Econometrics Journal (2021), volume **24**, pp. 134–161.
doi: 10.1093/ectj/utaa014

Machine learning estimation of heterogeneous causal effects: empirical Monte Carlo evidence

MICHAEL C. KNAUS, MICHAEL LECHNER
AND ANTHONY STRITTMATTER

University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland.
Email: michael.knaus@unisg.ch, michael.lechner@unisg.ch, anthony.strittmatter@unisg.ch

First version received: 31 October 2019; final version accepted: 28 April 2020.

Summary: We investigate the finite-sample performance of causal machine learning estimators for heterogeneous causal effects at different aggregation levels. We employ an empirical Monte Carlo study that relies on arguably realistic data generation processes (DGPs) based on actual data in an observational setting. We consider 24 DGPs, eleven causal machine learning estimators, and three aggregation levels of the estimated effects. Four of the considered estimators perform consistently well across all DGPs and aggregation levels. These estimators have multiple steps to account for the selection into the treatment and the outcome process.

Keywords: *Causal machine learning, conditional average treatment effects, selection-on-observables, Random Forest, Causal Forest, Lasso.*

JEL Codes: *C21.*

1. INTRODUCTION

Economists and many other professionals are interested in the causal effects of policies or interventions. This has triggered substantial advances in microeconometrics and statistics in understanding the identification and estimation of different average causal effects in recent decades (see, e.g., Imbens and Wooldridge, 2009; Athey and Imbens, 2017, and references therein). However, in most applications it is also interesting to look beyond the average effects in order to understand how the causal effects vary with observable characteristics. For example, finding those individuals who benefit most from active labour market policies, promotion campaigns or medical treatments is important for the efficient allocation of public and private resources.

In recent years, methods for the systematic estimation of heterogeneous causal effects have been developed in various research disciplines. These methods adapt standard machine learning methods to flexibly estimate heterogeneity along a potentially large number of covariates. The suggested estimators use regression trees (Su et al., 2009; Athey and Imbens, 2016), Random Forests (Wager and Athey, 2018; Athey et al., 2019), the least absolute shrinkage and selection operator (Lasso) (Qian and Murphy, 2011; Tian et al., 2014; Chen et al., 2017), support vector machines (Imai and Ratkovic, 2013), boosting (Powers et al., 2018), neural nets (Johansson et al., 2016; Shalit et al., 2017; Schwab et al., 2018) or Bayesian machine learning (Hill, 2011; Taddy et al., 2016).[1] Recently, the first applied studies using these methods appeared in economics (e.g., Bertrand et al., 2017; Davis and Heller, 2017; Knaus et al., 2020; Andini et al., 2018; Ascarza, 2018; Strittmatter, 2018).

In contrast to the rather mature literature about the estimation of average causal effects, the literature on the estimation of effect heterogeneity still lacks guidance for practitioners about which methods are well suited for their intended applications. Currently, theoretical asymptotic approximations are not available, are incomplete, or are based on non-overlapping assumptions, preventing comparisons of estimators. Furthermore, the information that is available on the finite-sample performance is of limited use to practitioners. Most comparisons are based on data generating processes (DGPs) that are very unrealistic for real applications. One exception is Wendling et al. (2018), who based their simulation study on data from medical records. However, they focus on the special case of binary outcomes and on data structures that are unusual in economics.

In this study, we categorize the major approaches from different fields. We distinguish between generic approaches that can be combined with a variety of off-the-shelf machine learning estimators and estimator-specific approaches that modify an existing method. The generic approaches are combined with the machine learning estimators Random Forest and Lasso. This leads to eleven causal machine learning estimators to be investigated. As opposed to standard simulation methods that rely on a synthetic DGP, we investigate the finite-sample performance of these estimators in an Empirical Monte Carlo Study (EMCS) approach (e.g., Huber et al., 2013; Lechner and Wunsch, 2013). An EMCS informs the DGPs as much as possible by real data and reduces the synthetic components in the DGP to a minimum. We consider six specifications of the heterogeneous causal effects, two sample sizes, and DGPs with and without selection into treatment.[2]

Our contribution to the aforementioned literature is three-fold. First, we provide a comprehensive comparison of different estimators and DGPs. Second, we consider the finite-sample properties of causal machine learning estimators for effect heterogeneity under DGPs that are arguably realistic, at least in some fields of economics. Third, this is the first simulation study that considers different aggregation levels of the heterogeneous effects. In particular, we consider an intermediate aggregation level between the most individualized causal effects and the average population effect. Such intermediate aggregation levels are important as feasible action rules for practitioners.

Our findings suggest that no causal machine learning estimator is superior for all DGPs and aggregation levels. However, four estimators show a relatively good performance in all settings: Random Forests combined with a doubly robust outcome modification (based on Chernozhukov et al., 2018), Causal Forest with local centring (Athey et al., 2019), Lasso combined with a covariate modification and efficiency augmentation (Tian et al., 2014), and Lasso with Orthogonal Learning (Nie and Wager, 2017; Foster and Syrgkanis, 2019). All those methods use multiple estimation steps to account for the selection into treatment and the outcome process. Several other estimators may be suitable in specific empirical settings, but their performance is unstable across different DGPs. Lasso estimators tend to be more unstable than Random Forests, which frequently prevents them from achieving a normal distribution.

[1] Hastie et al. (2009) introduce the underlying machine learning algorithms. Athey (2018) and Belloni et al. (2014a) provide an overview of how those methods might be used in the estimation of average causal effects and other parameters of interest.

[2] We focus on point estimation and leave the investigation of inference methods for future research.

In the next section, we introduce the notation and the estimation targets. In Section 3, we describe and categorize causal machine learning approaches to estimating heterogeneous causal effects. In Section 4, we explain the implementation of the estimators. In Section 5, we discuss the EMCS approach. In Section 6, we provide our simulation results. The final section concludes and hints at some topics for future research. The Online Appendix provides the full simulation results of all DGPs and other supplementary information. We provide code that implements the estimators under investigation in the R package CATEs on GitHub.

2. NOTATION AND ESTIMATION TARGETS

We describe the parameters of interest using Rubin's (1974) potential outcome framework. The dummy variable D_i indicates a binary treatment, for example participation in a training programme. Let Y_i^1 denote the outcome (e.g., employment) if individual i ($i = 1, ..., N$) receives the treatment ($D_i = 1$). Correspondingly, Y_i^0 denotes the outcome if individual i does not receive the treatment ($D_i = 0$). Each individual can either receive the treatment or not. This means that only one of the two potential outcomes (Y_i^d) is observable:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0. \tag{2.1}$$

Thus, the *individual treatment effect* (ITE) $\xi_i = Y_i^1 - Y_i^0$ of D_i on Y_i is never observed. However, the identification of expectations of ξ_i may be possible under plausible assumptions. For example, the identification of the average treatment effect (ATE), $\tau = E[\xi_i]$, or the average treatment effect on the treated (ATET), $\theta = E[\xi_i | D_i = 1]$, are standard in microeconometrics (see, e.g., Imbens and Wooldridge, 2009).

The focus of this study is on conditional average treatment effects (CATEs). CATEs take the expectations of ξ_i conditional on exogenous pre-treatment covariates.[3] We call the finest conditioning level that uses all available covariates X_i the *individualized average treatment effect* (IATE):

$$\tau(x) = E[\xi_i \mid X_i = x] = \mu^1(x) - \mu^0(x), \tag{2.2}$$

where $\mu^d(x) = E[Y_i^d \mid X_i = x]$ denotes the conditional expectation of the unobserved potential outcomes. IATEs provide an approximation of ITEs for the set of covariates that are at the disposal of the researcher in a specific application. However, researchers may additionally be interested in intermediate aggregation levels that are coarser than IATEs but finer than ATEs. In particular, groups based on a smaller set of predefined characteristics, G_i, may be of interest if the estimated IATEs need to be summarized for the research community, communicated to practitioners, or acted upon.[4] We call the effects defined on this aggregation level *group average treatment effects* (GATEs):

$$\tau(g) = E[\xi_i \mid G_i = g] = \int \tau(x) f_{X_i|G_i=g}(x) dx. \tag{2.3}$$

The identification of any aggregation level of individual treatment effects in observational studies is complicated by non-random treatment assignment. However, identification of the IATE and any coarser aggregation level is still possible if the observable covariates X_i contain all

[3] Covariates are also called features or predictors in parts of the machine learning literature.

[4] For example, if interest is in gender differences, $G_i \in \{female, male\}$.

confounders.[5] These are covariates that jointly affect the treatment probability and the potential outcomes. Although there are alternative ways to identify the various effects, here we focus on the case where all the confounders are contained in the data available to the researcher. This means that we operate throughout the paper under the following assumptions.

ASSUMPTION 2.1 *(Conditional independence): $Y_i^1, Y_i^0 \perp\!\!\!\perp D_i \mid X_i = x$, for all x in the support of X_i.*

ASSUMPTION 2.2 *(Common support): $0 < P[D_i = 1|X_i = x] = p(x) < 1$, for all x in the support of X_i.*

ASSUMPTION 2.3 *(Exogeneity of covariates): $X_i^1 = X_i^0$.*[6]

ASSUMPTION 2.4 *(Stable unit treatment value assumption, SUTVA): $Y_i = Y_i^1 D_i + Y_i^0(1 - D_i)$.*

Assumption 2.1 states that the potential outcomes are independent of the treatment conditional on the confounding covariates. According to Assumption 2.2, the conditional treatment probability (often called the propensity score) is bounded away from zero and one. Assumption 2.3 requires that the covariates are not affected by the treatment. Assumption 2.4 excludes spillover effects between treated and non-treated. Under Assumptions 2.1–2.4,

$$E[Y_i^d \mid X_i = x, D_i = 1 - d] = E[Y_i \mid X_i = x, D_i = d] = \mu(d, x) \tag{2.4}$$

$$\Rightarrow \tau(x) = \mu(1, x) - \mu(0, x), \tag{2.5}$$

and thus IATEs, GATEs and ATE are identified from observable data. We denote the conditional expectations of the outcomes in one treatment arm by $\mu(d, x) = E[Y_i|X_i = x, D_i = d]$, the conditional expectation of the outcome as $\mu(x) = E[Y_i|X_i = x]$, and the conditional treatment probability by $p(x) = P[D_i = 1|X_i = x]$.

