JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 541, 242–256: Theory and Methods
https://doi.org/10.1080/01621459.2021.1923511

Taylor & Francis
Taylor & Francis Group

Experimental Evaluation of Individualized Treatment Rules

Kosuke Imai [a] and Michael Lingzhi Li [b]

[a]Department of Government and Department of Statistics, Harvard University, Cambridge, MA; [b]Operation Research Center, Massachusetts Institute of Technology, Cambridge, MA

ABSTRACT

The increasing availability of individual-level data has led to numerous applications of individualized (or personalized) treatment rules (ITRs). Policy makers often wish to empirically evaluate ITRs and compare their relative performance before implementing them in a target population. We propose a new evaluation metric, the population average prescriptive effect (PAPE). The PAPE compares the performance of ITR with that of non-individualized treatment rule, which randomly treats the same proportion of units. Averaging the PAPE over a range of budget constraints yields our second evaluation metric, the area under the prescriptive effect curve (AUPEC). The AUPEC represents an overall performance measure for evaluation, like the area under the receiver and operating characteristic curve (AUROC) does for classification, and is a generalization of the QINI coefficient used in uplift modeling. We use Neyman's repeated sampling framework to estimate the PAPE and AUPEC and derive their exact finite-sample variances based on random sampling of units and random assignment of treatment. We extend our methodology to a common setting, in which the same experimental data are used to both estimate and evaluate ITRs. In this case, our variance calculation incorporates the additional uncertainty due to random splits of data used for cross-validation. The proposed evaluation metrics can be estimated without requiring modeling assumptions, asymptotic approximation, or resampling methods. As a result, it is applicable to any ITR including those based on complex machine learning algorithms. The open-source software package is available for implementing the proposed methodology. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2020
Accepted April 2021

KEYWORDS
Causal inference; Heterogeneous treatment effects; Machine learning; Precision medicine; Uplift modeling

1. Introduction

In today's data-rich society, the individualized (or personalized) treatment rules (ITRs), which assign different treatments to individuals based on their observed characteristics, play an essential role. Examples include personalized medicine and micro-targeting in business and political campaigns (e.g., Hamburg and Collins 2010; Imai and Strauss 2011). In the causal inference literature, a number of researchers have developed methods to estimate optimal ITRs using a variety of machine learning algorithms (see, e.g., Qian and Murphy 2011; Zhang et al. 2012; Fu, Zhou, and Faries 2016; Luedtke and van der Laan 2016a,b; Zhou et al. 2017; Athey and Wager 2018; Kitagawa and Tetenov 2018). In addition, applied researchers often use machine learning algorithms to estimate heterogeneous treatment effects and then construct ITRs based on the resulting estimates.

In this article, we consider a common setting, in which a policymaker wishes to experimentally evaluate the empirical performance of an ITR before implementing it in a target population. Such evaluation is also essential for comparing the efficacy of alternative ITRs. Specifically, we show how to use a randomized experiment for evaluating ITRs. We propose two new evaluation metrics. The first is the population average prescriptive effect (PAPE), which compares an ITR with a nonindividualized treatment rule that randomly assigns the same proportion of units to the treatment condition. The PAPE represents the difference between the average outcome under the ITR and that under the random treatment rule. The key idea is that a well-performing ITR should outperform the random treatment rule, which does not use any individual-level information.

Averaging the PAPE over a range of budget constraints yields our second evaluation metric, the area under the prescriptive effect curve (AUPEC). Like the area under the receiver and operating characteristic curve (AUROC) for classification, the AUPEC represents an overall summary measure of how well an ITR performs over the random treatment rule that treats the same proportion of units.

We estimate these evaluation metrics using Neyman's (1923) repeated sampling framework (see Imbens and Rubin 2015, chap. 6). An advantage of this approach is that it does not require any modeling assumption or asymptotic approximation. As a result, we can evaluate a broad class of ITRs including those based on complex machine learning algorithms. We show how to estimate the PAPE and AUPEC with a minimal amount of finite sample bias and derive the exact variance solely based on random sampling of units and random assignment of treatment.

CONTACT Michael Lingzhi Li mlli@mit.edu Operations Research Center, 77 Massachusetts Ave E40-103, Massachusetts Institute of Technology, Cambridge, MA 02139.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

We further extend this methodology to a common evaluation setting, in which the same experimental data is used to both estimate and evaluate ITRs. In this case, our finite-sample variance calculation is exact and directly incorporates the additional uncertainty due to random splits of data used for cross-validation. We implement the proposed methodology through an open-source R package evalITR available at *https://CRAN.R-project.org/package=evalITR*.

Our simulation study demonstrates the accurate coverage of the proposed confidence intervals in small samples (Section 5). We also apply our methods to the Project STAR (Student-Teacher Achievement Ratio) experiment and compare the empirical performance of ITRs based on several popular methods (Section 6). Our evaluation approach addresses theoretical and practical difficulties of conducting reliable statistical inference for ITRs.

1.1. Relevant literature

A large number of existing studies have focused on the derivation of optimal ITRs that maximize the population average value. For example, Qian and Murphy (2011) used penalized least square, whereas Zhao et al. (2012) showed how a support vector machine can be used to derive an optimal ITR. Another popular approach is based on doubly robust estimation (e.g., Dudík, Langford, and Li 2011; Zhang et al. 2012; Chakraborty, Laber, and Zhao 2014; Jiang and Li 2016; Athey and Wager 2018; Kallus 2018).

We propose a general methodology for empirically evaluating and comparing the performance of various ITRs including the ones proposed by these and other authors. While many of these methods come with uncertainty measures, even those that produce standard errors rely on asymptotic approximation, modeling assumptions, or resampling methods. In contrast, our methodology utilizes Neyman's repeated sampling framework and does not require any of these assumptions or approximations.

There also exists a related literature on policy evaluation. Starting with Manski (2004), many studies focus on the derivation of regret bounds given a class of ITRs. For example, Kitagawa and Tetenov (2018) showed that an ITR, which maximizes the empirical average value, is minimax optimal without a strong restriction on the class of ITRs, whereas Athey and Wager (2018) establish a regret bound for an ITR based on doubly robust estimation in observational studies (see also Zhou, Athey, and Wager 2018). In addition, Luedtke and van der Laan (2016a,b) proposed consistent estimators of the optimal average value even when an optimal ITR is not unique (see also Rai 2018).

Our goal is different from these studies. We focus on statistical inference using the Neyman's repeated sampling framework for the experimental evaluation of arbitrary ITRs including optimal or non-optimal and simple or complex ones. Our evaluation metric is also different from the existing metrics. In particular, to the best of our knowledge, we are the first to formally study the AUPEC as an AUROC-like summary measure for evaluation.

In contrast, much of the policy evaluation literature focus on the optimal average value, which is required to compute the regret of an ITR. Athey and Wager (2018) briefly discussed a quantity related to the PAPE in their empirical application, but this quantity evaluates an ITR against the treatment rule that randomly treats exactly one half of units rather than the same proportion as the one treated under the ITR. Although empirical studies in the campaign and marketing literatures have used "uplift modeling," which is based on the PAPE (e.g., Imai and Strauss 2011; Rzepakowski and Jaroszewicz 2012; Gutierrez and Gérardy 2016; Ascarza 2018; Fifield 2018), none develops formal estimation and inferential methods. We show that the AUPEC is a generalization of the QINI coefficient, which is a widely utilized statistic in uplift modeling (Radcliffe 2007; Diemert et al. 2018). Thus, our theoretical results for the AUPEC apply directly to the QINI coefficient as well.

Another related literature is concerned with the estimation of heterogeneous effects. Researchers have explored the use of tree-based methods (e.g., Imai and Strauss 2011; Athey and Imbens 2016; Wager and Athey 2018; Hahn, Murray, and Carvalho 2020), regularized regressions (e.g., Imai and Ratkovic 2013; Künzel et al. 2019), and ensemble methods (e.g., van der Laan and Rose 2011; Grimmer, Messing, and Westwood 2017). In practice, the estimated heterogeneous treatment effects based on these machine learning algorithms are used to construct ITRs.

However, as Chernozhukov et al. (2019) pointed out, most machine learning algorithms, which require data-driven tuning parameters, cannot be regarded as consistent estimators of the CATE unless strong assumptions are imposed. the authors proposed a methodology to estimate heterogeneous treatment effects without such assumptions. Similar to theirs, our methodology does not depend on any modeling assumption and accounts for the uncertainty due to splitting of data. The key difference is that we focus on the evaluation of ITRs. In addition, our variance calculation is based on randomization and does not rely on asymptotic approximation or resampling methods.

Finally, Andrews, Kitagawa, and McCloskey (2020) developed a conditional inference procedure, based on normal approximation, for the average value of the best-performing policy based on experimental or observational data. In contrast, we develop an unconditional exact inference for the difference in the average value between any pair of policies under a budget constraint. We also consider the evaluation of estimated policies based on the same experimental data using cross-validation, whereas Andrews, Kitagawa, and McCloskey focused on the evaluation of fixed policies.

2. Evaluation Metrics

In this section, we introduce our evaluation metrics. We first propose the PAPE that, unlike the population average value, adjusts for the proportion of units treated by an ITR. The idea is that an efficacious ITR should outperform a nonindividualized treatment rule, which randomly assigns the same proportion of units to the treatment condition. We extend the PAPE to the settings with a binding budget constraint. Finally, we propose the AUPEC as a univariate summary performance measure of an ITR under a range of budget constraint.

2.1. The Setup

Following the literature, we define an ITR as a deterministic map from the covariate space $\mathcal{X}$ to the binary treatment assignment

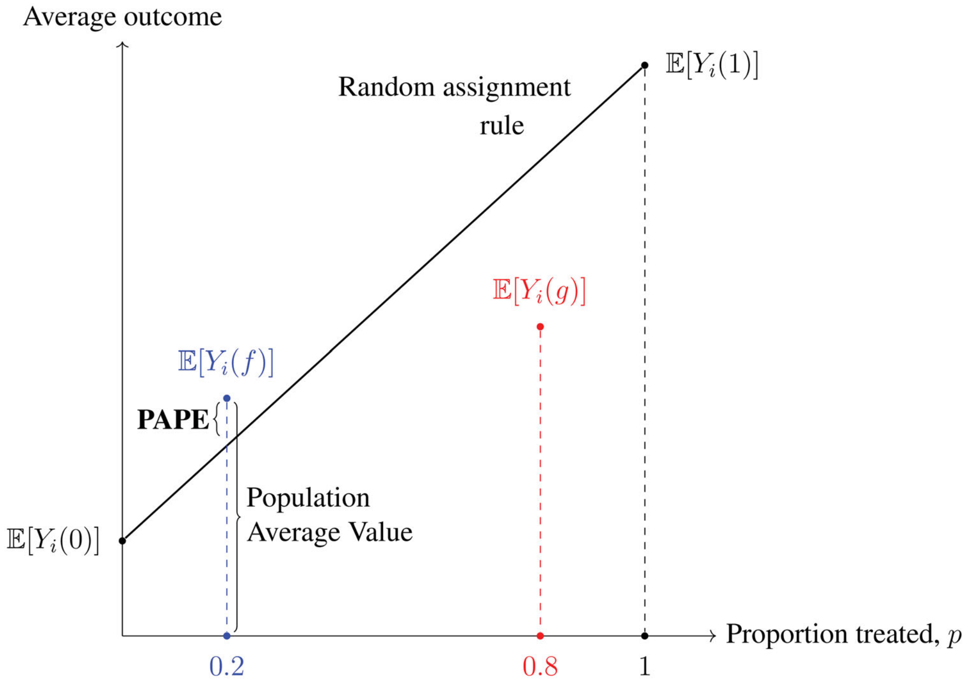

Figure 1. The Importance of Accounting for the Proportion of Treated Units. In this illustrative example, an ITR g (red) outperforms another ITR f (blue) in terms of the population average value, that is, $\mathbb{E}[Y_i(f)] < \mathbb{E}[Y_i(g)]$. However, unlike f, the ITR g is doing worse than the random treatment rule (black). In contrast, the PAPE measures the performance of an ITR as the difference in the average value between the ITR and random treatment rule.

