Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Treatment Effect Risk: Bounds and Inference

Nathan Kallus

To cite this article:
Nathan Kallus (2023) Treatment Effect Risk: Bounds and Inference. Management Science 69(8):4579-4590. https://doi.org/10.1287/mnsc.2023.4819

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 69, No. 8, August 2023, pp. 4579–4590
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Treatment Effect Risk: Bounds and Inference

Nathan Kallus[a]

[a] Cornell Tech, Cornell University, New York, New York 10044

Contact: kallus@cornell.edu, https://orcid.org/0000-0003-1672-0507 (NK)

Received: February 1, 2022
Revised: November 20, 2022
Accepted: March 22, 2023
Published Online in Articles in Advance: June 28, 2023

https://doi.org/10.1287/mnsc.2023.4819

Copyright: © 2023 INFORMS

Abstract. Because the average treatment effect (ATE) measures the change in social welfare, even if positive, there is a risk of negative effect on, say, some 10% of the population. Assessing such risk is difficult, however, because any one individual treatment effect (ITE) is never observed, so the 10% worst-affected cannot be identified, whereas distributional treatment effects only compare the first deciles within each treatment group, which does not correspond to any 10% subpopulation. In this paper, we consider how to nonetheless assess this important risk measure, formalized as the conditional value at risk (CVaR) of the ITE distribution. We leverage the availability of pretreatment covariates and characterize the tightest possible upper and lower bounds on ITE-CVaR given by the covariate-conditional average treatment effect (CATE) function. We then proceed to study how to estimate these bounds efficiently from data and construct confidence intervals. This is challenging even in randomized experiments as it requires understanding the distribution of the unknown CATE function, which can be very complex if we use rich covariates to best control for heterogeneity. We develop a debiasing method that overcomes this and prove it enjoys favorable statistical properties even when CATE and other nuisances are estimated by black box machine learning or even inconsistently. Studying a hypothetical change to French job search counseling services, our bounds and inference demonstrate a small social benefit entails a negative impact on a substantial subpopulation.

History: Accepted by J. George Shanthikumar, data science.

Funding: This work was supported by the Division of Information and Intelligent Systems [Grant 1939704].

Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mnsc.2023.4819.

Keywords: individual treatment effect • conditional average treatment effect • conditional value at risk • partial identification • debiased machine learning

1. Introduction

Policymakers and project managers regularly conduct randomized experiments ("A/B tests") to assess potential changes to policy or product. A key metric is the *average treatment effect (ATE)*, which is the difference in the population average outcome when everyone or no one is treated. ATEs are easily estimated by differences in the sample-average outcome within treatment groups, barring interference. Estimation from observational data are also possible under appropriate assumptions, for example, unconfoundedness (Imbens and Rubin 2015). Identifying an individual's outcome with their utility, as we will throughout this paper, the ATE is the difference in social welfare in these two counterfactual scenarios. By linearity, this coincides with the population average of each *individual*'s treatment effect, the difference in their own utility in the two counterfactual scenarios.

Treatment effects, however, can vary widely between individuals (Heckman et al. 1997, Crump et al. 2008). Thus, even if the ATE is positive, there is a *risk* that a large subpopulation has a negative effect, and the purpose of this paper is to assess this risk. *Distributional* treatment effects (DTEs), which compare the two counterfactual utility distributions, *cannot* capture this risk. Indeed, Imbens and Wooldridge (2009, p. 17) note "quantile effects are defined as differences between quantiles of the two marginal potential outcome distributions, and not as [differences between] quantiles of the unit level effect." They nonetheless advocate for focusing on DTEs, because policy choice "should be governed by preferences of the policymaker over these distributions." Such rational decision-making framing, however, presumes a policymaker facing a choice between lotteries drawing at random from individual outcomes. Instead, concerned with equity beyond social welfare, we should worry about the individuals, not the policymaker.

One way to gain further insight into heterogeneity is to consider conditional ATEs (CATEs) given pretreatment covariates. For example, if we observe a discrete sensitive attribute (e.g., race), we can simply compare the CATE in each attribute-value group.[1] However, it may not always be clear what are relevant attributes and whether we are omitting important ones. Given rich and continuous covariates, we can still reliably learn the CATE function by leveraging recent advances in causal machine learning (Imai and Ratkovic 2013, Athey and Imbens 2016, Wager and Athey 2018, Künzel et al. 2019, Kennedy 2020, Nie and Wager 2021). It can still be unclear, nonetheless, whether the covariates are relevant for fairness considerations, what groups are captured in this way, and how to summarize a complex CATE function into a single measure of risk.