3. CAUSAL MACHINE LEARNING OF EFFECT HETEROGENEITY

Equation (2.5) shows that the fundamental task is to estimate the difference of two conditional expectations. However, we never observe the differences at the individual level and have to estimate them in two different subpopulations. Thus, the estimation of IATEs is a non-standard machine learning problem. In this section, we present different approaches to target the estimation of IATEs. We distinguish between generic approaches and one estimator-specific approach. Generic approaches split the causal estimation problem into several standard prediction problems and may be combined with a large variety of supervised machine learning estimators. On the other hand, Causal Forest (Wager and Athey, 2018; Athey et al., 2019) is a modification of a specific machine learning estimator to move the target from the estimation of outcomes to the estimation of IATEs.

3.1. Generic approaches

A straightforward generic approach follows directly from equation (2.5). *Conditional mean regression* takes the difference of conditional expectations that are estimated in the two samples of

[5] X_i represents the union of confounders and heterogeneity variables for notational convenience. In principle, they may be completely or partly overlapping or non-overlapping (see, e.g., Knaus et al., 2020).

[6] *The potential confounders X_i^d are defined equivalently to potential outcomes.*

treated and non-treated separately using off-the-shelf machine learning methods to estimate the conditional outcome means $\hat{\mu}(d, x)$:[7]

$$\hat{\tau}_{CMR}(x) = \hat{\mu}(1, x) - \hat{\mu}(0, x). \tag{3.1}$$

Conditional mean regressions are straightforward to implement. Any supervised machine learning methods for conditional mean estimation can be used. However, their usual target is to minimize the mean squared error (MSE) in two separate prediction problems, and they are not tailored to estimate IATEs.[8] This suggests that they may be outperformed by more specialized methods for this causal problem. Three generic multi-step approaches that target IATE estimation are presented in the following, and a framework to summarize them is provided.

3.1.1. Modified outcome methods. Abadie (2005) introduced the idea of modifying the outcome to estimate CATEs on the treated in studies based on difference-in-differences. For IATEs in experimental and observational settings, the idea is formulated by Signorovitch (2007) and Zhang et al. (2012), respectively. The latter discuss two modifications of the outcome, which we summarize as modified outcome methods (MOMs). The first is based on inverse probability weighting (IPW) (e.g., Horvitz and Thompson, 1952; Hirano et al., 2003), where the modified outcome is

$$Y^*_{i,IPW} = Y_i \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))}. \tag{3.2}$$

The second is based on the doubly robust (DR) estimator of Robins and Rotnitzky (1995),

$$Y^*_{i,DR} = \mu(1, X_i) - \mu(0, X_i) + \frac{D_i(Y_i - \mu(1, X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \mu(0, X_i))}{(1 - p(X_i))}. \tag{3.3}$$

The crucial insight here is that $\tau(x) = E[Y^*_{i,IPW} \mid X_i = x] = E[Y^*_{i,DR} \mid X_i = x]$. This means that a regression with one of these modified outcomes and covariates X_i can be used to obtain estimates of IATEs, $\hat{\tau}_{IPW}(x)$ or $\hat{\tau}_{DR}(x)$. In practice, the researcher has no access to the true parameters $p(x)$ and $\mu(d, x)$, the so-called nuisance parameters. The conditional expectations need to be approximated in a first step and plugged into equations (3.2) and (3.3). Any suitable prediction method can be used to estimate the nuisance parameters as well as the IATEs.

The asymptotic properties of $E[Y^*_{i,DR}]$ as an estimator for ATEs are well understood (Belloni et al., 2014b; Farrell, 2015; Belloni et al., 2017; Chernozhukov et al., 2017; Chernozhukov et al., 2018). Furthermore, Abrevaya et al. (2015) and Lee et al. (2017) analyse the properties of estimating $\tau(z) = E[Y^*_{i,IPW} \mid Z_i = z]$ and $\tau(z) = E[Y^*_{i,DR} \mid Z_i = z]$ for a low-dimensional subset of covariates (Z_i), respectively. However, neither consider machine learning to estimate the nuisance parameters. This is considered for the doubly robust modification by Semenova and Chernozhukov (2017) using least squares series estimation, as well as by Fan et al. (2019) and Zimmert and Lechner (2019) using nonparametric regression in the final stage.

Simulation evidence from Powers et al. (2018) suggests that estimators based on $Y^*_{i,IPW}$ may exhibit high variance due to potentially extreme values of the propensity score in the denominator.

[7] This approach is also referred to as T-Learner (Künzel et al., 2019; Nie and Wager, 2017) or Q-Learning (Qian and Murphy, 2011).

[8] For an intuition of why this is not optimal: Biases that for the same value of x go in the same direction are less harmful than if they go in opposite directions. However, this cannot be accounted for by separate MSEs that are not directly linked (for a Modified Causal Forest estimator that takes up this theme directly, see Lechner (2018)).

Estimators based on $Y^*_{i,DR}$ might be more stable, because of the double-robustness property, but this has not been explored until now.

3.1.2. Modified covariate method. Tian et al. (2014) introduced the modified covariate method (MCM) for experiments, and Chen et al. (2017) extended it to observational studies. They showed that we can estimate IATEs by solving the objective function

$$\min_{\tau}\left[\frac{1}{N}\sum_{i=1}^{N} T_i \frac{D_i - p(X_i)}{p(X_i)(1-p(X_i))}\left(Y_i - \frac{T_i}{2}\tau(X_i)\right)^2\right], \tag{3.4}$$

where $T_i = 2D_i - 1 \in \{-1, 1\}$. The name MCM results from the practice of replacing the nonparametric function of the IATE with a linear working model, $\tau(x) = x\beta$. This enables us to rewrite the minimization problem (3.4) as

$$\hat{\beta}_{MCM} = \arg\min_{\beta}\left[\frac{1}{N}\sum_{i=1}^{N} T_i \frac{D_i - p(X_i)}{p(X_i)(1-p(X_i))}\left(Y_i - X_i^{MCM}\beta\right)^2\right], \tag{3.5}$$

where $X_i^{MCM} = T_i/2X_i$ are the modified covariates. The estimated IATEs are then obtained by $\hat{\tau}_{MCM}(x) = x\hat{\beta}_{MCM}$. The nuisance parameter $p(x)$ needs to be estimated in a first step using any suitable method.

In principle, rewriting 3.4 as

$$\min_{\tau}\left[\frac{1}{N}\sum_{i=1}^{N} T_i \frac{D_i - p(X_i)}{4p(X_i)(1-p(X_i))}\left(2T_iY_i - \tau(X_i)\right)^2\right] \tag{3.6}$$

allows us to apply any machine learning estimator that is able to solve weighted minimization problems. However, we are not aware of any study that notices and pursues this possibility.

MCM does not require us to specify any model of the so-called main effects $\mu(x)$ or $\mu(d, x)$. However, Tian et al. (2014) suggest that an estimate of $\mu(x)$ might be useful to increase efficiency. The efficiency-augmented version replaces the outcome Y_i in equations (3.4) to (3.6) by the residuals $Y_i - \mu(X_i)$. Thus, MCM with *efficiency augmentation* (EA) requires in addition that we estimate the nuisance parameter $\mu(x)$ in the first step. Tian et al. (2014) show that MCM with a linear working model provides the best linear predictor of the potentially non-linear $\tau(x)$. However, we are not aware of any further theoretical analyses of the statistical properties of this approach.

3.1.3. Orthogonal Learning. Foster and Syrgkanis (2019) describe Orthogonal Learning as a general framework that allows us to ignore the estimation of first-step nuisance parameters in second-step estimation. It requires that the loss function of the second step fulfils the condition of ‘Neyman orthogonality’ (Neyman, 1959; Chernozhukov et al., 2018). For the estimation of IATEs this holds for the following loss function:

$$\min_{\tau}\left\{\frac{1}{N}\sum_{i=1}^{N}\left[(Y_i - \mu(X_i)) - (D_i - p(X_i))\tau(X_i)\right]^2\right\}. \tag{3.7}$$

This special case of Orthogonal Learning is called A-learning (Chen et al., 2017) or R-learning (Nie and Wager, 2017) in different literatures. It is equivalent to MCM with EA for 50:50 randomization but otherwise solves a different minimization problem to estimate IATEs. As for

Table 1. Summary of generic approaches to estimate IATEs.

Approach	w_i	Y_i^*
MOM IPW	1	$Y_{i,IPW}^*$
MOM DR	1	$Y_{i,DR}^*$
MCM	$T_i\dfrac{D_i - p(X_i)}{4p(X_i)(1 - p(X_i))}$	$2T_iY_i$
MCM with EA	$T_i\dfrac{D_i - p(X_i)}{4p(X_i)(1 - p(X_i))}$	$2T_i(Y_i - \mu(X_i))$
Orthogonal Learning	$(D_i - p(X_i))^2$	$\dfrac{Y_i - \mu(X_i)}{D_i - p(X_i)}$

MCM, most implementations consider a linear working model for the IATE (Nie and Wager, 2017; Zhao et al., 2017) and solve

$$\hat{\beta}_{RL} = \arg\min_{\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}\left[(Y_i - \mu(X_i)) - X_i^{RL}\beta\right]^2\right\}, \tag{3.8}$$

where $X_i^{RL} = (D_i - p(X_i))X_i$ can be considered as an alternative way to modify covariates. The estimated IATEs are then obtained by $\hat{\tau}_{RL}(x) = x\hat{\beta}_{RL}$. Similar to MCM, (3.7) can be rewritten as

$$\min_{\tau}\left\{\frac{1}{N}\sum_{i=1}^{N}(D_i - p(X_i))^2\left[\frac{Y_i - \mu(X_i)}{D_i - p(X_i)} - \tau(X_i)\right]^2\right\} \tag{3.9}$$

and solved with any suitable method after estimating the nuisance parameters in a first step (see also Schuler et al., 2018).

3.1.4. Summary of generic approaches to estimating IATEs. One goal of this paper is to structure approaches that estimate IATEs in different literatures. One way to put the approaches above on more common ground is by noting that they can all be considered as solving a weighted minimization problem with modified outcomes:

$$\min_{\tau}\left\{\frac{1}{N}\sum_{i=1}^{N}w_i\left[Y_i^* - \tau(X_i)\right]^2\right\}. \tag{3.10}$$

Table 1 summarizes the weights, w_i, and outcome modifications, Y_i^*, underlying the different approaches. This common representation is helpful to see and understand the differences among the methods. The two MOMs require no additional weighting because $\tau(x) = E[Y_{i,IPW}^* \mid X_i = x] = E[Y_{i,DR}^* \mid X_i = x]$. For MCM, $\tau(x) = E[2T_iY_i|X_i = x] = E[2T_i(Y_i - \mu(X_i))|X_i = x]$ in the special case of 50:50 randomization of the treatment ($p(x) = 0.5$). Any deviating assignment mechanism requires reweighting with IPW weights to control for the deviation from 50:50 randomization. The modified outcome of Orthogonal Learning is equivalent to MCM with efficiency augmentation if $p(x) = 0.5$. However, the MCM modified outcome does not change with other propensity scores and preserves the interpretation as a mean comparison under 50:50 randomization, $E[(Y_i - \mu(X_i))/(D_i - p(X_i))|X_i = x]$ is lacking such an intuitive interpretation.