(e.g., Qian and Murphy 2011; Zhao et al. 2012),

$$f : \mathcal{X} \longrightarrow \{0, 1\}.$$

Let T_i denote the treatment assignment indicator variable, which is equal to 1 if unit i is assigned to the treatment condition, that is, $T_i \in \mathcal{T} = \{0, 1\}$. For each unit, we observe the outcome variable $Y_i \in \mathcal{Y}$ as well as the vector of pretreatment covariates, $\mathbf{X}_i \in \mathcal{X}$, where $\mathcal{Y}$ is the support of the outcome variable. We assume no interference between units and denote the potential outcome for unit i under the treatment condition $T_i = t$ as $Y_i(t)$ for $t = 0, 1$. Then, the observed outcome is given by $Y_i = Y_i(T_i)$.

Assumption 1 (No interference between units). The potential outcomes for unit i do not depend on the treatment status of other units. That is, for all $t_1, t_2, \ldots, t_n \in \{0, 1\}$, we have, $Y_i(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n) = Y_i(T_i = t_i)$.

Under this assumption, the existing literature almost exclusively focuses on the derivation of an optimal ITR that maximizes the following population average value (e.g., Qian and Murphy 2011; Zhao et al. 2012; Zhou et al. 2017),

$$\lambda_f = \mathbb{E}\{Y_i(f(\mathbf{X}_i))\}. \quad (1)$$

Next, we show that λ_f may not be the best evaluation metric in some cases.

2.2. The Population Average Prescriptive Effect

We now introduce our main evaluation metric, the PAPE. The PAPE is based on two ideas. First, it is reasonable to expect a good ITR to outperform a *nonindividualized* treatment rule, which does not use any information about individual units when deciding who should receive the treatment. Second, a budget constraint should be considered since the treatment is often costly. This means that a good ITR should identify units who benefit from the treatment most. These two considerations lead to the random treatment rule as a natural baseline for comparison, which assigns, with equal probability, the same proportion of units to the treatment condition.

Figure 1 illustrates the importance of accounting for the proportion of units treated by an ITR. In this figure, the horizontal axis represents the proportion treated and the vertical axis represents the average outcome under an ITR. The example shows that an ITR g (red), which treats 80% of units, has a greater average value than another ITR f (blue), which treats 20% of units, that is, $\mathbb{E}[Y_i(f)] < \mathbb{E}[Y_i(g)]$. Despite this fact, g is outperformed by the random treatment rule (black), which treats the same proportion of units, whereas f does a better job than the random treatment rule. This is indicated by the fact that the black solid line is placed above $\mathbb{E}[Y_i(g)]$ and below $\mathbb{E}[Y_i(f)]$.

To overcome this undesirable property of the average value, we propose an alternative evaluation metric that compares the performance of an ITR with that of the random treatment rule. The random treatment rule serves as a natural baseline because it treats the same proportion of units without any individual information. This is analogous to the predictive setting, in which a classification algorithm is often compared to random classification.

Formally, let $p_f = \Pr(f(\mathbf{X}_i) = 1)$ denote the population proportion of units assigned to the treatment condition under ITR f. Without loss of generality, we assume a positive average treatment effect $\tau = \mathbb{E}\{Y_i(1) - Y_i(0)\} > 0$ so that the random treatment rule assigns the exactly proportion $p_f > 0$ of the units to the treatment group. If the treatment is on average harmful (a testable condition using the experimental data), the best random treatment rule is to treat no one. In that case, the estimation of the average value is sufficient for the evaluation. We define the population average prescription effect (PAPE) of ITR f as the following difference in the average value between the ITR and random treatment rule,

$$\tau_f = \mathbb{E}\{Y_i(f(\mathbf{X}_i)) - p_f Y_i(1) - (1-p_f)Y_i(0)\}. \qquad (2)$$

One motivation for the PAPE is that administering a treatment is often expensive. Consider a costly treatment that does not harm anyone but only benefits a relatively small fraction of people. If we do not impose a budget constraint, then treating everyone is the best ITR but such a policy does not use any individual-level information. Thus, to further evaluate the efficacy of an ITR, we extend the PAPE to the settings with a budget constraint.

2.3. Incorporating a Budget Constraint

With a budget constraint, we cannot simply treat all units who are predicted to benefit from the treatment. Instead, an ITR must be based on a *scoring rule* that sorts units according to their treatment priority: a unit with a greater score has a higher priority to receive the treatment. Let $s : \mathcal{X} \longrightarrow \mathcal{S}$ be such a scoring rule where $\mathcal{S} \subset \mathbb{R}$. For simplicity, we assume that the scoring rule is bijective, that is, $s(\mathbf{X}) \neq s(\mathbf{X}')$ for any $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ with $\mathbf{X} \neq \mathbf{X}'$. This assumption is not restrictive as we can always redefine $\mathcal{X}$ such that the assumption holds.

We define an ITR based on a scoring rule by assigning a unit to the treatment group if and only if its score is higher than a threshold, c,

$$f(\mathbf{X}_i, c) = \mathbf{1}\{s(\mathbf{X}_i) > c\}.$$

Under a binding budget constraint p, we define the threshold that corresponds to the maximal proportion of treated units under the budget constraint, that is,

$$c_p(f) = \inf\{c \in \mathbb{R} : \Pr(f(\mathbf{X}_i, c) = 1) \leq p\}.$$

Our framework allows for any arbitrary scoring rule. A popular scoring rule is the conditional average treatment effect (CATE),

$$s(\mathbf{X}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{X}).$$

Researchers have studied the estimation of the CATE using various machine learning algorithms such as tree-based methods and regularized regressions.

We emphasize that the scoring rule need not be based on the CATE. In fact, policy makers rely on various indexes. Such examples include the MELD (Kamath et al. 2001) that determines liver transplant priority, and the IDA Resource Allocation Index that informs the World Bank about the provision of economic aid.

Given this setup, we generalize the PAPE to the setting with a budget constraint. As before, without loss of generality, we assume the treatment is on average beneficial, that is, $\tau = \mathbb{E}\{Y_i(1) - Y_i(0)\} > 0$, so that the constraint is binding for the random treatment rule treating at most 100 × p% of units. The PAPE with a budget constraint p is defined as follows:

$$\tau_{fp} = \mathbb{E}\{Y_i(f(\mathbf{X}_i, c_p(f))) - pY_i(1) - (1-p)Y_i(0)\}. \qquad (3)$$

A budget constraint facilitates the comparison of multiple ITRs on the same footing. Suppose that we compare two ITRs, f and g, using the difference in their average values,

$$\Delta(f,g) = \lambda_f - \lambda_g = \mathbb{E}\{Y_i(f(\mathbf{X}_i)) - Y_i(g(\mathbf{X}_i))\}. \qquad (4)$$

While this quantity is useful, like the average value, it also fails to take into account the proportion of units assigned to the treatment condition under each ITR.

We can address this issue by comparing the efficacy of two ITRs under the same budget constraint. Formally, we define the population average prescriptive effect difference (PAPD) under budget p as,

$$\begin{aligned}\Delta_p(f,g) &= \tau_{fp} - \tau_{gp} \\ &= \mathbb{E}\{Y_i(f(\mathbf{X}_i, c_p(f))) - Y_i(g(\mathbf{X}_i, c_p(g)))\}. \qquad (5)\end{aligned}$$

2.4. The Area Under the Prescriptive Effect Curve

Since the PAPE (Equation (3)) varies as a function of budget constraint p, it would be useful to develop a summary performance metric of an ITR over a range of p. We propose the AUPEC as a metric analogous to the area under the receiver operating characteristic curve (AUROC) for classification performance.

Figure 2 graphically illustrates the AUPEC. Similar to Figure 1, the vertical and horizontal axes represent the average outcome and the budget, respectively. The budget is operationalized as the maximal proportion treated. The red solid curve corresponds to the average value of an ITR f as a function of budget constraint p, that is, $\mathbb{E}\{Y_i(f(\mathbf{X}_i, c_p(f)))\}$, whereas the black solid line represents the average value of the random treatment rule. The AUPEC corresponds to the area under the red curve minus the area under the black line, which is shown as a red shaded area.

Thus, the AUPEC represents the average performance of an ITR relative to the random treatment rule over the entire range of budget constraint (one could also compute the AUPEC over a specific range of budgets). Unlike the previous work (e.g., Rzepakowski and Jaroszewicz 2012), we do not require an ITR to

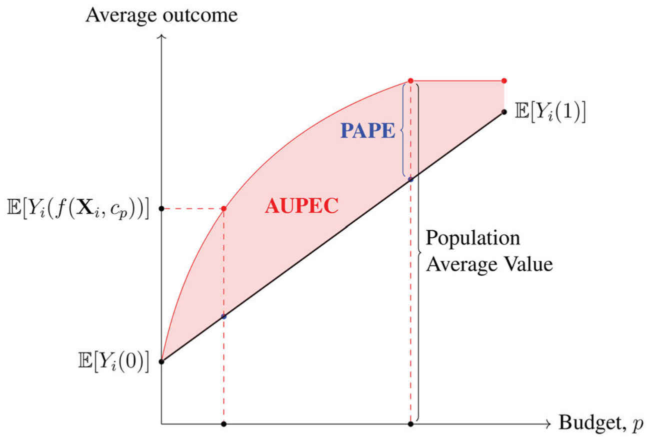

Figure 2. The AUPEC. The black solid line represents the average value under the random treatment rule while the red solid curve represents the average value under an ITR f. The difference between the line and the curve at a given budget constraint corresponds to the PAPE. The shaded area between the line and the curve represents the AUPEC of f.

assign the maximal proportion of units to the treatment condition though such a constraint can also be imposed if desired. For example, treating more than a certain proportion of units will reduce the average outcome if these additional units are harmed by the treatment. This is indicated by the flatness of the red line after p_f in Figure 2.

Formally, for a given ITR f, we define the AUPEC as,

$$\Gamma_f = \int_0^{p_f} \mathbb{E}\{Y_i(f(\mathbf{X}_i, c_p(f)))\}\mathrm{d}p + (1-p_f)\mathbb{E}\{Y_i(f(\mathbf{X}_i, c^*))\} - \frac{1}{2}\mathbb{E}(Y_i(0)+Y_i(1)), \tag{6}$$

where c^* is the predetermined minimum score such that one would be treated in the absence of a budget constraint, and $p_f = \Pr(f(\mathbf{X}_i, c^*) = 1)$ denotes the maximal proportion of units assigned to the treatment condition under the ITR with no budget constraint. The last term represents the area under the random treatment rule.

Different values of the threshold c^* are possible. If the goal is to treat only those who on average benefit from the treatment, then we could use the CATE as a scoring rule and set $c^* = 0$. Another example is the use of the MELD score as the scoring rule and choose an appropriate value of c^* so that a sufficiently healthy patient is never considered for a transplant. Finally, setting $c^* = -\infty$ would represent the setting where the maximum possible units should be treated regardless of the scoring rule.

We further note that the AUPEC is a generalization of the QINI coefficient widely utilized in literature for uplift modeling (Radcliffe 2007). Formally, the population-level QINI coefficient is commonly defined as

$$\mathrm{QINI} = n\left(\int_0^1 p\mathbb{E}\{Y_i(1)-Y_i(0) \mid f(\mathbf{X}_i, c_p(f)) = 1\}\mathrm{d}p - \frac{1}{2}\mathbb{E}(Y_i(1)-Y_i(0))\right).$$

After some algebra, we can rewrite this quantity in the following form:

$$n\left(\int_0^1 \mathbb{E}\{Y_i(f(\mathbf{X}_i, c_p(f)))\}\underset{\cdot}{\mathrm{p}} - \frac{1}{2}\mathbb{E}(Y_i(0)+Y_i(1))\right).$$

Thus, the QINI coefficient is (up to a constant factor n) a special case of AUPEC when $c^* = -\infty$ and $p_f = 1$. The choice of $c^* = -\infty$ may be reasonable in the applications where the treatment can be assumed to be never harmful, that is, $Y_i(1) \geq Y_i(0)$. In such cases, under no budget constraint one would treat the entire population.

