A Survey on Heterogenous Treatment Effects In the Era of Machine Learning

Mónica Robles Fontán, Carmen B. Rodríguez, George Sawyer

May 6th, 2024

Abstract

Examining treatment effect heterogeneity constitutes an essential component of precision medicine, which aids in identifying optimal treatment interventions and creating tailored health programs. This shift toward patient-centered healthcare has resulted in the development of statistical techniques motivated by randomized experiments with extensions to observational studies, which are increasingly common due to Electronic Health records (EHR) and large data repositories. This report surveys recent works in causal inference that target heterogeneous treatment effects (HTEs) estimands. We first provide an overview of the identification assumptions needed to estimate HTEs and then classify current methodologies for HTE estimation as non-algorithmic, existing or modified machine learning (ML) algorithms, and algorithmically agnostic methods.

1 Introduction

In making healthcare decisions and providing treatment, providers commonly rely on evidence-based guidelines that combine average treatment effects (ATEs) from randomized clinical trials (RCTs). However, RCTs do not account for the nuances of individual patients or subgroups of patients' characteristics, which might impact outcomes since these population averages could potentially average out treatment benefits and harms [1]. In recent years, precision or personalized medicine has become very popular because it recognizes that individuals may respond differently to the same treatment due to differences in their underlying biology. Thus, identifying the most effective treatments for an individual or subgroup is prioritized [2]. To that end, in a RCT we could construct a model from the data by incorporating an interaction term between the treatment exposure and predetermined baseline covariates that identify subgroups (e.g., age groups, gender, race/ethnicity). This interaction term directly measures the extent to which the treatment effect is *modified* by each covariate. However, when the population is broken down into subgroups, it can be unclear which subgroup level to use as the reference class (i.e., the reference class problem) [3]. This significant issue in applying group results to individual treatment choices stems from the fact that each patient has countless characteristics, allowing them to fit into an unlimited number of subgroups [3, 4].

In recent years, the availability of large registries and electronic health records has expanded the range of available observational data sources. This development has created new possibilities to move from traditional methods, such as obtaining ATEs from RCTs and instead evaluating heterogeneity in treatment effects (HTE) without having to pre-specify subgroups within a population, which may be a difficult statistical task, especially for observational studies, where confounding factors must be considered. HTEs are particularly important in biomedical studies, specifically health disparities research, to provide insights about potential differential access to health care, behavioral research, and social policy [5]. Estimating HTEs in these settings may provide helpful information for patient care, future medical research, and cost-effective interventions [2, 3, 6].

The most common parameter in the causal inference framework for estimating HTEs is the conditional average treatment effect (CATE), which measures the value of the ATE parameter within a subpopulation [7]. Following the discussion about subgroups above, we can think of two ways of estimating the CATE. First, assuming the exposure assignment in the given study is unconfounded given the rest of the individual's characteristics, one might be interested in estimating the ATE separately for each subgroup level or one-step-at-a-time approach using standard nonparametric estimators or semi-parametric estimators such as the targeted maximum likelihood estimator (TMLE)[7, 8, 9]. A second approach is to define the CATE as a function of the complete set of individual characteristics since, as mentioned, an individual could fit into multiple subgroups. As such, the CATE provides the conditional mean of the treatment effect for any combination of an individual's characteristics [7, 10, 11].

This report provides an overview of methods for examining and evaluating HTEs in the current literature using both approaches for estimating CATEs. The rest of the report is organized as follows. We start with the problem setup and a review of assumptions that allow for the identification of our causal parameter of interest, the CATE. We then present

currently proposed methodologies and algorithms for efficient estimation and their contributions and limitations. Finally, through simulation studies, we compare the performance of these methods in different data settings.

2 Identification assumptions for estimating HTEs

2.1 Framework, Notation and CATE function identification

We formalize the problem in terms of the potential outcomes framework. Suppose we have $O \sim \mathcal{P} \in \mathcal{M}$, where $\mathcal{P}$ is a population distribution with $O = \{(Y_i, A_i, L_i)\}_{i=1}^n$ be an independent and identically distributed random sample including the observed outcome $Y_i \in \mathbb{R}$, the treatment assignment variable $A_i \in \{0, 1\}$, and a set of covariates for each individual $L \in \mathbb{R}^p$. We posit the existence of potential outcomes $\{Y_i(0), Y_i(1)\}$, which denote the outcomes that would have been observed given treatment assignments $A_i = 0$, and $A_i = 1$, respectively, such that by consistency $Y_i = Y(a)$, whenever $A_i = a$ for $a \in \{0, 1\}$. With this definition, the average treatment effect (ATE) is

$$\text{ATE} := \mathbb{E}[Y(1) - Y(0)]$$

and we seek to understand how treatment effects vary with the observed covariates L_i and estimate the conditional average treatment effect (CATE) function

$$\tau(l) = \mathbb{E}[Y(1) - Y(0)|L = l]$$

We differentiate the CATE from individual treatment effects $D_i = Y_i(1) - Y_i(0)$; rather, the CATE is still an average treatment effect over a targeted subgroup of individuals characterized by their covariates L_i. In order to identify $\tau(l)$, we assume *unconfoundedness* of the treatment assignment, that is,

$$A_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}|L_i$$

Additionally, we write the propensity score as $\pi(l) = \mathbb{P}(A = 1|L = l)$ and throughout the report, we will define the conditional response as $\mu_a(l) = \mathbb{E}(Y(a)|L = l)$. Under unconfoundedness, the CATE can be written as:

$$\tau(l) = \mu_1(l) - \mu_0(l)$$

and therefore, one could estimate $\tau(l)$ by fitting $\hat{\mu}_1(\cdot)$ and $\hat{\mu}_1(\cdot)$ by separate non-parametric regressions on the treated and non-treated, respectively and estimate the CATE as the difference $\hat{\tau}(l) = \hat{\mu}_1(l) - \hat{\mu}_0(l)$. Although this approach is simple and consistent (if we obtain consistent estimators for $\hat{\mu}_a(\cdot)$), it may not perform well in finite samples. In the next section, we provide an overview on ways to identify the CATE with O and discuss algorithms and methods that have been developed for efficient estimation and inference of CATE.

