Multi-Calibrated Heterogeneous Treatment Effect Estimation

Yuan Yuan[*]
Daniels School of Business
Purdue University
yuanyuan@purdue.edu

Abstract

Accurately estimating heterogeneous treatment effects (HTE) is various domains such as personalized medicine, marketing campaigns, public policy and educational interventions. In this paper, we introduce an approach motivated by the concept of 'multi-calibration' from the recent advancements in machine learning. Our method seeks to reduce the bias in treatment effect estimates across a diverse set of subpopulations, which are identifiable by a model family, such as decision trees, linear models, and deep neural networks. A key benefit of our approach is its ability to provide HTE estimates with an upper limit on the estimation error for the Conditional Average Treatment Effect (CATE) for any subgroup determined by a model classifier. Specifically, for decision trees, our algorithm implements an iterative boosting technique, while for linear models and deep neural networks, it focuses on optimizing certain objective functions. To validate the efficacy of our approach, we apply it to synthetic datasets and demonstrate that it achieves enhanced multi-calibration and more equitable outcomes in subsequent decisions, surpassing the performance of other prevalent HTE estimation methods.

1 Introduction

Randomized controlled trials (RCTs), also known as controlled experiments or A/B testing, are widely recognized as a gold standard for estimating the causal impact of policies, medical treatments, computer algorithms, or business strategies (Imbens & Rubin, 2015). The average treatment effect is a key metric used to describe the impact of such interventions. With the increasing sample sizes of modern RCTs, it has become possible to analyze heterogeneous treatment effects, which are the average treatment effects conditional on subgroups determined by specific covariates (Athey & Imbens, 2016; Athey et al., 2019). This analysis enables the customization of downstream decisions, such as in personalized medicine or targeted advertising.

However, there is a pressing need to enhance the precision and calibration of heterogeneous treatment effect estimation. While conditional treatment effects provide valuable insights for decision-making based on subgroups, they may not fully capture the nuances within these subgroups, leading to suboptimal decisions at the individual level. For instance, in personalized medicine, relying solely on subgroup averages might lead to less effective treatments for patients whose specific characteristics deviate from the subgroup norm (Tibshirani, 1996).

To address these challenges, we propose a new concept: $(\mathcal{G}, \epsilon)$-multicalibration. This approach aims to predict heterogeneous treatment effects in a manner that aligns with the true conditional average treatment effects across all possible partitions $g \in \mathcal{G}$, where $\mathcal{G}$ represents a model family, such as greedy decision trees with a maximum depth of three and ϵ is an error bound. Our $(\mathcal{G}, \epsilon)$-multicalibrated method offers a more accurate and flexible estimation of heterogeneous treatment

[*] The author thanks Qinqin Zou's excellent research assistance. All results are preliminary and subject to change. This version: 12/28/2023.

effects, enabling precise predictions for diverse treatment groups and supporting targeted policies based on machine learning classifiers.

This enhanced framework for estimating individual treatment effects holds significant potential for applications like personalized medicine and targeted advertising. By allowing for more accurate intervention targeting, our approach could lead to improved patient outcomes in healthcare and more effective marketing strategies in business. Moreover, a more precise estimation of heterogeneous treatment effects can contribute to greater fairness and equity, preventing potential disparities in resource allocation and treatment arising from imprecise estimations.

2 Related Work

In this section, we review the relevant literature in the context of our study, which falls into three main areas: (1) calibration and multi-calibration; (2) heterogeneous treatment effects (HTE); and (3) machine learning for causal inference.

Calibration. Calibration is a fundamental concept in statistics and machine learning, focusing on the alignment of predicted probabilities with actual observed outcomes (DeGroot & Fienberg, 1983). Recent work has extended the notion of calibration to multi-calibration (Hébert-Johnson et al., 2018), which addresses the need for calibrated predictions across different subgroups given a model family. These concepts have been applied to various settings, including classification and regression tasks, to improve the reliability and interpretability of predictions.