To enable a comparison of efficacy across different datasets, the AUPEC can be normalized to be scale-invariant by shifting Γ_f by $\mathbb{E}(Y_i(0))$ and dividing by $\tau = \mathbb{E}(Y_i(1) - Y_i(0))$,

$$\Gamma_f^* = \frac{1}{\tau}\left[\int_0^{p_f} \mathbb{E}\{Y_i(f(\mathbf{X}_i, c_p(f)))\}\mathrm{d}p + (1-p_f)\mathbb{E}\{Y_i(f(\mathbf{X}_i, c^*))\} - \mathbb{E}(Y_i(0))\right] - \frac{1}{2}.$$

This normalized AUPEC is invariant to the affine transformation of the outcome variables, $Y_i(1), Y_i(0)$, while the standard AUPEC is only invariant to a constant shift. The normalized AUPEC takes a value in $[0, 1]$, and has an intuitive interpretation as the average percentage outcome gain using the ITR f compared to the random treatment rule, under the uniform prior distribution over the percentage treated.

3. Estimation and Inference

Having introduced the evaluation metrics, we show how to estimate them and compute standard errors under the repeated sampling framework of Neyman (1923). Here, we consider the setting, in which researchers are interested in evaluating the performance of fixed ITRs. In other words, throughout this section, ITR f is assumed to be known and has no estimation uncertainty. For example, one may first construct an ITR based on an existing dataset (experimental or observational), and then conduct a new experiment to evaluate its performance. In Section 4, we extend the methodology to the setting, in which the same experimental dataset is used to both construct and evaluate an ITR.

For the rest of this article, we assume that we have a simple random sample of n units from a super-population, $\mathcal{P}$. We conduct a completely randomized experiment, in which n_1 units are randomly assigned to the treatment condition with probability n_1/n and the rest of the $n_0(= \ n - n_1)$ units are assigned to the control condition. While it is straightforward to allow for unequal treatment assignment probabilities across units, for the sake of simplicity, we assume complete randomization. We formalize these assumptions below.

Assumption 2 (Random Sampling of Units). Each of n units, represented by a three-tuple consisting of two potential outcomes and pretreatment covariates, is assumed to be independently sampled from a super-population $\mathcal{P}$, that is,

$$(Y_i(1), Y_i(0), \mathbf{X}_i) \overset{\text{iid}}{\sim} \mathcal{P}.$$

Assumption 3 (Complete Randomization). For any $i = 1, 2, \ldots, n$, the treatment assignment probability is given by,

$$\Pr(T_i = 1 \mid Y_i(1), Y_i(0), \mathbf{X}_i) = \frac{n_1}{n},$$

where $\sum_{i=1}^n T_i = n_1$.

We now present the results under a binding budget constraint. The results for the average value and PAPE with no budget constraint appear in Appendix A.1 (supplementary material).

3.1. The Population Average Prescription Effect (PAPE)

To estimate the PAPE with a binding budget constraint p (Equation (3)), we consider the following estimator:

$$\hat{\tau}_{fp}(\mathcal{Z}) = \frac{1}{n_1}\sum_{i=1}^n Y_i T_i f(\mathbf{X}_i, \hat{c}_p(f)) + \frac{1}{n_0}\sum_{i=1}^n Y_i(1-T_i)(1-f(\mathbf{X}_i, \hat{c}_p(f))) - \frac{p}{n_1}\sum_{i=1}^n Y_i T_i - \frac{1-p}{n_0}\sum_{i=1}^n Y_i(1-T_i) \tag{7}$$

where $\hat{c}_p(f) = \inf\{c \in \mathbb{R} : \sum_{i=1}^{n} f(\mathbf{X}_i, c) \leq np\}$ represents the estimated threshold given the maximal proportion of treated units p. Unlike the case of no budget constraint (see Appendix A.1.2, supplementary material), the bias is not zero because the threshold $c_p(f)$ needs to be estimated without an assumption about the distribution of the score produced by the scoring rule. We derive an upper bound of bias and the exact variance.

Theorem 1 (Bias bound and exact variance of the PAPE estimator with a budget constraint). Under Assumptions 1–3, the bias of the proposed estimator of the PAPE with a budget constraint p defined in Equation (7) can be bounded as follows:

$$\begin{aligned}&\mathbb{P}_{\hat{c}_p(f)}(|\mathbb{E}\{\hat{\tau}_{fp}(\mathcal{Z}) - \tau_{fp} \mid \hat{c}_p(f)\}| \geq \epsilon)\\ &\quad \leq 1 - B(1 - p + \gamma_{fp}(\epsilon), n - \lfloor np \rfloor, \lfloor np \rfloor + 1)\\ &\qquad + B(1 - p - \gamma_{fp}(\epsilon), n - \lfloor np \rfloor, \lfloor np \rfloor + 1),\end{aligned}$$

where any given constant $\epsilon > 0$, $B(\epsilon, \alpha, \beta)$ is the incomplete beta function (if $\alpha \leq 0$ and $\beta > 0$, we set $B(\epsilon, \alpha, \beta) := H(\epsilon)$ for all ϵ where $H(\epsilon)$ is the Heaviside step function), and

$$\gamma_{fp}(\epsilon) = \frac{\epsilon}{\max_{c \in [c_p(f) - \epsilon,\, c_p(f) + \epsilon]} \mathbb{E}(\tau_i \mid s(\mathbf{X}_i) = c)}.$$

The variance of the estimator is given by

$$\begin{aligned}\mathbb{V}(\hat{\tau}_{fp}(\mathcal{Z})) = {}& \frac{\mathbb{E}(S_{fp1}^2)}{n_1} + \frac{\mathbb{E}(S_{fp0}^2)}{n_0} + \frac{\lfloor np \rfloor (n - \lfloor np \rfloor)}{n^2(n-1)}\\ &\times \left\{(2p-1)\kappa_{f1}(p)^2 - 2p\kappa_{f1}(p)\kappa_{f0}(p)\right\},\end{aligned}$$

where $S_{fpt}^2 = \sum_{i=1}^{n} (Y_{fpi}(t) - \overline{Y_{fp}(t)})^2/(n-1)$ and $\kappa_{ft}(p) = \mathbb{E}(\tau_i \mid f(\mathbf{X}_i, \hat{c}_p(f)) = t)$ with $Y_{fpi}(t) = \big(f(\mathbf{X}_i, \hat{c}_p(f)) - p\big) Y_i(t)$, and $\overline{Y_{fp}(t)} = \sum_{i=1}^{n} Y_{fpi}(t)/n$, for $t = 0, 1$.

Proof is given in Appendix A.2 (supplementary material). The last term in the uncertainty accounts for the variance due to estimating $c_p(f)$. The variance can be consistently estimated by replacing each unknown parameter with its sample analogue, that is, for $t = 0, 1$,

$$\begin{aligned}\widehat{\mathbb{E}(S_{fpt}^2)} &= \frac{1}{n_t - 1} \sum_{i=1}^{n} \mathbf{1}\{T_i = t\}(Y_{fpi} - \overline{Y_{fpt}})^2,\\ \widehat{\kappa_{ft}(p)} &= \frac{\sum_{i=1}^{n} \mathbf{1}\{f(\mathbf{X}_i, \hat{c}_p(f)) = t\} T_i Y_i}{\sum_{i=1}^{n} \mathbf{1}\{f(\mathbf{X}_i, \hat{c}_p(f)) = t\} T_i}\\ &\quad - \frac{\sum_{i=1}^{n} \mathbf{1}\{f(\mathbf{X}_i, \hat{c}_p(f)) = t\}(1 - T_i) Y_i}{\sum_{i=1}^{n} \mathbf{1}\{f(\mathbf{X}_i, \hat{c}_p(f)) = t\}(1 - T_i)},\end{aligned}$$

where $Y_{fpi} = (f(\mathbf{X}_i, \hat{c}_p(f)) - p) Y_i$ and $\overline{Y}_{fpt} = \sum_{i=1}^{n} \mathbf{1}\{T_i = t\} Y_{fpi}/n_t$. To estimate the term that appears in the denominator of $\gamma_{fp}(\epsilon)$ as part of the upper bound of bias, we may assume that the CATE, $\mathbb{E}(\tau_i \mid s(\mathbf{X}_i) = c)$, is continuous in c. Continuity is often assumed when estimating the CATE (e.g., Künzel et al. 2019; Wager and Athey 2018). We may also utilize an upper bound for CATE, if known, to estimate the bias conservatively.

Building on the above results, we also consider the comparison of two ITRs under the same budget constraint, using the following estimator of the PAPD (Equation (5)),

$$\begin{aligned}\widehat{\Delta}_p(f, g, \mathcal{Z}) = {}& \frac{1}{n_1} \sum_{i=1}^{n} Y_i T_i \{f(\mathbf{X}_i, \hat{c}_p(f)) - g(\mathbf{X}_i, \hat{c}_p(g))\} + \frac{1}{n_0}\\ &\times \sum_{i=1}^{n} Y_i (1 - T_i)\{g(\mathbf{X}_i, \hat{c}_p(g)) - f(\mathbf{X}_i, \hat{c}_p(f))\}.\end{aligned} \tag{8}$$

Theorem A3 in Appendix A.3 (supplementary material) derives the bias bound and exact variance of this estimator. In one of their applications, Zhou, Athey, and Wager (2018) applied the t-test to the cross-validated test statistic similar to the one introduced here under no budget constraint. However, no formal justification of this procedure is given under the cross-validation setting and it cannot be readily extended to the case with a budget constraint. In contrast, our methodology is applicable under these settings as well (see Section 4).

3.2. The Area Under the Prescriptive Effect Curve

Next, we consider the estimation and inference about the AUPEC (Equation (6)). Let n_f represent the maximum number of units in the sample that the ITR f would assign under no budget constraint, that is, $\hat{p}_f = n_f/n = \sum_{i=1}^{n} f(\mathbf{X}_i, c^*)/n$. We propose the following estimator of the AUPEC:

$$\begin{aligned}\widehat{\Gamma}_f(\mathcal{Z}) = {}& \frac{1}{n_1} \sum_{i=1}^{n} Y_i T_i\\ &\times \left\{\frac{1}{n}\left(\sum_{k=1}^{n_f} f(\mathbf{X}_i, \hat{c}_{k/n}(f)) + (n - n_f) f(\mathbf{X}_i, \hat{c}_{\hat{p}_f}(f))\right)\right\}\\ &+ \frac{1}{n_0} \sum_{i=1}^{n} Y_i (1 - T_i)\\ &\times \left\{1 - \frac{1}{n}\left(\sum_{k=1}^{n_f} f(\mathbf{X}_i, \hat{c}_{k/n}(f)) + (n - n_f) f(\mathbf{X}_i, \hat{c}_{\hat{p}_f}(f))\right)\right\}\\ &- \frac{1}{2n_1} \sum_{i=1}^{n} Y_i T_i - \frac{1}{2n_0} \sum_{i=1}^{n} Y_i (1 - T_i).\end{aligned} \tag{9}$$

The following theorem shows a bias bound and the exact variance of this estimator.