3.2. Causal Forest

Another strand of the literature modifies machine learning algorithms based on regression trees (Breiman et al., 1984) to estimate IATEs. We focus on Causal Forest, which is a special case of the Generalized Random Forest of Athey et al. (2019), as the most recent estimator in this line of research.[9] Causal Forests build on the idea of Random Forests (Breiman, 2001) and boil down to taking the difference of two weighted means in our case of binary treatments:

$$\hat{\tau}_{CF}(x) = \sum_{i=1}^{N} D_i w_i^1(x) Y_i - \sum_{i=1}^{N} (1 - D_i) w_i^0(x) Y_i, \tag{3.11}$$

where the weights $w_i^d(x)$ define an adaptive local neighbourhood around the covariate value of interest, x. These weights are obtained from the tailored splitting procedure that we describe in Section 4.1. Athey et al. (2019) show that this estimator can be consistent and asymptotically normal for a fixed covariate space.

4. IMPLEMENTATION OF ESTIMATORS

Section 3.1 describes generic approaches to splitting the estimation of IATEs into several prediction problems. The machine learning literature offers a large variety of potential methods, such that the investigation of every possible combination of approaches and machine learning methods is not feasible given our restrictions on computational costs. Thus, this study is restricted to two machine learning methods for the implementation of the prediction problems. They are chosen to be representative of more general approaches. First, we consider Random Forests (Breiman, 2001), which serve as a popular representative for methods that attempt *local* approximations of conditional mean functions. Second, we consider Lasso (Tibshirani, 1996) as a method that attempts a *global* approximation of conditional mean functions.[10] Both methods are increasingly popular in econometrics and are used in methodological contributions as well as in applications.

We consider the combinations of the generic approaches in Section 3.1 with Random Forest and Lasso.[11] Following Chernozhukov et al. (2018), we apply cross-fitting to all approaches that require the estimation of nuisance parameters. This means that the nuisance parameters and the IATEs are estimated in different samples to avoid overfitting.

Table 2 summarizes all estimators under investigation. We are currently not aware of a Random Forest implementation that supports weighted minimization. Thus, we implement MCM and Orthogonal Learning only with Lasso. We add further an *infeasible benchmark* estimator that has access to the true ITEs and uses them as outcome in a standard prediction problem.

Like all machine learning methods, Random Forests and Lasso involve a variety of choices in the implementation. The following sections briefly explain Random Forests and Lasso, present

[9] Previous works concerned with estimating IATEs by modifying tree-based methods are Su et al. (2009), Athey and Imbens (2016) and Wager and Athey (2018).

[10] By 'local' we mean that each point of the conditional mean function is approximated by the (weighted) average of neighbouring observations. By 'global' we mean the attempt to approximate the conditional mean by a flexible functional fitted to all data simultaneously.

[11] We consider only 'pure' combinations where all estimation steps are conducted with one of the two machine learning methods and neglect the possibility of estimating, e.g., the nuisance parameters via Random Forests and the IATEs via Lasso. In principle, this is possible but is not pursued owing to computational constraints. For the same reason, we do not pursue ensemble methods that combine different estimators for the nuisance parameters or for IATE estimation (see, e.g., Rolling et al., 2019; Schuler et al., 2018).

Table 2. List of considered causal machine learning estimators.

	Random Forest	Lasso	Cross-fitting
Infeasible benchmark	√	√	
Conditional mean regression	√	√	
MOM IPW	√	√	√
MOM DR	√	√	√
MCM		√	√
MCM with EA		√	√
Orthogonal Learning		√	√
Causal Forest	√		
Causal Forest with local centring	√		√

the details of the implementation and explain the use of cross-fitting. The resulting estimators target the estimation of IATEs. Additionally, we consider methods to estimate GATEs and ATE as a computationally cheap by-product of our analysis. To this end, we average the estimated IATEs, $\hat{\tau}(x)$, to GATEs by $\hat{\tau}(g) = N_g^{-1} \sum_{i=1}^{N} \mathbb{1}[G_i = g]\hat{\tau}(X_i)$, with $N_g = \mathbb{1}[G_i = g]$, and to the ATE by $\hat{\tau} = N^{-1} \sum_{i=1}^{N} \hat{\tau}(X_i)$.

4.1. Random Forest

The building blocks of Random Forests for conditional mean estimation are regression trees (Breiman et al., 1984). However, regression trees are unstable and exhibit a high variance. The Random Forest of Breiman (2001) addresses this issue by combining a large number of decorrelated regression trees. The decorrelation is achieved by growing each tree on a random subsample (generated either by bootstrapping or by subsampling) and by randomly choosing the covariates for each split decision.

The standard regression and probability forests split the trees to minimize the MSE of the observed outcomes. Those trees can then be used to form predictions for a realization of X_i. These predictions are formed as a weighted average of the observed outcomes, where the weights are larger the more often the observed outcome shares a final leaf with the realization of X_i. These kinds of forests are required for the conditional mean regression and the modified outcome approaches.

The Causal Forest of Athey et al. (2019) follows a similar structure. However, instead of splitting the sample according to observed outcomes, Causal Forests split the samples along the gradient of the mean difference with the pseudo-outcomes

$$\rho_i = (D_i - \bar{D}_P)(Y_i - \bar{Y}_P - (D_i - \bar{D}_P)\hat{\beta}_P)/Var_P(D_i), \tag{4.1}$$

where $\bar{D}_P$ and $\bar{Y}_P$ are averages of treatment indicator and outcome, $\hat{\beta}_P$ is the mean difference, and $Var_P(D_i)$ is the variance of the treatment in the parent node. This splitting rule is tailored to maximize heterogeneity and produces splits that are used to calculate the weights for the weighted mean comparison in equation (3.11).

We also consider the Causal Forest with *local centring*. This means that D_i and Y_i in equation (4.1) are replaced by $D_i - p(X_i)$ and $Y_i - \mu(X_i)$, respectively. The nuisance parameters are again estimated in a first step, and this partialling out should remove the confounding at a high level before building the Causal Forest. Athey et al. (2019) show that local centring can improve the performance of Causal Forests substantially in the presence of confounding.

We implement the regression forests for the conditional mean regression and the modified outcomes as well as the Causal Forests using the R package `grf` (Athey et al., 2019). We provide the forests with 105 baseline covariates for the prediction, and set the number of variables that are considered at each split to 70. The minimum leaf size is set to one, and one forest consists of 1,000 trees. We build honest trees such that the building of the tree and the estimation of the parameters are conducted in separate samples (Athey and Imbens, 2016). To this end, we split the sample randomly for each tree into three parts: 25% of the observations are used to build the tree, 25% are used to calculate the predictions, and 50% are left out.

Causal Forests are expected to perform well in low-dimensional settings, where the regression functions can be well approximated by trees (Wager and Athey, 2018). The generic approaches that combine standard regression Random Forests could theoretically work also in higher dimensions (Wager and Walther, 2015).

4.2. Lasso

The Lasso is a shrinkage estimator and can be considered as an OLS estimator with a penalty on the sum of the absolute coefficients. The standard least squares Lasso solves the following minimization problem:

$$\min_{\beta}\left[\sum_{i=1}^{N} w_i\,(Y_i - X_i\beta)^2\right] + \lambda \sum_{j=1}^{p} \left|\beta_j\right|, \tag{4.2}$$

where w_i are weights, p is the number of covariates, and λ is a tuning parameter to be optimally chosen.[12] We obtain the standard OLS coefficients if the penalty term is equal to zero and we have at least as many observations as covariates. For a positive penalty term, at least some coefficients are shrunk towards zero to satisfy the constraint. The Lasso serves as a variable selector because some coefficients are set to zero if the penalty term is sufficiently increased. By increasing the penalty term to a sufficiently large number, eventually all coefficients apart from the constant are zero. The idea of this procedure is to shrink those variables with little or no predictive power to zero and to use the remaining shrunk coefficients for prediction. The degree of shrinkage should be chosen to balance the bias-variance trade-off and is the crucial tuning parameter of the Lasso. The choice of the penalty term can be based on information criteria (Zou et al., 2007), data-driven procedures (Belloni et al., 2012, 2013) or cross-validation (Chetverikov et al., 2017).

We apply the R package `glmnet` to produce the predictions in the different approaches (Friedman et al., 2010). We provide the estimator a set of 1,749 potential covariates, including second-order interactions and fourth-order polynomials.[13] The tuning parameter is selected via 10-fold cross-validation.

Lasso-based estimators are expected to perform well in sparse settings where only a few variables among many potential variables are important to explain the heterogeneity in an approximately linear fashion. It is not required to know the important variables a priori.

[12] For the estimation of the propensity score, we use the equivalent logistic regression.

[13] We exclude binary variables that represent less than 1% of the observations. Furthermore, we keep only one variable of variable combinations that show correlations of magnitude larger than ±0.99 in order to speed up computation.

4.3. Cross-fitting

Some approaches require the estimation of the nuisance parameters in a first step. We follow Chernozhukov et al. (2018) and apply cross-fitting to remove bias due to the overfitting that is induced if nuisance and main parameters are estimated using the same observations. We implement a 50:50 version of their DML1 procedure. This means that we split the sample into two parts of the same size. In the first half, we estimate models for the nuisance parameters. We take these models to predict the nuisance parameters in the second half. These predicted nuisance parameters are then used in the estimation of the IATEs, $\hat{\tau}_1(x)$. We reverse the role of the two halves to obtain $\hat{\tau}_2(x)$. The estimates of the IATE are then calculated as $\hat{\tau}(x) = 1/2(\hat{\tau}_1(x) + \hat{\tau}_2(x))$.

