Estimation of Bounds on Potential Outcomes For Decision Making

Maggie Makar [1] **Fredrik Johansson** [2] **John Guttag** [1] **David Sontag** [1]

Abstract

Estimation of individual treatment effects is commonly used as the basis for contextual decision making in fields such as healthcare, education, and economics. However, it is often sufficient for the decision maker to have estimates of upper and lower bounds on the potential outcomes of decision alternatives to assess risks and benefits. We show that, in such cases, we can improve sample efficiency by estimating simple functions that bound these outcomes instead of estimating their conditional expectations, which may be complex and hard to estimate. Our analysis highlights a trade-off between the complexity of the learning task and the confidence with which the learned bounds hold. Guided by these findings, we develop an algorithm for learning upper and lower bounds on potential outcomes which optimize an objective function defined by the decision maker, subject to the probability that bounds are violated being small. Using a clinical dataset and a well-known causality benchmark, we demonstrate that our algorithm outperforms baselines, providing tighter, more reliable bounds.

1. Introduction

In many practical situations, a decision maker wishes to intervene or assign a treatment to ensure that an outcome of interest falls within a safe range. One example, which we use throughout the paper, is when a physician considers whether or not to prescribe anticoagulants to mitigate the risk of stroke, as measured by the International Normalized Ratio (INR). The INR reflects the time it takes for blood to clot. For previous stroke patients, a healthy INR is 2–3. Values lower than 2 signal elevated risk of an Ischemic stroke, and higher than 3 signal elevated risk of a Hemorrhagic stroke. To make an informed decision, the physician

needs to know if the potential outcomes under treatment and non-treatment fall within 2–3. Learning that the difference between the potential outcomes, i.e., the Individual Treatment Effect (ITE) is 1.5, does not immediately imply an optimal treatment decision; it could be that the patient's INR decreases from 4 to 2.5 or from 5.5 to 4. More information about the potential outcomes themselves is needed, but knowing their exact value is not necessary. It is sufficient to know that the patient's INR is somewhere between 2.1 and 2.9 if treated. For example, knowing that it is 2.853 does not provide additional insight. For these two reasons, we study the task of estimating reliable covariate-conditional bounds on potential outcomes using observational data.

Most existing methods for estimating causal effects and potential outcomes attempt to fit the expected outcomes as functions of observed covariates, typically relying on variants of Empirical Risk Minimization (ERM) strategies (Hill, 2011; Shalit et al., 2017; Alaa & van der Schaar, 2018; 2017). Some of these methods produce prediction intervals centered around the estimated expected response (outcome) surface, which can be used to bound the potential outcome from above and below. These intervals have approximately valid coverage for large samples, provided that the mean estimate is sufficiently unbiased. However, achieving this is not always feasible in small samples, leading to high false coverage rates (FCRs), defined as the rate at which outcomes are observed outside of the given prediction interval.

Instead of attempting to directly fit the potential outcomes, which may be complex and hard to estimate from small samples, we propose to fit simpler functions that bound the outcomes from above and below. Within this simpler function class, we identify estimates of the potential outcomes that maximize a utility (objective) function specified by the decision maker. Figure 1 shows the intuition behind our approach. For example, if the decision maker wants to ensure that the uncertainty in the potential outcome estimates is small on average, they could require that the average interval width (= upper bound - lower bound) is small. Alternatively, if they wish to ensure that no patient sub-population has excessively uncertain estimates (i.e., wide intervals) they could require that the maximum interval width is minimized.

We make the following main contributions: (i) We give results on the generalization properties of learned bounds on

[1]CSAIL, MIT [2]Chalmers University of Technology. Correspondence to: Maggie Makar <mmakar@mit.edu>.
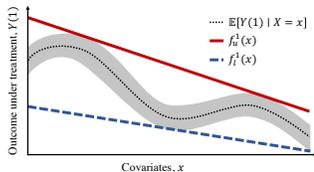
Figure 1. Illustration of the intuition behind our theoretical findings. While the true potential outcome (black/gray) belongs to a complex class, the upper (red) and lower (blue) bounds f_u^1, f_l^1 that correctly cover it belong to a simple linear function space.

potential outcomes and the conditions under which estimation of such bounds yields better sample complexity than fitting the expected outcomes using standard risk minimization methods. Our analysis highlights a trade-off between reliability (i.e., the probability that the bounds correctly cover the data) and the complexity of the learning task. (ii) We design an algorithm that finds the optimal bound estimates that maximize a given utility or objective function while providing reliable bounds. We explore different objective functions, analyzing the differences between the resulting bounds, and prove equivalence to quantile regression in a special case. (iii) We evaluate our algorithm on a semi-synthetic clinical dataset and a well-known causality benchmark. We show how it can guide treatment decisions, and that it achieves a better trade-off between bound violations and utility than baseline algorithms.

2. Related work

Research into methods for estimating conditional causal effects has focused primarily on estimating the expected potential outcomes or conditional average treatment effect (CATE) as functions of observed covariates (Dorie et al., 2019). For example, Alaa & van der Schaar (2018) showed that the CATE estimation problem is as hard as modelling the more complex of the two potential outcomes in the mini-max sense. Similarly, Nie & Wager (2017) show asymptotic bounds that rely on the complexity of the underlying function class of the CATE. More generally, recent work in CATE estimation has focused on the learning challenges associated with the difference between the treated and control populations, and on improving finite sample efficiency by sharing data between treatment groups (Johansson et al., 2016; Shalit et al., 2017; Alaa & van der Schaar, 2017; Hill, 2011). In contrast, we aim to improve sample efficiency by providing bounds on the causal estimands.

Other work focuses on estimating lower or upper bounds of Average Treatment Effect (ATE), to account for the possibility of unobserved confounding (Balke & Pearl, 1997; Bareinboim & Pearl, 2012; Pearl, 2009; Cai et al., 2008). Recently, this type of analysis was extended to include bounds on CATE, but again in the presence of hidden confounding (Kallus et al., 2019). This line of work falls under sensi-

tivity analysis (Rosenbaum, 2014), which is distinct from our work in that we aim to find bounds on the potential outcomes even in the absence of unobserved confounding.