While the ATE and the CATE refer to the overall population, we may also be interested in defining a treatment effect for a well-defined subpopulation. The Treatment Effect on the Treated (TT) represents the average treatment effect among those who are treated, formally:

$$TT = E[Y_i(1) - Y_i(0) \mid A = 1]$$

Similarly, the Treatment Effect on the Untreated (TUT) is the average difference by treatment status for those who are not treated:

$$TUT = E[Y_i(1) - Y_i(0) \mid A = 0]$$

If treatment effects are homogeneous in our population, we would expect these quantities to be equal to each other and the overall ATE. Differences in the TT and TUT indicate treatment effect heterogeneity [5]. These treatment effects defined within subpopulations become quite useful when discussing treatment effects on the strata of the population and matching algorithms.

In the next sub-sections, we will review the necessary assumptions for identifying the CATE when we are interested in the non-functional version (i.e., approach 1 described in the introduction).

2.2 Primer on Effect Modification

How do we model treatment efficacy that differs between genders, different racial groups, or other biological features? Is it more appropriate to compute average effects in the entire population or particular subgroups where the effect of the treatment on the outcome significantly differs? These questions target effect modification, formally, when the effect of the exposure differs across strata of another variable, the effect modifier [12]. Identification of Effect Modification, in the context described, can be seen as the first step towards characterizing the relationship between two treatments [12].

2.3 Identification through Stratification

A common approach to identifying effect modification is stratifying by the variable hypothesized to cause heterogeneous treatment effects, computing the causal effect within each stratum. Say we are looking at the impact of a new drug (treatment) on the incidence of viral load suppression in a patient with HIV. We might hypothesize that the drug efficacy differs between men and women. Using stratification, we would compute a causal effect of the treatment on HIV viral load suppression separately for men ($\tau(l) = \mathbb{E}[Y(1) - Y(0)|L = 1]$) and women ($\tau(l) = \mathbb{E}[Y(1) - Y(0)|L = 0]$).

In a randomized experiment, if treatment assignment in our overall population was random and unconditional, conditional exchangeability holds in the overall population. It follows that conditional exchangeability holds in every stratum of our population. This implies that our causal risk difference for a level of strata in our subgroup would be the same as the associational risk difference in that same strata. The following line shows this equality for the female stratum ($L = 1$ for females).

$$\Pr[Y(1) = 1|L = 1] - \Pr[Y(0) = 1|L = 1] = \Pr[Y = 1|A = 1, L = 1] - \Pr[Y = 1|A = 0, L = 1]$$

To identify effect modification in a randomized trial with unconditional randomization that leads to balanced covariates, researchers can conduct a stratified analysis, computing treatment effects within each stratum. In the observational setting, we also must assume *no unmeasured confounding*; all confounders have been measured and adjusted for. Subsequently, conditional exchangeability is satisfied ($Y(a) \perp\!\!\!\perp A \mid L$). To carry out a stratification analysis in the observational setting, researchers generally use standardization (or IP weighting) to compute causal effects within each stratum L, with causal effects dependent on the distribution of L.

The average causal effect can vary across different effect measures(i.e., the risk ratio, risk difference, odds ratio), and the presence of effect modification depends on the effect measures chosen. We define Additive Effect Modification (AEM) and Multiplicative Effect Modification (MEM) below

$$AEM = E[Y(1) - Y(0)|L = 1] \neq E[Y(1) - Y(0)|L = 0]$$

$$MEM = \frac{E[Y(1)|L = 1]}{E[Y(0)|L = 1]} \neq \frac{E[Y(1)|L = 0]}{E[Y(0)|L = 0]}$$

If the causal risk difference or ratio varies across levels of L, we observe additive effect modification and/or multiplicative effect modification, respectively. If AEM and MEM indicate effect modification in different directions, we observe *qualitative effect modification.* Notably, effect modification can be present for either the risk ratio or risk difference and not the other [12]. Thus, the identification of effect modification depends on the choice of estimand.

Transportability, the extrapolation of causal effects computed in a population to another population, can cause poor generalizability in the presence of HTEs [12]. Imagine we compute MEM for two different populations in which causal identifiability assumptions are satisfied and there are heterogeneous treatment effects across gender. Given that the gender ratio differs between the two populations, the risk ratio average causal effect will differ. In the presence of heterogeneous treatment effects, the signs of the effects will be the same, but the magnitude of the effect will likely differ. That said, the causal effect within a stratum of the population may exhibit better transportability to the same stratum of another population than the overall causal effect[12].

Another concern for estimating causal effects, such as the Odds Ratio, is *noncollapsability.* An effect measure is collapsible when the population effect measure can be expressed as a weighted average of the effects measured in each stratum [12]. When considering GLMs, such as logistic regression with an Odds Ratio estimand, we will often estimate stratum-specific causal effects that are contrary to the average causal effect in the population, yielding counterintuitive results [12, 13]. *Positivity* is an additional critical assumption that is required for the identification of causal effects using stratification. Causal effects cannot be computed for subsets $L = l$ for which there are only treated or untreated people.

2.4 Identification through Matching

The goal of matching is to obtain a subset of the population for which the distribution of L is the same for both the treated and untreated [12]. In the HIV example above, we would match treated males to untreated males and treated females to untreated females to obtain *matched pairs*. Assuming conditional exchangeability in the overall population, the treated and untreated people are exchangeable in the matched population. In this simple example, we defined L as a single binary covariate (gender), but often L is a multidimensional vector of many covariates that we are adjusting for.

Notably, the matched population ensures *positivity*, as strata that do not have both untreated and treated individuals cannot be matched to anyone and, therefore, are excluded from the population. While matching ensures positivity, if very few strata can be matched across the treated and the untreated, the sample size of matched pairs can be severely reduced. This particularly becomes an issue with a high dimensional L where the probability of finding matched pairs across many covariates becomes increasingly low.

3 Current Methodologies: Strengths and Limitations

We explore the strengths and weaknesses of various methods for estimating heterogeneous treatment effects via the CATE function and effect modification. To illustrate the development trajectory, we review these methods roughly chronologically.