4.4. Alternative estimation approaches

Table 2 above lists all estimators that we consider in this study. This list does not include all alternatives, because we are not able to consider all estimators that have been proposed or that would be possible combinations of the generic approaches and existing machine learning methods. We do not consider methods that are tailored to experimental studies (e.g., Imai and Ratkovic, 2013; Grimmer et al., 2017; McFowland et al., 2018). Furthermore, we consider only estimators that require at most one additional estimation step on top of the estimation of the nuisance parameters, owing to restrictions in computation power.[14] Thus, we are not able to consider the X-learner of Künzel et al. (2019), the three conditional outcome difference methods proposed by Powers et al. (2018), Orthogonal Random Forests (Oprescu et al., 2018), Modified Causal Forests (Lechner, 2018), methods based on neural nets (e.g., Johansson et al., 2016; Shalit et al., 2017; Schwab et al., 2018), Bayesian approaches like those based on Bayesian additive regression trees (BART) (Hill, 2011; Hahn et al., 2020) or Bayesian forests (Taddy et al., 2016), and potentially other approaches that we are currently not aware of.

Further, the generic approaches discussed in Section 3.1 could be implemented using different machine learning algorithms, such as Boosting, Elastic Nets, Neural Nets, Ridge or any other supervised machine learning algorithm that minimizes the required loss functions (for an overview see Hastie et al., 2009).

5. SIMULATION SET-UP

5.1. Previous empirical Monte Carlo study

The simulation study of Wendling et al. (2018) is close in spirit to our approach, in the sense that their and our DGP rely as much as possible on real data. They compare eight conditional outcome difference estimators for binary outcomes; that is, they focus on probability models. Their four DGPs are based on the covariates and the observational treatment assignment of four medical datasets. Thus, the IATEs and the resulting binary potential outcomes are the only components that need to be specified. The outcomes are simulated based on predictions of $\mu(0, x)$ and $\mu(1, x)$ from logistic neural networks (for more details, see Wendling et al., 2018).[15] This is a realistic approach in the medical context. However, it removes two important features from the

[14] Figure S12 of the Online Appendix illustrates how we leverage synergies in the implementation of the considered approaches.

[15] Nie and Wager (2017) use a similar EMCS for binary outcomes to assess the performance of different implementations of Orthogonal Learning. However, they do not estimate the IATE but specify it to depend on two covariates.

Table 3. Empirical Monte Carlo study.

1. Take the full sample and estimate the propensity score, $p^{full}(x)$, using the method and specification of choice.
2. Remove all treated and keep only the N_{nt} non-treated observations. This means that Y_i^0 is observed for all members of the remaining subpopulation.
3. Specify the true ITEs, ξ_i.
4. Calculate the potential outcome under treatment as $\hat{Y}_i^1 = Y_i^0 + \xi_i$ for all observations.
5. Set aside a random validation sample of N_v observations. Remove this validation sample from the main sample.
6. Calculate any other parameters of interest in the validation sample as benchmarks. For example equation GATEs as $\tau(g) = \left(\sum_{i=v}^{N_v} \mathbb{1}[G_v = g]\right)^{-1} \sum_{v=1}^{N_v} \mathbb{1}[G_v = g]\xi_v$ or ATEs as $\tau = N_v^{-1} \sum_{v=1}^{N_v} \xi_v$.
7. Draw a random sample of size N_s from the remaining $N_{nt} - N_v$ observations.
8. Simulate pseudo treatment indicators $D_i \sim Bernoulli(p^{sim}(x))$, where $p^{sim}(x)$ is a potentially modified version of $p^{full}(x)$, to control the ratio of treated and controls or other features of the selection process.
9. Use the observation rule in Equation (2.1) to create the observable outcome Y_i.
10. Use the N_s observations to estimate $\tau(x)$ with all estimators of interest.
11. Predict $\hat{\tau}(x)$ for all observations in the validation sample and use them to calculate $\hat{\tau}(g)$ as well as $\hat{\tau}$ for each estimator.
12. Repeat steps 7 to 11 R times.
13. Calculate performance measures.

true outcome generating process. First, the projection of the outcome on observable covariates removes the impact of unobservable variables. Second, the true error structure is lost by imposing a logistic error term. Our EMCS aims to preserve these features of the data at least for the non-treatment outcome. Wendling et al. (2018) find that conditional mean regressions (implemented with BART, see Hill, 2011) and causal boosting (Powers et al., 2018) perform consistently well, while causal Multivariate Adaptive Regression Splines (MARS) (Powers et al., 2018) and Causal Forests (Athey et al., 2019) are found to be competitive for complex IATE but perform poorly if the variance of the IATE is relatively low. We implement conditional mean regressions and Causal Forests, but omit causal boosting and MARS because of computational restrictions.

5.2. *Empirical Monte Carlo study*

Similar to Wendling et al. (2018) for the medical context, our study aims to approximate a real application in economic policy evaluation as closely as possible. The idea of an EMCS is introduced by Huber et al. (2013) and Lechner and Wunsch (2013). It aims to take as many components of the DGP as possible from real data. We build this EMCS on 96,298 observations of Swiss administrative social security data that have been used in previous evaluation studies (Behncke et al., 2010a,b; Huber et al., 2017). In particular, the EMCS mimics an evaluation of job search programmes as in Knaus et al. (2020). The interest is in the heterogeneous effects of such a programme on employment over the 33 months after the programme starts.[16]

Before we describe and motivate our EMCS approach, we list the general steps to evaluate estimators for IATEs, GATEs and ATEs in Table 3. We leave out a validation sample (10,000

[16] Online Appendix S1 provides more details about the outcomes and the rest of the dataset.

randomly drawn observations) to compare the estimated IATEs against the true ITEs, while previous EMCS compare in-sample estimates to true values of a known IATE. This modification is intended to focus on the out-of-sample predictive power of the estimated causal effects. The advantage of this procedure is that we can specify the ITEs as ground truth without knowing the IATE, as we describe below.

After removing the 10,000 observations of our validation sample, the remaining 78,844 observations form our 'population' from which we draw random subsamples of size 1,000 and 4,000 for estimation. We replicate this 2,000 times for the smaller and 500 times for the larger samples. The precision of the estimators and the computational costs increase with the sample size. Thus, we reduce the number of replications when we increase the sample size to restrict the latter. In case of $\sqrt{N}$-convergence, this will keep the simulation error approximately constant. Table 5 at the end of this section shows the variants of the DGP for different N_s, R, $p(x)$ and ξ_i. Before this, however, we explain the specification of the two latter functions.

5.2.1. Propensity score. The 'population' propensity score is estimated in the full sample with 7,454 treated and 88,844 controls. After this estimation step, all treated are removed from the sample. The specification of the propensity score is taken from Huber et al. (2017) and estimated using a standard logistic regression. We manipulate the constant to create a 50:50 split into treated and non-treated in the simulated samples.[17] Online Appendix S2.1 provides the details of the specification of the original propensity score and the distribution of the modified propensity score.

5.2.2. Specification of ITE. We are not able to observe the ITEs or any of their aggregates in a real world dataset. Therefore, we need either to estimate or to specify them. We choose the latter because estimation might favour similar estimators under investigation. Thus, our goal is to create a challenging synthetic ITE that uses components from real data. In observational studies, the estimators must be able to disentangle selection bias and effect heterogeneity. We make it hard for the estimators by using the 'population' propensity score $p^{HLM}(x)$ directly to calculate the ITEs. To this end, the propensity score is normalized and put into a sine function,

$$\omega(x) = sin\left(1.25\pi \frac{p^{HLM}(x)}{max(p^{HLM}(x))}\right) + \varepsilon_i, \tag{5.1}$$

where ε_i is random noise. This highly non-linear function of the propensity score is standardized to have mean zero and variance one before it is scaled by the parameter α:

$$\Omega(x) = \alpha \frac{\omega(x) - \bar{\omega}}{SD(\omega(x))}, \tag{5.2}$$

where $\bar{\omega}$ is the mean of $\omega(x)$, and $SD(\omega(x))$ is its standard deviation. Finally, we force the ITEs to respect two features of our outcome variable. This means that they are rounded to the next integer and that they must respect that $\hat{Y}_i^1$ falls between zero and 33.[18] Thus, the final ITEs take the form

[17] We remove the 342 observations with a modified propensity score below 5% and above 95%. We deviate at this point from the real dataset and make the problem better behaved than in reality in terms of common support (see the discussion in, e.g., Lechner and Strittmatter, 2019). We leave the investigation of performance in the presence of unbalanced ratios and insufficient common support for future studies and focus here on a relatively nice setting to start with.

[18] The histogram of the (observed) Y_i^0 is provided in Figure S1 of the Online Appendix. Given the censored and integer nature of the outcome, we considered also using Poisson Lasso to estimate the outcome regressions. However, the computation time compared with least squares Lasso is substantially longer, while the predictive performance is very similar for our outcomes. Thus, we chose the least squares version for the simulations.

$$\xi(x, y^0) = \begin{cases} \lfloor \Omega(x) \rceil & \text{if } 0 \leq y^0 + \lfloor \Omega(x) \rceil \leq 33 \\ -y^0 & \text{if } y^0 + \lfloor \Omega(x) \rceil < 0 \\ 33 - y^0 & \text{if } y^0 + \lfloor \Omega(x) \rceil > 33, \end{cases} \tag{5.3}$$

where $\lfloor \cdot \rceil$ indicates that we round to the nearest integer. The sine function of the propensity score creates the following selection of participants, which is partly motivated by empirical evidence. Figure S8 in the Online Appendix shows that unemployed persons with negative IATEs have a programme participation probability below 50% (propensity score below 0.5), whereas unemployed persons with positive IATEs have a programme participation probability above 50% (propensity score above 0.5). Accordingly, the participation probability tends to increase with the gains of participation. However, for unemployed persons with very high participation probabilities, the programme gains decrease or even become negative. Such a pattern is often called 'cream-skimming' in the labour economics literature (see, e.g., Bell and Orr, 2002). Cream-skimming means that caseworkers assign unemployed persons with good labour market opportunities (with a high Y_i^0) to the programme. However, the effectiveness of the programme is low for these unemployed persons, since they would have good labour market opportunities even without training participation. In the worst case, they suffer from negative lock-in effect (see, e.g., Card et al., 2018).

$\xi(x, y^0)$ is highly non-linear and complicated owing to the logistic function, the sine function and the rounding. Additionally, enforcement of the censoring makes it dependent on Y_i^0 which is taken directly from the data and thus depends on the covariates in an unknown fashion. Thus, we know the true ITEs but we do not know the functional form of the true IATE.