Another related line of work is the problem of conditional quantile treatment effect estimation (Koenker & Bassett Jr, 1978; Chernozhukov & Hansen, 2005). Like our method, quantile methods give can give approximate bounds on the potential outcomes. The distinction is that the main objective of our method is not to estimate the specific quantile of treatment effect, but rather to provide the simplest functions that bound the outcomes such that an objective function given by the decision maker is optimized; we do not wish in general to establish asymptotic convergence to a particular quantile of the treatment effect. However, as we prove later, quantile estimation is a special case of our setting for a certain objective function.

At the time of publication of this paper, new work extended conformal intervals (Lei et al., 2018) to settings similar to ours, where the outcomes are counterfactual (Lei & Candès, 2020). Our work is distinct from the work presented in (Lei & Candès, 2020) in three ways (1) we provide theoretical guarantees for the *finite* sample rather than asymptotic regime, (2) our theoretical analysis highlights a fundamental trade-off between the statistical complexity of the learning problem and the confidence with which the learned interval truly covers the potential outcomes. Finally, (3) our approach allows for a more general definition of interval optimality; we not assume that tightness of the bounds is the only important metric to be optimized, but it allows the decision maker to define their own desiderata for optimality (e.g., fairness).

Our work is related to offline policy learning (e.g., Swaminathan & Joachims (2015a;b)). The main difference between this work and ours is that we wish to obtain bounds for the potential outcomes, not just an optimal policy. This allows the decision maker to consider the estimated effect of the treatment against a backdrop of additional information that may not be recorded in the observational data.

3. Background

We consider learning of bounds on potential outcomes from finite-sample observational data, adopting the notation of the Neyman-Rubin potential outcomes framework (Rubin, 2005). For each unit i (e.g. patient), we observe a set of features $X_i \in \mathcal{X}$, with $\mathcal{X}$ a bounded subset of $\mathbb{R}^d$, an action (also known as treatment or intervention) $T_i \in \{0, 1\}$ and an outcome $Y_i \in \mathbb{R}$. We observe these variables through samples $(x_1, t_1, y_1), ..., (x_n, t_n, y_n) \overset{i.i.d.}{\sim} p(X, T, Y)$ and denote by $n_t = \sum_{i=1}^n \mathbb{1}\{t_i = t\}$ the number of observed samples for treatment group $t \in \{0, 1\}$, and let $p_t(X) = p(X \mid T = t)$. The observed outcome is one of the two *potential outcomes*, $Y(0)$ and $Y(1)$, under control ($T = 0$)

and treatment ($T = 1$), respectively. We use $\|a\|_p$ to denote the $p-$norm of a vector a. When the subscript is omitted, we refer to the 2-norm.

We seek to learn high-probability bounds on both potential outcomes, $Y(0)$ and $Y(1)$, conditioned on the set of observed features X. Since only one outcome is observed, the other is not identifiable without strong assumptions. To that end, we assume that the features X are sufficient to deconfound estimates of $Y(0), Y(1)$:

Assumption 1. *The features X, treatment T and potential outcomes $Y(0), Y(1)$ satisfy for some $\epsilon > 0$*

1. *Strong ignorability: $Y(0), Y(1) \perp\!\!\!\perp T \mid X$*
2. *Overlap: $\forall x, t : p(T = t \mid x) > \epsilon$*
3. *Consistency: $Y = Y(T)$*

Under Assumption 1, $p(Y(t) = y \mid X = x) = p(Y = y \mid T = t, X = x)$ (Imbens & Wooldridge, 2009). This means that the distribution of potential outcomes can be estimated through regression or other standard methods. When treatment and outcomes are confounded, estimates of causal effects exhibit bias. For example, if medication A was prescribed more often to terminally ill patients than the alternative treatment B, we might learn that the life expectancy on treatment A was lower than on B, regardless of its average causal effect. To undo this bias, it is common to use the propensity score $e(x, t) := p(T = t \mid X = x)$ to re-weight the cohort using importance weighting.

Definition 1. *The importance weighting function w_t for group $t \in \{0, 1\}$ is $w_t(x) := p(T = t)/e(x, t)$.*

We use w_i to denote $w_{t_i}(x_i)$ for a sample $(x_i, t_i) \sim p$. With w_t as in Definition 1, we have for an arbitrary function f on $\mathcal{X}$ (e.g., the expected outcome or a prediction loss), $\mathbb{E}_X[f(X)] = \mathbb{E}_{X|T}[w_t(X)f(X) \mid T = t]$. By Assumption 1, we have that the importance weights are bounded, meaning that for some $C_t < \infty$ and $t \in \{0, 1\}$:

$$\sup_{x \in \mathcal{X}} w_t(x) = \sup_{x \in \mathcal{X}} \frac{p(T = t)}{e(x, t)} = 2^{D_\infty(p\|p_t)} = C_t, \quad (1)$$

where $D_k(p\|q)$ is the k$^{\text{th}}$-order Rényi divergence, and the second equality follows by applying the Bayes rule, and the definition of the Rényi divergence. It will be convenient to denote $2^{D_k(p\|q)}$ by $d_k(p\|q)$. Since $2^{D_{k-1}(p\|p_t)} < 2^{D_k(p\|p_t)}$, we have $d_2(p\|p_t) < C_t$.

4. Generalization of bounds on potential outcomes

Our goal is to estimate four functions; lower and upper bounds for the potential outcome under treatment, $\boldsymbol{f^1}(x) = \{f_l^1(x), f_u^1(x)\}$, and similarly defined functions for the outcome under control $\boldsymbol{f^0}(x) = \{f_l^0(x), f_u^0(x)\}$. For these

estimates to be useful for decision-making, we want to make the assertion that for some small $\nu' > 0$, and for $t \in \{0, 1\}$, we have false coverage rate (FCR) bounded by ν',

$$\text{FCR}_{\boldsymbol{f^t}} := \Pr_{X, Y(t)} \left[Y(t) \notin [f_l^t(X), f_u^t(X)] \right] \leq \nu' . \quad (2)$$

Without loss of generality, we will focus on estimating a lower bound for the outcome under treatment $T = t$, meaning we will focus on finding some $f_l^t(x)$ such that for a small $\nu > 0$, we have that

$$\Pr_{X, Y(t)}[f_l^t(X) \leq Y(t)] \geq 1 - \nu. \quad (3)$$

Note that in expressions 2 and 3 the probabilities are defined over $p(X, Y(t)) \neq p(X, Y \mid T = t)$, due to confounding. However, under Assumption 1, this probability is identifiable from observed data.