3.1 HTE using linear modeling

The classical approach to modeling the HTE using linear regression includes an interaction between the treatment variable and the effect modifier of interest as such:

$$Y_i = \beta_0 + \beta_1 A_i + \beta_2 L_i + \beta_3 (A_i \cdot L_i) + C_i \beta + \epsilon_i$$

where A_i is a binary treatment variable, L_i is the effect modifier of interest, and C_i is a variable matrix representing all other covariates we adjust the linear regression for. Under this simple setup, β_3, the parameter estimate of the interaction term indicates how the effect of the treatment on the outcome varies across levels of L_i. In a randomized controlled trial, β_3 is a causal effect, but in the observational setting we need to satisfy assumptions under the Neyman Pearson counterfactual framework to interpret β_3 in a causal sense.[8] Notable features of estimating the HTE via linear methods are the assumption of a linear form of the HTE and that the effect modifier does not affect the way in which the treatment was selected [8].

Often, researchers want to let the treatment interact with many covariates and relax the linearity through semi-parametric modeling. Researchers can do this by fitting a Generalize Additive Model (GAMs) or spline-based method that retains the additive structure of the model while incorporating non-linearities. Below we provide a GAM with interaction terms to test HTE:

$$Y_i = \beta_0 + \beta_1 g_t(A_i) + \beta_2 g_c(L_i) + \beta_3 g_{tc}(A_i \cdot L_i) + \epsilon_i$$

where $g_i(\cdot)$ can be any link function or smoother. Notably, the link function can differ between the main terms and interaction.

Several drawbacks to these semi-parametric GAM approach include subjectivity around the correct specification of the $g_i(\cdot)$ functions, loss of degrees of freedom, and the necessity of a continuous measure to smooth over. Often, the treatment is a binary variable that cannot be smoothed over [8]. Additionally, switching from a linear parametric model to a semi-parametric GAM does not necessarily lead to satisfaction of the ignorability and common support assumptions needed for causal inference in the observational setting [8]. To examine the treatment assignment mechanism in observational settings, we must estimate the propensity score defined as the probability of receiving treatment given L [8].

3.2 HTE Propensity Score Methods

Heterogeneous treatment effect estimation utilizing the propensity score compares differences in treatment effects between strata of the propensity score. For example, researchers want to know if taking a drug during pregnancy that suppresses HIV viral load reduces the disease transmission rates to the pregnant person's child. We also assume household income levels are related to a person's ability to obtain these treatments. In this scenario, a person's ability to obtain and stick to a treatment is directly related to their income level. Propensity score-based methods connect the probabilities that assess income levels (in this case, the propensity score) and drugs that suppress the HIV virus (the treatment). Utilizing propensity score estimation, we can answer questions such as "Who benefits more?" from a particular treatment to lower the probability of disease transmission. If those with lower income levels would benefit more from treatment (due to a smaller propensity score), we have identified a vulnerable population to target [8].

Next we describe three propensity score based methods (one parametric and two non-parametric) to estimate the interaction between treatment and the propensity score for estimating HTEs.

Stratification and Multilevel Method

1. Estimate $\pi(l) = P(a = 1 \mid L)$, the propensity score for the treated for all observations given a set of observed covariates.

2. Construct balanced propensity score strata where there are no significant differences in the average value of the covariates and the propensity scores between the treated and untreated groups. This method ignores heterogeneity within a stratum which will almost certainly be present for $L = l$. The hope is that stratification by the propensity score effectively removes the majority of biases between the treated and untreated groups, and the observations within a strata are more homogeneous than the data prior to stratification.[5, 14]).

3. Estimate $E[Y(1) - Y(0) \mid \pi(l)] = E[Y(1) \mid \pi(l)] - E[Y(0) \mid \pi(l)]$, the propensity score stratum specific treatment effects within each stratum.

4. Using the strata-specific treatment effects obtained in the prior step, evaluate trends across strata using a variance-weighted least squares regression approach. This step differs from conventional PS approaches since the main objective is to observe "systematic patterns of HTEs across strata". [5].

Typically, propensity score methods focus on removing biases caused by covariate imbalances by averaging the estimated treatment effects across strata. The goal of the *Stratification and Multilevel Method* is to identify systematic trends of HTEs across strata. Using linear regression to model patterns across strata leads to simplicity and preservation of statistical power [5].

Limitations of this approach include "imposing a form of within-group homogeneity so that treated and untreated observations are considered interchangeable within strata"[5]. This within-group homogeneity stems from grouping the propensity score strata into a limited number of strata in which we assume no heterogeneity bias within either the treated or untreated groups. A second limitation is the strong linear functional form needed to detect a pattern of heterogeneity. Since we fit a model over a limited number of observations (strata), weaker functional forms (i.e., semi parametrics) are rarely justified [5].

Matching and Smoothing Method

1. Estimate $\pi(l) = P(a = 1 \mid L)$, the propensity score for the treated for all observations given a set of observed covariates.

2. Match the treated to the untreated observations using a matching algorithm. See section 2.4 for an overview of matching and the key assumptions.

3. Plot the difference between matched pairs (of the treated and the untreated) against a "continuous representation of the propensity score". This transforms the data so that differences in matched pairs make up the observed data we model in step 4 [5].

4. Apply a non-parametric model such as a local polynomial regression or a lowess smoother to the matched differences that indicate trends in HTEs [15, 16]. This propensity score-based method produces a graphical representation of a nonparametric smoothed curve for the trend in matched differences as a function of the propensity score [5].

The *Matching and Smoothing Method* is advantageous since it does not assume homogeneity within propensity score strata like the SM method. This approach allows researchers to retain individual-level information *before* making cross-individual comparisons [5].

Smoothing Difference Method

1. Estimate $\pi(l) = P(a = 1 \mid L)$, the propensity score for the treated for all observations given a set of observed covariates.

2. Fit a non-parametric regression of the outcome on the propensity scores for the treated and untreated groups separately. For example, we could smooth using local polynomial regression [15].

3. Take the difference in the non-parametric regression line between the treated and untreated at different levels of the propensity score (using a grid of values over the common support) [5].

The Smoothing Difference Method differs from the Matched smoothing method in that we apply the smoother to the treated and untreated groups first and then compare the two groups. The advantage of the matching and smoothing method is its ability to compare observation level differences between the treated and untreated. The advantages of the Smoothing Difference Method are its simplicity and reliance on fewer modeling decisions. Since the smoothing difference method only requires one modeling step (step 2) compared to the two steps for the matching and smoothing method (steps 2 and 4), the computation of confidence intervals is simpler for the former method [5].