The true ITEs depend on the observables X_i and additionally on some unobservables through Y_i^0.[19] This means that the estimators approximate $\xi(x, y^0)$ using observables and produce estimates $\hat{\tau}(x)$ of $\tau(x)$. The goal of the EMCS is to figure out which estimators approximate the ITE comparatively well in this arguably realistic setting. The relative performance of the estimators thus translates directly into the ability to approximate the unknown IATEs because estimators that minimize the MSE of the ITE also minimize the MSE of the IATE (see, e.g., Künzel et al., 2019). Note that the aggregations of IATEs to GATEs and ATE in step 6 of Table 3 can be considered as true values because the influence of Y_i^0 is averaged out for them asymptotically. This implies that the MSE of GATEs and ATEs would be approximately zero if $\hat{\tau}(x) = \tau(x)$, while the MSE of ITEs might still be positive in this case.

We consider three different values of α in equation (5.2) to vary the size of the ITEs: $\alpha = 0$ (ITE0), $\alpha = 2$ (ITE1) and $\alpha = 8$ (ITE2). Additionally, we create one specification without random noise and one with an error term in equation (5.1), $\varepsilon_i = 0$ and $\varepsilon_i \sim 1 - \textit{Poisson}(1)$, respectively. Table 4 reports the basic descriptive statistics of the resulting potential outcomes, ITEs, and GATEs. ITE0 without random noise creates a benchmark scenario that is most likely to be informative about which estimators are prone to confuse effect heterogeneity with selectivity. ITE1 leads to a scenario with moderate variance of the resulting ITEs. Their standard deviation amounts to about 14% of the non-treatment outcome. ITE2 produces larger ITEs with a standard deviation of about 6, which is roughly 50% of the standard deviation of the non-treatment outcome. Thus, they should be less difficult to detect.

[19] These unobservables do not invalidate the CIA in our simulated samples, as Y_i^0 and thus the unobservables are not part of the population propensity score. The alternative to ensure a valid CIA in an EMCS is to keep the true treatment allocation structure and to specify the potential outcomes as functions of the observables. This is the approach of Wendling et al. (2018) that is discussed in Section 5.1.

Table 4. Descriptive statistics of simulated outcomes and ITEs.

	Mean	Std. dev.	Skewness	Kurtosis	Percent censored
Without random noise ($\varepsilon_i = 0$):					
Y^0 in all DGPs	16.1	12.8	−0.1	1.4	–
Y^1 in ITE0	16.1	12.8	−0.1	1.4	–
Y^1 in ITE1	16.3	12.6	−0.1	1.4	–
Y^1 in ITE2	16.3	12.6	0.1	1.5	–
ITE0	0.0	0.0	–	–	0.0
ITE1	0.1	1.8	−0.3	2.3	39.2
ITE2	0.2	6.4	−0.4	2.5	43.7
GATE0	0.0	0.0	–	–	–
GATE1	−0.4	1.8	−0.1	2.0	–
GATE2	−1.8	6.2	−0.3	2.1	–
With random noise ($\varepsilon_i \sim 1 - Poisson(1)$):					
Y^1 in ITE0	16.2	12.7	−0.1	1.4	–
Y^1 in ITE1	16.3	12.6	−0.1	1.4	–
Y^1 in ITE2	16.5	12.3	0.0	1.5	–
ITE0	0.1	0.9	−1.2	5.1	26.6
ITE1	0.1	1.8	−1.0	4.9	36.7
ITE2	0.3	6.3	−1.0	4.8	41.1
GATE0	0.0	1.1	−1.2	3.5	–
GATE1	0.0	1.7	−0.7	3.2	–
GATE2	0.0	5.8	−0.6	3.7	–

Notes: Potential outcomes and ITEs are considered for all observations. GATEs are considered for the validation sample.

The ITEs without and with random noise are created to be similar in their first two moments. However, they differ substantially in their variation that can be explained by observables. The influence of Y_i^0 on the ITEs is substantial because between 27% and 44% of the observations are censored for the nonzero ITEs. This explains why ITE1 and ITE2 without random noise are not deterministic either, and therefore not perfectly predictable by X_i. Still, the out-of-sample R^2 of Random Forest and Lasso predictive regressions shown in Table S3 of the Online Appendix document that we can explain between about 50% and 70% of the ITEs with our covariates. With random noise, this explainable part decreases to close to zero for ITE0 and to 6.3% for ITE1. We consider the latter to be a more realistic scenario because the individual component is expected to be relatively large.[20] Thus, we select ITE1 and ITE2 with random noise as our baseline DGPs in addition to the benchmark scenario ITE0 without random noise.

The first column of Table 4 shows that we specify the mean of the ITEs, the ATE, close to zero.[21] Online Appendix S2.3 describes how we aggregate the ITEs into 64 groups with sizes between 32 and 420 observations to specify the true GATEs.

[20] For example, Djebbari and Smith (2008) provide evidence that the ITEs in their applications show only little systematic variation.

[21] Online Appendix S2 shows in detail how the ITEs and potential outcomes are distributed, how the ITEs are related to the propensity score and Y_i^0, as well as an interpretation of the simulated selection behaviour of caseworkers. Note that the lower standard deviations of Y_i^1 compared with Y_i^0 result from the censoring that moves mass away from the bounds.

Table 5. List of DGPs.

	N_s	α in 5.2	Propensity score	R	ε_i in 5.1
With selection and without random noise:					
ITE0*	1,000	$\alpha = 0$	$p^{HLM}(x)$	2,000	$\varepsilon_i = 0$
ITE1	1,000	$\alpha = 2$	$p^{HLM}(x)$	2,000	$\varepsilon_i = 0$
ITE2	1,000	$\alpha = 8$	$p^{HLM}(x)$	2,000	$\varepsilon_i = 0$
ITE0*	4,000	$\alpha = 0$	$p^{HLM}(x)$	500	$\varepsilon_i = 0$
ITE1	4,000	$\alpha = 2$	$p^{HLM}(x)$	500	$\varepsilon_i = 0$
ITE2	4,000	$\alpha = 8$	$p^{HLM}(x)$	500	$\varepsilon_i = 0$
With selection and random noise:					
ITE0	1,000	$\alpha = 0$	$p^{HLM}(x)$	2,000	$\varepsilon_i \sim 1 - Poisson(1)$
ITE1*	1,000	$\alpha = 2$	$p^{HLM}(x)$	2,000	$\varepsilon_i \sim 1 - Poisson(1)$
ITE2*	1,000	$\alpha = 8$	$p^{HLM}(x)$	2,000	$\varepsilon_i \sim 1 - Poisson(1)$
ITE0	4,000	$\alpha = 0$	$p^{HLM}(x)$	500	$\varepsilon_i \sim 1 - Poisson(1)$
ITE1*	4,000	$\alpha = 2$	$p^{HLM}(x)$	500	$\varepsilon_i \sim 1 - Poisson(1)$
ITE2*	4,000	$\alpha = 8$	$p^{HLM}(x)$	500	$\varepsilon_i \sim 1 - Poisson(1)$
With random assignment and without random noise:					
ITE0	1,000	$\alpha = 0$	0.5	2,000	$\varepsilon_i = 0$
ITE1	1,000	$\alpha = 2$	0.5	2,000	$\varepsilon_i = 0$
ITE2	1,000	$\alpha = 8$	0.5	2,000	$\varepsilon_i = 0$
ITE0	4,000	$\alpha = 0$	0.5	500	$\varepsilon_i = 0$
ITE1	4,000	$\alpha = 2$	0.5	500	$\varepsilon_i = 0$
ITE2	4,000	$\alpha = 8$	0.5	500	$\varepsilon_i = 0$
With random assignment and random noise:					
ITE0	1,000	$\alpha = 0$	0.5	2,000	$\varepsilon_i \sim 1 - Poisson(1)$
ITE1	1,000	$\alpha = 2$	0.5	2,000	$\varepsilon_i \sim 1 - Poisson(1)$
ITE2	1,000	$\alpha = 8$	0.5	2,000	$\varepsilon_i \sim 1 - Poisson(1)$
ITE0	4,000	$\alpha = 0$	0.5	500	$\varepsilon_i \sim 1 - Poisson(1)$
ITE1	4,000	$\alpha = 2$	0.5	500	$\varepsilon_i \sim 1 - Poisson(1)$
ITE2	4,000	$\alpha = 8$	0.5	500	$\varepsilon_i \sim 1 - Poisson(1)$

Notes: Asteriks mark the baseline DGPs.

In summary, we consider six different scenarios defined by different choices for the scale of the ITEs and the random noise variables. In addition to the DGP with selection into the treatment, we consider also the case of an experiment with 50:50 random assignment. These twelve different DGPs are considered for sample sizes of 1,000 and 4,000 observations, leading to a total number of 24 different settings. Table 5 summarizes all parameter settings in which the eleven estimators are compared.

5.2.3. Performance measures. We consider three major performance measures: mean squared error (MSE), absolute bias ($|Bias|$) and standard deviation (SD) for the prediction of each observation v in the validation sample:[22]

$$MSE_v = \frac{1}{R}\sum_{r=1}^{R}\left[\xi\left(x_v, y_v^0\right) - \hat{\tau}(x_v)_r\right]^2, \tag{5.4}$$

[22] The formulas are written for the ITE. The same measures are used for GATE and ATE.

$$|Bias_v| = \left| \underbrace{\frac{1}{R}\sum_{r=1}^{R}\hat{\tau}(x_v)_r}_{\bar{\hat{\tau}}(x_v)_r} - \xi\left(x_v, y_v^0\right) \right|, \tag{5.5}$$

$$SD_v = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{\tau}(x_v)_r - \bar{\hat{\tau}}(x_v)_r\right]^2}. \tag{5.6}$$

Most simulation studies are interested in only a few parameters, such that the performance measure for each parameter can be reported and interpreted. However, in our case we have 10,000 parameters, such that we need to summarize the performance over the whole validation sample by taking the averages $\overline{MSE}$, $\overline{|Bias|}$ and $\overline{SD}$.[23] Additionally, we apply the Jarque–Bera (JB) test to the distribution of predictions for each observation v in the validation sample and report the fraction of observations for which normality is rejected at the 5% confidence level.[24]

6. RESULTS

6.1. *IATE estimation*

Table 6 shows the main performance measures for the three baseline DGPs.[25] First, we compare estimators within similar approaches to identify the competitive versions. Then, we compare the competitive versions over all approaches to identify those estimators that show an overall good performance, and provide a general comparison of Random Forest- and Lasso-based methods.

6.1.1. Conditional mean regressions. The Random Forest version of conditional mean regressions clearly outperforms the Lasso version in terms of mean MSE. The differences are particularly striking in the smaller-sample estimation of ITE0, where the mean MSE of the Lasso version is more than three times as large as that of the Random Forest version. The substantially worse performance of the Lasso version is consistently observed over all baseline DGPs and sample sizes. This is driven mostly by a substantially lower mean SD of Random Forest-based conditional mean regressions, which is thus the dominant choice within the two considered versions of conditional mean regressions.