Theorem 2 (Bias and Variance of the AUPEC Estimator). Under Assumptions 1–3, the bias of the AUPEC estimator defined in Equation (9) can be bounded as follows:

$$\begin{aligned}&\mathbb{P}_{\hat{p}_f}(|\mathbb{E}(\widehat{\Gamma}_f(\mathcal{Z}) - \Gamma_f \mid \hat{p}_f)| \geq \epsilon)\\ &\quad \leq 1 - B(1 - p_f + \gamma_{p_f}(\epsilon), n - \lfloor np_f \rfloor, \lfloor np_f \rfloor + 1)\\ &\qquad + B(1 - p_f - \gamma_{p_f}(\epsilon), n - \lfloor np_f \rfloor, \lfloor np_f \rfloor + 1),\end{aligned}$$

where any given constant $\epsilon > 0$, $B(\epsilon, \alpha, \beta)$ is the incomplete beta function (if $\alpha = 0$ and $\beta > 0$, we set $B(\epsilon, \alpha, \beta) := H(\epsilon)$ for all ϵ where $H(\epsilon)$ is the Heaviside step function), and

$$\gamma_{p_f}(\epsilon) = \frac{\epsilon}{2 \max_{c \in [c^* - \epsilon,\, c^* + \epsilon]} \mathbb{E}(\tau_i \mid s(\mathbf{X}_i) = c)}.$$

The variance is given by

$$\mathbb{V}(\widehat{\Gamma}_f(\mathcal{Z})) = \frac{\mathbb{E}(S_{f1}^{*2})}{n_1} + \frac{\mathbb{E}(S_{f0}^{*2})}{n_0}$$
$$+\mathbb{E}\left[-\frac{1}{n}\left\{\sum_{z=1}^{Z}\frac{z(n-z)}{n^2(n-1)}\kappa_{f1}(z/n)\kappa_{f0}(z/n)\right.\right.$$
$$\left.+\frac{Z(n-Z)^2}{n^2(n-1)}\kappa_{f1}(Z/n)\kappa_{f0}(Z/n)\right\}$$
$$-\frac{2}{n^4(n-1)}\sum_{z=1}^{Z-1}\sum_{z'=z+1}^{Z} z(n-z')\kappa_{f1}(z/n)\kappa_{f1}(z'/n)$$
$$-\frac{Z^2(n-Z)^2}{n^4(n-1)}\kappa_{f1}(Z/n)^2 - \frac{2(n-Z)^2}{n^4(n-1)}$$
$$\left.\times\sum_{z=1}^{Z} z\kappa_{f1}(Z/n)\kappa_{f1}(z/n) + \frac{1}{n^4}\sum_{z=1}^{Z} z(n-z)\kappa_{f1}(z/n)^2\right]$$
$$+\mathbb{V}\left(\sum_{z=1}^{Z}\frac{z}{n}\kappa_{f1}(z/n) + \frac{(n-Z)Z}{n}\kappa_{f1}(Z/n)\right),$$

where Z is a Binomial random variable with size n and success probability p_f, and $S_{ft}^{*2} = \sum_{i=1}^{n}(Y_i^*(t) - \overline{Y^*(t)})^2/(n-1)$, $\kappa_{ft}(k/n) = \mathbb{E}(Y_i(1) - Y_i(0) \mid f(\mathbf{X}_i, \hat{c}_{k/n}(f)) = t)$, with $Y_i^*(t) = \left[\left\{\sum_{z=1}^{n_f} f(\mathbf{X}_i, \hat{c}_{z/n}(f)) + (n - n_f)f(\mathbf{X}_i, \hat{c}_{\hat{p}_f}(f))\right\}/n - \frac{1}{2}\right] Y_i(t)$ and $\overline{Y^*(t)} = \sum_{i=1}^{n} Y_i^*(t)/n$, for $t = 0, 1$.

Proof is given in Appendix A.4 (supplementary material). When $c^* = -\infty$ (i.e., the AUPEC equals the QINI coefficient), the estimator is unbiased, implying that the bias comes from estimating the proportion treated p_f under no budget constraints and $c^* > -\infty$. As before, $\mathbb{E}(S_{ft}^{*2})$ does not equal $\mathbb{V}(Y_i^*(t))$ due to the need to estimate the terms $c_{z/n}$ for all z, and the additional terms account for the variance of estimation. We can consistently estimate the upper bound of bias, for example, under by assuming that the CATE is continuous. We may also use an upper bound for CATE, if known, to estimate the bias conservatively.

To estimate the variance, we replace each unknown parameter with its sample analogue

$$\widehat{\mathbb{E}(S_{ft}^{*2})} = \frac{1}{n_t - 1}\sum_{i=1}^{n}\mathbf{1}\{T_i = t\}(Y_i^* - \overline{Y_t^*})^2,$$
$$\hat{\kappa}_{ft}(z/n) = \frac{\sum_{i=1}^{n}\mathbf{1}\{f(\mathbf{X}_i, \hat{c}_{z/n}(f)) = t\}T_iY_i}{\sum_{i=1}^{n}\mathbf{1}\{f(\mathbf{X}_i, \hat{c}_{z/n}(f)) = t\}T_i} - \frac{\sum_{i=1}^{n}\mathbf{1}\{f(\mathbf{X}_i, \hat{c}_{z/n}(f)) = t\}(1-T_i)Y_i}{\sum_{i=1}^{n}\mathbf{1}\{f(\mathbf{X}_i, \hat{c}_{z/n}(f)) = t\}(1-T_i)}, \tag{10}$$

for $t = 0, 1$. In the extreme cases with $z \to 1$ for $t = 1$ and $z \to n$ for $t = 0$, each denominator in Equation (10) is likely to be close to zero. In such cases, we instead use the estimator $\widehat{\kappa_{f1}(z_{\min}/n)}$ for all $z < z_{\min}$ where $z_{\min}$ is the smallest z such that Equation (10) for $\kappa_{f1}(z/n)$ does not lead to division by zero. Similarly, for $t = 0$, we use $\widehat{\kappa_{f0}(z_{\max}/n)}$ for all $z > z_{\max}$ where $z_{\max}$ is the largest z.

For the terms involving the binomial random variable Z, we first note that, when fully expanded out, they are the polynomials of $p_f = \mathbb{E}(f(\mathbf{X}_i))$. To estimate the polynomials, we can utilize their unbiased estimators as discussed in Stuard and Ord (1994), that is, $\hat{p}_f^z = s(s-1)\cdots(s-z+1)/\{n(n-1)\cdots(n-z+1)\}\}$ where $s = \sum_{i=1}^{n} f(\mathbf{X}_i)$ is unbiased for p_f^z for all $z \leq n$. When the sample size is large, this estimation method is computationally inefficient and unstable due to the presence of high powers. Hence, we may use the Monte Carlo sampling of Z from a Binomial distribution with size n and success probability $\hat{p}_f$. In our simulation study, we show that this Monte Carlo approach is effective even when the sample size is small (see Section 5).

Finally, a consistent estimator for the normalized AUPEC is given by

$$\widehat{\Gamma}_f^*(\mathcal{Z}) = \frac{1}{\sum_{i=1}^{n} Y_iT_i/n_1 - Y_i(1-T_i)/n_0}$$
$$\times\left\{\frac{1}{nn_1}\sum_{i=1}^{n} Y_iT_i\left(\sum_{z=1}^{n_f} f(\mathbf{X}_i, \hat{c}_{z/n}(f)) + (n-n_f)f(\mathbf{X}_i, \hat{c}_{\hat{p}_f}(f))\right)\right.$$
$$-\frac{1}{nn_0}\sum_{i=1}^{n} Y_i(1-T_i)$$
$$\left.\times\left(\sum_{z=1}^{n_f} f(\mathbf{X}_i, \hat{c}_{z/n}(f)) + (n-n_f)f(\mathbf{X}_i, \hat{c}_{\hat{p}_f}(f))\right)\right\} - \frac{1}{2}. \tag{11}$$

The variance of $\widehat{\Gamma}_f^*(\mathcal{Z})$ can be estimated using the Taylor expansion of quotients of random variables to an appropriate order as detailed in Stuard and Ord (1994).

4. Estimating and Evaluating ITRs Using the Same Experimental Data

We next consider a common situation, in which researchers use the same experimental data to both estimate and evaluate an ITR via cross-validation. This differs from the setting we have analyzed so far, in which a fixed ITR is given for evaluation. We first extend the evaluation metrics introduced in Section 2 to the current setting with estimated ITRs. We then develop inferential methods under Neyman's repeated sampling framework by accounting for both estimation and evaluation uncertainties. Below, we consider the scenario, in which researchers face a binding budget constraint. Appendix A.5 (supplementary material) presents the results for the case with no budget constraint.

4.1. Evaluation Metrics

Suppose that we have the data from a completely randomized experiment as described in Section 3. We first estimate an ITR f by applying a machine learning algorithm F to training data $\mathcal{Z}^{tr}$. Then, under a budget constraint of the maximal proportion of treated units p, we use test data to evaluate the resulting estimated ITR $\hat{f}_{\mathcal{Z}^{tr}}$. As before, we assume that this constraint is binding, that is, $p < p_F$ where $p_F = \Pr(\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i) = 1)$ represents the proportion of treated units under the ITR without a budget constraint.

Formally, an machine learning algorithm F is a deterministic map from the space of training data $\mathcal{Z}$ to that of scoring rules $\mathcal{S}$,

$$F : \mathcal{Z} \to \mathcal{S}.$$

Then, for a given training dataset $\mathcal{Z}^{tr}$, the estimated ITR is given by,

$$\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, c_p(\hat{f}_{\mathcal{Z}^{tr}})) \;=\; \mathbf{1}\{\hat{s}_{\mathcal{Z}^{tr}}(\mathbf{X}_i) > c_p(\hat{f}_{\mathcal{Z}^{tr}})\},$$

where $\hat{s}_{\mathcal{Z}^{tr}} = F(\mathcal{Z}^{tr})$ is the estimated scoring rule and $c_p(\hat{f}_{\mathcal{Z}^{tr}}) = \inf\{c \in \mathbb{R} : \Pr(\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, c) = 1 \mid \mathcal{Z}^{tr}) \leq p\}$ is the threshold based on the maximal proportion of treated units p. The CATE is a natural choice for the scoring rule, that is, $s_{\mathcal{Z}^{tr}}(\mathbf{X}_i) = \mathbb{E}(\tau_i \mid \mathbf{X}_i)$. We need not assume that $\hat{s}_{\mathcal{Z}^{tr}}(\mathbf{X}_i)$ is consistent for the CATE.

To extend the PAPE (Equation (3)), we first define the population proportion of units with $\mathbf{X}_i = \mathbf{X}$ who are assigned to the treatment condition under the estimated ITR as,

$$\begin{aligned}\bar{f}_{Fp}(\mathbf{X}) &= \mathbb{E}_{\mathcal{Z}^{tr}}\{\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, c_p(\hat{f}_{\mathcal{Z}^{tr}})) \mid \mathbf{X}_i = \mathbf{X}\} \\ &= \Pr\{\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, c_p(\hat{f}_{\mathcal{Z}^{tr}})) = 1 \mid \mathbf{X}_i = \mathbf{X}\}.\end{aligned}$$

While $c_p(\hat{f}_{\mathcal{Z}^{tr}})$ depends on the specific ITR generated from the training data $\mathcal{Z}^{tr}$, the population proportion of treated units averaged over the sampling of training data, $\bar{f}_{Fp}(\mathbf{X}_i)$, only depends on p.

Lastly, the PAPE of the estimated ITR under budget constraint p is defined as,

$$\begin{aligned}\tau_{Fp} = \mathbb{E}\{&\bar{f}_{Fp}(\mathbf{X}_i)Y_i(1) \\ &+ (1 - \bar{f}_{Fp}(\mathbf{X}_i))Y_i(0) - pY_i(1) - (1-p)Y_i(0)\}.\end{aligned}$$

This evaluation metric corresponds to neither that of a specific ITR estimated from the whole experimental dataset nor its expectation. Rather, we are evaluating the efficacy of a learning algorithm that is used to estimate an ITR based on the same experimental data.

We can also compare *estimated* ITRs by further generalizing the definition of the PAPD (Equation (5)) to the current setting. Specifically, we define the PAPD between two machine learning algorithms, F and G, under budget constraint p as,

$$\begin{aligned}\Delta_p(F, G) = \mathbb{E}_{\mathbf{X},Y}[&\{\bar{f}_{Fp}(\mathbf{X}_i) - \bar{f}_{Gp}(\mathbf{X}_i)\}Y_i(1) \\ &+ \{\bar{f}_{Gp}(\mathbf{X}_i) - \bar{f}_{Fp}(\mathbf{X}_i)\}Y_i(0)]. \qquad (12)\end{aligned}$$

Finally, we consider the AUPEC of an estimated ITR. Specifically, the AUPEC of a machine learning algorithm F is defined as,

$$\begin{aligned}\Gamma_F = \mathbb{E}_{\mathcal{Z}^{tr}}\Bigg[&\int_0^{p_{\hat{f}}} \mathbb{E}\{Y_i(\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, c_p(\hat{f}_{\mathcal{Z}^{tr}})))\}\mathrm{d}p \qquad (13) \\ &+ (1 - p_{\hat{f}})\mathbb{E}\{Y_i(\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, c^*))\}\Bigg] - \frac{1}{2}\mathbb{E}(Y_i(0) + Y_i(1)),\end{aligned}$$

where $p_{\hat{f}} = \Pr(\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i) = 1)$ is the maximal population proportion of units treated by the estimated ITR, $\hat{f}_{\mathcal{Z}^{tr}}$.