3.3 HTE using machine learning: Algorithmic and Meta-learners

3.3.1 Algorithmic/ Base algorithms

LASSO for HTE

Lasso regression is a common variable selection method that shrinks regression coefficients to 0 using the $\mathcal{L}_1$ norm regularization. [17] Imai and Ratkovic propose a modified LASSO model to select the most data-sensitive treatment covariate interactions. By manipulating the values of the covariates, the CATE can be predicted to show HTEs [8].

To demonstrate the Imai and Ratkovic Lasso framework for HTE assume the following model where L_i are 3 covariates and $A*L_1$, $A*L_2$, and $A*L_3$ represent interactions between the binary treatment and the 3 covariates respectively.

$$Y = \beta_0 + \beta_1 A + \beta_2 A \cdot L_1 + \beta_3 A \cdot L_2 + \beta_4 A \cdot L_3 + \beta_5 L_1 + \beta_6 L_2 + \beta_7 L_3 + \epsilon$$

We then posit a loss function with two tuning parameters, one for each group of predictors: the treatment effect + the interaction effects (λ_1) and the main effects of each of the potential confounders(λ_2):

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} \left[\epsilon_i^2 + \lambda_1(|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4|) + \lambda_2(|\beta_5| + |\beta_6| + |\beta_7|)\right]$$

From $\hat{\beta}$ we can compute the HTE by predicting our conditional average treatment effect with missing imputed covariate models [8].

$$HTE = \tau(L = l) = E(Y|\hat{\beta}, A = 1, L = l) - E(Y|\hat{\beta}, A = 0, L = l)$$

For example, suppose the Lasso penalization shrunk the β_1 term to 0. Then our predicted CATE for the HTE would be:

$$CATE = \tau(L = l) = (\hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5 + \hat{\beta}_6 + \hat{\beta}_7) - (\hat{\beta}_0 + \hat{\beta}_5 + \hat{\beta}_6 + \hat{\beta}_7) = \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$$

Tree Based Methods: Causal Trees and Causal Forests

Regression Trees work by recursively splitting the covariate space into smaller subsets from root node to leaves. The algorithm starts by selecting the best covariate and corresponding split point that decreases the dispersion of the target attribute value. After the first split, the process is repeated in each subset (node) until a stopping criterion is reached. The subsets left at the bottom of the tree after the stopping criteria have been reached are the leaf nodes. Typically, the goal is to minimize the variation of the outcome within each leaf [8].

To target HTEs, regression trees can be constructed so the variance of the treatment effect within each leaf is small while the variance between leaves is large [8, 18]. The variance of the treatment effect within each leaf concerns the goodness of model fit, while the interleaf variation concerns HTEs. This intuitively makes sense since we want each leaf to classify outcomes that have similar covariates in the same way, but also expect there to be heterogeneity across stratum of covariance. Athey proposes a loss function that incorporates model accuracy and HTEs:

$$Loss = \underbrace{(\frac{1}{N^{tr}} + \frac{1}{N^{es}}) \sum_{k \in \Pi} (\frac{S^2_{S^{tr}(1)}(k)}{p(k)} + \frac{S^2_{S^{tr}(0)}(k)}{1 - p(k)})}_{A} - \underbrace{\frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\tau}^2(L_i : S^{tr}, \Pi)}_{B}$$

- N^{tr}: Size of training subsample
- N^{es}: Size of estimation subsample
- $S^2_{S^{tr}(1)}(k)$: Variance of treatment in leaf 1
- $S^2_{S^{tr}(0)}(k)$: Variance of control in leaf 1
- $p(k)$ proportion of treatment cases in leaf 1
- Π: partitioning framework of the covariate space

This Loss is considered an "honest estimation" procedure since the algorithm uses different sub-samples for covariate space partitioning and treatment effect estimation. "A" represents the degree of overall uncertainty in the estimation, while "B" refers to the extent of variation in the treatment effect across leaves, the HTE. [8, 18]. Causal Trees yield the following estimate of the HTE:

$$CATE = \tau(L = l) = E[Y(A = 1) - y(A = 0) \mid (L \in k(l; \hat{\Pi})]$$

Causal Trees can be expanded into causal forests where a batch of trees is fit and the best leaf-specific HTE is computed by averaging outputs across them. [8, 18]. Since causal forests report average effects across multiple trees, they reduce the impact of any one suboptimal tree.

BART

BART is a Bayesian tree model version of the Generalized Additive Model [19]. Similar to our prior discussion of GAMs, causal trees allow us to relax the assumption of linearity between covariates and the outcome. Tree-based smoothers, without regularization, can result in a limited number of complex trees that dominate average treatment effect estimates. BART addresses this issue by ensuring each tree is a weak learner by tuning its prior probability to limit tree complexity [8]. The BART functional form is specified below:

$$Y = \sum_{h=1}^{m} \hat{g}(L, A, Tree_h) + \epsilon$$

For each Tree, $Tree_h$, BART imposes a prior probability $\alpha(1+d)^{-\beta}$, where d controls the depth/complexity of the tree. The larger d is, the smaller the prior probability is for that particular tree. This reduces the probability of fitting overly complex trees [8]. After training the BART algorithm, we can estimate the CATE:

$$CATE = \hat{\tau}(L = l) = \sum_{h=1}^{m} \hat{g}(A = 1, L = l, Tree_h) - \sum_{h=1}^{m} \hat{g}(A = 0, L = l, Tree_h)$$

3.3.2 Meta-learners

Meta-learners or meta-algorithms emerged as promising developments for estimating CATE non-parametrically by combining outputs from supervised machine learning (ML) methods or base learners. Therefore, meta-learners are flexible in that the base learners can be chosen to fit any data structure, thus allowing the estimation of CATE in different settings. The estimation of CATE through meta-learners usually consists of estimating the conditional means of the treatment and the outcome, and using predictions to find $\hat{\tau}(l)$. We can broadly group meta-learners into two methods based on i) conditional mean regression methods, and ii) pseudo-outcome methods. Since the meta-learners have different ways to estimating the CATE, as described in [8, 20, 21], for the same base learner, the predicted CATE may be different across meta-learners. In this report, we review five meta-learners that are currently used in the literature for CATE estimation.