6.1.2. Modified outcome methods. The ranking of the MOM estimators depends on the sample size. Table 6 shows that Random Forests are superior to the Lasso versions in the smaller samples. In particular the DR modification with Random Forests performs well owing to the relatively low mean SD. In contrast, the Lasso equivalent is by far the worst estimator in the smaller samples. It has a mean SD up to twice as large as that of the next worst estimator and consequently shows a very high mean MSE. One potential reason is that only the DR estimators require the estimation

[23] For example, $\overline{MSE} = N_v^{-1}\sum_{v=1}^{N_v} MSE_v$.

[24] Online Appendix S4 discusses and provides alternative performance measures.

[25] The full tables with more performance measures are provided in Online Appendix Tables S12 for ITE0, S16 for ITE1, and S17 for ITE2.

Table 6. Simulation results of ITE estimation for baseline DGPs.

	1,000 observations				4,000 observations			
	$\overline{MSE}$	$\lvert\overline{Bias}\rvert$	$\overline{SD}$	JB	$\overline{MSE}$	$\lvert\overline{Bias}\rvert$	$\overline{SD}$	JB
ITE0 with selection and without random noise								
Random Forest:								
Infeasible	No variation in dependent variable							
Conditional mean regression	3.69	0.62	1.78	6%	2.79	1.33	1.49	4%
MOM IPW	10.52	2.05	2.16	18%	5.76	1.89	1.74	12%
MOM DR	**2.00**	0.40	1.35	7%	1.16	0.85	1.03	8%
Causal Forest	3.52	0.75	1.69	12%	2.31	1.21	1.29	6%
Causal Forest with local centring	3.42	0.34	1.81	10%	2.05	1.13	1.40	9%
Lasso:								
Infeasible	No variation in dependent variable							
Conditional mean regression	11.21	0.69	3.19	91%	6.14	1.92	2.30	35%
MOM IPW	11.31	1.09	2.99	100%	5.02	1.56	1.93	100%
MOM DR	45.39	0.60	6.31	100%	**0.51**	0.51	0.62	35%
MCM	13.03	1.50	3.05	100%	5.79	1.65	1.94	100%
MCM with efficiency augmentation	2.08	0.42	1.36	100%	**0.48**	0.49	0.62	97%
Orthogonal Learning	**2.03**	0.45	1.33	100%	**0.47**	0.49	0.61	97%
ITE1 with selection and random noise								
Random Forest:								
Infeasible	2.98	1.29	0.15	71%	2.93	1.30	0.11	26%
Conditional mean regression	7.04	1.45	1.78	8%	6.05	1.44	1.49	4%
MOM IPW	12.92	2.26	2.20	16%	8.42	1.76	1.77	11%
MOM DR	**5.08**	1.36	1.33	8%	4.17	1.32	1.01	9%
Causal Forest	6.86	1.49	1.68	12%	5.61	1.48	1.29	6%
Causal Forest with local centring	6.50	1.35	1.79	12%	5.10	1.32	1.39	10%
Lasso:								
Infeasible	3.00	1.28	0.21	100%	2.93	1.28	0.16	83%
Conditional mean regression	14.26	1.46	3.16	90%	9.20	1.43	2.30	36%
MOM IPW	15.69	1.56	3.12	100%	8.03	1.46	2.01	100%
MOM DR	48.76	1.40	6.32	100%	**3.66**	1.34	0.64	96%
MCM	15.31	1.72	3.10	100%	8.14	1.51	1.96	100%
MCM with efficiency augmentation	5.27	1.37	1.36	100%	**3.62**	1.33	0.63	98%
Orthogonal Learning	5.16	1.38	1.29	100%	**3.65**	1.34	0.63	98%
ITE2 with selection and random noise								
Random Forest:								
Infeasible	38.46	4.43	0.52	66%	37.86	4.43	0.41	31%
Conditional mean regression	43.74	4.58	1.74	8%	42.26	4.54	1.46	5%
MOM IPW	46.83	4.83	2.23	16%	42.69	4.54	1.80	11%
MOM DR	**41.45**	4.50	1.32	10%	**40.03**	4.45	1.03	11%
Causal Forest	43.87	4.61	1.66	12%	42.34	4.58	1.29	7%
Causal Forest with local centring	42.84	4.50	1.78	12%	41.05	4.46	1.40	9%
Lasso:								
Infeasible	38.66	4.42	0.71	100%	37.84	4.40	0.53	84%
Conditional mean regression	50.11	4.52	3.15	92%	44.31	4.46	2.33	34%

Table 6. Continued.

	1,000 observations				4,000 observations			
	$\overline{MSE}$	$\lvert\overline{Bias}\rvert$	$\overline{SD}$	JB	$\overline{MSE}$	$\lvert\overline{Bias}\rvert$	$\overline{SD}$	JB
	ITE2 with selection and random noise							
MOM IPW	49.82	4.50	3.20	100%	43.21	4.43	2.17	97%
MOM DR	537.16	4.55	5.04	100%	40.11	4.48	0.76	97%
MCM	49.25	4.47	3.18	100%	42.63	4.41	2.07	100%
MCM with efficiency augmentation	41.99	4.51	1.41	100%	**40.04**	4.47	0.75	99%
Orthogonal Learning	42.13	4.54	1.35	100%	40.25	4.49	0.74	98%

Notes: $\overline{MSE}$ shows the mean MSE of all 10,000 observations in the validation sample, $\lvert\overline{Bias}\rvert$ denotes the mean absolute bias, $\overline{SD}$ the mean standard deviation, and JB the fraction of observations for which the Jarque–Bera test is rejected at the 5% level. Bold numbers indicate the best-performing estimators in terms of $\overline{MSE}$ and estimators within two standard (simulation) errors of the lowest $\overline{MSE}$.

of $\mu(d, x)$ as a nuisance parameter. These predictions are then based on only 250 observations when using cross-fitting, while $\mu(x)$ and $p(x)$ are based on 500 observations. The instability of the Lasso as outcome predictor in small samples seems to spill over to the IATE estimation.[26] The results for the larger sample size indicate that the poor performance is a small-sample issue. The DR modification with Lasso outperforms the other versions of MOM in ITE0 and ITE1 and is also close to its Random Forest equivalent for ITE2. The good performance is driven mainly by relatively low mean SDs.

As expected from the results of Powers et al. (2018), the IPW modification has a relatively high mean SD and is therefore not competitive. This is despite the fact that our DGP does not lead to extreme propensity scores and thus creates a relatively favourable setting for IPW. Therefore, the DR modification seems to be in general the dominant choice as long as the Lasso version is not used in small samples.

6.1.3. MCM and Orthogonal Learning. The results of the three estimators with modified covariates are similar to the results for the MOM. The MCM is clearly outperformed by its efficiency augmented version and Orthogonal Learning which both use the outcome regression in addition to the propensity score as a nuisance parameter. For MCM, the efficiency augmentation more than halves the mean SD in all baseline DGPs. Thus, the additional computational effort is fruitful when using MCM. For all DGPs, efficiency augmented MCM and Orthogonal Learning perform very similarly along all dimensions. This finding is in line with the synthetic simulation in Appendix D of Chen et al. (2017), who also find very similar results for these two estimators.

6.1.4. Causal Forests. The Causal Forest is specialized to maximize heterogeneity in experimental settings but it is not built to explicitly account for selection. Thus, it is prone to choose splits that do not sufficiently remove selection bias. However, Causal Forests with local centring address this problem by partialling out the selection effects in a first step. They are specialized

[26] Chernozhukov et al. (2018) observe a similar problem of global approximations for the estimation of average effects. See also Waernbaum and Pazzagli (2017) for conditions under which a poor approximation of the outcome leads to a worse performance of DR estimators compared with IPW. Similarly, Kang and Schafer (2007) demonstrate that double robust methods can perform poorly if both nuisance parameters are misspecified.

to maximize effect heterogeneity and to account for selection bias. Consequently, they uniformly perform better than Causal Forests. This is driven by a relatively low mean absolute bias, but a higher mean SD partly offsets this advantage. The differences between the two Causal Forest versions are moderate, but the version with local centring is the dominant choice if the goal is to minimize the mean MSE. However, the improvement comes at the cost of estimating an additional two nuisance parameters before estimating the Causal Forest.[27]

6.1.5. Overall comparison. The results in Table 6 show that no estimator is uniformly superior for all sample sizes and DGPs. However, we can categorize the estimators into those that show a relatively good performance in all settings, volatile ones with outstanding performance only in particular settings, and those that are never competitive.

The first category comprises Random Forest MOM DR, MCM with efficiency augmentation, Orthogonal Learning, and Causal Forest with local centring. These four estimators are in a similar range over all DGPs and sample sizes and belong consistently to the five best estimators. Thus, they seem to be reasonable choices to estimate IATEs. Causal Forest with local centring is the only one of those four that never shows the best performance in terms of mean MSE. This is driven by a larger mean SD that works against the very competitive mean absolute bias. The feature that unifies all four best-performing estimators is that they use propensity score and outcome regressions as nuisance parameters in the estimation process.

The MOM DR with Lasso belongs to the second category because it is very competitive for larger samples but the worst choice in smaller samples. Thus, it remains a risky choice for applications because the critical sample size for good performance may depend on the particular dataset.

Finally, conditional mean regressions, MOM IPW with Lasso and Causal Forest should not be considered in settings like ours if minimizing MSE has a high priority. However, if computational constraints are binding, conditional mean regressions with Random Forests and Causal Forests can be attractive options.

6.1.6. Random Forest versus Lasso. A direct comparison of Random Forest and Lasso is possible for conditional mean regressions and MOM. For the smaller sample size, Random Forest clearly outperforms the Lasso-based versions. This is driven by the substantially lower mean SD of Random Forest-based estimators. The reason is that the global approximations of Lasso are rather unstable for small samples. This instability is reduced for larger samples, and the Lasso-based MOM performs better than the Random Forest equivalents for ITE0 and ITE1.