Heterogeneous Treatment Effects (HTE). Estimating HTE has been a key challenge in causal inference and econometrics, as it allows for more tailored policy recommendations and decision-making (Heckman et al., 2006). Recently, machine learning techniques have been employed to capture complex effect heterogeneity and estimate heterogeneous treatment effects. Athey & Imbens (2016) proposed a decision tree-based algorithm to split the population into disjoint partitions where there are large differences in average treatment effects conditional on subgroups. Wager & Athey (2018); Athey et al. (2019) extend this tree approach to random forests, which show better asymptotic properties. Künzel et al. (2019) propose the 'x-learner' approach, which addresses the conventional issue of using separate prediction models for treatment and control groups without considering the propensity of treatment (i.e., the probability of treatment given covariates).

Machine Learning for Causal Inference The intersection of machine learning and causal inference has received increasing attention in recent years, with a focus on leveraging powerful machine learning techniques to improve causal effect estimation (Athey, 2017). This includes the development of causal trees and forests (Athey et al., 2019), which extend classical decision trees and random forests to capture treatment effect heterogeneity. Other methods, such as doubly robust estimation (Bang & Robins, 2005) and targeted maximum likelihood estimation (Van Der Laan & Rubin, 2006), have also been proposed to address the challenges of causal inference in high-dimensional settings.

Our study builds upon these areas of research by proposing the $(\mathcal{G}, \epsilon)$-multicalibrated approach for estimating individual treatment effects (ITE), which addresses the limitations of existing methods and offers a more flexible and accurate framework for ITE estimation. By combining the strengths of multi-calibration, heterogeneous treatment effect estimation, and machine learning for causal inference, our approach aims to contribute to the ongoing advancements in these fields.

3 Multi-calirbated individual treatment effect

Consider a super-population with samples $(Y_i(1), Y_i(0), X_i) \sim \mathcal{D}$. Here, $Y_i(1)$ and $Y_i(0)$ represent the potential outcomes for individual i under treatment and control conditions, respectively, and X_i are the associated covariates for individual i. Let $\mathcal{X}$ be the covariate space.

Let $\tau_i = Y_i(1) - Y_i(0)$ be the true individual treatment effect for each individual i. The the fundamental problem of causal inference, i.e. we can only observe one outcome for each unit ($Y_i(1)$ or $Y_i(0)$), poses challenges to accurately estimate the treatment effects. We let $\hat{\tau}(X)$ denote the estimated treatment effect given feature X. We denote $\hat{\tau}_i = \hat{\tau}(X_i)$ as the estimated treatment effect for individual i.

We propose the notion of $(\mathcal{G}, \epsilon)$-multicalibration as a measure of the precision of individual treatment effect predictions. For each possible partition of the covariate space, defined by a binary classifier $g \in \mathcal{G}$, where $\mathcal{G}$ is a model family that maps and $g : \mathcal{X} \to \{-1, 1\}$, and given a predetermined tolerance threshold ϵ, and the estimator $\hat{\tau}$, we have the definition of $(\mathcal{G}, \epsilon)$-multicalibration.

Definition 1 ($(\mathcal{G}, \epsilon)$ multicalibration)**.** *An estimation function of individual treatment effect $\hat{\tau}$ is said to be $(\mathcal{G}, \epsilon)$ multicalibrated if, for each classifier g in model family $\mathcal{G}$ and $g : \mathcal{X} \to \{-1, 1\}$, we have:*

$$R(g, \hat{\tau}) = \mathbb{E}_{\mathcal{D}}[g(X_i)(\tau_i - \hat{\tau}_i)] < \epsilon \tag{1}$$

Note that because g is a binary classifier we have

$$\begin{aligned} R(g) = {} & (\tau(g, 1) - \mathbb{E}_{\mathcal{D}}[\hat{\tau}_i | g(X_i) = 1])\, \mathbb{P}[g(X_i) = 1] \\ & - (\tau(g, -1) - \mathbb{E}_{\mathcal{D}}[\hat{\tau}_i | g(X_i) = -1])\, \mathbb{P}[g(X_i) = -1], \end{aligned} \tag{2}$$

where $\tau(g, 1) = \mathbb{E}_{\mathcal{D}}[\tau_i | g(X_i) = 1]$ and $\tau(g, -1) = \mathbb{E}_{\mathcal{D}}[\tau_i | g(X_i) = -1]$.