3.3.3 Conditional mean regression methods

S-learner: In this meta-learner, the treatment assignment A is included as a feature similar to covariates L, without any special role. This learner considers a single model for the combined response function,

$$\mu(a, l) = \mathbb{E}[Y|A = a, L = l]$$

using a base learner on the entire dataset, and therefore the CATE for the S-learner is denoted as

$$\hat{\tau}_S(l) = \hat{\mu}(l, A = 1) - \hat{\mu}(l, A = 0)$$

As described by several authors, one pitfall of this estimator is that there is no guarantee that the treatment assignment will be included by the ML algorithm. If we use any regularization, only the covariates that are most predictive of Y are retained, and if the treatment does not have strong predictive power, it could be completely ignored.[8, 20, 21] Additionally, the S-learner could be biased towards zero and have a convergence rate lower than $1/\sqrt{n}$ [22].

T-learner: This learner emerged as an improvement to the **S-learner**. More specifically, the T-learner forcefully preserves A and takes two steps. First, we use a base learner to predict Y for the untreated using the observations in the untreated group alone, namely $\{L_i, Y_i\}_{A_i=0}$, and so $\hat{\mu}_0 = \mathbb{E}[Y(0)|L = l]$. Similarly, we estimate the treatment response function as $\hat{\mu}_1 = \mathbb{E}[Y(1)|L = l]$. Then, in the second step, the T-learner is obtained as

$$\hat{\tau}_T(l) = \hat{\mu}_1(l) - \hat{\mu}_0(l).$$

The T-learner does not target $\tau(l)$ directly; thus, it might not perform well in certain settings and might be biased.[8, 20, 21] For instance, in the setting where there is an imbalance of treatment assignment, we might be able to estimate μ_0 very well but not μ_1. This example motivated the X-learner, which we will discuss in the next section.

3.3.4 Pseudo-outcome methods

The pseudo-outcome-based methods estimate conditional means of the outcome and propensity scores. However, instead of estimating the CATE from the nuisance estimations, the methods use these to construct *pseudo-outcome* that serves as an initial approximation of the CATE.

X-learner Motivated by the T-learner limitations, Künzel et al. in [20] proposed the X-learner. The X-Learner has two stages and a propensity score model. The first one is identical to the **T-learner**; we split the data into treated and untreated and fit a ML model for the treated and untreated:

$$\hat{\mu}_0 = \mathbb{E}[Y(0)|L = l]$$
$$\hat{\mu}_1 = \mathbb{E}[Y(1)|L = l]$$

Next, we impute the individual treatment effects for the untreated and treated observations using the data in an *"X"-like* way. More specifically, we impute the individual treatment effects for the treated using the control-outcome estimator $\hat{\mu}_0$, and vice-versa. That is, we define

$$D_i^1 = Y_i(1) - \hat{\mu}_0(L_i, A = 1),$$
$$D_i^0 = \hat{\mu}_1(L_i, A = 0) - Y_i(0)$$

and $\tau_a(l) = \mathbb{E}[D^a|L = l]$. We can use any supervised learning or regression methods to estimate $\hat{\tau}_1(l)$ and $\hat{\tau}_0(l)$. Lastly, we estimate the CATE using a weighted average of the two estimates in the previous step:

$$\hat{\tau}(l) = g(l)\hat{\tau}_0(l) + (1 - g(l))\hat{\tau}_1(l),$$

where $g \in [0, 1]$, and it is commonly defined as the propensity score $\hat{\pi}$:

$$\hat{\tau}(l) = \hat{\pi}\hat{\tau}_0(l) + (1 - \hat{\pi})\hat{\tau}_1(l),$$

One strength of the X-learner is that it can borrow information across models in the imputation step to build better estimators of $\hat{\tau}_0(l)$, and $\hat{\tau}_1(l)$. Additionally, through the weighted average approach, the X-learner accounts for imbalanced data, which is common in current data sources[23], via the $g(\cdot)$ function. In the case where we have a smaller number of treated observations, μ_1 may have more uncertainty than μ_0, and therefore, we would want the estimate of $\tau(l)$ to weigh more or rely more on $\tau_1(l)$. The X-learner does exactly that by pulling $\hat{\tau}(l)$ towards $\hat{\tau}_1(l)$. Through theoretical and simulation results, Künzel et al. in [20] show that in this setting, the X-learner outperforms the T-learner.

DR-Learner The DR-Learner was introduced by Edward Kennedy in 2020 motivated by the fact that treatment effects often have complex non-linear forms [24]. The idea behind the DR-Learner is to use flexible regression functions for the pseudo-outcome, so as to capture the complexities inherent in the conditional means. The DR-Leaner uses sample-splitting to for the estimation of the CATE. In particular, let (S_1^n, S_2^n) be two independent samples of size n of O. The first step is the nuisance training: construct estimates $\hat{pi}, \hat{\mu}_1, \hat{\mu}_0$ for the propensity score and the conditional means using S_1^n. The second step is to construct the pseudo-outcome and to regress it on the covariates L from S_2^n to obtain $\hat{\tau}(l)$. The proposed pseudo-outcome is

$$\hat{\varphi}(O) = \frac{A - \hat{\pi}(L)}{\hat{\pi}(L)\{1 - \hat{\pi}(L)\}}\left\{Y - \hat{\mu}_A(L)\right\} + \hat{\mu}_1(L) - \hat{\mu}_0(L)$$

which regressed on L yields $\hat{\tau}(l) = \hat{\mathbb{E}}\{\hat{\varphi}(O)|L = l\}$. A third optional step is to perform cross-fitting by repeating the first two steps, swapping the roles of the samples, and using the average of the resulting estimators as the final estimate of the CATE.

The strengths of the DR-Learner stem from the fact that $\hat{\varphi}$ is the efficient influence function for the ATE. Observe that the ATE is efficiently estimated through the doubly robust estimator by averaging $\hat{\varphi}$, which the DR-Leaner regresses on. Kennedy also shows that the DR-Learner attains oracle efficiency and that the CATE estimate through this method is doubly robust. The major weakness of the DR-Learner is that the pseudo-outcome incorporates the propensity score estimate, which could be close to boundaries causing instability in the final estimate.

