effects.

In order to identify and estimate $\text{CATE}(\mathbf{x})$ using the observed data $(\mathbf{X_i}, Y_i, W_i)$, we must assume strong ignorability such that treatment assignment is independent of the potential outcomes conditional on the covariates (Wager and Athey, 2018).

$$Y_i^0, Y_i^1 \perp W_i \,|\, \mathbf{X_i} \tag{1.3}$$

We assume that assumption 1.3 is guaranteed under proper randomization. As $\mathbf{X}$ is typically high-dimensional, many methods for estimation of individualized CATEs entail leveraging a supervised learning approach. There are many potential advantages of being able to estimate CATEs. Estimation of CATEs can allow for building of personalized treatment regimes, hypothesis generation, or development of better understanding of the (biological, social, or other) causal mechanisms leading to the outcome (Künzel et al., 2019; Hernan and Robins, 2020). In the realm of personalized medicine, targeting the CATE can also help researchers identify specific subgroups that are more likely to benefit from a treatment.

1.4 Contributions of This Thesis

Various methods for estimation and inference of "individualized" CATEs have been proposed with good asymptotic properties, but the sample size required for good statistical properties may depend heavily on the underlying data and heterogeneity. However, many practitioners are implementing these methods in clinical trials with smaller or moderate sample sizes where the performance of these methods is not clear (Afshar et al., 2024; Seitz et al., 2023; Goligher et al., 2023). Previous simulation studies often consider sample sizes of 5,000 participants or greater. Studies that do consider smaller sample sizes do not include metrics about confidence interval coverage (Hoogland et al., 2021) or do include coverage but not under the setting of heterogeneity (Wager and Athey, 2018). Thus, these methods are being used in trials with smaller sample sizes than those which have been the focus of previous investigations of empirical properties of these methods.

The goal of this work is to investigate the *finite-sample properties* of popular methods for estimation and inference of individualized CATEs across a range of scenarios and sample sizes with focus on sample sizes that are more commonly used in clinical trials to get a better understanding of when one can achieve valid CATE inference using RCT data in practice.

In line with our goal, these are a few questions we seek to address:

1. How can we reliably detect HTE in clinical trials?
2. What guidance can we offer to the practitioner in terms of what sample size is necessary to expect valid performance of different estimators?
3. What guidance can we offer to the practitioner in terms of which method should be chosen to get better performance at smaller sample sizes?

CHAPTER 2

Methods

2.1 Traditional HTE Methods

To begin, we review traditional methods for assessing treatment effect heterogeneity. As these are commonly performed using regression models, we focus on the linear regression setting for simplicity. The traditional approach entails consideration of heterogeneity "one variable at a time" which can be facilitated by two very similar approaches: subgroup analysis and interaction analysis.

Put simply, subgroup analysis entails estimating a treatment effect in a subgroup defined by patient characteristics. Conventional subgroups tend to be defined on the basis of sex, age, body mass index, or other measured baseline variables (Yarnell and Fralick, 2024). This approach is commonly used because it is straightforward to understand and use clinically apparent variables to form subgroups. Although this approach is frequently employed in current literature, investigations based on subgroup analysis could be considered too simplistic and often do not yield practical insights (Yarnell and Fralick, 2024). Subgroups can also be formed according to baseline risk of the outcome (independent of treatment assignment), or formed according to predicted treatment effect (Goligher et al., 2023), although these are less straightforward to prespecify.

HTE can also be assessed by performing inference for the coefficient of an interaction term in a model. For example if we wish to test whether X_2 is a true effect modifier in the linear regression model:

$$E[Y|X = x, W = w] = \beta_0 + \beta_1 w + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1 w + \beta_5 x_2 w \tag{2.1}$$

we would test whether $\beta_5 = 0$. Testing for interaction across a covariate can provide a more formal approach to study differential effects than comparing confidence intervals for the treatment effects in subgroups defined by that covariate for overlap. It also avoids forced categorization of continuous covariates and should result in more efficient estimates of auxiliary parameters. As mentioned earlier, research best practices stipulate that hypothesis testing should be part of a pre-specified analysis plan because pre-specification helps reduce the detection of spurious associations (Type I errors), but as described in Section 1.3, pre-specifying all analyses could mean missing unexpected treatment effect heterogeneity - ignoring the increasingly hard to pre-specify potential interactions between two or more covariates with the treatment.

2.2 General Framework for CATE Estimation

Given the limitations of traditional HTE methods, modern frameworks for assessing HTE based on estimation of CATE parameters have become increasingly popular. In particular, leveraging supervised learning models to estimate CATE($\mathbf{x}$) for all unique $\mathbf{x}$ (or a subset of unique $\mathbf{x}$) observed among trial participants can yield an insightful exploration of heterogeneity which avoids many of the aforementioned limitations. This approach is sometimes termed estimating "individualized treatment effects" but we use the term CATE throughout to avoid confusion with individual treatment effects, which are **not** identifiable due to the fundamental problem of causal inference.

The first step to estimating CATE parameters in a trial is to fit an outcome model. Here, we describe a general framework for CATE estimation using an unspecified supervised learning algorithm to predict outcomes. Later we will provide more details and explore performance of specific learners. In supervised learning, we would like to learn from data based on a set of **features** (observed random variables such as covariates and treatment assignment) to predict an **outcome** (in this case, we assume a continuous treatment response). To avoid overfitting, we take a random subset of the full data and designate it as *training* data where we have both the outcome and the features for the study participants. We then build a model so we can predict the outcome for other participants that only have their feature set (and not the outcome) observed. This proportion of the data with only the outcome is the *test* data (Hastie et al., 2017).

The proportion of data in the testing and training subsets may be selected by the practitioner, so for simplicity, we use a 1:1 split throughout. The **training** set is used to fit the model for Y conditional on $\mathbf{X}$, where the model should explicitly or implicitly consider interactions between treatment and covariates since the goal is to investigate HTE. In most supervised learning problems, we then evaluate model performance on the **testing** data (which can be considered a hypothetical second independent sample). In our case, the testing (holdout) set serves as the set where CATEs are estimated but for which the true CATEs would only be known in a simulation.

To do this, we calculate predicted potential outcomes by first taking the testing data set and coercing all $W = 1$ and then calculating model-based predictions. For example, in the setting of linear regression, model-based predictions follow the form:

$$\widehat{Y}^1(\mathbf{x}) = \mathbf{x}^{*\mathbf{1}}\widehat{\beta} \tag{2.2}$$

where $\hat{\beta}$ is estimated using the training data and $\mathbf{x}^{*\mathbf{w}}$ denotes a design matrix with treatment coerced to w, covariates set to the observed $\mathbf{x}$, and interactions between treatment and covariates. We then repeat this procedure to calculate all model-based potential outcomes as if all participants were in the control group: $\hat{Y}^0(\mathbf{x}) = \mathbf{x}^{*\mathbf{0}}\hat{\beta}$. Finally, we can calculate the model-based conditional average treatment effects:

$$\widehat{\text{CATE}}(\mathbf{x}) = \hat{Y}^1(\mathbf{x}) - \hat{Y}^0(\mathbf{x}) \tag{2.3}$$

The CATE is a function of $\mathbf{x}$. We are interested in estimating the CATE for all observed values of the covariate matrix in our holdout set where $\mathbf{X} = \mathbf{x}$. Using a contrast of predicted potential outcomes from a single outcome model is a version of g-computation sometimes referred to as S-learning (Künzel et al., 2019). While other frameworks such as T-learning, X-learning, and doubly-robust learners exist, these are not the focus of this thesis (and choice of framework should have less impact when focusing on the randomized trial setting).

We then form confidence intervals for $\widehat{\text{CATE}}(\mathbf{x})$ at each observation. The relevant methods for forming confidence intervals based on each supervised learning model are described in their respective sections below. In the simulation study (described in Chapter 3) we are able to assess whether the confidence interval captures the true CATE for each individual in a simulated clinical trial. In this thesis, we will consider two outcome model methods for estimation and inference: linear regression and random forests (implemented through the causal forest procedure) which we describe in the following sections.

2.3 Estimation of CATEs with Linear Regression

Linear least squares models make strong assumptions about the structure of the relationship between the features and the outcome (Hastie et al., 2017). In reality, the association between features and outcome rarely follows an exactly straight-line relationship, but we may prefer the stability and optimality of ordinary least squares for the trade off of estimating a first-order approximation to the relationship.

Ordinary linear regression is a semiparametric method with structure imposed on the mean model. Given a $N \times K$ design matrix $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{k-1})$ where $\mathbf{x}_0$ is generally a vector of ones for the intercept, we predict a length N outcome vector, $\hat{\mathbf{Y}}$ most commonly by using the method of ordinary least squares (OLS). In ordinary least squares, we estimate the set of coefficients β by minimizing the residual sum of squares $||\mathbf{y} - \mathbf{X}\beta||^2$. Then the fitted ("predicted") value is called $\hat{\mathbf{y}}$ which, by definition, is equal to $\mathbf{X}\hat{\beta}$. In this way, $\hat{\beta}$ is the least squares estimator. Assuming $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ where $E[\varepsilon] = 0$ and $E[\varepsilon|\mathbf{X}] = 0$, $\hat{\beta}$ is unbiased (Puntanen and Styan, 1989). The Gauss Markov theorem asserts that the ordinary least squares estimator is the unique minimum variance unbiased estimator among all unbiased linear estimators. Any linear combination, $\mathbf{a}^T\beta$ of the $\hat{\beta}$s is the best linear unbiased estimator - no other estimator (linear in $\mathbf{y}$ and unbiased for $\mathbf{a}^T\beta$) has lower variance than $\mathbf{a}^T\hat{\beta}$. After fitting a linear model, we can estimate the CATE for a specific $\mathbf{x}$ by calculating all model-based potential outcomes under the two scenarios of treatment and control in Equation 2.2.