This dependence on the sample size is not observed for the Lasso-specific estimators MCM with efficiency augmentation and Orthogonal Learning. Both show competitive performance regardless of the sample size. This is particularly surprising given the highly non-linear ITEs. However, all Lasso-based methods are far from being normally distributed. For at least 30% of the validation observations, the JB test rejects normality. For many estimators it is even rejected for all observations in the validation sample, while we would expect only a fraction of 5% to be rejected under normality. Columns 9 in the tables of Online Appendix S4.1 show that this is due to excess kurtosis, which indicates that the Lasso-based methods are prone to produce outliers. It is mitigated for the sample size of 4,000, but still the JB test is rejected for a large majority. This

[27] Together with the conditional mean regression based on Random Forests, the Causal Forest is thus attractive if computation time is a concern. Online Appendix S4.4 shows that both require a very similar computation time and are the fastest Random Forest-based estimators under consideration.

reflects the theoretical results of Leeb and Pötscher (2005, 2008) that shrinkage estimators like Lasso exhibit non-normal finite-sample distributions.

In contrast, all Random Forest-based estimators appear to be approximately normally distributed. This is also reflected by a mean skewness close to zero and a kurtosis close to three. Decently performing Random Forest-based estimators might be therefore preferable to slightly better performing Lasso-based estimators. The former produce fewer outliers and seem therefore more reliable and robust in empirical applications, as well as more amenable to statistical inference.

6.2. *GATE and ATE estimation*

Table 7 shows the main performance measures of GATE estimation for the three baseline DGPs.[28] We observe patterns similar to those for the IATE estimation, and the categorization of estimators in Section 6.1.5 remains by and large the same. The four estimators that show a consistently good performance for IATEs are also good choices for the estimation of GATEs.

For GATE estimation, we observe a new candidate with outstanding performance in a particular setting. MCM performs remarkably well for ITE2, showing the second lowest mean MSE. The mean absolute bias of MCM is already competitive for the estimation of IATEs of ITE2 in Table 6. However, the mean SD is more than twice as large compared with the best estimators, which prevents a competitive performance in terms of mean MSE. The averaging of these noisy but relatively unbiased estimators seems to produce a competitive estimator for the higher aggregation level. Still, MCM performs poorly for the other DGPs.

The averaging improves also the performance of locally centred Causal Forests relative to its uncentred version. The results for the estimation of IATEs show that the advantage in terms of mean absolute bias is partly offset by a higher variability. The aggregation step reduces this difference such that the lower bias translates into a substantially lower mean MSE.

The aggregation also leads to a substantial reduction in the excess kurtosis of all Lasso estimators (see tables of Online Appendix S4.2). However, the JB test is still rejected for most observations. Note that we observe for all estimators a substantial amount of bias, although the influence of Y_i^0 and the irreducible noise is averaged out to a large extent. This indicates that the estimators are not able to completely remove the selection bias, which is particularly problematic if we are interested in statistical inference. The results for ITE0 without noise and with random assignment in Online Appendix S4.2 provide evidence in this direction.

The results for the estimation of ATEs in Online Appendix S4.3 are mostly in line with those for GATEs. Again, MCM is highly competitive and provides the best-performing estimators for ITE2. In addition, the benefits of averaging the locally centred Causal Forest are observed. The bias is halved compared with the uncentred version, while both versions show similar SDs.

The skewness and kurtosis show that the ATE estimators are mostly normally distributed with mean skewness close to zero and mean kurtosis close to three. The obvious exception is MOM DR with Lasso, for which also averaging the IATEs does not mitigate the bad performance due to extreme outliers.

Finally, we note that the comparison of the mean MSE for the two sample sizes indicates that GATE and ATE estimators show a substantially faster convergence rate compared with the respective IATE estimators. This indicates that the additional averaging of noisily estimated

[28] The full tables with more performance measures are provided in Online Appendix Tables S26 for ITE0, S30 for ITE1, and S31 for ITE2. The results for all DGPs are provided in Online Appendix S4.2.

Table 7. Simulation results of GATE estimation for baseline DGPs.

	1,000 observations				4,000 observations			
	$\overline{MSE}$	$\lvert\overline{Bias}\rvert$	$\overline{SD}$	JB	$\overline{MSE}$	$\lvert\overline{Bias}\rvert$	$\overline{SD}$	JB
GATEs from ITE0 with selection and without random noise								
Random Forest:								
Conditional mean regression	1.77	0.55	1.19	22%	1.07	0.49	0.86	8%
MOM IPW	4.44	1.59	1.16	47%	1.12	0.67	0.66	6%
MOM DR	**0.87**	0.38	0.85	20%	0.30	0.25	0.48	14%
Causal Forest	1.44	0.70	0.96	17%	0.74	0.64	0.53	0%
Causal Forest with local centring	1.08	0.33	0.99	8%	0.35	0.22	0.54	3%
Lasso:								
Conditional mean regression	3.35	0.55	1.70	34%	1.45	0.47	1.06	6%
MOM IPW	3.15	0.78	1.50	100%	1.19	0.53	0.87	86%
MOM DR	38.85	0.59	6.20	100%	0.30	0.31	0.45	38%
MCM	4.56	1.20	1.62	100%	1.65	0.75	0.92	94%
MCM with efficiency augmentation	1.04	0.41	0.93	66%	**0.27**	0.26	0.45	55%
Orthogonal Learning	1.04	0.44	0.92	73%	**0.27**	0.28	0.44	25%
GATEs from ITE1 with selection and random noise								
Random Forest:								
Conditional mean regression	2.30	0.85	1.18	20%	1.53	0.76	0.86	6%
MOM IPW	3.84	1.41	1.17	41%	0.99	0.54	0.67	11%
MOM DR	**1.16**	0.59	0.83	20%	**0.49**	0.42	0.48	17%
Causal Forest	2.04	1.01	0.95	19%	1.26	0.93	0.53	2%
Causal Forest with local centring	1.38	0.56	0.97	17%	0.58	0.44	0.54	5%
Lasso:								
Conditional mean regression	3.68	0.78	1.69	41%	1.66	0.60	1.06	14%
MOM IPW	3.03	0.65	1.53	100%	1.17	0.47	0.89	77%
MOM DR	39.33	0.79	6.20	100%	0.59	0.50	0.46	42%
MCM	3.98	0.93	1.65	97%	1.31	0.52	0.93	91%
MCM with efficiency augmentation	1.40	0.62	0.93	80%	0.55	0.47	0.45	39%
Orthogonal Learning	1.45	0.68	0.91	75%	0.61	0.51	0.45	48%
GATEs from ITE2 with selection and random noise								
Random Forest:								
Conditional mean regression	5.28	1.57	1.15	20%	3.97	1.44	0.85	22%
MOM IPW	3.75	1.25	1.18	41%	1.72	0.95	0.68	14%
MOM DR	**3.50**	1.29	0.84	23%	2.19	1.08	0.49	23%
Causal Forest	5.33	1.71	0.94	25%	4.17	1.60	0.53	8%
Causal Forest with local centring	3.74	1.28	0.98	11%	2.43	1.12	0.55	9%
Lasso:								
Conditional mean regression	5.73	1.34	1.71	42%	2.59	0.95	1.09	11%
MOM IPW	4.00	1.01	1.59	100%	1.92	0.82	0.96	45%
MOM DR	30.36	1.43	4.71	100%	2.71	1.23	0.52	69%
MCM	3.65	0.66	1.67	100%	**1.58**	0.61	0.97	81%
MCM with efficiency augmentation	3.94	1.35	0.95	72%	2.63	1.22	0.51	67%
Orthogonal Learning	4.27	1.43	0.93	72%	2.95	1.29	0.50	50%

Notes: $\overline{MSE}$ shows the mean MSE of all 10,000 observations in the validation sample, $|\overline{Bias}|$ denotes the mean absolute bias, $\overline{SD}$ the mean standard deviation, and JB the fraction of observations for which the Jarque–Bera test is rejected at the 5% level. Bold numbers indicate the best-performing estimators in terms of $\overline{MSE}$ and estimators within two standard (simulation) errors of the lowest $\overline{MSE}$.

IATEs results in faster convergence, and the ATEs may be estimable with close to parametric rates. However, we do not overemphasize this finding as it is only based on two sample sizes.

6.3. *Alternative DGPs*

The discussions in the previous subsections focus on the results of the three baseline DGPs. This subsection summarizes the major insights from the alternative DGPs. The results are provided in Online Appendix S4, where we also discuss the details and peculiarities of the specific DGPs and aggregation levels. In general, the four estimators that are identified as the best performing for the baseline DGPs also belong to the best-performing ones for the alternative DGPs.

For the estimation of *IATEs*, we observe new candidates that are only successful in particular DGPs with selection into treatment. For example, Random Forest MOM IPW performs very well for ITE2 without noise. However, these and other peculiarities discussed in the Appendix hold only for either mean MSE or median MSE, while the four best-performing candidates are usually competitive in both measures. Additionally, we assess whether our findings remain robust if we ignore the natural bounds of our outcome variable when creating the DGP. The results in Online Appendix S4.1.5 show that this is the case in a DGP that allows treated outcomes outside the natural bounds of the original outcome when defining the ITEs.

We also consider all previously discussed DGPs with *random treatment assignment*. This means that the estimation problem becomes easier because selection bias is no longer a concern. By and large, the results are in line with the respective results for the DGPs with selection into treatment; in particular, the conclusions about the best-performing estimators are not changed. As expected, the mean MSEs for the DGP with random assignment are lower for most of the estimators and thus closer to the infeasible benchmark. This is always driven by a lower mean absolute bias, while the mean SDs are very similar to the equivalent DGPs with selection. This suggests that the methods are not able to completely remove the selection bias.

Two other differences from the baseline DGPs are noteworthy. First, MOM DR with Lasso shows competitive performance even in small samples. This indicates that the bad performance is related to large errors made in the outcome *and* the selection equation in small samples, which is in line with the simulation evidence of Kang and Schafer (2007) for DR ATE estimators. Second, Causal Forest and its locally centred version show a nearly identical performance. This illustrates that local centring is only beneficial when there is selection into treatment.

The results for *GATE and ATE* estimators confirm the observation in the baseline DGPs that IATE estimators with low bias but high variance can provide competitive estimators if they are averaged to higher aggregation levels. In particular, averaging MCM IATEs shows in many alternative DGPs a relatively good performance. However, MCM performs worst in some other DGPs in a non-systematic way. Thus, the results show that noise can be averaged out for these higher aggregation levels, but there is no guarantee of this. The estimators that are already successful for the IATEs remain the most reliable choices for GATEs and ATEs.

In general, we find that the differences between the estimators become smaller, the more the IATEs are aggregated. In particular, the SDs become more similar by averaging IATEs such that the differences between the estimators are driven mainly by bias.