4.2. Estimation and Inference

Rather than simply splitting the data into training and test sets (in such a case, the inferential procedure for fixed ITRs is applicable), we maximize efficiency by using cross-validation to estimate the evaluation metrics introduced above. First, we randomly split the data into K subsamples of equal size $m = n/K$ by assuming, for the sake of notational simplicity, that n is a multiple of K. Then, for each $k = 1, 2, \ldots, K$, we use the kth subsample as a test set $\mathcal{Z}_k = \{\mathbf{X}_i^{(k)}, T_i^{(k)}, Y_i^{(k)}\}_{i=1}^m$ with the data from all $(K-1)$ remaining subsamples as the training set $\mathcal{Z}_{-k} = \{\mathbf{X}_i^{(-k)}, T_i^{(-k)}, Y_i^{(-k)}\}_{i=1}^{n-m}$.

Algorithm 1 Estimating and Evaluating an Individualized Treatment Rule (ITR) using the Same Experimental Data via Cross-Validation

Input: Data $\mathcal{Z} = \{\mathbf{X}_i, T_i, Y_i\}_{i=1}^n$, Machine learning algorithm F, Evaluation metric τ_f, Number of folds K

Output: Estimated evaluation metric $\hat{\tau}_F$, Estimated variance of $\hat{\tau}_F$

```
Split data into $K$ random subsets of equal size $(\mathcal{Z}_1, \ldots, \mathcal{Z}_k)$
$k \leftarrow 1$
while $k \leq K$ do
    $\mathcal{Z}_{-k} = [\mathcal{Z}_1, \ldots, \mathcal{Z}_{k-1}, \mathcal{Z}_{k+1}, \ldots, \mathcal{Z}_K]$
    $\hat{f}_{-k} = F(\mathcal{Z}_{-k})$   ▷ Estimate ITR $f$ by applying $F$ to $\mathcal{Z}_{-k}$
    $\hat{\tau}_k = \hat{\tau}_{\hat{f}_{-k}}(\mathcal{Z}_k)$   ▷ Evaluate estimated ITR $\hat{f}$ using $\mathcal{Z}_k$
    $k \leftarrow k + 1$
end while
return $\hat{\tau}_F = \frac{1}{K}\sum_{k=1}^K \hat{\tau}_k$, $\widehat{\mathbb{V}(\hat{\tau}_F)} = v(\hat{f}_{-1}, \ldots, \hat{f}_{-k}, \mathcal{Z}_1, \cdots, \mathcal{Z}_K)$
```

To simplify notation, we assume that the number of treated (control) units is identical across different folds and denote it as m_1 ($m_0 = m - m_1$). For each split k, we estimate an ITR by applying a learning algorithm F to the training data $\mathcal{Z}_{-k}$

$$\hat{f}_{-k} \;=\; F(\mathcal{Z}_{-k}). \qquad (14)$$

We then evaluate the performance of the learning algorithm F by computing an evaluation metric of interest τ based on the test data $\mathcal{Z}_k$. Repeating this K times for each k and averaging the results gives a cross-validation estimator of the evaluation metric. Algorithm 1 formally presents this estimation procedure. The variance formula for the estimated evaluation metric is omitted as it is generally a complex function $v(\cdot)$ (see Theorem 3).

We develop the inferential methodology for the evaluation based on the cross-validation procedure described above under Neyman's repeated sampling framework. We focus on the case with a binding budget constraint. The results with no budget constraint appear in Appendix A.5 (supplementary material). We begin by introducing the cross-validation estimator of the PAPE with a binding budget constraint p,

$$\hat{\tau}_{Fp} \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{\tau}_{\hat{f}_{-k},p}(\mathcal{Z}_k), \qquad (15)$$

where $\hat{\tau}_{fp}$ is defined in Equation (7).

Like the fixed ITR case, the bias of the proposed estimator is not exactly zero. However, we are able to show that the bias can be upper bounded by a small quantity while the exact randomization variance can still be derived.

Theorem 3 (Bias bound and exact variance of the cross-validation PAPE estimator with a budget constraint). Under Assumptions 1–3, the bias of the cross-validation PAPE estimator with a budget constraint p defined in Equation (15) can be bounded as follows,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{Z}^{tr}}[\mathbb{P}_{\hat{c}_p(\hat{f}_{\mathcal{Z}^{tr}})}(|\mathbb{E}\{\hat{\tau}_{Fp} - \tau_{Fp} \mid \hat{c}_p(\hat{f}_{\mathcal{Z}^{tr}})\}| \geq \epsilon)] \\
&\quad \leq 1 - B(1 - p + \gamma_p(\epsilon), m - \lfloor mp \rfloor, \lfloor mp \rfloor + 1) \\
&\qquad + B(1 - p - \gamma_p(\epsilon), m - \lfloor mp \rfloor, \lfloor mp \rfloor + 1),
\end{aligned}$$

where any given constant $\epsilon > 0$, $B(\epsilon, \alpha, \beta)$ is the incomplete beta function (if $\alpha = 0$ and $\beta > 0$, we set $B(\epsilon, \alpha, \beta) := H(\epsilon)$ for all ϵ where $H(\epsilon)$ is the Heaviside step function), and

$$\gamma_p(\epsilon) = \frac{\epsilon}{\mathbb{E}_{\mathcal{Z}^{tr}}\{\max_{c \in [c_p(\hat{f}_{\mathcal{Z}^{tr}}) - \epsilon,\, c_p(\hat{f}_{\mathcal{Z}^{tr}}) + \epsilon]} \mathbb{E}_{\mathcal{Z}}(\tau_i \mid \hat{s}_{\mathcal{Z}^{tr}}(\mathbf{X}_i) = c)\}}.$$

The variance of the estimator is given by

$$\begin{aligned}
\mathbb{V}(\hat{\tau}_{Fp}) = {} & \frac{\mathbb{E}(S^2_{\hat{f}p1})}{m_1} + \frac{\mathbb{E}(S^2_{\hat{f}p0})}{m_0} + \frac{\lfloor mp \rfloor (m - \lfloor mp \rfloor)}{m^2(m-1)} \\
& \times \left\{(2p-1)\kappa_{F1}(p)^2 - 2p\kappa_{F1}(p)\kappa_{F0}(p)\right\} \\
& - \frac{K-1}{K}\mathbb{E}(S^2_{Fp}),
\end{aligned}$$

where $S^2_{\hat{f}pt} = \sum_{i=1}^m (Y_{\hat{f}pi}(t) - \overline{Y_{\hat{f}p}(t)})^2/(m-1)$, and $S^2_{Fp} = \sum_{k=1}^K (\hat{\tau}_{\hat{f}_{-k},p}(\mathcal{Z}_k) - \overline{\hat{\tau}_{\hat{f}_{-k},p}(\mathcal{Z}_k)})^2/(K-1)$, and $\kappa_{Ft}(p) = \mathbb{E}(\tau_i \mid \hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, \hat{c}_p(\hat{f}_{\mathcal{Z}^{tr}})) = t)$, with $Y_{\hat{f}pi}(t) = \{\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, \hat{c}_p(\hat{f}_{\mathcal{Z}^{tr}})) - p\}Y_i(t)$, $\overline{Y_{\hat{f}p}(t)} = \sum_{i=1}^n Y_{\hat{f}pi}(t)/n$, and $\overline{\hat{\tau}_{\hat{f}_{-k},p}(\mathcal{Z}_k)} = \sum_{k=1}^K \hat{\tau}_{\hat{f}_{-k},p}(\mathcal{Z}_k)/K$, for $t = 0, 1$.

Proof is given in Appendix A.6 (supplementary material). The estimation of the term $\mathbb{E}(\widetilde{S}^2_{\hat{f}t})$ is done similarly as before. For $\kappa_{Ft}(p)$, we replace it with its sample analog

$$\begin{aligned}
\widehat{\kappa_{Ft}(p)} = {} & \frac{1}{K}\sum_{l=1}^K \frac{\sum_{i=1}^m \mathbf{1}\{\hat{f}_{-k}(\mathbf{X}_i, \hat{c}_p(\hat{f}_{-k})) = t\}T_i^{(k)}Y_i^{(k)}}{\sum_{i=1}^m \mathbf{1}\{\hat{f}_{-k}(\mathbf{X}_i, \hat{c}_p(\hat{f}_{-k})) = t\}T_i^{(k)}} \\
& - \frac{\sum_{i=1}^m \mathbf{1}\{\hat{f}_{-k}(\mathbf{X}_i, \hat{c}_p(\hat{f}_{-k})) = t\}(1 - T_i^{(k)})Y_i^{(k)}}{\sum_{i=1}^m \mathbf{1}\{\hat{f}_{-k}(\mathbf{X}_i, \hat{c}_p(\hat{f}_{-k})) = t\}(1 - T_i^{(k)})}. \qquad (16)
\end{aligned}$$

To estimate the term that appears in the denominator of $\gamma_p(\epsilon)$ as part of the upper bound of bias, we assume that the CATE, that is, $\mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{s}_{\mathcal{Z}^{tr}}(\mathbf{X}_i) = c)$, is continuous in c, and replace the maximum with a point estimate. We may also utilize an upper bound for CATE, if known, to estimate the bias conservatively. Building on this result, Appendix A.7 (supplementary material) shows how to compare two estimated ITRs by estimating the PAPD (Equation (12)).

Finally, we consider the following cross-validation estimator of the AUPEC for an estimated ITR (Equation (13)):

$$\widehat{\Gamma}_F = \frac{1}{K}\sum_{k=1}^K \widehat{\Gamma}_{\hat{f}_{-k}}(\mathcal{Z}_k), \qquad (17)$$

where $\widehat{\Gamma}_f$ is defined in Equation (9). This can be seen as an overall statistic that measures the prescriptive performance of the machine learning algorithm F on the dataset under cross-validation. The following theorem derives a bias bound and the exact variance of this cross-validation estimator.

Theorem 4 (Bias bound and exact variance of the cross-validation AUPEC estimator). Under Assumptions 1–3, the bias of the AUPEC estimator defined in Equation (17) can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{Z}^{tr}}[\mathbb{P}_{\hat{p}_{\hat{f}}}(|\mathbb{E}(\widehat{\Gamma}_F - \Gamma_F \mid \hat{p}_{\hat{f}})| \geq \epsilon)] \\
&\quad \leq \mathbb{E}\{1 - B(1 - p_{\hat{f}} + \gamma_{p_{\hat{f}}}(\epsilon), m - \lfloor mp_{\hat{f}} \rfloor, \lfloor mp_{\hat{f}} \rfloor + 1) \\
&\qquad + B(1 - p_{\hat{f}} - \gamma_{p_{\hat{f}}}(\epsilon), m - \lfloor mp_{\hat{f}} \rfloor, \lfloor mp_{\hat{f}} \rfloor + 1)\},
\end{aligned}$$

where any given constant $\epsilon > 0$, $B(\epsilon, \alpha, \beta)$ is the incomplete beta function (if $\alpha = 0$ and $\beta > 0$, we set $B(\epsilon, \alpha, \beta) := H(\epsilon)$ for all ϵ where $H(\epsilon)$ is the Heaviside step function), and

$$\gamma_{p_{\hat{f}}}(\epsilon) = \frac{\epsilon}{2\mathbb{E}_{\mathcal{Z}^{tr}}\{\max_{c \in [c^* - \epsilon,\, c^* + \epsilon]} \mathbb{E}(\tau_i \mid \hat{s}_{\mathcal{Z}^{tr}}(\mathbf{X}_i) = c)\}}.$$