It will be useful to restate our objective in terms of the (signed) residual of a function f, defined next.

Definition 2. *For an arbitrary function f, the signed residuals for $x, y \in \mathcal{X} \times \mathcal{Y}$: $\underline{r}_f(x, y) = y - f(x)$.*

Expression (3) can be restated as $\Pr[\underline{r}_{f_l^t}(X, Y(t)) \geq 0] \geq 1 - \nu$. To be more cautious, we might wish to leave a "buffer zone" or a margin, and instead demand that $\underline{r}_{f_l^t}(x, y) \geq \gamma$ for some $\gamma > 0$. In this setting, a violation occurs when $\underline{r}_{f_l^t}(x, y) < \gamma$. Larger values of γ would imply higher reliability: we are more confident that we are unlikely to observe a violation of the bounds, i.e., unlikely to overestimate the outcome under treatment t. With that, direct parallels could be drawn between our setup and that of maximum-margin algorithms: we want to ensure that the signed residual is larger than 0 by a margin of γ. The larger γ is, the more confident we are that our lower bound holds. We can now define the unobserved risk that we wish to study:

Definition 3. *For $f_l^t \in \mathcal{F}$, $\gamma > 0$, we define the risk of overestimation over the full unknown distribution:*

$$\underline{R}_{f_l^t}(\gamma) = \mathbb{E}_{X, Y(t)} \left[\mathbb{1}\{\underline{r}_{f_l^t}(X, Y(t)) < \gamma\} \right].$$

To account for confounding due to biased (non-randomized) treatment assignment, we consider a re-weighted risk:

$$\underline{R}_{f_l^t}^w(\gamma) = \mathbb{E}_{X, Y|T} \left[w(x)\mathbb{1}\{\underline{r}_{f_l^t}(X, Y) < \gamma\} \mid T = t \right]$$

Under Assumption 1, $\underline{R}_{f_l^t}(\gamma) = \underline{R}_{f_l^t}^{w_1}(\gamma)$. Since our notions of confidence are closely related to the margin, γ, it will be more useful to reason about the magnitude of margin violations, which is defined next.

Definition 4. *For $z = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, known $\boldsymbol{w}_t$, $f_l^t \in \mathcal{F}$, and $\gamma > 0$, we define the average weighted magnitude of training set violations as*

$$\underline{D}^{\boldsymbol{w}_t}(z, f_l^t, \gamma) = \sum_{x, y \in z} w_t(x) \max\{0, \gamma - \underline{r}_{f_l^t}(x, y)\}$$

In the remainder of this section, we give bounds on expected margin violation as a function of $\underline{D}^{\boldsymbol{w}_t}$. We restrict our analyses to sturdy function classes, as defined in (Shawe-Taylor & Williamson, 1999) with with range = $[a, b]$. Informally, sturdy function classes have images that are compact subsets of $\mathbb{R}$. We rely on the covering number as a measure of complexity of the analyzed function classes. We use fat-shattering dimensions to study how fast the complexity of a function class can grow with the sample size. Explicit definitions of these three concepts are presented in the supplement (definitions A1, A2 and A3 respectively).

4.1. Generalization of reliable estimators

We start by studying the risk of overestimation for re-weighted estimators. To make our main finding easy to follow, we focus on the class of linear functions in a kernel defined feature space. Theorem A1 in the supplement gives the analogous bounds for more general function spaces.

Theorem 1. *Let $\mathcal{F}$ be the class of linear functions in a kernel defined feature space, $z = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, and C_t be as defined in expression* (1). *For $f_l^t \in \mathcal{F}$, and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_t}(z, f_t^l, \gamma) = D > 0$. With a probability $1 - \delta$ over the draw of random samples, we have that:*

$$\underline{R}_{f_l^t}(\gamma) \leq \frac{4C_t(k_t + \log \frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log \frac{1}{\delta})}{n_t}}$$
(4)

where, for $t \in \{0,1\}$,

$$k_t = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2n_t) + \frac{D}{\gamma} \log \frac{\exp(n_t + D/\gamma - 1)}{D/\gamma} \right\rceil .$$

The proof is outlined in the supplement. **Remarks:**

1. Theorem 1 states that the expected rate of overestimation is bounded by terms at most linear in k_t—the sum of the log covering number of $\mathcal{F}$ *as defined by the margin* γ, and the ratio of the violations on the training data to γ. The fact that the covering number is controlled by the margin parameter γ shows that the complexity of this learning task relies on how certain we wish to be that the lower bound is not overestimated; more certainty requires a larger γ which implies a smaller log covering number. This approach departs from previous literature which instead shows that the sample complexity of risk minimization relies on the covering number of a class containing the true function (Alaa & van der Schaar, 2018). In applications where it is sufficient to have reliable *bounds* on the potential outcomes to make good decisions, this finding can be crucial—especially if the outcomes are difficult to estimate accurately using small samples. Note that the covering number can be bounded by the fat-shattering dimension at a scale proportional to γ.

2. Both terms in k_t decrease as γ increases, which means that the risk of overestimation decreases as γ increases. This property is important because it implies that we can control the risk of overestimation by requiring a large margin. To see that, note that larger γ shrinks the space of viable functions, which decreases the γ-covering number. The second term includes the ratio of the sum of violations on the training set, D, which decreases as γ increases, to γ. Hence the second term also decreases as γ increases.

Corollary A1 in the supplement, builds on theorem 1 to get a bound on the generalization error for bounds on the ITE.