2.3.1 Confidence Intervals for the CATE with Linear Regression

Next, we aim to form confidence intervals for the CATE. In the setting where we assume error homoscedascity ($\text{Cov}[\varepsilon] = \sigma^2\mathbf{I}$), the covariance of the least squares estimator $\hat{\beta}$ is equal to $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Recognize that σ^2 is not typically known, but can be estimated. Next, let $\mathbf{C}$ be a contrast matrix comparing the coefficients for the treated and control subjects. For example, in Equation 2.1 where there are 2 effect modifiers, $X_1 = x_1$ and $X_2 = x_2$ and the treatment indicator, W,

$$\hat{E}[Y|\mathbf{X} = \mathbf{x}, W = w] = \hat{\beta}_0 + \hat{\beta}_1 w + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_2 + \hat{\beta}_4 x_1 w + \hat{\beta}_5 x_2 w \tag{2.4}$$

the relevant contrast matrix $\mathbf{C}$ is

$$\begin{bmatrix} 0_n & 1_n & 0_n & 0_n & x_1 & x_2 \end{bmatrix}$$

such that the model based variance of $\widehat{\text{CATE}}(\mathbf{x})$ is given by $\text{Var}(\widehat{\text{CATE}}(\mathbf{x})) = \text{diag}(\mathbf{C}\text{Cov}(\hat{\beta})\mathbf{C}^{\mathbf{T}})$, where diag(.) is a function that takes a matrix and outputs a vector corresponding to the diagonal elements of the matrix. Finally, we form the 95% Wald-based confidence interval for the CATE in the following manner, letting $z_{1-\alpha/2}$ be the $1-\alpha/2$ quantile of the Normal distribution:

$$\widehat{\text{CATE}}(\mathbf{x}) \pm z_{1-\alpha/2} * \sqrt{\text{Var}(\widehat{\text{CATE}}(\mathbf{x}))} \tag{2.5}$$

2.4 Estimation of CATEs with Causal Forests

Before describing causal forests, we will first cover random forests which form the theoretical underpinning for causal forests.

2.4.1 Random Forests

Random forests are a type of tree-based method for regression and classification developed in the literature between 1995 and 2001 (Ho, 1995; Breiman, 2001). A single decision tree is formed by segmenting the predictor space into a set of simple bins which are summarized in the format of a tree where each "branch" is a decision (Hastie et al., 2017). While one tree may be straightforward to interpret, by itself a single tree typically does not perform well in terms of prediction accuracy. So, we employ an ensemble approach to combine many "building block" models consisting of individual decision trees to obtain a single model which, at the sacrifice of interpretability, often has greatly improved prediction accuracy (James et al., 2023).

Regression trees consist of a series of splitting rules, with the most important factor at the *root* of the tree. Regression trees are typically displayed upside down compared to a real tree. The root is at the top and the *leaves* (terminal nodes) are at the bottom of the tree. In between the root and the leaves, the points at which the predictor space is split

are called *internal nodes* and the segments that connect these nodes are called *branches* (James et al., 2023). This structure lends itself to a nice graphical representation (Figure 2.1).

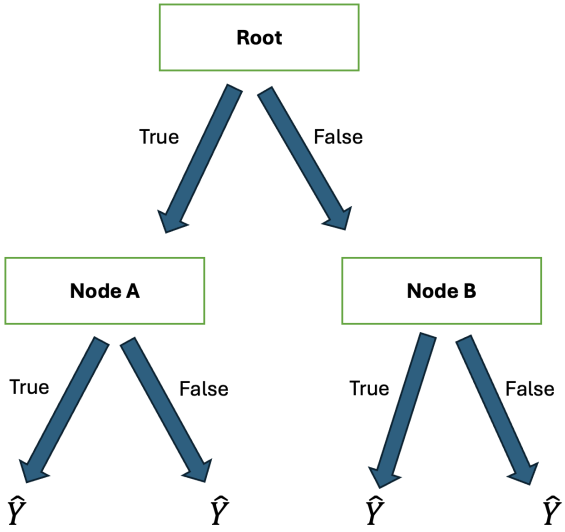

Figure 2.1: Example of a regression tree

The process of forming a regression tree involves two main steps. First, divide the predictor space into K distinct non-overlapping partitions. Next, for each observation that falls into the region R_j, predict the new outcome. Prediction is typically done by using the mean of the response for the set of training observations included in R_j (James et al., 2023). Since we seek to minimize the residual sum of squares between the fitted and observed outcome in a computationally tractable way, we employ recursive binary splitting. Recursive binary splitting is considered a *greedy* approach because, at each junction for a decision, we make the best possible split at that particular step and do not consider future potential splits that *could* have resulted in a better tree. This procedure *top down* in the sense that we begin at the top of the tree (where all observations belong to one region) and make the first split into two branches before proceeding to recursively make the next splits continuing on in this manner until we meet a stopping criteria.

Notice, regression trees assume a different model format from linear regression, where $R_1, \ldots, R_M$ represent a partition of the feature space (Hastie et al., 2017):

$$f(X) = \sum_{m=1}^{M} c_m * I(X \in R_m) \tag{2.6}$$

We may seek to adjust hyperparameters of tree building in order to minimize overfitting to training data (which would result in poor model performance on unseen data). Imagine that we first build one very large tree T_0 and then obtain a subtree, T, by "pruning" the larger tree. This way, instead of considering the entire sample space of all possible subtrees, we can consider only a smaller subset of candidate subtrees indexed by α, a nonnegative tuning

parameter (James et al., 2023). Akin to the idea of using lasso for regularization in linear regression, the value of α controls the balance between the subtree's fit to the training data and it's complexity. When $\alpha = 0$, $T = T_0$. As α increases there is a penalty for more terminal nodes which encourages a smaller subtree. We can select a value for α using cross validation, which is done in the simulation study described in Section 3.3.3.

One can use bootstrap aggregation, or **bagging**, to improve the performance of decision trees by lowering the between-sample variance (Hastie et al., 2017). Using the bootstrap to resample from our dataset with replacement, we generate B different bootstrapped training data sets. Then, we fit our model (in this case, the regression tree) on the bth bootstrapped training data set. Bagging then consists of taking the average across all predictions - which is particularly useful for regression trees because although individual trees have high variance (and low bias), averaging across all B trees reduces the variance (Hastie et al., 2017; James et al., 2023).

Random forests consist of averaged predictions across a large number of regression trees. Following the idea of bagging, the random forest method involves building a large set of de-correlated trees and then averaging across them. Trees have the capacity to "capture complex interaction structures" - resulting in a noisy but approximately unbiased model that can be greatly improved by averaging (Hastie et al., 2017). During bagging, we can take advantage of the fact that each tree is identically distributed to apply expectation rules and recognize that the bias of all bagged trees and the bias of an individual tree are the same - so the only way to reduce mean squared error (MSE) is by reducing variance (Hastie et al., 2017).

Random forests improve upon the variance reduction from bagging by *decorrelating* the trees. During tree growing, we "shake up" the included predictors by randomly selecting input variables. The procedure is as follows. Before each decision split, randomly subset the p set of predictors to m input variables (where $m \leq p$) as the only candidates for splitting. Like α, m is considered a tuning variable and is often selected such that $m = p/3$ or $\sqrt{p}$ (Hastie et al., 2017). By disallowing the tree from considering all possible predictors at each split, we're able to dilute the influence of a few strong predictors. For example, if there was one very strong predictor then it's probable that each tree's root would have the same strong split - which would result in many correlated trees (and resulting predictions). We realize the benefits of bagging when averaging across *uncorrelated or only weakly correlated* trees, which is why the insight of using only m predictors is so advantageous for random forests (James et al., 2023).

To estimate the test error of a bagged model, we turn to *out-of-bag* (OOB) error estimation. On average, each bagged tree includes about $2/3$ of the observations, so the remaining $1/3$ unused observations are called *out of bag* observations (James et al., 2023). Since OOB estimates are very similar to N-fold cross validation and can be performed in the sequence of fitting random forests, we can conclude training the random forest once OOB error reaches stability (Hastie et al., 2017). In summary, we conclude fitting each tree when the residuals are minimized and we conclude model training once OOB error is low enough compared to a threshold.

2.4.2 Causal Forests

Causal forests are built upon random forests, which as detailed in Section 2.4.1, consist of averaged predictions across a large number of regression trees. Causal forests therefore build on random forests' utility in flexibly modeling interactions in high dimensions, but causal forests are adapted to be directly useful for causal inference by directly estimating heterogeneity in casual effects (Wager and Athey, 2018). Causal forests are composed of "causal trees that estimate the effect of treatment at the leaves of the trees". This nonparametric method, originally published in 2018, was developed to perform well in the setting of numerous covariates where previous methods (like nearest neighbor matching) had failed.

Causal forests were developed to estimate HTE "in the potential outcomes framework with unconfoundedness". Causal forests promise "provably valid statistical inference" due to properties of asymptotic unbiasedness and asymptotic normality which allows us to form confidence intervals and perform hypothesis testing (Wager and Athey, 2018). Therefore, the key contribution of causal forests as compared to random forests is the capacity for asymptotic inference (because the asymptotic properties of random forests were previously unknown).

2.4.3 Confidence Intervals for the CATE with Causal Forests

Due to the challenge of navigating the potential outcomes framework when it comes to prediction for causal inference, we need to leverage asymptotic theory to extend support for statistical inference to causal forests (Wager and Athey, 2018). In order to conduct asymptotic inference with causal forests, some conditions are required:

- Trees "used to build the forest must be grown on subsamples of the training data" (Wager and Athey, 2018).
- Individual trees must be honest: at each point in the training set i, the response Y_i must be used for **either** "estimating within-leaf treatment effect" or to determine where to place splits, **but not both** (Wager and Athey, 2018).

In the potential outcomes framework with ignorability (unconfoundedness), consistency and asymptotic normality of casual forests allow for the formation of confidence intervals for estimating HTE. Honest causal trees are required to obtain confidence intervals for the CATE (Athey and Imbens, 2016). Asymptotic variance for causal forests can be consistently estimated using the infinitesimal jackknife procedure (which assumes that the number of bootstrapped trees is sufficiently larger to drown out any Monte Carlo variability of the forest so we only measure randomness in $\widehat{\text{CATE}}(\mathbf{x})$ in the training sample). Once we have an estimate for variance, it is straightforward to form confidence intervals using Equation 2.5. This process can be carried out by the `grf` package in R.