The variance is given by

$$\begin{aligned}
\mathbb{V}(\widehat{\Gamma}_F) = \mathbb{E}\Bigg[& -\frac{1}{m}\Bigg\{\sum_{z=1}^Z \frac{k(n-z)}{m^2(m-1)}\kappa_{F1}(z/m)\kappa_{F0}(z/m) \\
& + \frac{Z(m-Z)^2}{m^2(m-1)}\kappa_{F1}(Z/m)\kappa_{F0}(Z/m)\Bigg\} \\
& - \frac{2}{m^4(m-1)}\sum_{z=1}^{Z-1}\sum_{z'=z+1}^{Z} z(m-z')\kappa_{F1}(z/m)\kappa_{F1}(z'/m) \\
& - \frac{Z^2(m-Z)^2}{m^4(m-1)}\kappa_{F1}(Z/m)^2 \\
& - \frac{2(m-Z)^2}{m^4(m-1)}\sum_{z=1}^Z k\kappa_{F1}(Z/m)\kappa_{F1}(z/m) \\
& + \frac{1}{m^4}\sum_{z=1}^Z z(m-z)\kappa_{F1}(z/m)^2\Bigg] \\
& + \mathbb{V}\left(\sum_{z=1}^Z \frac{z}{m}\kappa_{F1}(z/m) + \frac{(m-Z)Z}{m}\kappa_{F1}(Z/m)\right) \\
& + \frac{\mathbb{E}(S^{*2}_{\hat{f}1})}{m_1} + \frac{\mathbb{E}(S^{*2}_{\hat{f}0})}{m_0} - \frac{K-1}{K}\mathbb{E}(S^{*2}_F),
\end{aligned}$$

where Z is a Binomial random variable with size m and success probability $p_{\hat{f}}$, $S^{*2}_{\hat{f}t} = \sum_{i=1}^m (Y^*_{\hat{f}i}(t) - \overline{Y^*_{\hat{f}}(t)})^2/(m-1)$, $S^2_F = \sum_{k=1}^K (\widehat{\Gamma}_{\hat{f}_{-k}}(\mathcal{Z}_k) - \overline{\widehat{\Gamma}_{\hat{f}_{-k}}(\mathcal{Z}_k)})^2/(K-1)$, and $\kappa_{Ft}(z/m) = \mathbb{E}(\tau_i \mid \hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, \hat{c}_{z/m}(\hat{f}_{\mathcal{Z}^{tr}}) = t)$ with $\overline{Y^*_{\hat{f}}(t)} = \sum_{i=1}^m Y^*_{\hat{f}i}(t)/m$, $\overline{\widehat{\Gamma}_{\hat{f}_{-k}}(\mathcal{Z}_k)} = \sum_{k=1}^K \widehat{\Gamma}_{\hat{f}_{-k}}(\mathcal{Z}_k)/K$, and $Y^*_{\hat{f}i}(t) = \left[\left\{\sum_{z=1}^{m_f} \hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, \hat{c}_{z/m}(\hat{f}_{\mathcal{Z}^{tr}})) + (m - m_f)\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i, \hat{c}_{z/m}(\hat{f}_{\mathcal{Z}^{tr}}))\right\}/ m - \frac{1}{2}\right] Y_i(t)$ for $t = 0, 1$.

Proof is similar to that of Theorem 2. The estimation of $\mathbb{E}(S^{*2}_{\hat{f}1})$, $\mathbb{E}(S^{*2}_{\hat{f}0})$, and $\mathbb{E}(S^{*2}_F)$ is the same as before, and the $\kappa_{Ft}(p)$ term can be estimated using Equation (16).

5. A Simulation Study

We conduct a simulation study to examine the finite sample performance of our methodology, for both fixed and estimated

ITRs. We find that the empirical coverage probability of the confidence interval, based on the proposed variance, approximates its nominal rate even in a small sample. We also find that the bias is minimal even when the proposed estimator is not unbiased and that our variance bounds are tight.

5.1. Data-Generation Process

Our data-generating process (DGP) is based on the one used in the 2017 Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge (Hahn, Dorie, and Murray 2018). A total of 8 covariates $\mathbf{X}$ are taken from the Infant Health and Development Program, which originally had 58 covariates and 4302 observations. In our simulation, the population distribution of covariates is assumed to equal the empirical distribution of this data set. Therefore, we obtain each simulation sample via bootstrap. We vary the sample size: $n = 100, 500,$ and 2000.

We use the same outcome model as the one used in the competition,

$$\mathbb{E}(Y_i(t) \mid \mathbf{X}_i) = \mu(\mathbf{X}_i) + \tau(\mathbf{X}_i)t, \qquad (18)$$

where $\pi(\mathbf{X}) = 1/[1 + \exp\{3(x_1 + x_{43} + 0.3(x_{10} - 1)) - 1\}]$, $\mu(\mathbf{X}) = -\sin(\Phi(\pi(\mathbf{X}))) + x_{43}$, and $\tau(\mathbf{X}) = \xi(x_3 x_{24} + (x_{14} - 1) - (x_{15} - 1))$ with $\Phi(\cdot)$ representing the standard Normal CDF and x_j indicating a specific covariate in the data set. One important difference is that we assume a complete randomized experiment whereas the original DGP generated the treatment using a function of covariates to emulate an observational study. As in the competition, we focus on two scenarios regarding the treatment effect size by setting ξ equal to 2 ("large") and 1/3 ("small"). Although the original DGP included four different error distributions, we use the iid error, $\sigma(\mathbf{X}_i)\epsilon_i$ where $\sigma(\mathbf{X}) = 0.25\sqrt{\mathbb{V}(\mu(\mathbf{X}) + \pi(\mathbf{X})\tau(\mathbf{X}))}$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

For fixed ITRs, we can directly compute the true values of our causal quantities of interest using the outcome model specified in Equation (18) and evaluate each quantity based on the entire original dataset. This computation is valid because we assume the population distribution of covariates is equal to the empirical distribution of the original data set. For the estimated ITR case, however, we do not have an analytical expression for the true value of a causal quantity of interest. Therefore, we obtain an approximate true value via Monte Carlo simulation. We generate 10,000 independent datasets based on the same DGP, and train the specified algorithm F on each of the datasets using 5-fold cross-validation (i.e., $K = 5$). Then, we use the sample mean of our estimated causal quantity across 10,000 datasets as our approximate truth.

We evaluate Bayesian additive regression trees (BART) (Chipman, George, and McCulloch 2010; Hahn, Murray, and Carvalho 2020), which had the best overall performance in the original competition. We compare this model with two other popular methods: causal forest (Athey, Tibshirani, and Wager 2019) as well as the LASSO, which includes all main effects and two-way interaction effects between the treatment and all covariates (Tibshirani 1996). All three models are trained on the original data from the 2017 ACIC Data Challenge. The number of trees was tuned through the 5-fold cross-validation for BART and causal forest. The regularization parameter was tuned similarly for LASSO. All models were cross-validated on the PAPE. For implementation, we use R 3.4.2 with bartMachine (version 1.4.2) for BART, grf (version 0.10.2) for Causal Forest, and glmnet (version 2.0.13) for LASSO. Once the models are trained, an ITR is derived based on the magnitude of the estimated CATE $\hat{\tau}(\mathbf{X})$, that is, $f(\mathbf{X}_i) = \mathbf{1}\{\hat{\tau}(\mathbf{X}_i) > 0\}$.

5.2. Results

We first present the results for fixed ITRs followed by those for estimated ITRs. Table 1 presents the bias and standard deviation of each estimator for fixed ITRs as well as the coverage probability of its 95% confidence intervals based on 1000 Monte Carlo trials. The results are shown separately for the large and small treatment effects scenarios. We estimate the PAPE τ_f for BART without a budget constraint as well as the PAPE with a budget constraint of 20% as the maximal proportion of treated units, $\tau_f(c_{0.2})$. In addition, we estimate the AUPEC Γ_f and compute the difference in the PAPE or PAPD between BART and causal forest ($\Delta(f, g)$), and between BART and LASSO ($\Delta(f, h)$).

Under both scenarios and across sample sizes, the bias of our estimator is small. Moreover, the coverage rate of 95% confidence intervals is close to their nominal rate even when the sample size is small. Although we can only bound the

Table 1. The Results of the simulation study for fixed individualized treatment rules (ITRs).

		$n = 100$			$n = 500$			$n = 2000$		
Estimator	Truth	Coverage	Bias	s.d.	Coverage	bias	s.d.	Coverage	Bias	s.d.
Small treatment effect										
$\hat{\tau}_f$	0.066	94.3%	0.005	0.124	96.2%	0.001	0.053	95.1%	0.001	0.026
$\hat{\tau}_f(c_{0.2})$	0.051	93.2	−0.002	0.109	94.4	0.001	0.046	95.2	0.002	0.021
$\widehat{\Gamma}_f$	0.053	95.3	0.001	0.106	95.1	0.001	0.045	94.8	−0.001	0.024
$\widehat{\Delta}_{0.2}(f, g)$	−0.022	94.0	0.006	0.122	95.4	0.002	0.051	96.0	0.000	0.026
$\widehat{\Delta}_{0.2}(f, h)$	−0.014	93.9	−0.001	0.131	94.9	−0.000	0.060	95.3	−0.000	0.030
Large treatment effect										
$\hat{\tau}_f$	0.430	94.7%	−0.000	0.163	95.7%	0.000	0.064	94.4%	−0.000	0.031
$\hat{\tau}_f(c_{0.2})$	0.356	94.7	0.004	0.159	95.7	0.002	0.072	95.8	0.000	0.035
$\widehat{\Gamma}_f$	0.363	94.3	−0.005	0.130	94.9	0.003	0.058	95.7	0.000	0.029
$\widehat{\Delta}_{0.2}(f, g)$	−0.000	96.9	0.008	0.151	97.9	−0.002	0.073	98.0	−0.000	0.026
$\widehat{\Delta}_{0.2}(f, h)$	0.000	94.7	−0.004	0.140	97.7	−0.001	0.065	96.6	0.000	0.033

NOTE: The table presents the bias and standard deviation of each estimator as well as the coverage of its 95% confidence intervals under the "Small treatment effect" and "Large treatment effect" scenarios. The first three estimators shown here are for BART f: Population Average Prescription effect (PAPE; $\hat{\tau}_f$), PAPE with a budget constraint of 20% treatment proportion ($\hat{\tau}_f(c_{0.2})$), AUPEC; $\widehat{\Gamma}_f$. We also present the results for the difference in the PAPE between BART and Causal Forest g ($\widehat{\Delta}_{0.2}(f, g)$) and between BART and LASSO h ($\widehat{\Delta}_{0.2}(f, g)$) under the budget constraint.

Table 2. The results of the simulation study for cross-validated ITR.

	$n = 100$				$n = 500$				$n = 2000$			
Estimator	Truth	Coverage	Bias	s.d.	Truth	Coverage	Bias	s.d.	Truth	Coverage	Bias	s.d.
Small treatment effect												
$\hat{\lambda}_F$	0.073	96.4%	0.001	0.216	0.095	96.7%	0.002	0.100	0.112	97.2%	0.002	0.046
$\hat{\tau}_F$	0.021	94.6	−0.002	0.130	0.030	95.5	−0.002	0.052	0.032	94.4	−0.000	0.027
$\hat{\tau}_F(c_{0.2})$	0.023	95.4	−0.003	0.120	0.034	95.4	−0.002	0.057	0.043	96.8	0.001	0.029
$\widehat{\Gamma}_F$	0.009	98.2	0.002	0.117	0.029	96.8	−0.001	0.048	0.039	95.9	0.001	0.001
Large treatment effect												
$\hat{\lambda}_H$	0.867	96.9%	−0.007	0.261	0.875	96.5%	−0.003	0.125	0.875	97.3%	0.001	0.062
$\hat{\tau}_F$	0.338	93.6	−0.000	0.171	0.358	93.0	0.000	0.093	0.391	95.3	0.001	0.041
$\hat{\tau}_F(c_{0.2})$	0.341	94.8	−0.002	0.170	0.356	96.2	−0.005	0.075	0.356	95.8	0.001	0.037
$\widehat{\Gamma}_F$	0.344	98.5	0.001	0.126	0.362	98.9	0.005	0.053	0.363	99.0	0.001	0.026

NOTES: The table presents the true value (truth) of each quantity along with the bias and standard deviation of each estimator as well as the coverage of its 95% confidence intervals under the "Small treatment effect" and "Large treatment effect" scenarios. All of the results shown here are for LASSO.

variance when estimating the PAPD between two ITRs (i.e., $\Delta_{0.2}(f, g)$ and $\Delta_{0.2}(f, h)$), the coverage stays close to 95%, implying that the bound for covariance has little effect on the variance estimation.