R-learner The R-Learner was introduced by Nie and Wager around the same time that Kennedy introduced the DR-Learner [21]. Nie and Wager were motivated by the fact that as of the time of publication, there was no definite way of using machine learning to estimate treatment effects in observational studies. In particular, the authors propose a loss function learned from the conditional means of outcomes and treatment, that is then optimized to find the estimate of causal estimand of interest. This is different from other machine learning approaches that usually perform estimation based on arbitrary loss functions. Similar to the DR-Learner, the R-Learner is performed in two steps. The first step consists of dividing the data O into Q evenly sized folds and estimating the propensity score, π, and the conditional

mean of the outcome, μ, by cross-fitting over the Q folds. In the second step, the treatment effects are estimated as follows.

$$\hat{\tau}(\cdot) = \arg\min_{\tau}[\hat{L}_n\{\tau(\cdot)\} + \Lambda_n\{\tau(\cdot)\}]$$

where $\Lambda_n\{\tau(\cdot)\}$ is a regularizer of the complexity of $\tau(\cdot)$ and the loss function is

$$\hat{L}_n\{\tau(\cdot)\} = \frac{1}{n}\sum_{i=1}^{n}\left[\{Y_i - \hat{\mu}^{\{-i\}}(L_i)\} - \{A_i - \hat{\pi}^{\{-i\}}(L_i)\}\tau(L_i)\right]^2$$

where $\hat{\mu}^{\{-i\}}, \hat{\pi}^{\{-i\}}$ are predictions made without using the fold to which the ith observation belong to. The form of the loss function comes from using a decomposition introduced by Robinson in 1988,

$$Y_i - \mu(L_i) = \{A_i - \pi(L_i)\}\tau(L_i) + Y_i(A_i) - \{\mu_0(L_i) + A_i\tau(L_i)\}$$

which motivated the R in R-Learner.

Observe that the loss function can also be expressed as

$$\hat{L}_n\{\tau(\cdot)\} = \frac{1}{n}\sum_{i=1}^{n}\{A_i - \hat{\pi}^{\{-i\}}(L_i)\}^2\left[\underbrace{\frac{Y_i - \hat{\mu}^{\{-i\}}(L_i)}{A_i - \hat{\pi}^{\{-i\}}(L_i)}}_{\hat{\varphi}(L_i)} - \tau(L_i)\right]^2$$

where $\hat{\varphi}(L_i)$ is the pseudo-outcome. Now it is clear that R-learner is equivalent to regressing the pseudo-outcome on the covariates L and weighting by $\{A_i - \hat{\pi}^{\{-i\}}(L_i)\}^2$. The R-learner has the same problem as the DR-Leaner. The pseudo-outcome is estimated using the propensity score in the denominator, which might incorporate instability in the final estimate of the estimand of interest.

3.3.5 Targeted Learning of HTEs: Broadly speaking, another meta-learner

In this section, we provide an overview of a recent method developed by Wei et al. in 2023 [9], in which they adopted a model-free, semiparametric approach to estimated HTE across multiple subgroups utilizing a one-step iterative targeted maximum-likelihood estimation (iTMLE) and adjusting for potential confounding. This method is motivated by the one-step-a-time approach, as described in the introduction, where we divide individuals into subgroups based on relevant pre-treatment characteristics (e.g., gender, age groups) and then estimate the ATE in each subgroup. Instead, their proposed method targets multiple subgroups simultaneously. In contrast to the methods described in the previous sections, their method does not target the CATE. Instead, we are interested causal effect measures: the relative risk (RR), the odds ratio (OR), and the absolute risk difference (ADR) across different subgroups. These are commonly used measures with good interpretability for biomedical and epidemiological studies applications as Wei et al. show through a case study in their article. [9]

We follow the same notation as in previous sections, introducing the set $\{X_j\}_{j=1}^{d}$ as the set of d pre-specified subgroups of interest based on L, which can overlap. We can define $\alpha_{a,j} = P(Y(a) = 1|L \in X_j)$, $a \in \{0,1\}$, $j = 1, ..., d$ as the disease risk of treatment arm a for subgroup j, and $\boldsymbol{\alpha}_a = (\alpha_{a,1}, ..., \alpha_{a,d})^\intercal$. Similarly, we can define the RR as $\boldsymbol{\alpha}_{RR} = (\alpha_{RR,1}, ..., \alpha_{RR,d})^\intercal$, where $\alpha_{RR,j} = \alpha_{1,j}/\alpha_{0,j}$. Therefore, the target estimand in this setting is $\boldsymbol{\alpha}_a$, which are identifiable under the same assumptions described in 2.

The primary idea behind iTMLE is to replace the univariate clever covariate used in the classic one-step TMLE [25] with a multi-dimensional vector of clever covariates in the logistic regression step. Instead of targeting disease risk under treatment arm a for each subgroup individually, we can target all subgroups simultaneously. The "naive" initial extension approach may give rise to a problem wherein the algorithm suffers from numerical instability due to low subgroup cell values and propensity scores. This is because the probabilities, which are in the denominator of the clever covariate estimator, become small in such cases. This may also inflate the estimator's variance for $\hat{\boldsymbol{\alpha}}_a$. Therefore, the authors suggest a robust method in which, at each iteration of the algorithm, the initial estimator enters each iteration updated, which leads to increased efficiency and estimation bias (results shown in supplemental material by Wei et al. [9]). The clever covariate is also self-normalized at each iteration to provide more stable estimates. We do not provide a simulation example in this report using this method but refer the reader to Wei et al. original article in arXiv [26] for a more detailed discussion and our project GitHub repository for an example.