4.2. Generalization of reliable, informative estimators

Theorem 1 establishes that the probability of overestimation decreases as we increase the margin γ. However, arbitrarily large values of γ could result in excessively "cautious" estimates with low risk of overestimation, at the expense of being too loose to be useful in guiding decisions. In this work, we consider bounds to be informative or have high utility if they imply low uncertainty in the value of the true potential outcomes. We restrict ourselves to definitions of uncertainty that rely on the interval width (IW) of bounds $\boldsymbol{f} := (f_u, f_l)$

$$\text{IW}_{\boldsymbol{f}}(x) := f_u(x) - f_l(x) .$$
(5)

Smaller $\text{IW}_{\boldsymbol{f}}(x)$ implies that bounds are tighter, which implies less uncertainty in the value of the potential outcomes. Intuitively, for f_u and f_l to give small $\text{IW}_{\boldsymbol{f}}$, they need to *close* to each other. We define these "close" functions and the classes to which they belong as follows:

Definition 5. *Let $p \geq 1$, and $\mathcal{X} := \{x : ||x|| \leq r\}$. We say that two classes of bounded linear functionals $\mathcal{F}_l, \mathcal{F}_u$ are* informative *if $\mathcal{F}_l \subseteq \{\mathcal{X} \ni x \mapsto \langle f_l, x \rangle, ||f_l|| \leq A\}$ and $\mathcal{F}_u \subseteq \{\mathcal{X} \ni x \mapsto \langle f_u, x \rangle, \forall f_l \in \mathcal{F}_l; ||f_u - f_l|| < B, \forall x \in \mathcal{X} : f_l(x) \leq f_u(x)\}$.*

In words, $\mathcal{F}_l$ is the set of functions with norm $\leq A$, while $\mathcal{F}_u$ is the set of functions that are close to functions in $\mathcal{F}_l$, specifically, within B distance from each $f_l \in \mathcal{F}_l$. In addition, we specify that $f_l(x) \leq f_u(x)$ for every $x \in \mathcal{X}$.

The next theorem extends theorem 1 to these informative function classes, allowing us to study the risk of overestimation for tight intervals. To improve readability, log terms which do not affect the interpretation of the statement have been suppressed. The full statement is presented in Theorem A2.

Theorem 2. *Let $\mathcal{F}_l^t$, $\mathcal{F}_u^t$, A, B, and r be as defined in definition 5, z, and D as defined in theorem 1, and C_t be as defined in expression* (1). *For $f_l^t \in \mathcal{F}_l^t$, $f_u^t \in \mathcal{F}_u^t$ and any $\gamma > 0$, with a probability $1 - \delta$ over the draw of random samples, the bound* (4) *in Theorem 1 applies with, for $t \in$*

$\{0, 1\}$,

$$k_t \approx \left\lceil \left(\frac{r(A+B)}{\gamma}\right)^2 + \frac{D}{\gamma} \log \frac{e(n_t + D/\gamma - 1)}{D/\gamma} \right\rceil .$$

Theorem 2 gives us an idea of how to learn informative bounds that reliably cover the potential outcomes. It suggests that one way to reduce generalization error is to minimize A, the norm of f_l^t, B the distance between f_l^t and f_u^t, and D, the sum of violations on the training data.

5. Learning reliable, informative bounds

We present the Bounded Potential outcomes algorithm (BP) for learning informative bounds on potential outcomes under the constraint that they are violated with low probability. The algorithm is flexible in that it can maximize different utilities or notions of informativeness that the decision maker might have. For brevity, we focus on utility as defined by small IW. BP leverages our theoretical findings by explicitly constraining the violations on the training data, and minimizing some loss function, ℓ, of the interval widths.

The appropriate loss function will vary between applications. We consider optimizing three loss functions of IW over $p(x)$: $\ell^{(1)}$ represents the desire to achieve a tight prediction bound on average, captured in the mean absolute interval width. $\ell^{(2)}$ penalizes the mean squared interval width, placing a higher penalty on points with very wide bounds. The third $\ell^{(\infty)}$ minimizes the worst (widest) interval by penalizing the maximum interval width.

We consider learning under the following conditions. Let $\phi : \mathcal{X} \to \mathbb{R}$ be the feature map corresponding to a reproducing kernel $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. For treatments $t \in \{0, 1\}$ and bounds $b \in \{l, u\}$ (lower/upper), let $f_b^t(x_i) := \langle \theta_b^t, \phi(x_i) \rangle + \rho_b^t$. In this setting, all three losses $(\ell^{(1)}, \ell^{(2)}, \ell^{(\infty)})$ are convex in θ. Let sample weights w_{t_i} be defined as in Definition 1, and define $\widetilde{w}_{t_i} := w_{t_i} / \sum_{j : t_j = t_i} w_{t_j}$. Finally, let $\Lambda(f)$ denote a term that measures complexity of f, e.g., the squared norm of parameters.

We describe two versions of BP: BP-D, a decoupled version where the bounds for the treated and control groups are fitted separately, and BP-C, a coupled version where the two are fitted simultaneously.

5.1. BP-D: decoupled treatment groups

First, we consider estimating bounds f_u, f_l on a single potential outcome $Y(t)$, independently of others. We minimize the weighted loss $\ell_{\widetilde{w}}^{(p)}(f)$ and desire for bounds to be violated only with small probability over $p(x)$. We let the loss $\ell_{\widetilde{w}}^{(p)}(f)$ be defined by either the mean absolute interval width, $\ell_{\widetilde{w}}^{(1)}(f) \sum_{i : t_i = t} \widetilde{w}_{t_i} |\text{IW}_f(x_i)|$, the mean squared

interval width, $\ell_{\widetilde{w}}^{(2)}(f) = \sum_{i : t_i = t} \widetilde{w}_{t_i} (\text{IW}_f(x_i))^2$, or the maximum interval width, $\ell_{\widetilde{w}}^{(\infty)}(f) = \sup_{i : t_i = t} (\text{IW}_f(x_i))$.

$$
\begin{aligned}
\underset{f = \{f_u, f_l\}}{\text{minimize}} \quad & \ell_{\widetilde{w}}^{(p)}(f) + \alpha \Lambda(f) \\
\text{subject to} \quad & \sum_{i : t_i = t} \widetilde{w}_{t_i} \max(y_i - f_u(x_i), 0) \leq \beta_u \\
& \sum_{i : t_i = t} \widetilde{w}_{t_i} \max(f_l(x_i) - y_i, 0) \leq \beta_l \\
& f_l(x_i) \leq f_u(x_i) , \forall i : t_i = t .
\end{aligned}
\tag{6}
$$

Note that the constraints are defined with respect to the magnitude of the violations, which does not immediately translate into a specific FCR. We address this issue in section 5.3. Problem (6) can be solved separately for the two treatment groups, as is done in two-learners or the treatment variable could be added in as a feature and the two treatment groups can be jointly trained, as is done in single-learners (Künzel et al., 2019). Next, we highlight some important characteristics of this estimator.