For estimated ITRs, Table 2 presents the results of LASSO under cross-validation. For BART and causal forest, obtaining an accurate Monte Carlo estimate of the true causal parameter values under cross-validation takes a prohibitively large amount of time. While the out-of-bag estimates of such true values can be computed, they have been shown to create bias under certain scenarios (Janitza and Hornung 2018). These true values are generally greater for a larger sample size because LASSO performs better with more data.

The proposed cross-validated estimators are approximately unbiased even for $n = 100$. The coverage is generally around or above the nominal 95% value, reflecting the conservative estimate of the variance. For the PAPE without and with budget constraint, that is, $\hat{\tau}_F$ and $\hat{\tau}_F(c_{0.2})$, the coverage is close to the nominal value. This indicates that the bias of the proposed conservative variance estimator is relatively small even though the number of folds for cross-validation is only $K = 5$. The performance of the proposed methodology is good even when the sample size is as small as $n = 100$. When the sample size is $n = 500$, the standard deviation of the cross-validated estimator is roughly half of the corresponding $n = 100$ fixed ITR estimator (this is a good comparison because each fold has 100 observations). This confirms the theoretical efficiency gain that results from cross-validation.

6. An Empirical Application

We apply the proposed methodology to the data from the Tennessee's Student/Teacher Achievement Ratio (STAR) project, which was a longitudinal study experimentally evaluating the impacts of class size in early education on various outcomes (Mosteller 1995). Another application based on a canvassing experiment is shown in Appendix A.8 (supplementary material).

6.1. Data and Setup

The STAR project randomly assigned over 7000 students across 79 schools to three different groups: small class,regular class,and regular class with a full-time teacher's aide. The experiment began when students entered kindergarten and continued through third grade. To create a binary treatment, we focus on the first two groups: small class and regular class without an aid. The treatment effect heterogeneity is important because reducing class size is costly, requiring additional teachers and classrooms. Policy makers who face a budget constraint may be interested in finding out which groups of students benefit most from a small class size so that the priority can be given to those students.

We follow the analysis strategies of the previous studies (e.g., Ding and Lehrer 2011; McKee, Sims, and Rivkin 2015) that estimated the heterogeneous effects of small classes on educational attainment. These authors adjust for school-level and student-level characteristics, but do not consider within-classroom interactions. Unfortunately, addressing this limitation is beyond the scope of this article. We use a total of 10 pretreatment covariates $\mathbf{X}_i$ that include four demographic characteristics of students (gender, race, birth month, and birth year) and six school characteristics (urban/rural, enrollment size, grade range, number of students on free lunch, number of students on school buses, and percentage of white students). Our treatment variable is the class size to which they were assigned at kindergarten: small class $T_i = 1$ and regular class without an aid $T_i = 0$. For the outcome variables Y_i, we use three standardized test scores measured at third grade: math, reading, and writing SAT scores.

The resulting dataset has a total of 1911 observations. The estimated average treatment effects (ATE) based on the entire data set are 6.78 (s.e. = 1.71), 5.78 (s.e. = 1.80), and 3.65 (s.e. = 1.63), for the reading, math, and writing scores, respectively. For the fixed test data, the estimated ATEs are similar; 5.10 (s.e. = 3.07), 2.78 (s.e.= 3.15), and 1.48 (s.e. = 2.96).

We evaluate the performance of ITRs using two settings considered above. First, we randomly split the data into the training data (70%) and test data (30%). We estimate an ITR from the training data and then evaluate it as a fixed ITR using the test data. This follows the setup considered in Sections 2 and 3. Second, we consider the evaluation of estimated ITRs based on the same experimental data. We utilize Algorithm 1 with 5-fold cross-validation (i.e., $K = 5$). For both settings, we use the same three machine learning algorithms. For Causal Forest, we set `tune.parameters = TRUE`. For BART, tuning was done on the number of trees. For LASSO, we tuned the regularization parameter while including all interac-

Table 3. The estimated population average prescription effect (PAPE) for BART, Causal Forest, and LASSO with and without a budget constraint.

	BART			Causal forest			LASSO		
	Est.	s.e.	Treated	Est.	s.e.	Treated	est.	s.e.	Treated
Fixed ITR									
No budget constraint									
Reading	0	0	100%	−0.38	1.14	84.3%	−0.41	1.10	84.4%
Math	0.52	1.09	86.7	0.09	1.18	80.3	1.73	1.25	78.7
Writing	−0.32	0.72	92.7	−0.70	1.18	78.0	−0.30	1.26	80.0
Budget constraint									
Reading	−0.89	1.30	20	0.66	1.23	20	−1.17	1.18	20
Math	0.70	1.25	20	2.57	1.29	20	1.25	1.32	20
Writing	2.60	1.17	20	2.98	1.18	20	0.28	1.19	20
Estimated ITR									
No budget constraint									
Reading	0.19	0.37	99.3%	0.31	0.77	86.6%	0.32	0.53	87.6%
Math	0.92	0.75	84.7	2.29	0.80	79.1	1.52	1.60	75.2
Writing	1.12	0.86	88.0	1.43	0.71	67.4	0.05	1.37	74.8
Budget constraint									
Reading	1.55	1.05	20	0.40	0.69	20	−0.15	1.41	20
Math	2.28	1.15	20	1.84	0.73	20	1.50	1.48	20
Writing	2.31	0.66	20	1.90	0.64	20	−0.47	1.34	20

NOTES: We estimate the PAPE for fixed and estimated individualized treatment rules (ITRs). The fixed ITRs are based on the training (70%) and test data (30%), whereas the estimated ITRs are based on 5-fold cross-validation. In addition, the average treatment effect estimates using the entire dataset. For each of the three outcomes, the point estimate, the standard error, and the average proportion treated are shown. The budget constraint considered here implies that the maximum proportion treated is 20%.

tion terms between covariates and the treatment variable. All tuning was done through the 5-fold cross-validation procedure on the training set using the PAPE as the evaluation metric. We then create an ITR as $\mathbf{1}\{\hat{\tau}(\mathbf{X}) > 0\}$ where $\hat{\tau}(\mathbf{X})$ is the estimated CATE obtained from each fitted model. As mentioned in Section 3, we center the outcome variable Y in evaluating the metrics to minimize the variance of the estimators.

6.2. Results

The upper panel of Table 3 presents the estimated PAPEs, their standard errors, and the proportion treated for fixed ITRs. We find that without a budget constraint, none of the machine learning algorithms significantly improves upon the random treatment rule. With a budget constraint of 20%, however, the ITRs based on causal forest appear to improve upon the random treatment rule at least for the math and writing scores. In contrast, the ITR based on BART only performs well for the writing score, whereas the performance of LASSO is not distinguishable from that of random treatment rule across all scores. The results are largely similar for the estimated ITRs, as shown in the lower panel of Table 3. The only difference is that BART performs slightly better while the performance of Causal Forest is slightly better for the fixed ITRs and worse for the estimated ITRs.

In the case of BART and Causal Forest, the standard errors for estimated ITRs are generally smaller than those for fixed ITRs, reflecting the efficiency gain due to cross-validation. For LASSO, however, the standard errors for fixed ITRs are often smaller than those for estimated ITRs. This is due to the fact that the ITRs estimated by LASSO are significantly variable across different folds under cross-validation. This variability results in the poor performance of LASSO in this application. In contrast, Causal Forest is most stable, generally leading to the best performance and the smallest standard errors. Lastly, we note that when compared to the estimated ATEs, some estimated PAPEs are of substantial size.

Table 4. The Estimated PAPD between Causal Forest, BART, and LASSO under a budget constraint of 20%.

	Causal Forest				BART	
	vs. BART		vs. LASSO		vs. LASSO	
	est.	95% CI	est.	95% CI	est.	95% CI
Fixed ITR						
Math	1.55	[−0.35, 3.45]	1.83	[−0.50, 4.16]	0.28	[−2.39, 2.95]
Reading	1.86	[−0.79, 4.51]	1.31	[−1.49, 4.11]	−0.55	[−4.02, 2.92]
Writing	0.38	[−1.66, 2.42]	2.69	[−0.27, 5.65]	2.32	[−0.53, 5.15]
Estimated ITR						
Reading	−1.15	[−3.99, 1.69]	0.55	[−1.05, 2.15]	1.70	[−0.90, 4.30]
Math	−0.43	[−2.57, 3.43]	0.34	[−1.32, 2.00]	0.77	[−1.99, 3.53]
Writing	−0.41	[−1.63, 0.80]	2.37	[0.76, 3.98]	2.79	[1.32, 4.26]

NOTE: The point estimates and 95% confidence intervals are shown.

Table 4 directly compares these three ITRs based on Causal Forest, BART, and LASSO by estimating the PAPD under the same budget constraint (i.e., 20%) as above. Causal Forest outperforms BART and LASSO in essentially all cases though the difference is not statistically significant. Under the cross-validation setting, Causal Forest and BART are statistically significantly more effective than LASSO in identifying students with grater treatment effects on their writing scores.

Finally, Figure 3 presents the estimated PAPE for the writing score across a range of budget constraints with the pointwise confidence intervals based on the variance formulas, for the fixed and estimated ITR cases. The difference between the solid red and black lines is the estimated PAPE, while the area between the red line and the black line corresponds to the area under the prescriptive effect curve (AUPEC). In each plot, the horizontal axis represents the budget constraint as the maximum proportion treated. In both the fixed and estimated ITR settings, BART and causal forest identify students who benefit positively from small class sizes when the maximum proportion treated is relatively small. In contrast, LASSO has a difficult time in finding these individuals. Additionally, we find that the standard error of the estimated AUPEC is greater for estimated ITRs than for fixed ITRs even though in the case