CHAPTER 3

Simulations

3.1 Data Generating Mechanisms

Suppose that the set of covariates is $\{X_1,...,X_j,X_{j+1},...,X_{j+k},X_{j+k+1},...,X_{j+k+p}\}$ where the first j covariates are true Normal effect modifiers, the next k covariates are true Bernoulli effect modifiers and the remaining p covariates are not effect modifiers nor predictive of the outcome, which we label as nuisance covariates moving forward. Individual covariates can be distributed either as a standard Normal variable or as a Bernoulli random variable with p = 0.5. W is a binary treatment indicator (treatment or control group) with probability 0.5 of assignment to each arm and Y is a continuous outcome. Both linear and nonlinear DGMs were considered. Simulations for linear DGMs were conducted in R under the following set of scenarios:

Total trial enrollment (n): 500, 1,000, 2,000, 4,000

Number of true Normal effect modifiers: j = 0, 1, 4, 8

Number of true Bernoulli effect modifiers: k = 0, 1, 4, 8

Number of nuisance variables, Bernoulli distributed: p = 0, 10, 20, 40

For any simulated trial enrollment, participants were split 1:1 into treatment and control groups. For the control group, the simulated potential outcome, Y^0 is assigned following a standard Normal distribution (and thus has an expectation of 0). For those in the treatment group, the potential outcome consists of the treatment effect plus the random noise that subjects experience in the control group (so the only difference between treatment and control is the treatment effect, CATE_i). In linear settings, the true data generating mechanism was simple linear addition such that

$$\text{CATE}_i = \sum_{i=1}^{j+k} X_i \tag{3.1}$$

We also considered a nonlinear data generating mechanism similar to one of the functions used to validate causal forests (Wager and Athey, 2018). In particular, $\text{CATE}_i = \sum_{i=1}^{j} f(X_i)$ where

$$f(x) = 1 + \frac{1}{1+e^{-2(x-\frac{1}{4})}} \tag{3.2}$$

We set $k = 0$ in all nonlinear settings, but otherwise explored nearly identical varying values of n, j, and p. In particular, in the linear data generating mechanism setting, this resulted in 252 scenarios since we exclude scenarios with 0 true predictors (Normal and Bernoulli) and 0 nuisance variables: $4^4 - 4 = 252$. We examined 48 scenarios in the nonlinear

data generating mechanism, excluding the situations with Bernoulli effect modifiers as well as the scenario where there were 0 Normal effect modifiers (as there would be no distinct nonlinear data generating mechanism in the scenario with only nuisance variables): 4 trial enrollment sizes $*$ 3 true effect modifiers $*$ 4 nuisance variables $=$ 48. After data generation, the data is further split 1:1 into training and holdout sets. The training set is used to build the model and the holdout set was used to evaluate performance. All simulations were run to 2000 replicates to balance computational time with precision in inference.

3.2 The Estimand

Treatment effects are conceptualized under the potential outcomes framework (as described earlier in Section 1.2). While the individual treatment effects are unobservable due to the fundamental problem of causal inference (ie. we can't observe the outcome under two distinct realities), we can estimate CATE_i for any individual in a holdout set. In our simulation, we are estimating individualized CATEs, CATE_i, which refer to average contrasts in potential outcomes for hypothetical subpopulations belonging to the same strata with baseline covariates $\mathbf{x}_i$. As the holdout set varies in each simulation, we have a total of n estimands of interest in each scenario.

3.3 Methods for Comparison

3.3.1 Ordinary least-squares (OLS) Regression

After data is generated as described above, a linear regression model is fit on the training data using the `lm()` function. CATEs and corresponding confidence intervals are estimated as described in Section 2.3.1. Model performance for estimating the CATE (95% CI coverage, bias, and variance) is assessed using the holdout data. When the mean model is not specified correctly, the ability of OLS to recover unbiased parameter estimates is compromised. We anticipated that simple linear regression would perform better in the additive linear setting but would not recover nominal confidence interval coverage in the complex non-linear setting.

3.3.2 Causal forest (CF) with Default Settings

The same data that was generated and used to train the OLS model is also used to fit a causal forest using the `grf::causal_forest()` function under the default settings. The code chunk below demonstrates fitting a causal forest and extracting the estimates of interest. Performance is assessed in the holdout set, as before.

```
cf_covariates <- training_dat %>% dplyr::select(starts_with("X"))
cf_outcome <- training_dat$outcome
cf_trtassign <- training_dat$W

## run a causal forest
tau.forest <- grf::causal_forest(X = cf_covariates,
     Y = cf_outcome, W = cf_trtassign)

#create testing matrix with same
    # number cols as training cf_covariates in the same order
```

```r
cf_test_covariates <- testing_dat %>% dplyr::select(starts_with("X"))

# Estimate treatment effects for the test sample
tau.hat <- predict(tau.forest, newdata = cf_test_covariates,
        estimate.variance = TRUE)

# store causal forest bias and standard errors
sigma.hat <- sqrt(tau.hat$variance.estimates) # sigma hat for each individual in n_test
CF.se[row.get, i] <- sigma.hat

# Column 'predictions' contains estimates of the conditional average treatment effect (CATE)
cf.CATE[row.get, i] <- tau.hat$predictions
## subtracting 2 vectors which should be composed of entries in the same positions in
# their respective data frames and storing them in the correct position

CF.bias[row.get, i] <- cf.CATE[row.get, i] - true_CATE.test
```

Listing 3.1: Causal forest R simulation example

3.3.3 Causal forest with Hyperparameter Optimization

In this variation of causal forest, we use cross validation to select optimal hyperparameters for forming causal forests across the first 100 reps of the simulation. After 100 reps, all chosen hyperparameter values are averaged and used for the following 1,900 simulation replicates in place of the default hyperparameters. Ideally hyperparameter optimization would occur on every replicate, but this approach was too computationally expensive for this thesis.

```r
if(cf.tune == TRUE & i <= 100 ){
   tau.forest.tune <- grf::causal_forest(X = cf_covariates, Y = cf_outcome, W = cf_trtassign,
   tune.parameters = "all")

   cf_test_covariates <- testing_dat %>% dplyr::select(starts_with("X"))
   tau.hat.tune <- predict(tau.forest.tune, newdata = cf_test_covariates, estimate.variance =
   TRUE)
   sigma.hat <- sqrt(tau.hat.tune$variance.estimates)
   CF.se.tune[row.get, i] <- sigma.hat
   cf.CATE.tune[row.get, i] <- tau.hat.tune$predictions
   cf.ci.L.tune <- cf.CATE.tune[row.get, i] - qnorm(0.975) * sigma.hat
   cf.ci.U.tune <- cf.CATE.tune[row.get, i] + qnorm(0.975) * sigma.hat
   covered.CF.tune[row.get, i] <- ifelse(true_CATE.test >= cf.ci.L.tune & true_CATE.test <= cf.
   ci.U.tune, 1, 0)

   # record the bias
   CF.tune.bias[row.get, i] <- cf.CATE.tune[row.get, i] - true_CATE.test

   # store details about the cf tuned hyper parameters
   cf.tune.HP[i, ] = c(unlist(tau.forest.tune$tuning.output$params), error = tau.forest.tune$
   tuning.output$error)
   }

 if(cf.tune == TRUE & i == 101){
   # extract the averages above
   samp.frac <- mean(cf.tune.HP[,"sample.fraction"], na.rm = TRUE)
   mt <- mean(cf.tune.HP[,"mtry"], na.rm = TRUE)
   mns <- mean(cf.tune.HP[,"min.node.size"], na.rm = TRUE)
   hs <- mean(cf.tune.HP[,"honesty.fraction"], na.rm = TRUE)
   hpl <- ifelse(sum(cf.tune.HP[,"honesty.prune.leaves"], na.rm = TRUE) >= 15, 1, 0)
   a <- mean(cf.tune.HP[,"alpha"], na.rm = TRUE)
   ip <- mean(cf.tune.HP[,"imbalance.penalty"], na.rm = TRUE)
 }

 if(cf.tune == TRUE & i > 100){
    # use the extacted values which I only had to calculate once
```

```
    tau.forest.tune <- grf::causal_forest(X = cf_covariates, Y = cf_outcome, W = cf_trtassign,
                                          sample.fraction = samp.frac,
                                          mtry = mt,
                                          min.node.size = mns,
                                          honesty.fraction = hs,
                                          honesty.prune.leaves = hpl,
                                          alpha = a,
                                          imbalance.penalty = ip
                                          )
# ... record the same parameters as above (omitted for brevity)
}
```

Listing 3.2: Causal forests with optimized hyperparameters

3.4 Performance Metrics

3.4.1 Bias

The bias metric refers to the difference between the individual's true $CATE_i$, the `true_CATE.test` which is dictated by the data generating mechanism and observed random variable values and the individual level estimated $\widehat{\text{CATE}}_i$. For a scenario with *nrep* number of simulation iterations, and $\widehat{\text{CATE}}_{ih}$ as the CATE estimate for person i in iteration h:

$$\text{Bias}(\text{CATE}_i) = \frac{1}{nrep} \sum_{h=1}^{nrep} \widehat{\text{CATE}}_{ih} - \text{CATE}_i \tag{3.3}$$

A density plot for average bias measures for one simulated scenario is included in Figure 4.1. The aggregated bias plots in Sections 4.2 and 4.3 display the mean (points) and standard errors (bands) of the bias metrics across *all* individuals within each scenario, i.e., aggregated mean bias in a particular scenario is given by:

$$\frac{1}{n} \sum_{i=1}^{n} \text{Bias}(\text{CATE}_i) \tag{3.4}$$

3.4.2 Coverage

Coverage is calculated as the proportion of simulation replicates where the estimated confidence interval for CATE_i contains the true parameter value for each CATE_i.

$$\frac{1}{nrep} \sum_{h=1}^{nrep} I(\text{CATE}_i \in \{\hat{L}_{ih}, \hat{U}_{ih}\}) \tag{3.5}$$

where $\hat{L}_{ih}$ and $\hat{U}_{ih}$ represent the calculated lower and upper bounds of the confidence interval for CATE_i in iteration h. We would expect a 95% confidence interval formed in the same manner under hypothetical repeated trials to capture the truth (unknown outside of simulation) 95% of the time to achieve nominal coverage. A density plot for coverage for *one* simulated scenario is included in Figure 4.2. Later, the aggregated coverage plots in Sections 4.2 and 4.3 display the mean (points) and standard errors (bands) of the coverage metrics across *all* individuals in each scenario, as described above.