1. BP-D minimizes the lower bound in Theorem 2. Note that BP-D is specified over the set of linear functions with kernel defined feature spaces. With Λ defined as the 2-norm of the vector θ, and because of the last constraint ($f_l \leq f_u$), the functions returned by BP-D fall within the set of functions defined in definition 5 with high probability, and hence theorem 2 is applicable here. Recall that theorem 2 states that for this estimated function to be optimal, they need to minimize $A = ||\theta||$, $B =$ distance ($p-$norm) between the upper and the lower bounds and $D =$ the magnitude of the training set violations while maximizing γ. Problem (6) directly minimizes the A, B (for $p = 1, 2, \infty$ depending on ℓ) and D. As for γ: suppose we fix the bias to be $\tilde{\rho}_b^t$, then $\gamma_b^t = \tilde{\rho}_b^t - \rho_b^t$, where the latter is the bias returned by solving problem (6). Because problem (6) minimizes ρ_b^t, it maximizes γ_b^t for a fixed $\tilde{\rho}_b^t$. Ideally, we would not fix $\tilde{\rho}_b^t$ in advance, but let it be decided by the data. We address this issue in section 5.3.

2. BP-D with $\ell^{(1)}$-loss is equivalent to quantile regression. When minimizing the mean absolute interval width, our problem reuces to a quantile regression with non-crossing constraints (Takeuchi et al., 2006) of quantiles q and $1 - q$ for for some choice of $q \in (0, .5)$.

Theorem 3. *Assume that* (6) *is strictly convex and has a strictly feasible solution. Then, for any fixed quantile $q \in (0.5, 1)$, there are parameters $\beta_u, \beta_l \geq 0$ such that the minimizers f_u^*, f_l^* of* (6) *with absolute loss and the minimizers of the quantile loss for quantiles $(q, 1 - q)$, with non-crossing constraints, are equal.*

A proof is given in the appendix.

BP-D allows us to learn reliable and informative bounds but it does not make use of the "unlabeled" data from the opposite treatment group. This is addressed next.

5.2. BP-C: coupled treatment groups

In the coupled problem, we make use of samples from the counterfactual treatment group in two ways. First, we apply constraints that ensure that the lower and upper bounds do not cross also for counterfactual outcomes. Second, the loss functions are defined with respect to the full marginal distribution of subjects (including counterfactual treatment assignments). We define the coupled version of the mean absolute loss $\ell^{(1)} = \sum_{i=1}^{n} \sum_{t=0}^{1} \widetilde{w}_{ti} |\mathrm{IW}_{\boldsymbol{f}^t}(x_i)|$, mean squared interval width, $\ell^{(2)} = \sum_{i=1}^{n} \sum_{t=0}^{1} \widetilde{w}_{ti} \mathrm{IW}_{\boldsymbol{f}^t}(x_i)^2$, and maximum interval width, $\ell^{(\infty)} = \sup_{i=1}^{n} \sum_{t=0}^{1} \mathrm{IW}_{\boldsymbol{f}^t}(x_i)$. The coupled problem becomes:

$$\underset{\{\boldsymbol{f}^t = \{f_u^t, f_l^t\}\}}{\text{minimize}} \quad \ell_{\widetilde{w}}^{(p)}(\boldsymbol{f}^0, \boldsymbol{f}^1) + \alpha \cdot (\Lambda(\boldsymbol{f}^0) + \Lambda(\boldsymbol{f}^1))$$

$$\text{subject to} \quad \sum_{i:t_i=t} \widetilde{w}_{ti} \max(y_i - f_u^t(x_i), 0) \leq \beta_u, \forall t$$

$$\sum_{i:t_i=t} \widetilde{w}_{ti} \max(f_l^t(x_i) - y_i, 0) \leq \beta_l, \forall t$$

$$f_l^t(x_i) \leq f_u^t(x_i), \forall t, i : t_i = t. \quad (7)$$

Given Assumption 1, specifically, the assumption of overlap this encourages the counterfactual outcome intervals to be small even if the corresponding treatment assignment is not observed. By coupling the two objectives, we allow information to be shared between the treated and non-treated populations in a semi-supervised way. We caution, however, that in the absence of overlap, the coupled loss might be overly optimistic about in regions of non-overlap, returning intervals that do not cover the true data. With f_l, f_u linear in the representation ϕ and $\Lambda(f)$ defined as the L2 norm of the function weights, expressions (6) and (7) are both convex programs which can be readily solved by a general solver. Our code is available at <github.com/mymakar/bpo.git>.

5.3. Cross-Validating BP

BP-C/D requires a regularization parameter, α, a level of tolerance to violations, $\beta_{u,l}$, and σ, which controls the kernel (e.g., the length scale for Gaussian kernels or the polynomial degree for polynomial kernels). Suppose that we solve problem (6) or (7) and get some estimate for the bias $\tilde{\rho}_b^t$, we specify an additional parameter $\gamma > 0$, and take the final estimate $\rho_l^t := \tilde{\rho}_l^t - \gamma$ and $\rho_u^t := \tilde{\rho}_u^t + \gamma$. This allows us to set γ based on the data rather than specify it apriori.

BP constrains the magnitude of the violations rather than the FCR directly. This allows the algorithm to directly reflect

the theory and makes the optimization problem easier. The disadvantage is that the magnitude of violations does not directly translate into a specific FCR. We address this issue by designing a cross-validation algorithm that picks the hyperparameters of the model to achieve a required FCR, ν. The algorithm takes as an input the training data, ν, ℓ, the required loss to minimize, and M, the set of hyperparameters to consider. We then split the data into training and validation. For each set of parameters $m \in M$, we use the training set to solve problem (6) or (7). We estimate $\hat{\nu}_m$ and $\widehat{\ell}_m$, the FCR and loss corresponding to m on the held-out set. We discard of all the hyperparameters with a corresponding $\hat{\nu}_m > \nu$, and define $M' = \{m : \hat{\nu}_m \leq \nu\}$. We set the optimal hyperparameters $m^* := \min_{m \in M'} \widehat{\ell}_m$. The procedure is summarized in Algorithm 1 in the supplement.