Intuitively, the largest possible value for R would occur under the condition that $g(X_i)$ is $+1$ when $\tau_i - \hat{\tau}_i > 0$ and $g(X_i)$ is -1 otherwise. This definition is equivalent to $R(g^*, \hat{\tau}) < \epsilon$ where $g^* \in \arg\sup_{g \in \mathcal{G}} R(g, \hat{\tau})$.

In empirical settings with sample size N, a major causal inference challenge is that we do not know the true effect τ_i for any individual. This is why we run randomized experiments (or A/B tests) to estimate treatment effects. We consider a randomized experiment that assigns individuals either to treatment ($Z_i = 1$) or control ($Z_i = 0$). $\mathbf{Z}$ represents random assignment and follows a probability distribution $\mathbb{P}_{\mathbf{Z}}$. We assume no interference and thus $y_i(1), y_i(0), X_i | Z_i$ (this is called Stable Unit Treatment Values Assumption or "SUTVA"). $Y_i = Z_i y_i(1) + (1 - Z_i) y_i(0)$, which means the observed outcome. Note that here only X_i, Y_i, and Z_i are known, but the potential outcomes are unknown.

As $\tau(g, 1)$ and $\tau(g, -1)$ are not observed, we compute the empirical risk given the finite sample:

$$\tilde{R}(g) = \frac{1}{N} \left(\sum_{i=1}^{N} (\tilde{\tau}(g, 1) - \hat{\tau}_i)\, \mathbb{1}[g(X_i) = 1] - \sum_{i=1}^{N} (\tilde{\tau}(g, -1) - \hat{\tau}_i)\, \mathbb{1}[g(X_i) = -1] \right) \tag{3}$$

where $\tilde{\tau}(g, m)$, where $m = +1$ or -1, is the sample estimation (e.g., difference in means) for the average treatment effect given all samples where g classifies to label m.

In our study, we focus on the Horvitz–Thompson Estimator[2], that is:

$$\tilde{\tau}(g, m) = \frac{\sum_{i=1}^{N} Y_i Z_i \mathbb{1}[g(X_i) = m \text{ and } Z_i = 1]}{p \sum_{i=1}^{N} Y_i Z_i \mathbb{1}[g(X_i) = m]} - \frac{\sum_{i=1}^{N} Y_i (1 - Z_i) 1[g(X_i) = m \text{ and } Z_i = 0]}{(1 - p) \sum_{i=1}^{N} Y_i (1 - Z_i) 1[g(X_i) = m]} \tag{4}$$

If we can ensure that for $g^* \in \arg\sup_{g \in \mathcal{G}} R(g, \hat{\tau})$, $R(g^*) < \epsilon$, and the empirical satisfaction $\tilde{R}(g) < \epsilon$, then our method is successful in achieving $(\mathcal{G}, \epsilon)$-multicalibration.

4 Multicalibration algorithm

4.1 Boosting algorithm

Next, we present a boosting algorithm that finds a solution satisfying the $(\mathcal{G}, \epsilon)$-multicalibration condition, as discussed in the previous section. We start with an initial individual treatment effect estimator $\hat{\tau}^0(\cdot)$, where $\hat{\tau}_i^0$ is the individual treatment effect estimation for unit i. The algorithm is shown in Algorithm 1.

[2] In the causal inference literature empirically it is a way common way to replace this Horvitz–Thompson type estimator with Hájek estimator, i.e. $\tilde{\tau}(g, m) = \frac{\sum_{i=1}^{N} Y_i Z_i \mathbb{1}[g(X_i) = m \text{ and } Z_i = 1]}{\sum_{i=1}^{N} \mathbb{1}[g(X_i) = m \text{ and } Z_i = 1]} - \frac{\sum_{i=1}^{N} Y_i (1 - Z_i) \mathbb{1}[g(X_i) = m \text{ and } Z_i = 0]}{\sum_{i=1}^{N} \mathbb{1}[g(X_i) = m \text{ and } Z_i = 1]}$.