4 Simulations

In this section, we perform simulations to compare the meta-learners discussed in the previous section under different scenarios. In particular, we constructed three scenarios following Kunzel et al [20]. For each simulation we use a general framework $O_i = (L_i, A_i, Y_i)_{i=1}^{n}$, where

- L_i represents a covariate vector with dimension $d = 5$ generated as

$$L_i \overset{iid}{\sim} N(0, \Sigma)$$

$$\text{where } \Sigma = \begin{pmatrix} 1 & 0.5 & 0.25 & 0.125 & 0.0625 \\ 0.5 & 1 & 0.5 & 0.25 & 0.125 \\ 0.25 & 0.5 & 1 & 0.5 & 0.25 \\ 0.125 & 0.25 & 0.5 & 1 & 0.5 \\ 0.0625 & 0.125 & 0.25 & 0.5 & 1 \end{pmatrix},$$

- $A_i \sim \text{Bern}(\pi(L_i))$ is the treatment assignment and $\pi(L_i)$ is the propensity score (assignment probability),

- and $Y_i = Y_i(A_i)$ where

$$Y_i(A_i) = \mu_{A_i}(L_i) + \varepsilon_i(A_i).$$

The first and simple scenario, which we denote by A, is a scenario with balanced cases and no confounding. We assume a complex linear CATE function,

$$\pi(l) = 0.5$$
$$\mu_A(l) = l^T \boldsymbol{\beta}_A$$

where $\boldsymbol{\beta}_A \sim \text{Unif}([0, 30]^5)$.

In the second scenario, denoted by B, we assume no treatment effect. It has the same generating process, however we generate a single $\boldsymbol{\beta} \sim \text{Unif}([0, 30]^5)$ and assume that the mean function μ_A is the same for both treatment levels.

The third scenario, denoted by C, incorporates complexity by assuming an unbalanced case framework,

$$\pi(l) = 0.01$$
$$\mu_0(l) = l^T \boldsymbol{\beta} + 5\mathbb{I}(l_1 > 0.5)$$
$$\mu_1(l) = \mu_0(l) + 8\mathbb{I}(l_2 > 0.1)$$

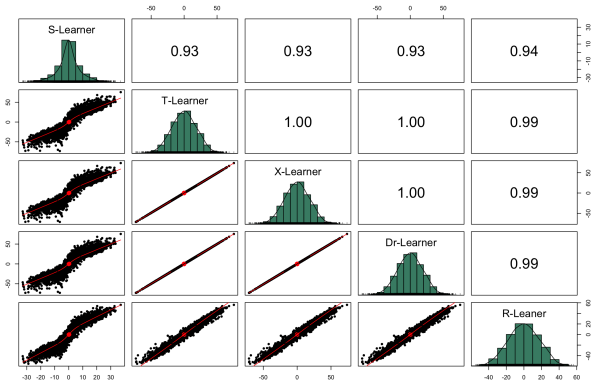
where $\boldsymbol{\beta} \sim \text{Unif}([-5, 5]^5)$.

To find meta-learners, we fit models using the super learner framework using the SuperLearner package in R. We find the conditional mean of the outcome and conditional mean of the treatment assignment by stacking i) a simple mean model, ii) a generalized additive model, iii) a random forest, and iv) an elastic net regression. To perform predictions, we discard models whose coefficients are 0 in the SuperLearner. We used the same learning process for the three scenarios to guarantee an equivalent comparison for the meta-learners. To build the meta-learners we used code provided by Salditt et al. [27].

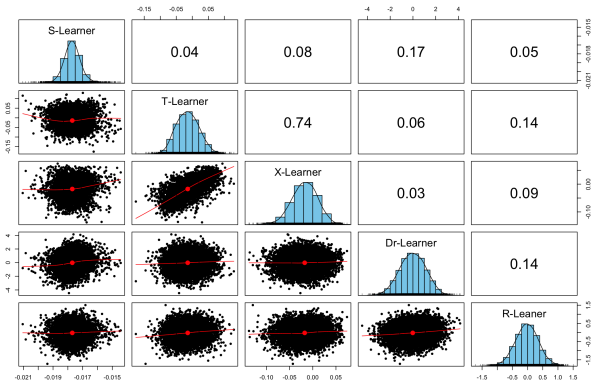
Figure 1 presents results for the simulation by scenario. All meta-learners perform relatively well for a balanced treatment scenario (A). In this simulation, the true ATE is -0.0142, which aligns well with the mean of each of the distributions. The S-Learner performs really well with the smallest deviation from the mean. For the no treatment effect scenario (B), again, the S-Learner performs well, but the differences between meta-learners is more evident. For scenario (C), where there is an unbalanced treatment assignment, the DR-Learner and the R-Learner show their improvement upon simpler methods. Of particular interest is the T-Learner, which cannot separate the effects of the treatment and the control group. This is expected because the T-Learner fits two separate models to the treatment and the control group, and then combines them (without any weighting) to obtain the CATE. The pseudo-outcome-based models incorporate weighting, increasing certainty in the ATE final estimate.

5 Future Extensions

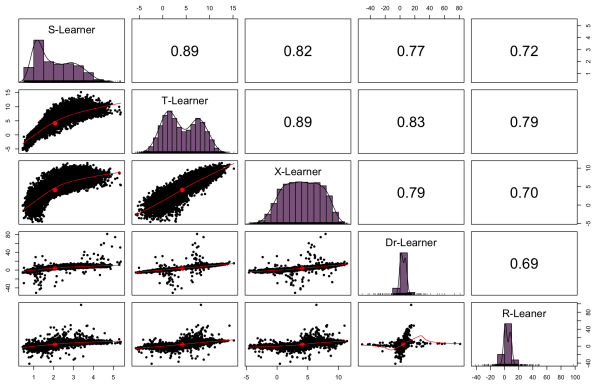
HTEs have become a growing area of research with a high prevalence of methodologies available for their estimation. As we have surveyed in this report, there has been a shift from conventional methods like linear modeling (i.e., interaction terms and propensity scores) to methods that leverage machine learning and algorithmic agnostic models. A discussion in the field of HTE is how to interpret discordant results from different methods, as we showed through simulations; depending on the setting, at least for algorithmic agnostic models, we



(a) Scenario A: Balanced treatment & no unconfounding



(b) Scenario B: No HTE



(c) Scenario C: Unbalanced treatment

Figure 1: Distributions and pairwise correlations of individual treatment effects under simulation scenarios

might get results that do not align. Therefore, it is recommended that the choice of method be made thinking about the theoretical reasoning for the question of interest and not solely on model performance. [8, 28, 29] Generalizing HTE methods to multiple treatments is still not fully explored. In this setting, we might want to compare the treatment effect from a control group to multiple treatment groups, and thus, for a future extension, it will be interesting to study the performance of each meta-learner and if there is a dependency on the base learner, we use to estimate the nuisance functions.

Supporting information. We provide all R code for the simulation examples on the GitHub repository: `https://github.com/cbrodriguez01/bst258finalproject`.