6. Experiments

We compare our model to other interval estimation methods. First is classical confidence-interval based approaches. We use **XX-CCI** to refer to this approach, where XX will be replaced by the name of the base model (e.g., if it is a Gaussian Process, we use GP-CCI). While popular, confidence intervals are known to have poor coverage in finite samples (Sargent et al., 1992; Lei et al., 2018). Conformal intervals, the second interval estimation method we compare against, were introduced as an alternative with better finite sample coverage (Lei et al., 2018). Conformal intervals are estimated by splitting the training data into two parts. The first part is used to train the outcome model, where parameters are picked via the usual cross-validation techniques. We estimate the residuals on the second subset of the training data. If the required FCR is q, we take the $1 - q^{th}$ quantile of the residuals to be a "shifting" parameter (akin to γ in our setting). The conformal intervals for a test sample are taken to be the estimated outcome $\pm$ the shifting parameter. We use **XX-CI** to refer to this approach. Finally, we introduce γ-intervals, which we refer to as **XX-γ**. Similar to conformal intervals, we split the data into two, fitting the best model on the first half and then picking the smallest shifting parameter γ that achieves the required FCR on the second half. We use **BP-V-Lp** to refer to our models, where V refers to the D (decoupled) or C (coupled) version and Lp refers to the norm of the loss (1, 2, or ∞). Recall that the 1-norm is similar to quantile regressions (QR) (by theorem 3).

We evaluate the performance of our models and the baselines on a held-out test set with respect to two criteria: the achieved FCR, as defined in equation (2) and the utility as measured by the mean IW and the max IW, as defined in equation (5). Additional cross-validation details for our model and the baselines are included in the supplement.

We analyze settings where we expect BP to outperform base-

Table 1. IST results. Table shows results averaged over 20 simulations, confirming conclusions from figure 2.

Model	FCR	Mean IW	Max IW
		$Y(1)$ results	
BP-D-L2	0.007 (0.36)	1.04 (0.05)	2.15 (0.19)
BP-D-L$_\infty$	0.007 (0.37)	1.16 (0.06)	1.16 (0.06)
QR/BP-D-L1	0.007 (0.43)	1.07 (0.09)	2.25 (0.26)
KR$-\gamma$	0.004 (0.81)	1.96 (0.09)	1.96 (0.09)
KR-CI	0.0 (0.0)	2.41 (0.07)	2.41 (0.07)
		$Y(0)$ results	
BP-C-L2	0.007 (0.59)	1.35 (0.17)	1.62 (0.26)
BP-D-L2	0.005 (0.51)	1.37 (0.13)	1.72 (0.2)

lines. Most baselines make restrictive assumptions about the distributions of the residuals. When such assumptions break, the resulting intervals are no longer tight or do not correctly cover the outcomes. We briefly outline such assumptions:

1. **Symmetry**. This assumption states that in order to get a 5% FCR, we need to ensure that the lower and upper bounds are violated by at most 2.5% each. In some cases, the tightest bounds would be achieved by non-symmetrical bounds, e.g., the lower bound is violated by 1% whereas the upper bound is violated by 4%. Violations to the symmetry assumption occur, for example, when the model is misspecified, which leads to biased estimates. In that case, tight bounds should reflect the direction of bias: if the estimates are biased downwards (meaning lower than the true value), it is more important that the upper bounds are not violated, whereas violations to the lower bound are more permissible (since the estimate itself is a lower value than the true outcome).

2. **Well-behaved residual distribution**: This assumption states that the residuals concentrate around a single, central value. Such an assumption is also violated when there is model misspecification, or if the outcome noise is heteroskedastic.

We stress that our approach does not make these assumptions. Our analysis will focus on setting where violations to the symmetry assumption might occur. Additional analysis in section 7.3 in the supplement shows the our approach is superior when the well-behavedness assumption is violated (in the presence of heteroskedasticity). In addition, section 7.5 in the supplement includes shows that in settings where the two assumptions are unlikely to be violated, BP still outperforms other kernel-based methods.

6.1. IST data

We begin with a simple illustrative example that highlights the strengths of BP vis-a-vis baselines and the properties of different utility functions in a practical setting. We aim to answer the following: (1) How do different losses reflecting different notions of utility affect the estimates? (2) How does the coupled objective make use of counterfactual data?

We study the task of a physician deciding whether or not to prescribe Heparin, an anticoagulant, to reduce the risk of Ischemic and Hemorrhagic strokes. Patients with an elevated risk of forming blood clots can reduce their risk of an Ischemic stroke by taking Heparin. However, some patients experience excessive bleeding if placed on Heparin increasing their risk of a Hemorrhagic stroke. In this setting, to make an informed decision, the physician only needs to know if the INR under treatment *roughly* falls within the healthy range of 2–3 as described in the introduction. The exact value of INR provides little additional insight.

We use data from a randomized control trial measuring the effects of Heparin (International Stroke Trial Collaborative Group, 1997). We restrict our analysis to the patients who received Heparin (treatment, $n_1 = 4530$) or no anticoagulant (control, $n_0 = 4534$). To introduce confounding, we drop 70% of the older (age > 70), untreated population. Note that the distribution of age in the trial is skewed, with a mean of 71.8 and a skewness of -0.79, which means that young patients are under-represented. Figure 4 in the supplement shows the distribution of ages for the treated and control groups in the training set. Because INR was not measured in the original data, we simulate the INR under treatment according to $\mathbb{E}[Y_i(1) \mid age_i] = S(-5, age_i') + 2.5$, where $S(a, x)$ denotes the sigmoid function with coefficient a, and age' is the age rescaled between -10, 10. This setup ensures that the majority of the population (older than 60) falls within the normal range if treated, while the few young patients younger than 60 have high INR if treated. Similarly, the outcome under control is determined by $\mathbb{E}[Y_i(0) \mid age_i] = S(-5, age_i' - 4) + 1.5$. This reflects the setting where patients older than 70 (who are under-represented in the untreated population) would have too low of an INR if not placed on Heparin. Noise for both $Y(1)$ and $Y(0)$ is drawn from a Gaussian distribution with mean 0 and variance 0.1.