3.4.3 Standard Error

Standard error estimates are computed using either standard methods for linear regression (see Section 2.3.1) or based on built in calculations described in the `grf` paper (Athey et al., 2019). The standard error is the square root of the model-based variance estimate:

$$\text{SE}(\widehat{\text{CATE}}_i) = \sqrt{\text{Var}(\widehat{\text{CATE}}_i)} \tag{3.6}$$

Since this is a simulation, empirical standard error estimates could also be recorded but we chose to report model based standard errors in order to facilitate comparisons with confidence interval coverage as discussed in Chapter 4. A density plot for standard errors for *one* simulated scenario is included in Figure 4.3. Later, the aggregated standard error plots in Sections 4.2 and 4.3 display the mean (points) and standard errors (bands) of the standard error metrics across *all* individuals within each scenario.

CHAPTER 4

Results

4.1 Results for One Scenario

To evaluate the performance of each method, we record and plot sample statistics from each simulation scenario as described in Section 3.4. Before introducing the aggregated results in Sections 4.2 and 4.3, we will describe results from one example simulation scenario.

Consider a particular simulation scenario where there are $n = 1,000$ participants, $j = 1$ Normal true effect modifier, $k = 1$ Bernoulli true effect modifier, and $p = 10$ nuisance variables under the true linear data generating mechanism (with 2,000 replicates). First, outside of the simulation loop, the covariates are generated randomly following their indicated distributions. In this example, 1,000 values of $X_1 \sim N(0,1)$ and $X_2 \sim \text{Bernoulli}(0.5)$ are generated and stored. The true CATE_i *for each individual* in the simulated trial is calculated by summing $x_1 + x_2$, resulting in $n = 1,000$ treatment effects. Nuisance variables are created in the same manner as the true effect modifiers, but given their status as nuisance variables, they are not involved in the data generating process. Instead, they are only included when we fit the model. For simplicity, these variables are also Bernoulli distributed.

In each replicate of the simulation loop, we generate the true outcome under control for all participants which follows a standard Normal distribution (with the implication that the potential outcome has expectation = 0). Then, the true response under treatment is computed by adding the control outcome value plus the treatment effect for each participant. Next, the $1,000$ participants are randomly assigned 1:1 to treatment and control groups. Following treatment assignment, the simulated trial data is then split 1:1 into testing and training sets. Assignments are set using row indexes, named `row.get` in the code chunks so that a record can be saved for each of the $nrep = 2000$ replicates. Using the row index, a vector of the true individual treatment effects in the test data set is saved as `true_CATE.test`, which will allow us to assess confidence interval coverage. Then the model is fit either with linear regression or two variants of causal forests: with default settings or with hyperparameter optimization, following the applicable procures described above in Sections 2.3.1 and 2.4.3.

4.1.1 Bias

As described in Section 3.4.1, the bias is saved for each individual in the testing data set for each simulation. Since each individual "participant" (indexed by `row.get`) has their true treatment effect assigned *outside* the simulation loop, we can generate sample statistics for bias across all 2,000 simulation replicates, where each individual will be in the holdout set of about half of the iterations per scenario. Results for the illustration example are shown in Figure 4.1. While all methods have bias centered around 0, linear regression is much more concentrated at 0 (as expected under

this linear data generating mechanism).

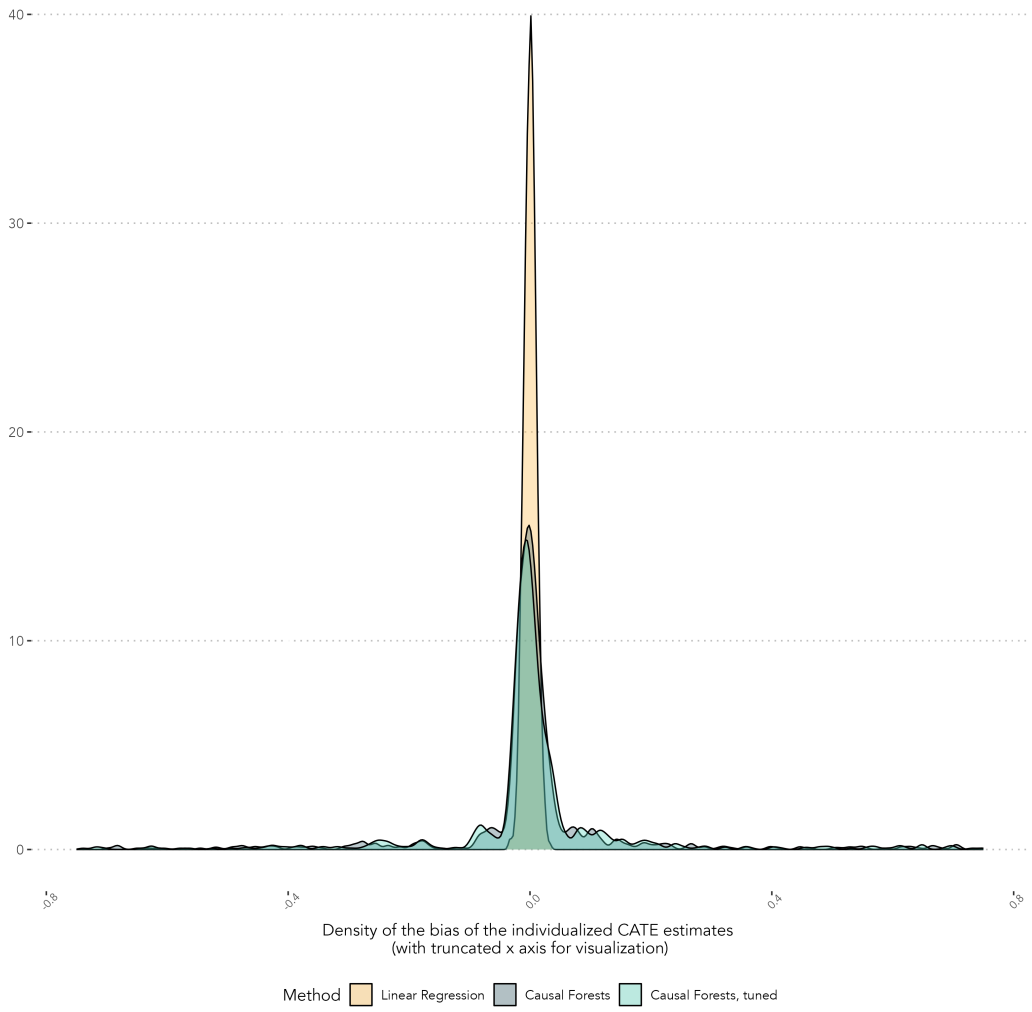

Figure 4.1: Density plots for bias of the $\widehat{\text{CATE}}_i$ in one highlighted simulation scenario. A red vertical reference line is included for 0.

4.1.2 95% Confidence Interval Coverage

In our simulation, coverage indicators (described in Section 3.4.2) are stored in one column per replicate. So, for a simulation with n = 1,000 participants and 2,000 replicates, we will have a coverage results matrix that is 1,000 rows and 2,000 columns. Based on probability theory, we would expect that a 95% confidence interval should capture the true value (which would be unknown outside of the context of a simulation) 95% of the time to achieve nominal coverage. For this reason, 1,000 row means are calculated and plotted as a density in 4.2. Note that causal forests exhibit a strong left skew regardless of tuning, indicating individuals for whom most intervals do not capture the true value.

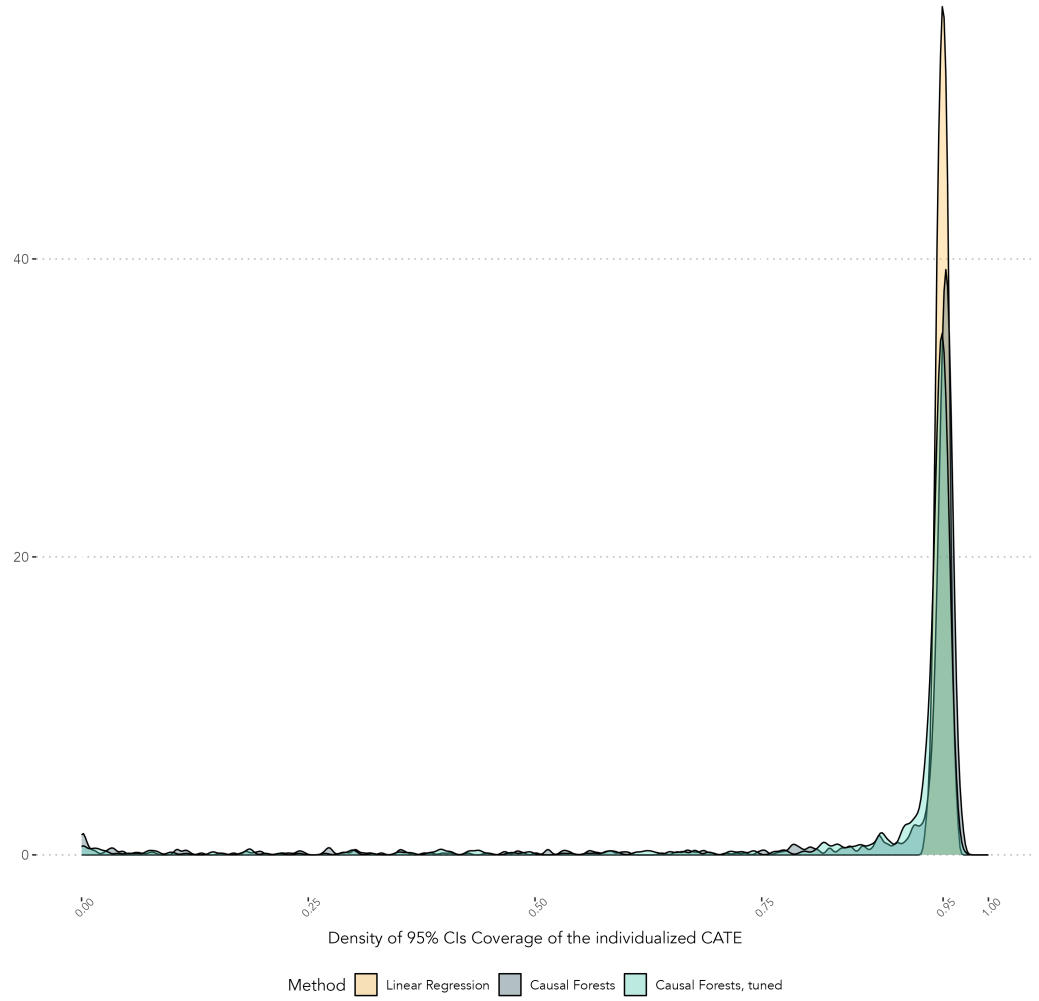

Figure 4.2: Density plots for confidence interval coverage proportions. Notice that some casual forest CIs for the CATE_i have very low coverage proportions.