Algorithm 1 Multi-Calibrated Boosting Algorithm

```
Require: $X_i$, $Y_i$, $Z_i$: training data
Require: $\epsilon$: expected value for $\epsilon$
Require: $\alpha$: learning rate
Initialize $\hat{\tau}_i^0 = \frac{\sum_i Y_i Z_i}{\sum_i Z_i} - \frac{\sum_i Y_i(1-Z_i)}{\sum_i (1-Z_i)}$ for all $i$, {Compute the average treatment effect}
for $t = 1, 2, \cdots$ do
    Find $g^t, m^t = \arg\max_{g \in \mathcal{G}} \tilde{R}(g)$ {Find a mapping $g$ that maximizes the absolute difference between observed and predicted values}
    if $\tilde{R}(g^t) < \epsilon$ then
        $\hat{\tau}^* = \hat{\tau}^t$
        Break {Check for convergence: if the absolute difference weighted by the probability is below the tolerance, terminate}
    end if
    Set $d_{+1}^t = \frac{\sum_{i=1}^N (\tilde{\tau}(g^t,+1)-\hat{\tau}_i)\mathbb{1}[g^t(X_i)=1]}{\sum_{i=1}^N \mathbb{1}[g^t(X_i)=1]}$
    Set $d_{-1}^t = \frac{\sum_{i=1}^N (\tilde{\tau}(g^t,-1)-\hat{\tau}_i)\mathbb{1}[g^t(X_i)=-1]}{\sum_{i=1}^N \mathbb{1}[g^t(X_i)=-1]}$ {Compute the expected difference between observed and predicted values for the chosen group}
    Update $\hat{\tau}_i^{t+1} = \begin{cases} \hat{\tau}_i^t + \alpha d_{+1}^t & \text{if } g^t(X_i) = +1 \\ \hat{\tau}_i^t + \alpha d_{-1}^t & \text{if } g^t(X_i) = -1 \end{cases}$ {Update the predicted values for the next iteration}
end for
return $\hat{\tau}_i^*$
```

This boosting algorithm iteratively refines the treatment effect estimation, ensuring that the $(\mathcal{G}, \epsilon)$-multicalibration condition is satisfied. If the algorithm converges, it returns the current $\hat{\tau}_i^t$; otherwise, it indicates that the condition cannot be satisfied.

Theorem 1 (Convergence)**.** *With $y_i(1)$ and $y_i(0)$ bounded in $[-B, B]$, the Multi-Calibrated Boosting Algorithm converges within $O(1/\epsilon^2)$.*

Proof. We first set a loss function that is denoted by $\mathcal{L}$.

Let $\hat{\boldsymbol{\tau}}$ be the N-dimensional vector of estimated treatment effects. Let $\boldsymbol{\gamma}$ be another N-dimensional vector where $\gamma_i = Y_i(Z_i/p - (1 - Z_i)/(1 - p))$, note that $\mathbb{E}[\gamma_i] = \tau_i$.

$$\mathcal{L} = \frac{1}{N} \|\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}}\|_2^2.$$

When initialized, $\mathcal{L}^0 = \frac{1}{N}(B + \bar{\tau})^2$, where $\bar{\tau}$ is the average treatment effect estimation from the sample.

Assume that after iteration t (with estimation $\hat{\tau}^t$, and we use $\boldsymbol{\tau}$ to denote the N-dimensional vector of the training set), the algorithm has not converged, which means that there exists $g \in \mathcal{G}$ such that $\tilde{R}(g) > \epsilon$.

Let $g^{t+1} = \arg\sup_{g \in \mathcal{G}} |\tilde{R}(g)|$, and $\boldsymbol{G}$ be the N-dimensional vector where $G_i = g^{t+1}(X_i)$, and we aim to find $\hat{\tau}^{t+1}$ to maximally reduce the loss (and accordingly we define the vector$\hat{\tau}^{t+1}$).

Let $\boldsymbol{d}$ be an N-dimensional vector such that $\hat{\boldsymbol{\tau}}^{t+1} = \hat{\boldsymbol{\tau}}^t + \Delta \odot \boldsymbol{G}$.