We assume that the physician is restricted to linear models. In this setting the models are inherently misspecified, which means that the residuals violate the symmetry and well behaved-ness assumptions. We fit a kernel regression with a linear kernel (**KR**) for the baselines. We repeat our simulation 20 times and report averages. In each simulation, we randomly sample 3000 patients for training and validation and 3000 held out for testing. Following Chernozhukov et al. (2016), we use half the training data to estimate the nuisance parameter, that is the propensity scores, and the other half to fit the potential outcomes. For propensity scores, we fit a logistic regression. We pick the regularization parameter for the propensity score model and all the response surface models via 3-fold cross-validation as described in detail in the supplement. For all experiments, we set the required FCR to be ≤ 0.01, i.e., $\leq 1\%$.
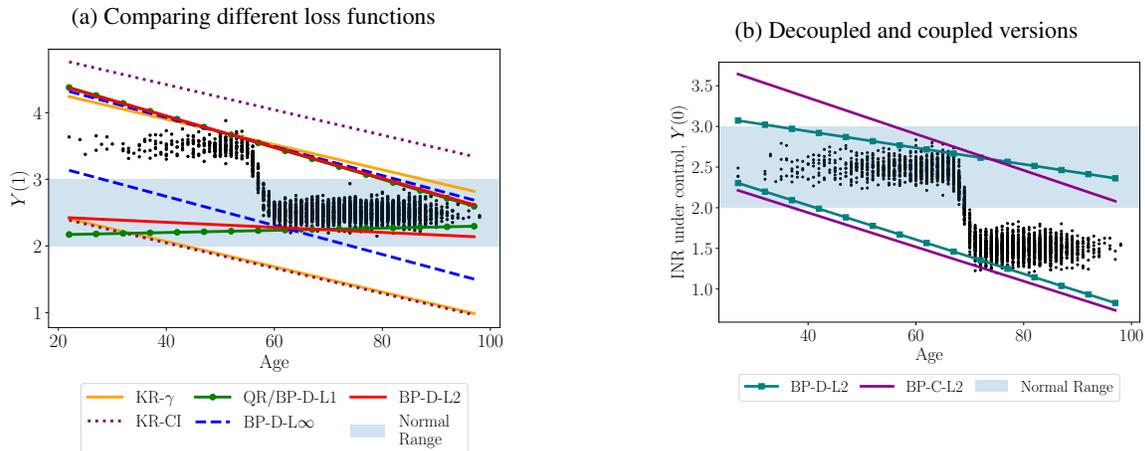
Figure 2. IST results. Plots show results from a single simulation. Black dots show potential outcomes on the test set, lines show fitted values, and shaded region shows healthy range. Plot 2a show that BP-D-L∞ is a "fair" objective, ensuring that the younger (≤ 60) population has tight intervals, sacrificing tight intervals for older population. QR (equivalent to BP-D-L1) ensures intervals are tight for older population but returns wider intervals for the younger population. BP-D-L2 gives an estimate "in-between" the two objectives, penalizing large intervals more aggressively than QR/BP-D-L1. Baselines (KR-CI/KR-γ) return bounds that are loose for both populations. Plot 2b shows that penalizing the counterfactual interval widths enables the coupled objective, BP-C-L2, to return a tighter fit for $Y(0)$ in the area where few untreated examples exist in the training data (age> 70).

Table 1 (top) shows that BP-D-L∞ achieves the smallest max IW. BP-D-L2 and QR (equivalent to BP-D-L1) achieve the smallest mean IW, with the former achieving a smaller max IW. Figure 2a explains why. BP-D-L∞ achieves the smallest max IW since it penalizes large intervals in the younger population while sacrificing by fitting a wider interval for age ≥ 60. Such an objective is most appropriate when issues of fairness might be at play, such as if a physician wants to ensure that younger patients are never given abnormally large intervals compared to the older group. QR/ BP-D-L1 achieves a tight mean IW for the older population but sacrifices for the younger population. Such an objective is appropriate when we want estimates that are as tight as possible on average, even if that entails computing wide estimates for small subpopulations. BL-D-L2 is in between the two extremes of BP-D-L∞ and BP-D-L1/QR; its mean IW is slightly higher than that of BP-D-L1 (for the younger population) and lower than that of BP-D-L∞, its max IW is lower than that of BP-D-L1 but higher than that of BP-D-L∞. This is because the L2 loss penalizes large IWs more aggressively than L1. Most notably, KR-CI and KR-γ return loose estimates compared to BP/QR. This is because KR-CI assumes symmetry of the residuals, returning overly loose upper bounds. KR-γ implicitly assumes non-fat tailedness by shifting the estimates by the same constant for all individuals. More generally, the baselines fail because they aim to first estimate the outcome as best as possible, and then estimate the intervals post-training. Ultimately, the model is picked based on what reduces the mean squared error, not what reduces over/under-estimation.

A physician who prescribes Heparin only when they are certain that a patient's INR would fall in the normal range

(i.e., both upper and lower bounds fall in the normal range) would not prescribe heparin to anyone if they rely on KR-γ, KR-CI, or BP-D-L∞ estimates. The latter has the advantage of providing tighter bounds for the younger patient group, whereas the former three also fails on that task.

Table 1 (bottom) shows that the decoupled version achieves a smaller mean and max IW compared to the coupled version, though the difference is not statistically significantly different. Figure 2b gives insight into the difference between the two versions. The coupled objective returns tighter intervals for the majority of the population, that is patients with age > 70, who are under-represented in the control group. This happens because the coupled objective has an incentive to minimize the interval width for older, untreated patients since wider counterfactual interval for the old treated patients is penalized, whereas the decoupled objective is unaware of these patients.