4.1.3 Standard Error

As described in Section 3.4.3, the standard error is saved for each individual in the testing data set for each replicate of the simulation. Results for the example scenario are in Figure 4.3.

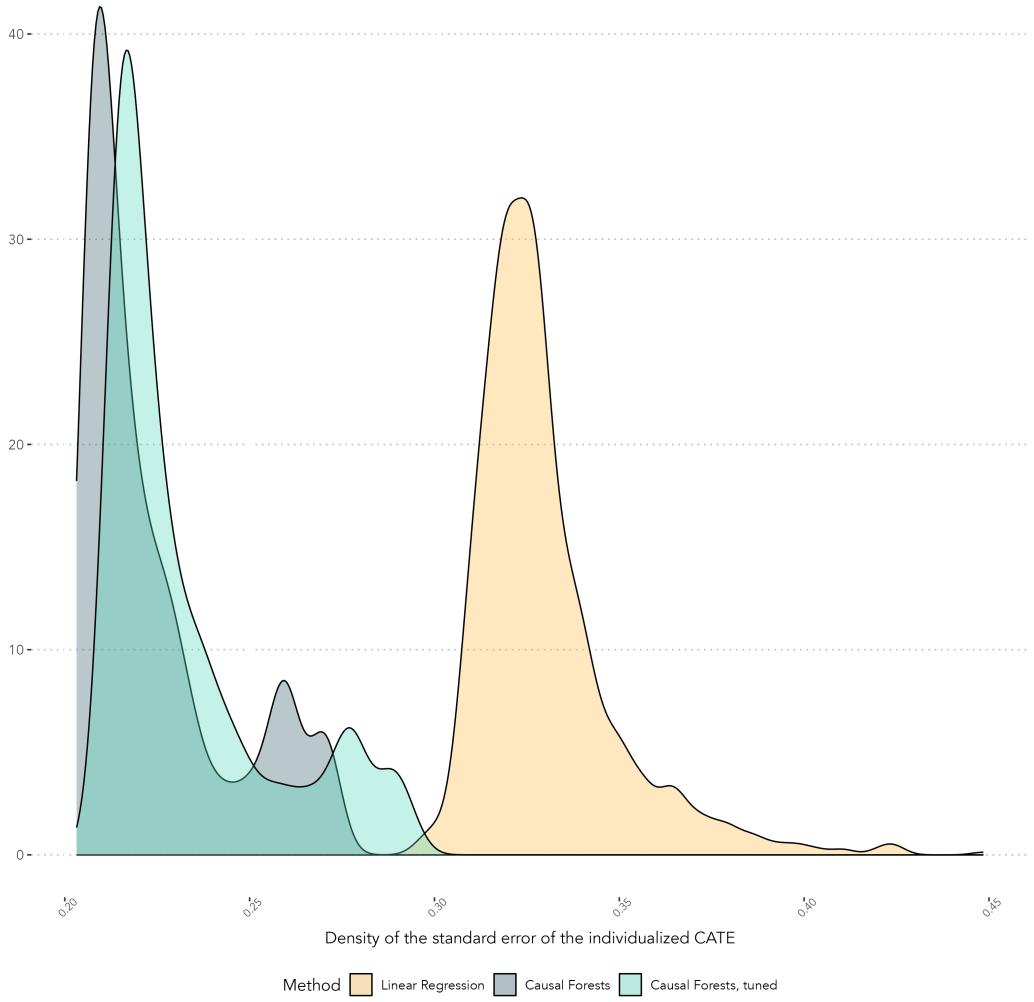

Figure 4.3: Density plots for $\text{SE}(\widehat{\text{CATE}_i})$.

4.2 Aggregated Results for the Linear DGM

4.2.1 Bias

Under a linear data generating mechanism, all methods exhibit unbiasedness. However, as the number of true effect modifiers increase, the range for the standard error of the bias of the causal forest based CATE estimators also increases as shown in Figure 4.4, with some attenuation in the standard error of the bias when limited tuning is implemented. Linear regression, when correctly specified, exhibits extremely small error bands, essentially difficult to visualize on the plot in Figure 4.4. Interestingly, we observed that the standard errors of the causal forest bias estimators decrease from the setting where there are 0 Normal effect modifiers (top left subplot Figure 4.4) compared to when there is 1 Normal effect modifier (top right subplot). Perhaps this is due to more relative ease for tree building with the presence of a single Normally distributed random variable.

Figure 4.4: Bias for the CATE under a linear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to 8 Bernoulli true effect modifiers (row wise facets) in each subplot. Labeled subplots show aggregate bias metrics for the additional combination of 0, 1, 4, or 8 Normally distributed true effect modifiers.

4.2.2 95% Confidence Interval Coverage

In the null setting (where there are 0 true effect modifiers) all three methods are near 95% average 95% CI coverage (Figure 4.5). We note that in the null case there is no true data generating mechanism (neither linear or nonlinear), but the nuisance variables are included in the linear regression model in a linear manner. OLS has the best fidelity to nominal coverage across all sample sizes (500 to $4,000$) and numbers of nuisance variables considered ($10-40$). Causal forests with hyperparameter tuning had slightly below average nominal coverage, and in most cases causal

forests with default settings are slightly above average nominal coverage. Coverage proportions over the nominal level could be more concerning if this is due to overly wide confidence intervals that are not scientifically meaningful.

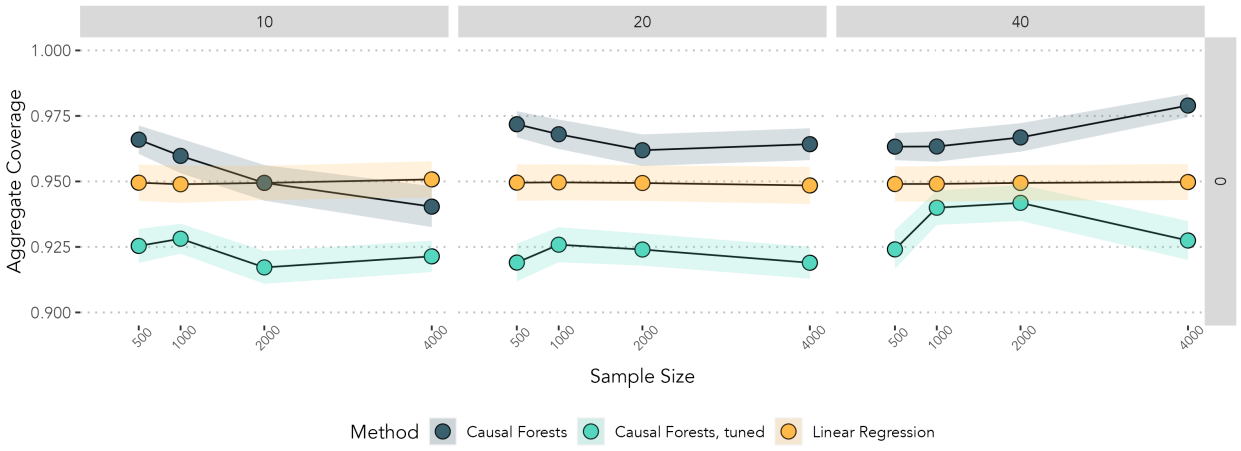

Figure 4.5: Coverage in the null setting for under a linear data generating mechanism. Situations did not include any true effect modifiers but did include 10 to 40 extraneous nuisance variables included in the model. Note the y axis range is from 0.9 to 1 in order to show detail.

In the next multi-panel plot displaying all combinations of effect modifiers and nuisance variables under a linear DGM (Figure 4.6), we notice that there are some scenarios where linear regression and causal forests are comparable in coverage, but many more scenarios where aggregate CI coverage is much lower for causal forests. The inclusion of limited hyperparameter tuning does slightly improve coverage proportions. Linear regression, as expected, exhibits nominal coverage at all sample sizes, nuisance variables, and combinations of true effect modifiers. In general for causal forests, 95% CI coverage tends to decrease when there are more nuisance variables (across the subplot columns).

When there are zero, one, or four Normal effect modifiers, causal forests tend to have aggregated coverage proportions improve as the sample size increases. We also notice a similar patter to the bias plots (Figure 4.4), where the inclusion of one Normal effect modifier tends to improve causal forest estimator performance compare to when there are zero Normal effect modifiers. Beyond the fact that the average coverage is much below 0.95 for casual forests, we note that the standard error bands for the casual forest aggregated coverage metric are very wide.