The reduction of loss would be:

$\mathcal{L}^{t+1} - \mathcal{L}^t = \frac{1}{N}\|\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}}^{t+1}\|_2^2 - \frac{1}{N}\|\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}}^t\|_2^2 = (\boldsymbol{d} \odot \boldsymbol{G})^T(2\boldsymbol{\gamma} - 2\hat{\boldsymbol{\tau}}^t - \boldsymbol{d} \odot \boldsymbol{G})$.

To maximize the reduction, the optimal $\boldsymbol{d}$ should be:

$$\boldsymbol{d}^* = (\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}})^T \boldsymbol{G}.$$

Then,

$$\mathcal{L}^{t+1} - \mathcal{L}^{t} = \frac{1}{N}\|\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}}^{t}\|_2^2.$$

Note that, by the Cauchy-Schwarz inequality:

$$\sqrt{\frac{1}{N}(\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}})^T(\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}})} = \sqrt{(\frac{1}{N}\boldsymbol{G}^T\boldsymbol{G}) \cdot (\frac{1}{N}(\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}})^T(\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}}))} \geq \tilde{R}(g) = \frac{1}{N}[\boldsymbol{G}^T(\boldsymbol{\gamma} - \hat{\boldsymbol{\tau}})] > \epsilon.$$

Then we get

$$\mathcal{L}^{t+1} - \mathcal{L}^{t} > \epsilon^2$$

which means the loss reduces more than ϵ^2 per iteration. Since the loss should not be less than 0 and starts from a bounded positive value, it would converge within $O(1/\epsilon^2)$. □

We next introduce a lemma:

Lemma 1. *With poly(ϵ) samples,*

$$\sup_{g \in \mathcal{G}} |R(g) - \tilde{R}(g)| \leq \epsilon$$

This allows us to prove the sample complexity of the algorithm.

Theorem 2 (Sample complexity)**.** *With* $\text{poly}(\epsilon)$ *samples, the Multi-Calibrated Boosting Algorithm outputs $\hat{\tau}^*$ that is ($\mathcal{G}, 2\epsilon$) multi-calibrated.*

Proof. Since by Lemma 1, $\sup_{g \in \mathcal{G}} |R(g) - \tilde{R}(g)| \leq \epsilon$, thus

$$\sup_{g \in \mathcal{G}} |R(g)| \leq \epsilon + \sup_{g \in \mathcal{G}} |\tilde{R}(g)| < 2\epsilon$$

Therefore, $\hat{\tau}^*$ that is ($\mathcal{G}, 2\epsilon$) multi-calibrated. □

4.2 General binary classifiers

In addition to the boosting tree algorithm, we also consider general binary classifiers, denoted by $q(X_i) \in [0, 1]$. We can rewrite the expectation as follows:

$$\begin{aligned}\mathbb{E}[g(X_i)(\tilde{\tau}_i - \hat{\tau}_i)] &= \mathbb{E}[(\tilde{\tau}_i - \hat{\tau}_i)|g(X_i) = +1]\mathbb{P}[g(X_i) = +1] - \mathbb{E}[(\tilde{\tau}_i - \hat{\tau}_i)|g(X_i) = -1]\mathbb{P}[g(X_i) = -1] \\ &= \sum_{i=1}^{N}(2q(X_i) - 1)(Y_i(\frac{Z_i}{p} - \frac{1 - Z_i}{1 - p}) - \hat{\tau})\end{aligned} \tag{5}$$

That is, finding the supremum of $\mathbb{E}[g(X_i)(\tilde{\tau}_i - \hat{\tau}_i)]$ over all possible q is equivalent to finding the classifier where the outcome is $\tilde{\tau}_i - \hat{\tau}_i = (\frac{Z_i}{p} - \frac{1-Z_i}{1-p})Y_i - \hat{\tau}_i$.

This observation provides a way to adapt any binary classifier based on loss optimization for the purpose of multi-calibration. By optimizing the classifier to predict the outcome $(\tilde{\tau}_i - \hat{\tau}_i)$, we can find a classifier that maximizes the expected difference between the estimated treatment effect and the true treatment effect. This allows us to leverage the power of various binary classifiers to achieve multi-calibrated estimators.