It is therefore appealing to focus directly on the distribution of *individual* treatment effects (ITEs). In this paper, we will consider the conditional value at risk (CVaR) of this distribution, which gives the average effect among the worst-affected 10%, 20%, etc. The challenge is that no ITE can ever be observed: the so-called fundamental problem of causal inference. Nonetheless, regardless of whether covariates are meaningful for fairness considerations, if they control for heterogeneity, CATE may predict ITE well.

In this paper, we study the *tightest-possible* upper and lower bounds on the CVaR of the ITE distribution given by CATE. We first characterize these best-possible bounds had we known the CATE distribution exactly. Then, having specified the most we can know from infinite data (i.e., distributions), we consider inference on these from actual data. One challenge is that CATE can be complex and we may wish to use flexible machine-learning estimates thereof. Another challenge is our bounds depend on quantiles of an unknown function (CATE). We design debiased estimators and confidence intervals (CIs) that overcome these challenges by being exceedingly robust to CATE learning: enabling inference even under slow estimation rates (i.e., local robustness; Chernozhukov et al. 2022), remaining consistent even under mis-estimation of some nuisances, that is, double robustness (Robins et al. 1994), and surprisingly remaining valid as bounds even when CATE is mis-estimated (i.e., double validity; Dorn et al. 2021). We conclude by using our tools to illustrate treatment effect risk in a case study of job search assistance benefits.

2. Problem Setup and Definitions

Each individual in the population is associated with two potential outcomes, $Y^*(0), Y^*(1) \in \mathbb{R}$, corresponding to individual utility under "treat all" and "treat none," respectively, and baseline covariates (observable characteristics), $X \in \mathcal{X}$. The ITE, ATE, and CATE are, respectively,

$$\delta = Y^*(1) - Y^*(0), \quad \overline{\tau} = \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] = \mathbb{E}\delta = \mathbb{E}\tau(X),$$
$$\tau(X) = \mathbb{E}[\delta \mid X] = \mu(X,1) - \mu(X,0),$$
$$\text{where } \mu(X,a) = \mathbb{E}[Y^*(a) \mid X].$$

We assume $\mathbb{E}\delta^2 < \infty$ throughout.

Of interest is the average effect among the $(100 \times \alpha)$%-worst affected, formalized by $\mathrm{CVaR}_\alpha(\delta)$, where for any variable Z, we define its cumulative distribution function (CDF), α-quantile, and α-CVaR,[2] respectively, as

$$F_Z(z) = \mathbb{P}(Z \le z), \tag{1}$$
$$F_Z^{-1}(\alpha) = \inf\{z : F_Z(z) \ge \alpha\}, \tag{2}$$
$$\begin{aligned}\mathrm{CVaR}_\alpha(Z) = \mathbb{E}[Z \mid Z \le F_Z^{-1}(\alpha)] - (\alpha^{-1}F_Z(F_Z^{-1}(\alpha)) - 1)\\ \times (F_Z^{-1}(\alpha) - E[Z \mid Z \le F_Z^{-1}(\alpha)]).\end{aligned} \tag{3}$$

This formally defines the expectation among the $(100 \times \alpha)$% smallest values in the population described by Z. It is the average at/below the α-quantile when there is exactly α-fraction at/below the α-quantile ($F_Z(F_Z^{-1}(\alpha)) = \alpha$, such as would occur if Z were continuous). Otherwise, we must remove a fraction of the atom at the α-quantile so as to make an exactly α-sized subpopulation to average over (i.e., the second term in Equation (3)). Also, $\mathrm{CVaR}_1(Z) = \mathbb{E}Z$.

Rockafellar and Uryasev (2000) give an optimization reformulation of CVaR: letting $(u)_- = u \wedge 0$,

$$\begin{aligned}\mathrm{CVaR}_\alpha(Z) &= \sup_\beta(\beta + \alpha^{-1}\mathbb{E}(Z - \beta)_-)\\ &= F_Z^{-1}(\alpha) + \alpha^{-1}\mathbb{E}(Z - F_Z^{-1}(\alpha))_-.\end{aligned} \tag{4}$$

We consider data from a randomized experiment or observational study. Each individual is associated with a treatment $A \in \{0,1\}$, and we observe the *factual* outcome $Y = Y^*(A)$ (never $Y^*(1-A)$). The data are $(X_i, A_i, Y_i) \sim (X, A, Y)$, $1 \le i \le n$. We assume unconfoundedness throughout: $Y^*(a) \perp\!\!\!\perp A \mid X$.[3] Randomized experiments (our focus) ensure this by design (often with $X \perp\!\!\!\perp A$). Our results nonetheless extend to observational settings assuming unconfoundedness. Under unconfoundedness, ATE and CATE are identifiable, i.e., are functions of the (X, A, Y) distribution: $\mu(X,a) = \mathbb{E}[Y \mid X, A = a]$, $\tau(X) = \mu(X,1) - \mu(X,0)$, $\overline{\tau} = \mathbb{E}\tau(X)$ $(= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$ if $X \perp\!\!\!\perp A)$. Define also the propensity score

$$e(X) = \mathbb{P}(A = 1 \mid X).$$

We now illustrate treatment effect risk and its *un*identifiability, which motivates us to consider the tightest-possible *identifiable* bounds (Section 3) and inference thereon (Section 4).