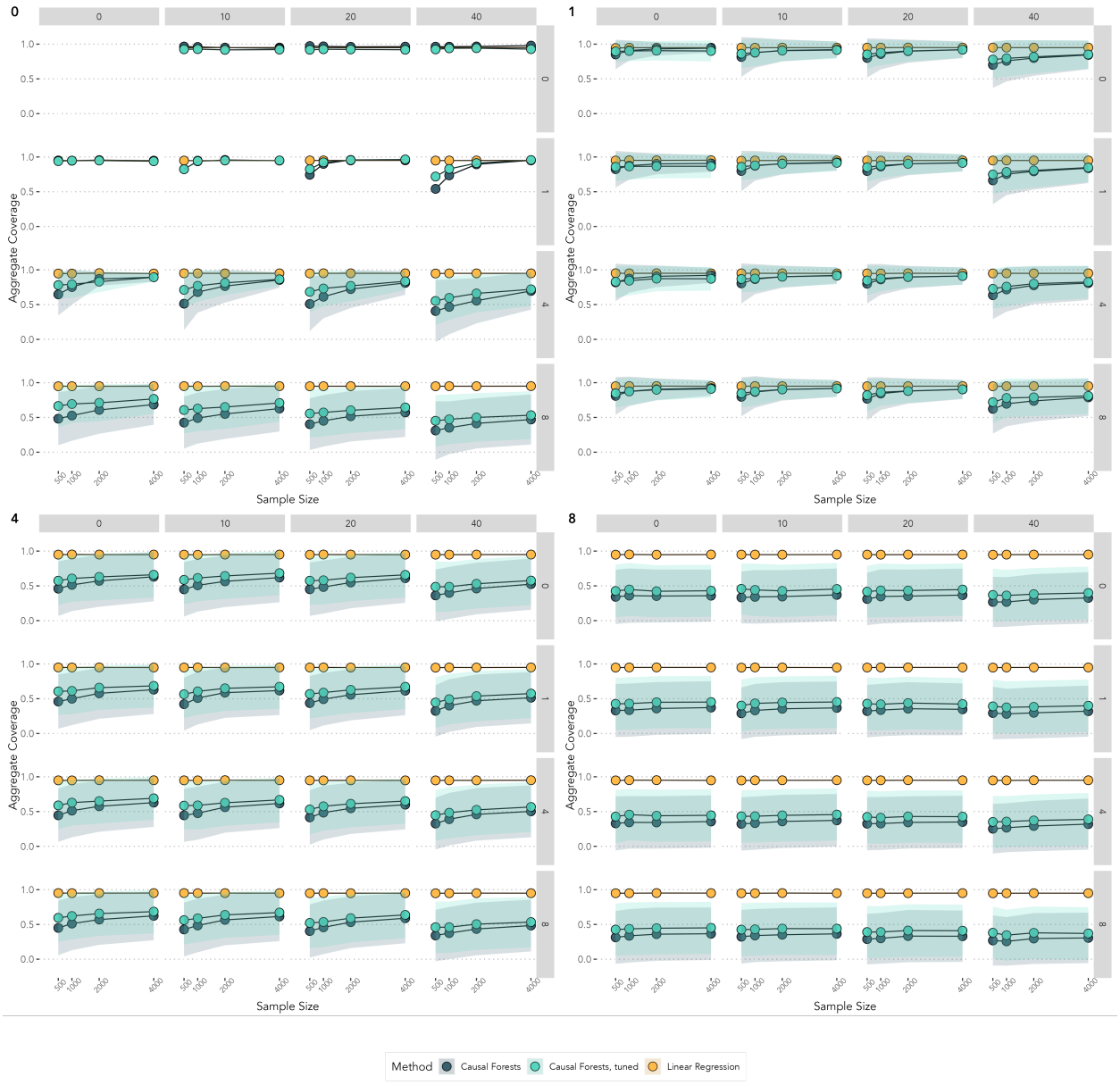

Figure 4.6: 95% confidence interval coverage under a linear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to 8 Bernoulli true effect modifiers (row wise facets) in each subplot. Labeled subplots show aggregate bias metrics for the additional combination of 0, 1, 4, or 8 Normally distributed true effect modifiers.

4.2.3 Standard Error

Estimator variance is reported on the scale of standard error. Under a *linear* data generating mechanism and 0 nuisance variables (leftmost column of each subplot) and when there are 4 or 8 Normal effect modifiers (bottom two subplots), OLS exhibits lower standard error than either causal forest method considered. When there are 0 or 1 Normal effect modifiers (top two subplots) and 0 nuisance variables, linear regression and causal forests have very similar model-

based standard errors.

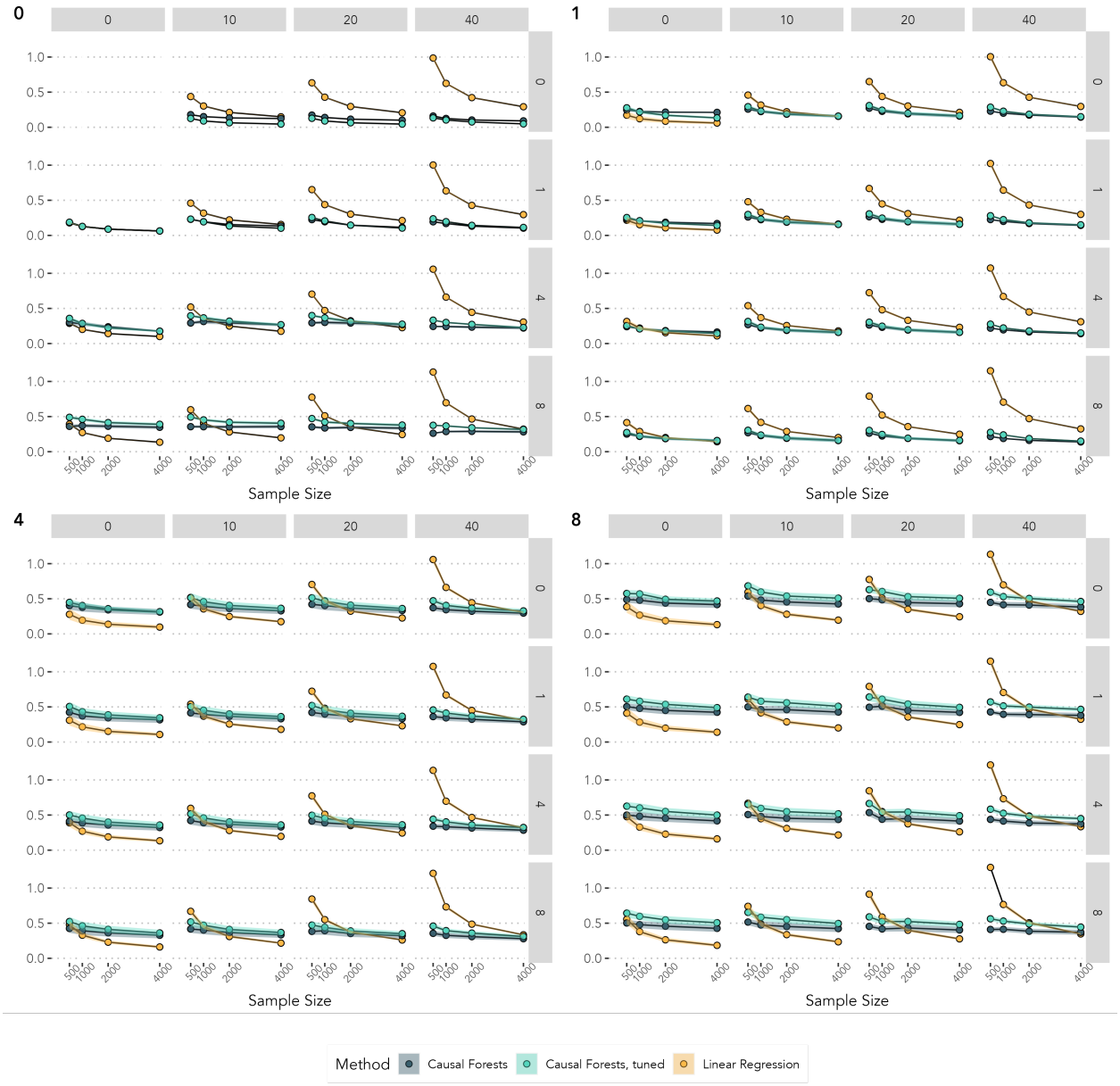

Figure 4.7: Standard error for the individual CATE estimator under a linear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to 8 Bernoulli effect modifiers (row wise facets) in each subplot. Labeled subplots show aggregate standard error metrics for the additional combination of 0, 1, 4, or 8 Normally distributed effect modifiers.

Across all 4 subplots in Figure 4.7, once there are 20 or 40 nuisance variables, linear regression loses its optimality advantage as we notice higher variance at lower sample sizes. The bottom rightmost plot represents a simulation scenario with 8 Normal effect modifiers, 8 Bernoulli effect modifiers, and 40 Bernoulli nuisance variables, which means we were fitting a model which would be vastly over parameterized based on the rule of thumb to include

no more than 1 predictor for each 10 participants. The model would have 1 intercept $+8+8+40$ base parameters $+8+8+40$ interaction terms $= 113$ total parameters for a 500 participant study with sample splitting performed. No reasonable practitioner should fit such a linear model, so we do recognize that not all scenarios considered are realistic. Finally, we will comment that although the model-based standard errors for linear regression tend to be larger, the 95% CI coverage hits the nominal level in all situations (see Figure 4.6) so these standard errors might be necessary to achieve nominal coverage.

4.3 Aggregated Results for the Nonlinear DGM

Recall that the nonlinear data generating mechanism situations did not include scenarios with Bernoulli effect modifiers, as described in Section 3.1.

4.3.1 Bias

Once again, the methods considered are unbiased. Standard errors of the bias grow immensely under a nonlinear data generating mechanism as the number of Normally distributed effect modifiers grows. Notice the the standard error bands and y axes in Figure 4.8, especially contrasting between the subplot in the middle with four Normal effect modifiers and the subplot on the right with eight Normal effect modifiers where the y axis suddenly extends from ± 2 to ± 20. As the sample size increases, the standard errors of the bias decrease for causal forests and furthermore, in many scenarios, the standard errors of the bias are smaller for tuned causal forests. The standard errors for the bias of (misspecified) linear regression remain constant across sample sizes and number of nuisance variables (they only seem to be affected by the number of Normal effect modifiers).