In practice, we use stochastic gradient descents to find the q that leads to the greatest $R(g)$, update $\hat{tau}_i$ as in the gradient boosting algorithm, and do this iteratively until convergence.

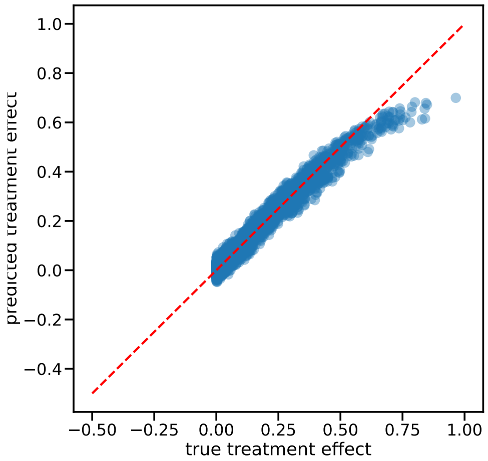

Figure 1: Scatter plot of units' individual true effects versus predicted effects. The red line has a slope of 1.

5 Experiments

5.1 Setup

We simulated a dataset with 5,000 observations to examine the causal effects of a binary treatment variable. The covariate matrix, denoted as X_i, was generated by drawing random samples from a uniform distribution between 0 and 1 for each of the five covariates $(X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5})$ in the dataset, $X_{ik} \sim$ Unif$[0, 1]$ where $k = 1, 2, \cdots, 5$.

We then defined two potential outcomes, $Y_i(1)$ and $Y_i(0)$, to represent the outcomes under treatment and control conditions, respectively. The potential outcome under treatment, Y_{1i}, was calculated as the product of the first three covariates $(X_{i1}X_{i2}X_{i3})$ plus the product of the last two covariates $(X_{i4}X_{i5})$: $Y_i(1) = X_{i1}X_{i2}X_{i3} + X_{i4}X_{i5}$.

The potential outcome under control, $Y_i(0)$, was defined as the product of the last two covariates $(X_{i4}X_{i5})$: $Y_i(0) = X_{i4}X_{i5}$.

The treatment assignment variable, Z_i, was generated as a binary variable with values 0 or 1, with equal probability, to simulate random assignment to treatment or control groups: $Z_i \sim$ Bern(0.5).

Finally, the observed outcome, Y_i, was determined by the treatment assignment, such that if an individual was assigned to the treatment group $(Z_i = 1)$, their observed outcome was Y_{1i}, and if they were assigned to the control group $(Z_i = 0)$, their observed outcome was Y_{0i}

$$Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i) \tag{6}$$

This setup allows for the estimation of causal effects by comparing outcomes between the treated and control groups while accounting for potential confounders captured by the covariates.

5.2 Preliminary Results

We use a decision tree with a maximum depth of 3, with a learning rate of 0.4 and iteration times of 100, as an example, we present the prediction results in Fig 1. The red line indicates the line with a slope of 1. As observed in the figure, most dots (each indicating a unit) have a predicted individual treatment effect close to their true effect. There are only a few points that have large treatment effects where we slightly underestimate the individual effect. This is because our boosting algorithm tends to avoid identifying subgroups that contain a small number of units.

References

Athey, Susan. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

Athey, Susan and Imbens, Guido. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Athey, Susan, Tibshirani, Julie, and Wager, Stefan. Generalized random forests. 2019.

Bang, Heejung and Robins, James M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

DeGroot, Morris H and Fienberg, Stephen E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Hébert-Johnson, Ursula, Kim, Michael, Reingold, Omer, and Rothblum, Guy. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Heckman, James J, Urzua, Sergio, and Vytlacil, Edward. Understanding instrumental variables in models with essential heterogeneity. *The review of economics and statistics*, 88(3):389–432, 2006.

Imbens, Guido W and Rubin, Donald B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Künzel, Sören R, Sekhon, Jasjeet S, Bickel, Peter J, and Yu, Bin. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Van Der Laan, Mark J and Rubin, Daniel. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Wager, Stefan and Athey, Susan. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.