7. CONCLUSION

This is the first comprehensive simulation study in economics to investigate the finite-sample performance of a large number of causal machine learning estimators. We rely on arguably realistic DGPs that have potentially more external validity than the mostly synthetic DGPs considered so far in the limited simulation literature for these estimators. We consider DGPs with and without selection into treatment. Our main goal is to estimate individualized average treatment effects. Additionally, we report the performance of estimators that aggregate individualized average treatment effects to an intermediate and the population level.

We do not find any single causal machine learning estimator that consistently outperforms all other estimators. However, we do find a group of four estimators that show competitive performance for all DGPs. This group comprises the Causal Forest with local centring, Random Forest-based MOM DR, MCM with efficiency augmentation, and Orthogonal Learning. These four estimators seem to be good choices in settings that are similar to our empirical Monte Carlo design. The best-performing estimators explicitly use both the outcome and the treatment equations in a multiple-step procedure. The estimators that use the Lasso have heavy tails in the smaller samples.

The best-performing estimators for the individualized average treatment effects also produce the most reliable estimators for higher aggregation levels. However, in some settings noisily estimated individualized average treatment effects with low bias also produce competitive estimators for higher aggregates because the noise is averaged out.

Despite relying as much as possible on arguably realistic DGPs, the external validity of every simulation study is uncertain. Future research will show if our findings hold in other empirical settings. Furthermore, it may be possible to improve the performance of each method with more tailored implementations. Finally, we have focused in this study on the finite-sample performance of point estimates and leave the investigation of inference procedures to future research.

ACKNOWLEDGEMENTS

Financial support from the Swiss National Science Foundation (SNSF) is gratefully acknowledged. The study is part of the project ‘Causal Analysis with Big Data’ (grant number SNSF 407540_166999) of the Swiss National Research Programme ‘Big Data’ (NRP 75). A previous version of the paper was presented at the *Annual Conference of the International Association for Applied Econometrics* in Nikosia, the *Machine Learning in Economics* workshop at the Barcelona GSE Summer Forum, the Economics Departments of the University of Wisconsin at Madison, the Universities of California at Berkeley, Davis, Irvine, Los Angeles, and San Diego, at Amazon, Seattle, *Econometrics in the Castle: Machine Learning in Economics and Econometrics* in Munich, and the *Statistical Week* in Linz. We thank participants, in particular Bryan Graham, Andres Santos, Alejandro Schuler, Stefan Wager and Michael Zimmert, for helpful comments and suggestions. The usual disclaimer applies.

REFERENCES

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*, 1–19.

Abrevaya, J., Y.-C. Hsu and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business and Economic Statistics 33*, 485–505.

Andini, M., E. Ciani, G. de Blasio, A. D'Ignazio and V. Salvestrini (2018). Targeting with machine learning: an application to a tax rebate program in Italy. *Journal of Economic Behavior and Organization 156*, 86–102.

Ascarza, E. (2018). Retention futility: targeting high risk customers might be ineffective. *Journal of Marketing Research 55*, 80–98.

Athey, S. (2018). The impact of machine learning on economics. In A. K. Agrawal, J. Gans and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*. Chicago, IL: University of Chicago Press, 507–47.

Athey, S. and G. W. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*, 7353–60.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Athey, S., J. Tibshirani and S. Wager (2019). Generalized random forests. *Annals of Statistics 47*, 1148–78.

Behncke, S., M. Frölich and M. Lechner (2010a). A caseworker like me – Does the similarity between the unemployed and their caseworkers increase job placements? *Economic Journal 120*, 1430–59.

Behncke, S., M. Frölich and M. Lechner (2010b). Unemployed and their caseworkers: should they be friends or foes? *Journal of the Royal Statistical Society: Series A (Statistics in Society) 173*, 67–92.

Bell, S. H. and L. L. Orr (2002). Screening (and creaming?) applicants to job training programs: the AFDC homemaker – home health aide demonstrations. *Labour Economics 9*, 279–301.

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*, 2369–429.

Belloni, A., V. Chernozhukov, I. Fernández-Val and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*, 233–98.

Belloni, A., V. Chernozhukov and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Belloni, A., V. Chernozhukov and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*, 608–50.

Belloni, A., V. Chernozhukov and C. Hansen (2013). Inference for high-dimensional sparse econometric models. In Acemoglu, D., M. Arellano and E. Dekel (Eds.), *Advances in Economics and Econometrics: Tenth World Congress* (Econometric Society Monographs, pp. 245–95). Cambridge: Cambridge University Press.

Bertrand, M., B. Crépon, A. Marguerie and P. Premand (2017). Contemporaneous and post-program impacts of a public works program: evidence from Côte d'Ivoire, Working Paper, World Bank. https://elibrary.worldbank.org/doi/abs/10.1596/28460

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

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen (1984). *Classification and Regression Trees*. CRC Press, New York.

Card, D., J. Kluve and A. Weber (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association 16*, 894–931.

Chen, S., L. Tian, T. Cai and M. Yu (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics 73*, 1199–209.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen and W. Newey (2017). Double/Debiased/Neyman machine learning of treatment effects. *American Economic Review Papers and Proceedings 107*(5), 261–65.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018). Double/Debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–68.

Chetverikov, D., Z. Liao and V. Chernozhukov (2017). On cross-validated Lasso, arXiv:1605.02214.

Davis, J. M. and S. B. Heller (2017). Using causal forests to predict treatment heterogeneity: an application to summer jobs. *American Economic Review 107*(5), 546–50.

Djebbari, H. and J. Smith (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics 145*, 64–80.

Fan, Q., Y.-C. Hsu, R. P. Lieli and Y. Zhang (2019). Estimation of conditional average treatment effects with high-dimensional data, arXiv:1908.02399.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*, 1–23.

Foster, D. J. and V. Syrgkanis (2019). Orthogonal statistical learning, arXiv:1901.09036.

Friedman, J., T. Hastie and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Grimmer, J., S. Messing and S. J. Westwood (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis 25*, 413–34.

Hahn, P. R., J. S. Murray and C. Carvalho (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, advance publication, 31 January 2020. doi:10.1214/19-BA1195.

Hastie, T., R. Tibshirani and J. Friedman (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction (2nd ed.)*. New York: Springer.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*, 217–40.

Hirano, K., G. W. Imbens and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–89.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–85.

Huber, M., M. Lechner and G. Mellace (2017). Why do tougher caseworkers increase employment? The role of program assignment as a causal mechanism. *Review of Economics and Statistics 99*, 180–83.

Huber, M., M. Lechner and C. Wunsch (2013). The performance of estimators based on the propensity score. *Journal of Econometrics 175*, 1–21.

Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics 7*, 443–70.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. In Balcan, M. F. and K. Q. Weinberger (Eds.), *Journal of Economic Literature*. New York: PMLR, *47*, 5–86.

Johansson, F., U. Shalit and D. Sontag (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning*, 3020–29.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*, 523–39.

Knaus, M. C., M. Lechner and A. Strittmatter (2020). Heterogeneous employment effects of job search programmes: a machine learning approach, J. Human Resources 0718-9615R1; published ahead of print March 26, 2020.

Künzel, S. R., J. S. Sekhon, P. J. Bickel and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences 116*, 4156–65.

Lechner, M. (2018). Modified causal forests for estimating heterogeneous causal effects, arXiv:1812.09487.

Lechner, M. and A. Strittmatter (2019). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews 38*, 193–207.

Lechner, M. and C. Wunsch (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics 21*, 111–21.

Lee, S., R. Okui and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*, 1207–25.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: facts and fiction. *Econometric Theory 21*, 21–59.

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of Hodges estimator. *Journal of Econometrics 142*, 201–11.

McFowland, E., S. Somanchi and D. B. Neill (2018). Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection, arXiv:1803.09159.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics 57*, 213–34.

Nie, X. and S. Wager (2017). Quasi-oracle estimation of heterogeneous treatment effects, arXiv:1712.04912.

Oprescu, M., V. Syrgkanis and Z. S. Wu (2019). Orthogonal random forest for causal inference. In International Conference on Machine Learning, pp. 4932–41.

Powers, S., J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie and R. Tibshirani (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine 37*, 1767–87.

Qian, M. and S. A. Murphy (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics 39*, 1180.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–29.

Rolling, C. A., Y. Yang and D. Velez (2019). Combining estimates of conditional treatment effects. *Econometric Theory 35*, 1089–110.

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

Schuler, A., M. Baiocchi, R. Tibshirani and N. Shah (2018). A comparison of methods for model selection when estimating individual treatment effects, arXiv:1804.05146.

Schwab, P., L. Linhardt and W. Karlen (2018). Perfect match: a simple method for learning representations for counterfactual inference with neural networks, arXiv:1810.00656.

Semenova, V. and V. Chernozhukov (2017). Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions, arXiv:1702.06240.

Shalit, U., F. D. Johansson and D. Sontag (2017). Estimating individual treatment effect: generalization bounds and algorithms, In International Conference on Machine Learning, pp. 3076–85.

Signorovitch, J. E. (2007). *Identifying informative biological markers in high-dimensional genomic data and clinical trials*, PhD thesis, Harvard University.

Strittmatter, A. (2018). What is the value added by using causal machine learning methods in a welfare experiment evaluation?, arXiv:1812.06533.

Su, X., C.-L. Tsai, H. Wang, D. M. Nickerson and B. Li (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research 10*, 141–58.

Taddy, M., M. Gardner, L. Chen and D. Draper (2016). A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation. *Journal of Business and Economic Statistics 34*, 661–72.

Tian, L., A. A. Alizadeh, A. J. Gentles and R. Tibshirani (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association 109*, 1517–32.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 267–88.

Waernbaum, I. and L. Pazzagli (2017). Model misspecification and bias for inverse probability weighting and doubly robust estimators. 01–25, arXiv:1711.09388.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*, 1228–42.

Wager, S. and G. Walther (2015). Adaptive concentration of regression trees, with application to random forests, arXiv:1503.06388.

Wendling, T., K. Jung, A. Callahan, A. Schuler, N. H. Shah and B. Gallego (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine 37*, 3309–24.

Zhang, B., A. A. Tsiatis, E. B. Laber and M. Davidian (2012). A robust method for estimating optimal treatment regimes. *Biometrics 68*, 1010–18.

Zhao, Q., D. S. Small and A. Ertefaie (2017). Selective inference for effect modification via the Lasso, arXiv:1705.08020.

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding, arXiv:1908.08779.

Zou, H., T. Hastie and R. Tibshirani (2007). On the 'degrees of freedom' of the Lasso. *Annals of Statistics 35*(5), 2173–92.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix

Co-editor Victor Chernozhukov handled this manuscript.