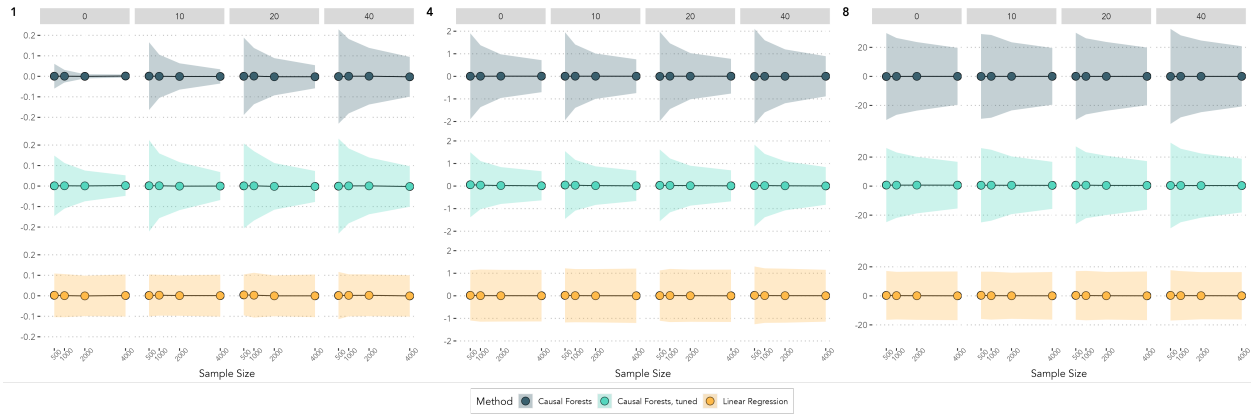

Figure 4.8: Bias for the individualized CATEs under a nonlinear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets). Labeled subplots show aggregate bias metrics for 1, 4, or 8 Normally distributed true effect modifiers. The y axis expands for the 4 and 8 effect modifiers subplots.

4.3.2 95% Confidence Interval Coverage

Simulation scenario results under a nonlinear DGM are plotted in Figure 4.9. We observe only reasonable performance in settings where there is one Normally distributed effect modifier, with the superior method difficult to determine due to overlapping standard error bands and variations across sample sizes and number of nuisance variables for the aggregated mean points that are nearest to the nominal level. The gap in average 95% CI coverage between default causal forests and causal forests that underwent hyperparameter optimization closes as more nuisance parameters are added. Since the linear model is misspecified, the aggregated CI coverage performance is quite poor: 95% CI coverage only meets the nominal level in a few limited scenarios. These include when there are more than 0 nuisance variables and only one Normally distributed effect modifier (Figure 4.9). When there is only one Normally distributed effect modifier and no nuisance variables, causal forests under the default hyperparameters achieve 95% nominal coverage on average but OLS does not. In fact, in this setting, OLS's average 95% CI coverage decreases as n increases from 500 to $4,000$.

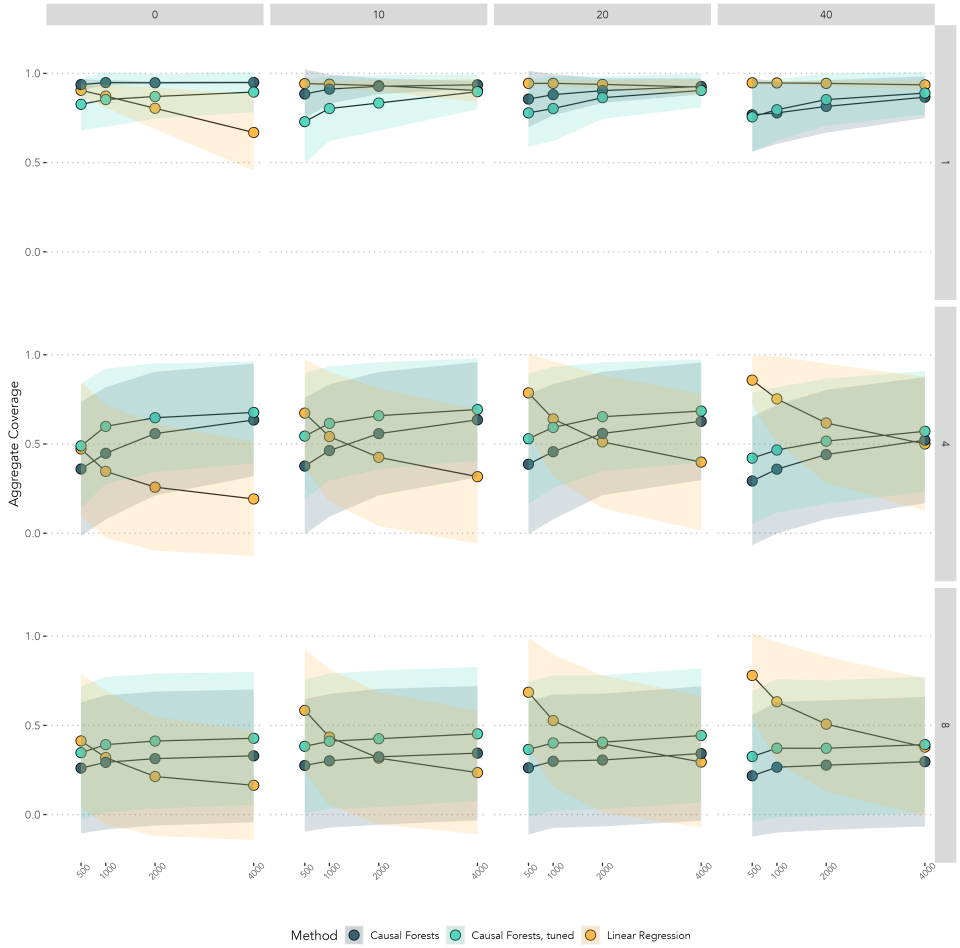

Figure 4.9: Coverage for under a nonlinear data generating mechanism with 1, 4, or 8 Normally distributed effect modifiers (top, middle, and bottom row) with 0 to 40 extraneous nuisance variables (columns).

When there are 8 Normally distributed effect modifiers (lower right subplot in Figure 4.6 and bottom row of Figure 4.9), average causal forest 95% CI coverage (whether or not hyperparameters optimization is included) fails to cross 50% regardless of the data generating mechanism in this simulation study. We note that tuning hyperparameters does improve coverage. Overall, causal forests methods tend to improve as sample size increases, but outside of simple scenarios, they seem to require more than $n = 4000$ ($n = 2000$ training data) to achieve nominal coverage.

4.3.3 Standard Error

We wondered how model misspecification for linear regression would affect the stability of the estimator's variance. As expected, we lose efficiency in a setting where linear regression is misspecified (especially with smaller sample sizes and many nuisance variables as shown in Figure 4.10). As more binary nuisance variables are included in a setting where there is nonlinear data generating mechanism (Figure 4.10) we see standard errors from linear regression increase until they are much higher than causal forest's CATE estimator's standard error (especially apparent in smaller sample sizes).

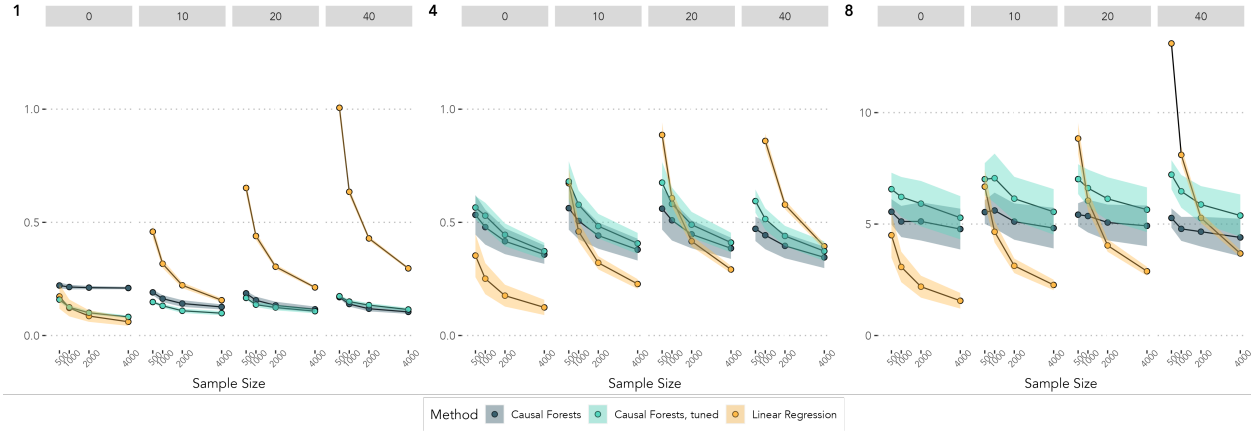

Figure 4.10: Standard error for the CATE estimator under a nonlinear data generating mechanism with 0 to 40 extraneous nuisance variables (column wise facets) and 0 to Bernoulli effect modifiers. Labeled subplots show aggregate standard error metrics for 1, 4, or 8 Normal effect modifiers.

Notice the y axis has has a span of 0 to 14 in the right subplot of Figure 4.10 reflecting the fact that including 8 Normal effect modifiers under a nonlinear data generating mechanism greatly increases the variances of the CATE estimator. Interestingly, for causal forests the standard errors remains relatively similar across the 0 to 40 nuisance variables and the simulated trial enrollment, although we do note that standard error decreases slightly (as expected) when the trial enrollment is larger.

CHAPTER 5

Discussion

For the discussion, we revisit the questions posed in Section 1.4:

1. How can we reliably detect HTE in clinical trials?
2. What guidance can we offer to the practitioner in terms of what sample size is necessary to expect valid performance of different estimators?
3. What guidance can we offer to the practitioner in terms of which method should be chosen to get better performance at smaller sample sizes?

First, how can we reliably detect HTE in clinical trials? We think of reliability in two dimensions - the ability to detect the *absence* of HTE when it is truly absent (specificity). Conversely, we also view reliability as being able to detect HTE when it is present (sensitivity). From Figure 4.5 which covers the situations with 0 true treatment effect modifiers and 10, 20, or 40 linearly included nuisance variables, notice that linear regression achieves nominal CI coverage at all sample sizes and nuisance variables considered. Causal forests with hyperparameter tuning tended to moderately undershoot 95% CI coverage while casual forests with default settings tended to overshoot CI coverage - which could be more concerning if confidence intervals are overly wide. For reliability in terms of specificity of detecting HTE through forming confidence intervals for $\widehat{\text{CATE}}_i$s reaching nominal coverage, linear regression performs as promised at these sample sizes and under a linear inclusion of nuisance variables in the model.