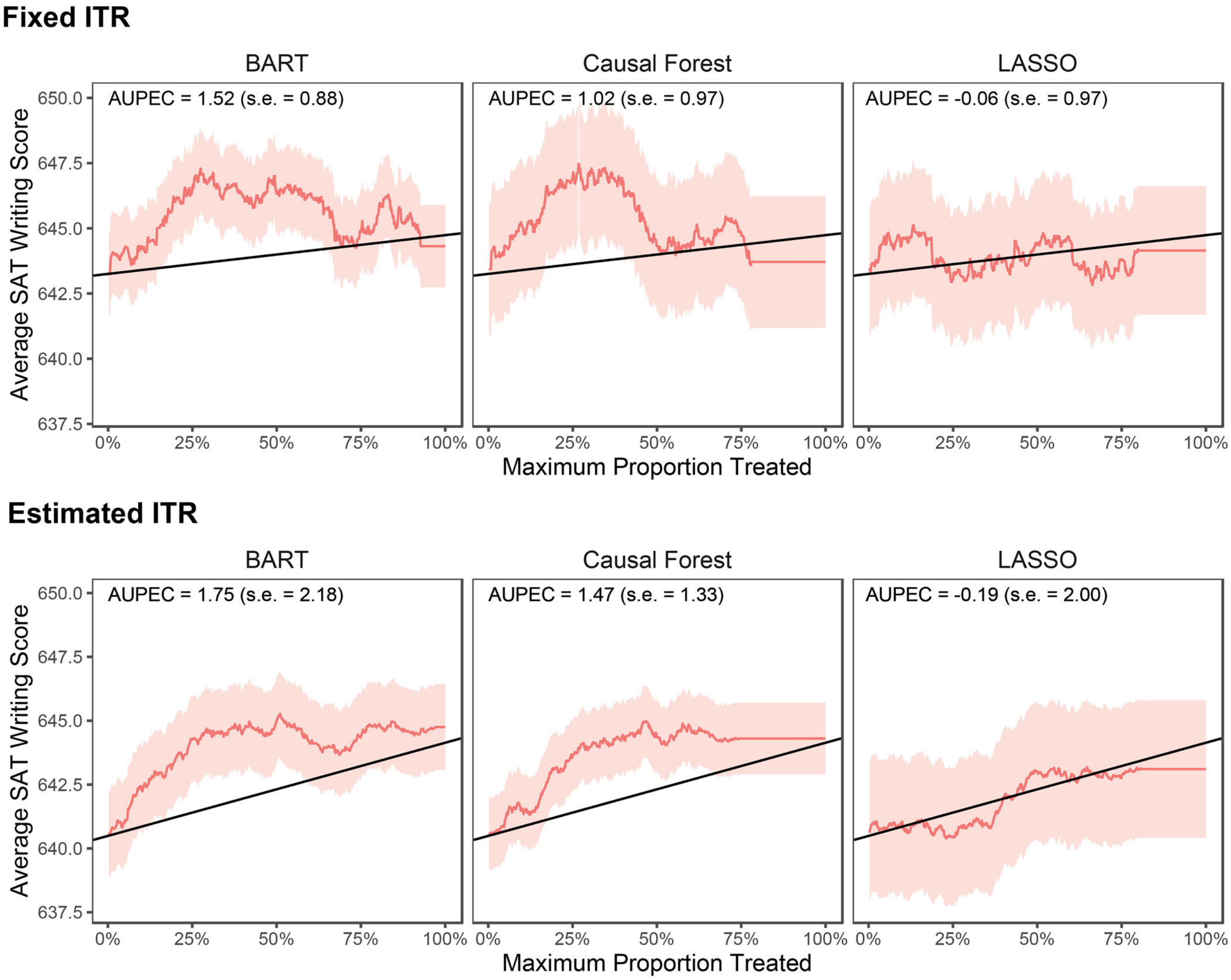

Figure 3. Estimated AUPEC. The results are presented for the fixed (upper panel) and estimated (bottom panel) individualized treatment rule (ITR) settings. A solid red line in each plot represents the PAPE for SAT writing scores across a range of budget constraint (horizontal axis) with the pointwise 95% confidence intervals. The area between this line and the black line representing random treatment is the AUPEC. The results are presented for the individualized treatment rules based on BART (left column), Causal Forest (middle column), and LASSO (right column), whereas each row presents the results for a different outcome.

of BART and Causal Forest, the opposite pattern is found for the PAPE with a given budget constraint. This finding is due to the high variance of the estimated AUPEC for different folds of cross-validation.

As the budget constraint is relaxed, the ITRs based on BART and Causal Forest yields the population average value similar to the one under the random treatment rule. This results in the inverted V-shape AUPEC curves observed in the left and middle plots of the figure. These inverted V-shape curves illustrate two separate phenomena. First, the students with the highest predicted CATE under BART and Causal Forests do have a higher treatment effect than the average, yielding an uplift in the PAPE curve compared to the random treatment rule. This shows that BART and Causal Forests are able to capture some causal heterogeneity that exists in the STAR data. Indeed, both BART and the Causal Forest estimated the CATE to be higher for nonwhite students and those who attend schools with a high percentage of students receiving free lunch. According to the variable importance statistic, these two covariates play an essential role in explaining causal heterogeneity (see, e.g., Finn and Achilles 1990; Jackson and Page 2013; Nye, Hedges, and Konstantopoulos 2000, similar findings).

However, as we further relax the budget constraint, the methods start treating students who are predicted to have a smaller (yet still positive) value of the CATE. These students tend to benefit less than the ATE, resulting in a smaller value of the PAPE as the budget increases. Eventually, both BART and Causal Forest start "over-treating" students who are estimated to have a small positive CATE but actually do not benefit from a small class. This results in the overall insignificant PAPE when no budget constraint is imposed.

7. Concluding Remarks

As the application of individualized treatment rules (ITRs) becomes more widespread in a variety of fields, a rigorous performance evaluation of ITRs plays an essential role before policy makers deploy them in a target population. We believe that the inferential approach proposed in this article provides a robust and widely applicable tool to policy makers. The proposed methodology also opens up opportunities to utilize the existing randomized controlled trial data for the efficient evaluation of ITRs as done in our empirical application. In addition, although we do not focus on the estimation of ITRs in this article, the proposed evaluation metrics can be used to tune hyper-parameters when cross validating machine learning algorithms. In future research, we plan to consider the extensions of the proposed methodology to other settings, including nonbinary treatments, dynamic treatments, and treatment allocations in the presence of interference between units.

Acknowledgments

We thank to Naoki Egami, Colin Fogarty, Zhichao Jiang, Susan Murphy, Nicole Pashley, Stefan Wager, three anonymous reviewers, and especially the Associate Editor for many helpful comments. We thank Dr Hsin-Hsiao Wang who inspired us to write this article. The proposed methodology is implemented through an open-source R package, evalITR, which is freely available for download at the Comprehensive R Archive Network (CRAN; *https://CRAN.R-project.org/package=evalITR*).

Funding

Imai thanks to the Alfred P. Sloan Foundation (# 2020-13946) for partial support of this research.

Supplementary Materials

The proofs for the theoretical results are included in the supplementary material.

ORCID

Kosuke Imai http://orcid.org/0000-0002-2748-1022
Michael Lingzhi Li http://orcid.org/0000-0002-2456-4834

References

Andrews, I., Kitagawa, T., and McCloskey, A. (2020), "Inference on Winners," Tech. rep., Harvard University, Department of Economics. [243]

Ascarza, E. (2018), "Retention Futility: Targeting High-Risk Customers Might be Ineffective," *Journal of Marketing Research*, 55, 1, 80–98. [243]

Athey, S., and Imbens, G. (2016), "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, pp. 7353–7360. [243]

Athey, S., Tibshirani, J., Wager, S. (2019), "Generalized Random Forests," *Annals of Statistics* 47, 1148–1178. [251]

Athey, S., and Wager, S. (2018), "Efficient Policy Learning," arXiv no. 1702.02896 . [242,243]

Chakraborty, B., Laber, E., and Zhao, Y.-Q. (2014), "Inference about the Expected Performance of a Data-Driven Dynamic Treatment Regime," *Clinical Trials*, 11, 408–417. [243]

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2019), "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments," Tech. rep. [243]

Chipman, H. A., George, E. I., McCulloch, R. E. (2010), "Bart: Bayesian Additive Regression Trees," *The Annals of Applied Statistics* 4, 266–298. [251]

Diemert, E., Betlei, A., Renaudin, C., and Amini, M.-R. (2018), "A Large Scale Benchmark for Uplift Modeling," In KDD, London, UK. [243]

Ding, W., and Lehrer, S. F. (2011), "Experimental Estimates of the Impacts of Class Size on Test Scores: Robustness and Heterogeneity," *Education Economics* 19, 229–252. [252]

Dudík, M., Langford, J., and Li, L. (2011), "Doubly Robust Policy Evaluation and Learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11)*, Marina Vannucci, ed., Rice University, Noah Harding, Omnipress. pp. 1097–1104. [243]

Fifield, B. (2018), "Empirical and Computational Methods for Electoral Politics," Ph.D. thesis, Princeton University. [243]

Finn, J. D., and Achilles, C. M. (1990), "Answers and Questions About Class Size: A Statewide Experiment," *American Educational Research Journal*, 27, 557–577. [254]

Fu, H., Zhou, J., and Faries, D. E. (2016), "Estimating Optimal Treatment Regimes via Subgroup Identification in Randomized Control Trials and Observational Studies," *Statistics in Medicine*, 35, 19, 3285–3302. [242]

Grimmer, J., Messing, S., and Westwood, S. J. (2017), "Estimating Heterogeneous Treatment Effects and The Effects of Heterogeneous Treatments With Ensemble Methods," *Political Analysis*, 25, 413–434. [243]

Gutierrez, P., and Gérardy, J.-Y. (2016). "Causal Inference and Uplift Modelling: A Review of the Literature," in *Proceedings of Machine Learning Research*, International Conference on Predictive Applications and APIs, Boston, MA, Vol. 67, pp. 1–13. [243]

Hahn, P. R., Dorie, V., and Murray, J. S. (2018), "Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017," Tech. rep., School of Mathematical and Statistical Sciences. [251]

Hahn, P. R., Murray, J. S., Carvalho, C. M. (2020), "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," *Bayesian Analysis*, 15, 965–1056. [243,251]

Hamburg, M. A., and Collins, F. S. (2010), "The Path to Personalized Medicine," *New England Journal of Medicine*, 363, 301–304. [242]

Imai, K., and Ratkovic, M. (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *Annals of Applied Statistics*, 7, 443–470. [243]

Imai, K., and Strauss, A. (2011), "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign," *Political Analysis*, 19, 1–19. [242,243]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press. [242]

Jackson, E., and Page, M. E. (2013), "Estimating the Distributional Effects of Education Reforms: A Look at Project Star," *Economics of Education Review*, 32, 92–103. [254]

Janitza, S., and Hornung, R. (2018), "On the Overestimation of Random Forest's Out-of-Bag Error," *PLoS ONE* 13, 8, e0201904. [252]

Jiang, N., and Li, L. (2016), "Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48 of *Proceedings of Machine Learning Research*, pp. 652–661. [243]

Kallus, N. (2018), "Balanced Policy Evaluation and Learning," in *Advances in Neural Information Processing Systems* (Vol. 31), eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc. pp. 8895–8906 [243]

Kamath, P. S., Wiesner, R. H., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., D'Amico, G., Dickson, E. R., and Kim, W. R. (2001), "A Model to Predict Survival in Patients with End-Stage Liver Disease," *Hepatology*, 33, 464–470. [245]

Kitagawa, T., and Tetenov, A. (2018), "Who Should Be Treated?: Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616. [242,243]

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019), "Meta-Learners for Estimating Heterogeneous Treatment Effects Using Machine Learning," *Proceedings of the National Academy of Sciences*, 116, 4156–4165. [243,247]

Luedtke, A. R., and van der Laan, M. J. (2016a), "Optimal Individualized Treatments in Resource-Limited Settings," *International Journal of Biostatistics*, 12, 283–303. [242,243]

——— (2016b), "Statistical Inference for the Mean Outcome Under a Possibly Non-Unique Optimal Treatment Strategy," *Annals of Statistics*, 44, 2, 713–742. [242,243]

Manski, C. F. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246. [243]

McKee, G., Sims, K. R. E., and Rivkin, S. G. (2015), "Disruption, Learning, and The Heterogeneous Benefits of Smaller Classes," *Empirical Economics*, 48, 1267–1286. [252]

Mosteller, F. (1995), "The Tennessee Study of Class Size in The Early School Grades," *The Future of Children*, 5, 113–127. [252]

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9. (translated in 1990)," *Statistical Science*, 5, 465–480. [242,246]

Nye, B. A., Hedges, L. V., and Konstantopoulos, S. (2000), "Do The Disadvantaged Benefit More from Small Classes? Evidence from The Tennessee Class Size Experiment," *American Journal of Education*, 109, 1–26. [254]

Qian, M., and Murphy, S. A. (2011), "Performance Gurantees for Individualized Treatment Rules," *Annals of Statistics*, 39, 1180–1210. [242,243,244]

Radcliffe, N. J. (2007), "Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models," *Direct Marketing Analytics Journal* 1, 14–21. [243,246]

Rai, Y. (2018), "Statistical Inference for Treatment Assignment Policies," Tech. rep., University of Wisconsin, Department of Economics. [243]

Rzepakowski, P., and Jaroszewicz, S. (2012), "Decision Trees for Uplift Modeling with Single and Multiple Treatments," *Knowledge and Information Systems*, 32, 303–327. [243,245]

Stuard, A., and Ord, J. K. (1994), *Kendall's Advanced Theory of Statistics, Distribution Theory* (Vol. 1), New York: Halsted Press. [248]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [251]

van der Laan, M. J., and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York: Springer. [243]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [243,247]

Zhang, B., Tsiatis, A. A., Davidian, M., and Laber, E. (2012), "Estimating Optimal Treatment Regimes From a Classification Perspective," *Statistics*, 1, 103–114. [242,243]

Zhao, Y., Zeng, D., Rush, J. A., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [243,244]

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017), "Residual Weighted Learning for Estimating Individualized Treatment Rules," *Journal of the American Statistical Association*, 112, 169–187. [242,244]

Zhou, Z., Athey, S., and Wager, S. (2018), "Offline Multi-Action Policy Learning: Generalization and Optimization," Tech. rep. [243,247]