References

[1] Liangyuan Hu et al. "A Flexible Approach for Assessing Heterogeneity of Causal Treatment Effects on Patient Survival Using Large Datasets with Clustered Observations". eng. In: *International Journal of Environmental Research and Public Health* 19.22 (Nov. 2022), p. 14903. ISSN: 1660-4601. DOI: 10.3390/ijerph192214903.

[2] Miriam Kasztura et al. "Cost-effectiveness of precision medicine: a scoping review". eng. In: *International Journal of Public Health* 64.9 (Dec. 2019), pp. 1261–1271. ISSN: 1661-8564. DOI: 10.1007/s00038-019-01298-x.

[3] Derek C. Angus and Chung-Chou H. Chang. "Heterogeneity of Treatment Effect: Estimating How the Effects of Interventions Vary Across Individuals". In: *JAMA* 326.22 (Dec. 2021), pp. 2312–2313. ISSN: 0098-7484. DOI: 10.1001/jama.2021.20552. URL: https://doi.org/10.1001/jama.2021.20552 (visited on 04/25/2024).

[4] Issa J Dahabreh, Rodney Hayward, and David M Kent. "Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence". In: *International journal of epidemiology* 45.6 (2016), pp. 2184–2193.

[5] Yu Xie, Jennie E. Brand, and Ben Jann. "Estimating Heterogeneous Treatment Effects with Observational Data". eng. In: *Sociological Methodology* 42.1 (Aug. 2012), pp. 314–347. ISSN: 0081-1750. DOI: 10.1177/0081175012452652.

[6] Julie M Zissimopoulos et al. "Sex and race differences in the association between statin use and the incidence of Alzheimer disease". In: *JAMA neurology* 74.2 (2017), pp. 225–232.

[7] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. "Estimating conditional average treatment effects". In: *Journal of Business & Economic Statistics* 33.4 (2015), pp. 485–505.

[8] Anning Hu. "Heterogeneous treatment effects analysis for social scientists: A review". In: *Social Science Research* 109 (Jan. 2023), p. 102810. ISSN: 0049-089X. DOI: 10.1016/j.ssresearch.2022.102810. URL: https://www.sciencedirect.com/science/article/pii/S0049089X22001211 (visited on 04/27/2024).

[9] Waverly Wei et al. "Efficient targeted learning of heterogeneous treatment effects for multiple subgroups". en. In: *Biometrics* 79.3 (2023), pp. 1934–1946. ISSN: 1541-0420. DOI: 10.1111/biom.13800. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13800 (visited on 04/23/2024).

[10] James J Heckman and Edward Vytlacil. "Structural equations, treatment effects, and econometric policy evaluation 1". In: *Econometrica* 73.3 (2005), pp. 669–738.

[11] Jinyong Hahn. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: *Econometrica* (1998), pp. 315–331.

[12] Miguel A Hernán and James M Robins. *Causal inference*. 2010.

[13] Sander Greenland. "Interpretation and choice of effect measures in epidemiologic analyses". In: *American journal of epidemiology* 125.5 (1987), pp. 761–768.

[14] Paul R Rosenbaum and Donald B Rubin. "Reducing bias in observational studies using subclassification on the propensity score". In: *Journal of the American statistical Association* 79.387 (1984), pp. 516–524.

[15] Jianqing Fan et al. "A study of variable bandwidth selection for local polynomial regression". In: *Statistica Sinica* (1996), pp. 113–127.

[16] William S Cleveland. "Robust locally weighted regression and smoothing scatterplots". In: *Journal of the American statistical association* 74.368 (1979), pp. 829–836.

[17] Kosuke Imai and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation". In: (2013).

[18] Susan Athey and Stefan Wager. "Estimating treatment effects with causal forests: An application". In: *Observational studies* 5.2 (2019), pp. 37–51.

[19] Hugh A Chipman, Edward I George, and Robert E McCulloch. "BART: Bayesian additive regression trees". In: (2010).

[20] Sören R. Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences* 116.10 (Mar. 2019). Publisher: Proceedings of the National Academy of Sciences, pp. 4156–4165. DOI: 10.1073/pnas.1804597116. URL: https://www.pnas.org/doi/full/10.1073/pnas.1804597116 (visited on 04/27/2024).

[21] X Nie and S Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (June 2021), pp. 299–319. ISSN: 0006-3444. DOI: 10.1093/biomet/asaa076. URL: https://doi.org/10.1093/biomet/asaa076 (visited on 04/23/2024).

[22] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.

[23] Ruta Brazauskas and Brent R Logan. "Observational studies: matching or regression?" In: *Biology of Blood and Marrow Transplantation* 22.3 (2016), pp. 557–563.

[24] Edward H. Kennedy. "Towards optimal doubly robust estimation of heterogeneous causal effects". In: *Electronic Journal of Statistics* 17.2 (Jan. 2023). Publisher: Institute of Mathematical Statistics and Bernoulli Society, pp. 3008–3049. ISSN: 1935-7524, 1935-7524. DOI: 10.1214/23-EJS2157. URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-17/issue-2/Towards-optimal-doubly-robust-estimation-of-heterogeneous-causal-effects/10.1214/23-EJS2157.full (visited on 04/27/2024).

[25] Mark van der Laan and Susan Gruber. "One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels". In: *The international journal of biostatistics* 12.1 (2016), pp. 351–378.

[26] Waverly Wei et al. "Efficient Targeted Learning of Heterogeneous Treatment Effects for Multiple Subgroups". In: *Biometrics* 79.3 (Nov. 2022), 1934â1946. ISSN: 1541-0420. DOI: 10.1111/biom.13800. URL: http://dx.doi.org/10.1111/biom.13800.

[27] Marie Salditt, Theresa Eckes, and Steffen Nestler. "A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners". In: *Administration and Policy in Mental Health and Mental Health Services Research* (2023), pp. 1–24.

[28] Vincent Dorie et al. “Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition”. In: (2019).

[29] Liangyuan Hu, Jiayi Ji, and Fan Li. “Estimating heterogeneous survival treatment effect in observational data using machine learning”. eng. In: *Statistics in Medicine* 40.21 (Sept. 2021), pp. 4691–4713. ISSN: 1097-0258. DOI: 10.1002/sim.9090.