In terms of sensitivity, linear regression also performs exceptionally under a linear data generating mechanism. However, a linear DGM is likely not plausible in most real life situations, so many practitioners will be more interested in comparing methods under a nonlinear DGM. In the nonlinear setting, causal forests tend to outperform linear regression in 95% CI coverage, especially as sample sizes grow (see Figure 4.9), unless there is just one Normally distributed effect modifier and some nuisance variables, in which case linear regression has comparable CI coverage (top row of Figure 4.9). In terms of bias, all methods evaluated are unbiased, but causal forests have much larger SE(Bias). As expected, as sample size increases, SE(Bias) decreases - except for misspecified linear regression which has relatively constant SE(Bias) width across N (see Figure 4.8).

Practitioners may also want to consider how many treatment effect modifiers an expert could reasonably evaluate in these analyses. If there are more than 4 and there are reasons to expect violations to assumptions of linearity, none of the methods considered seem to reliably deliver nominal CI coverage under the settings considered. This is concerning, as estimating individualized treatment effects with a moderate sample size and using models that consider many potential effect modifiers is becoming a more common practice.

For the second question, in terms of guidance we can offer to the practitioner for sample sizes necessary to expect valid performance of different estimators, we can recommend linear regression at all linear settings based on the results of our study. However, the plausibility of linearity is often dubious in practice. In the nonlinear settings considered in this study, the recommendation is less straightforward and will require the practitioner and study team to consider the sample size, number of covariates, and a forthright assessment of different potential subsets of covariates which may act as true effect modifiers.

Third, to address the setting restricted to a smaller sample size, we cannot recommend casual forests without significant caveats. We can say that if the sample size is smaller ($n = 500$ to $1,000$, see Figures 4.9), misspecified linear regression can outperform causal forest CI coverage on average, although the standard error bands overlap. Some of these limitations were outlined in the seminal work developing causal forests (Wager and Athey, 2018). This is an avenue for further active research.

Implementing limited hyperparameter tuning improves SE(Bias) and 95% CI coverage proportion as compared to causal forests with default parameters. However, tuning is computationally expensive, so we did not include tuning in every simulation replicate. It's possible that causal forests (both varieties) have lower 95% CI coverage because the model based standard errors are too small. Future work should consider a more extensive tuning procedure.

If there's reason to believe the underlying model is linear, practitioners are likely better off using linear regression. There are limited scenarios considered in which we would feel comfortable considering causal forests (with or without tuning) - specifically when there is one Normal effect modifier and ideally when there is a larger sample size. However, causal forests don't require the practitioner to specify the model form, which could be considered a positive if one does not want to impose structure, but could also be considered a negative in the sense that it generally means more data points are needed to achieve valid inference. For very large trials with sample sizes beyond those considered in this work, causal forests are likely still a prudent choice. It's also possible to build linear regression models that are robust to some types of model misspecification; practitioners may choose to include splines to accommodate departures from linearity, which were not included in this study but which would be valuable to include in future work (Harrell, 2015).

On one hand, we do find reasons to agree with the statement that many "claims about improved predictive discrimination from ML are exaggerated" (Kapoor and Narayanan, 2023) in the individualized treatment effect context. On the other hand, current literature indicates promise for causal forest methods as well as increases in their popularity (Inoue et al., 2024). For example, a survival causal forest algorithm did detect HTE in a retrospective validation analysis of a cardiac trial with a known effect modifier. The authors conclude "Carefully applied and validated predictive models hold promise in identifying heterogeneous treatment effects and are useful for hypothesis generation regarding the role of phenotypic characteristics in modifying the benefit of experimental interventions in clinical trials" (Desai et al., 2024). Overall, these methods hold much promise but should be applied carefully in trials with modest sample sizes.

5.1 Limitations

These methods would not be feasible in early phase (i.e., small enrollment) safety clinical trials. Also, due to the computational expense and timing constraints, only a limited hyperparameter tuning was considered in this simulation study. As in any simulation study, only a limited group of settings were able to be studied. Furthermore, this study did not include metaalgorithms (metalearners), which build on algorthms like random forests or Bayesian Additive Regression Trees (BART) to estimate the CATE (Künzel et al., 2019). Metalearners have been touted to have the capacity to significantly improve on the performance of causal forests (Künzel et al., 2019). Future work, including a publication with additional simulation approaches and parameters, is planned to expand on this important and interesting topic.

BIBLIOGRAPHY

Afshar, M., Graham Linck, E. J., Spicer, A. B., Rotrosen, J., Salisbury-Afshar, E. M., Sinha, P., Semler, M. W., and Churpek, M. M. (2024). Machine Learning–Driven Analysis of Individualized Treatment Effects Comparing Buprenorphine and Naltrexone in Opioid Use Disorder Relapse Prevention. *Journal of Addiction Medicine*, 18(5):511–519.

Athey, S. and Imbens, G. (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360. Publisher: Proceedings of the National Academy of Sciences.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178. Publisher: Institute of Mathematical Statistics.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brookes, S. T., Whitely, E., Egger, M., Smith, G. D., Mulheran, P. A., and Peters, T. J. (2004). Subgroup Analyses in Randomized Trials: Risks of Subgroup-Specific Analyses;: Power and Sample Size for the Interaction Test. *Journal of Clinical Epidemiology*, 57(3):229–236.

Collins, F. S. and Varmus, H. (2015). A New Initiative on Precision Medicine. *The New England Journal of Medicine*, 372(9):793–795.

Desai, R. J., Glynn, R. J., Solomon, S. D., Claggett, B., Wang, S. V., and Vaduganathan, M. (2024). Individualized Treatment Effect Prediction with Machine Learning — Salient Considerations. *NEJM Evidence*, 3(4):EVIDoa2300041. Publisher: Massachusetts Medical Society.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., and Granger, C. B. (2015). *Fundamentals of Clinical Trials*. Springer International Publishing AG, Cham, SWITZERLAND.

Goligher, E. C., McNamee, J. J., Dianti, J., Fan, E., Ferguson, N. D., Slutsky, A. S., and McAuley, D. F. (2023). Heterogeneous Treatment Effects of Extracorporeal CO2 Removal in Acute Hypoxemic Respiratory Failure. *American Journal of Respiratory and Critical Care Medicine*, 208(6):739–742. Publisher: American Thoracic Society - AJRCCM.

Harrell, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer International Publishing, Cham.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer, second edition.

Hernan, M. A. (2004). A Definition of Causal Effect for Epidemiological Research. *Journal of Epidemiology & Community Health*, 58(4):265–272. Publisher: BMJ Publishing Group Ltd.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Chapman and Hall/CRC. Hernan MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Hoogland, J., IntHout, J., Belias, M., Rovers, M. M., Riley, R. D., E. Harrell Jr, F., Moons, K. G. M., Debray, T. P. A., and Reitsma, J. B. (2021). A Tutorial on Individualized Treatment Effect Prediction from Randomized Trials with a Binary Endpoint. *Statistics in Medicine*, 40(26):5961–5981. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9154.

Hudgens, M. G. and Halloran, M. E. (2008). Toward Causal Inference with Interference. *Journal of the American Statistical Association*, 103(482):832–842. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Imbens, G. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction / Guido W. Imbens, Stanford University, Donald B. Rubin, Harvard University.* Cambridge University Press, Cambridge.

Inoue, K., Adomi, M., Efthimiou, O., Komura, T., Omae, K., Onishi, A., Tsutsumi, Y., Fujii, T., Kondo, N., and Furukawa, T. A. (2024). Machine Learning Approaches to Evaluate Heterogeneous Treatment Effects in Randomized Controlled Trials: A Scoping Review. *Journal of Clinical Epidemiology*, 176:111538. Publisher: Elsevier.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R*. Springer, second edition.

Kapoor, S. and Narayanan, A. (2023). Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns*, 4(9):100804.

Kent, D. M., Rothwell, P. M., Ioannidis, J. P., Altman, D. G., and Hayward, R. A. (2010). Assessing and Reporting Heterogeneity in Treatment Effects in Clinical Trials: A Proposal. *Trials*, 11:1–11.

Kent, D. M., Steyerberg, E., and Klaveren, D. v. (2018). Personalized Evidence Based Medicine: Predictive Approaches to Heterogeneous Treatment Effects. *BMJ*, 363:DOI: 10.1136/bmj.k4245. Publisher: British Medical Journal Publishing Group Section: Clinical Review.

Kosorok, M. R. and Laber, E. B. (2019). Precision Medicine. *Annual Review of Statistics and Its Application*, 6(1):263–286.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Piantadosi, S. (2017). *Clinical Trials: A Methodologic Perspective*. Wiley Series in Probability and Statistics. Wiley.

Puntanen, S. and Styan, G. P. H. (1989). The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator. *The American Statistician*, 43(3):153–161. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Seitz, K. P., Spicer, A. B., Casey, J. D., Buell, K. G., Qian, E. T., Graham Linck, E. J., Driver, B. E., Self, W. H., Ginde, A. A., Trent, S. A., Gandotra, S., Smith, L. M., Page, D. B., Vonderhaar, D. J., West, J. R., Joffe, A. M., Doerschug, K. C., Hughes, C. G., Whitson, M. R., Prekker, M. E., Rice, T. W., Sinha, P., Semler, M. W., and Churpek, M. M. (2023). Individualized Treatment Effects of Bougie versus Stylet for Tracheal Intubation in Critical Illness. *American journal of respiratory and critical care medicine*, 207(12):1602–1611.

Varadhan, R., Segal, J. B., Boyd, C. M., Wu, A. W., and Weiss, C. O. (2013). A Framework for the Analysis of Heterogeneity of Treatment Effect in Patient-Centered Outcomes Research. *Journal of Clinical Epidemiology*, 66(8):818–825.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Willke, R. J., Zheng, Z., Subedi, P., Althin, R., and Mullins, C. D. (2012). From Concepts, Theory, and Evidence of Heterogeneity of Treatment Effects to Methodological Approaches: A Primer. *BMC Medical Research Methodology*, 12:1–12. Publisher: BioMed Central Ltd.

Yarnell, C. J. and Fralick, M. (2024). Heterogeneity of Treatment Effect — An Evolution in Subgroup Analysis. *NEJM Evidence*, 3(4):EVIDe2400054.