6.2. ACIC data

Next, we evaluate our approach in a more challenging, high-dimensional task: semi-simulated data from the Atlantic Causal Inference Conference Competition (Dorie et al., 2017). In this task, 58 variables were extracted from the Collaborative Perinatal Project, a study on pregnant women and their children. The treatment assignment and the response surfaces were simulated. We focus on the simulation with limited overlap and high heterogeneity where the treatment response surface is polynomial and the response surface is exponential. We sample 200 data points for the training/validation of the main models, and 1000 for our test set. We sample 1000 data point for training/validation of the
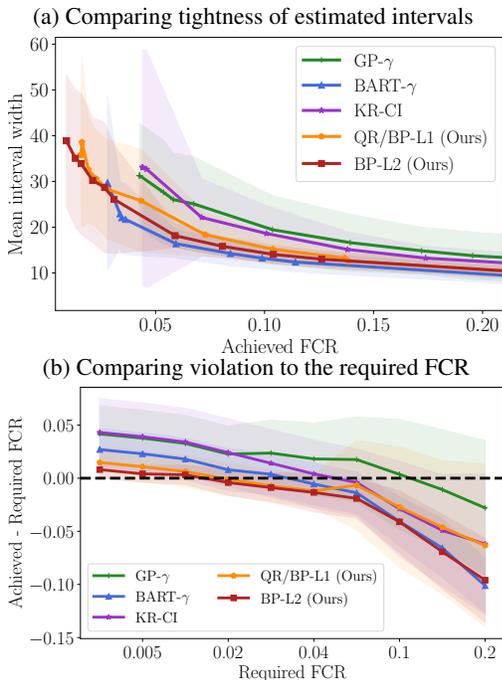
(a) Comparing tightness of estimated intervals

(b) Comparing violation to the required FCR

Figure 3. ACIC results. Plots show results averaged over 20 simulations. Plot 3a shows the mean interval width for different values of the achieved FCR on a held-out test set. Plot 3b shows the violation of the required FCR (= achieved - required) at different values of required FCR. Models above the dotted black line are in violation of the required FCR. The two plots show that BP achieves a mean interval width comparable to that of BART but at a lower violation of the required FCR. BP outperforms all kernel-based methods in terms of mean interval width and violation to the required FCR.

propensity score models. Propensity scores are estimated using 3 fold cross-validation.

To fit the potential outcomes, we use an RBF kernel for our BP/QR models. We also use an RBF kernel for the kernel regression models. We only present KR-CI, excluding KR-γ since it performs comparably to KR-CI. In addition, we include single-learners (Künzel et al., 2019) with Gaussian processes as the base-estimators (GP), and Bayesian Additive Regression Trees (BART; (Hill, 2011)). For the latter 2 models, we compute the classical confidence intervals (GP-CCI, and BART-CCI), and a variant of the γ-intervals (GP-γ, and BART-γ). Here, γ is used as a scaling rather than a shifting parameter; for an estimated outcome $\hat{y}$, and estimated standard deviation $\hat{\kappa}$, the lower/upper bounds are estimated as: $\hat{y} \pm \gamma \cdot \hat{\kappa}$, and the optimal γ is picked based on cross-validation as described previously.

We focus on getting the tightest bounds, so we only present results from BP-C-L2. We measure the performance of the models at required FCR = {0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2 }.

In this setting, since the small sample size potentially restricts the ability to fit the true functions, which may belong to a complex function class. This can be thought of as a "forced" model misspecification since the limited data does not afford us the ability to fit the true function, and limits us to simpler function classes. This is once again, a setting where we expect our models to outperform baselines that make strong assumptions about the residuals.

Figure 3a shows the mean achieved FCR on the $x-$ axis, and the mean IW on the $y-$axis for our model and baselines averaged over 20 simulations. First, we see that the mean IWs for all the models decrease as the achieved FCR increases. This confirms our theoretical findings that a trade-off between confidence that the bounds cover the potential outcomes and complexity of the function class; lower required FCRs (i.e., higher confidence that the bounds cover the true date) are associated with simpler function classes, which sacrifices accuracy, leading to higher mean IW. Second, we see that our models achieve interval widths that are tighter than all other kernel-based methods, and comparable to BART at every value of achieved FCR. However, figure 3b shows that our models achieve smaller violation compared to BART. This implies that our models are better able to exploit the trade-off between confidence and complexity.

Results from GP-CCI, and BART-CCI are excluded from the plots, and presented in the section 7.4 in the supplement since they achieve very large violations ($\approx.6$ for GP-CCI, and $\approx.2$ for BART-CCI for roughly all required FCRs). This conforms with previous studies that show that CCI methods tend to have poor coverage in finite samples (Sargent et al., 1992; Lei et al., 2018).

7. Conclusion

In this paper, we establish that the sample complexity of learning bounds on potential outcomes depends on how confident we wish to be that the bounds cover the true potential outcomes. For applications where it is sufficient to have reliable bounds on the potential outcomes to make good decisions, and the outcomes are complex functions, our findings indicate how to simplify the learning problem. Based on these findings, we introduced an algorithm that maximizes a notion of usefulness, specified by the decision maker, subject to constraints that guarantee validity of the bounds with high probability. Using semi-synthetic data, we showed that our algorithm can guide physicians in making treatment decisions for stroke patients. We also showed that our method outperforms baselines, estimating tight prediction intervals without violating a required level of false coverage rate.

Acknowledgements

References

Alaa, A. and van der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 129–138. PMLR, 10–15 Jul 2018.

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems 30*, pp. 3424–3432. Curran Associates, Inc., 2017.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Bareinboim, E. and Pearl, J. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pp. 100–108, 2012.

Cai, Z., Kuroki, M., Pearl, J., and Tian, J. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701, 2008.

Chernozhukov, V. and Hansen, C. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2016.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*, 2017.

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Imbens, G. W. and Wooldridge, J. M. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

International Stroke Trial Collaborative Group. The international stroke trial (ist): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349 (9065):1569–1581, 1997.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Kallus, N., Mao, X., and Zhou, A. Interval estimation of individual-level causal effects under unobserved confounding. *To appear in AISTATS*, 2019.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lei, L. and Candès, E. J. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects, 2017.

Pearl, J. *Causality*. Cambridge university press, 2009.

Rosenbaum, P. R. Sensitivity analysis in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.

Rubin, D. B. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469): 322–331, 2005.

Sargent, R. S., Kang, K., and Goldsman, D. An investigation of finite-sample behavior of confidence interval estimators. *Operations Research*, 40(5):898–913, 1992.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3076–3085, Sydney, Australia, 2017. PMLR.

Shawe-Taylor, J. and Williamson, R. C. Generalization performance of classifiers in terms of observed covering numbers. In *Computational Learning Theory*, pp. 274–285, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3231–3239. Curran Associates, Inc., 2015a.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823, 2015b.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. Nonparametric quantile estimation. *Journal of machine learning research*, 7(Jul):1231–1264, 2006.