Example 1 (Simple Example). Consider two hypotheses:

$$H_1 : \begin{pmatrix} Y^*(0) \\ Y^*(1) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\right),$$
$$A \sim \text{Bernoulli}(1/2),$$
$$H_2 : \begin{pmatrix} Y^*(0) \\ Y^*(1) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right),$$
$$A \sim \text{Bernoulli}(1/2).$$

Under both H_1 and H_2, we have $(A, Y) \sim \text{Bernoulli}(1/2) \times \mathcal{N}(0,1)$. However, under H_1, $\delta \sim \mathcal{N}(0,2)$ and

$\mathrm{CVaR}_{0.1}(\delta)=-3.5$, whereas under H_2, $\delta=0$ and $\mathrm{CVaR}_{0.1}(\delta)=0$. Therefore, $\mathrm{CVaR}_{0.1}(\delta)$ cannot be identified from (A, Y).

Remark 1 (Covariate-Conditional Policies). Treat (i.e., roll-out to) all or none is often the choice faced by project managers, but given covariates, we can learn covariate-conditional treatment policies (Qian and Murphy 2011, Zhao et al. 2012, Kallus 2018, Kitagawa and Tetenov 2018, Athey and Wager 2021, Kallus and Zhou 2021). Learning aside, treating only when $\tau(X)>0$ ensures all covariate-defined groups have nonnegative group-average effects.[4] The sum of these nonnegative effects can also be estimated using flexible nonparametric methods (Luedtke and van der Laan 2016a, b, 2017). Personalizing on all available covariates is, however, generally infeasible because of operational, nonstationarity, and/or ethical/reputational concerns. Nonetheless, given any policy $\pi:\mathcal{X}\to\{0,1\}$, we may simply redefine ITE as $Y(\pi(X))-Y(0)$, and our results still apply. This is especially relevant when π personalizes on some covariates and the rest explain heterogeneity conditionally thereon.

Remark 2 (Risk of Observed vs. Unobserved Variables). CVaR is an example of coherent risk measures (Artzner et al. 1999), which are used to assess distributions beyond expectations and are equivalent to distributionally robust worst-case expectations (Ruszczyński and Shapiro 2006). For example, CVaR is the worst-case expectation among distributions with Radon-Nikodym derivative to the given distribution bounded by $1/\alpha$. Other distributional divergences can also define ambiguity sets (Ben-Tal et al. 2013, Bertsimas et al. 2018, Esfahani and Kuhn 2018). Alternative approaches limit the *complexity* of subpopulations (Kearns et al. 2018, Lahoti et al. 2020). In finance (Krokhmal et al. 2002), distributionally robust supervised learning (Bagnell 2005), demographics-free fair learning (Lahoti et al. 2020), and CVaR-DTEs (Kallus et al. 2019), the variable whose risk is of interest is *always observed*. For example, model loss on each training example is observed. In contrast, we consider risk of an *unobserved variable*; hence, we study bounds in Section 3. For inference, we are uniquely concerned with risk of an *unknown function*; hence, we develop learning robust methods in Section 4.

3. Bounds

3.1. Upper Bound: CATE-CVaR

An upper bound on $\mathrm{CVaR}_\alpha(\delta)$ is crucial: If negative or substantially below ATE, the change poses certifiable risk or inequity to an $(100\times\alpha)\%$ subpopulation.

Theorem 1 (Upper Bound by CATE-CVaR). *We have*

$$\mathrm{CVaR}_\alpha(\delta)\le \mathrm{CVaR}_\alpha(\tau(X)). \tag{5}$$

Moreover, for any X distribution and integrable $\tau:\mathcal{X}\to\mathbb{R}$, there exists a (X,δ)-distribution with the given X distribution, $\tau(X)=\mathbb{E}[\delta\mid X]$, and Equation (5) holding with equality.

Because $\tau(X)$ represents our *best guess* for δ (in squared error), imputing the unknown δ with $\tau(X)$ seems reasonable. Theorem 1 shows this in fact provides an upper bound.[5] If $\tau(X)$ is continuous, $\mathrm{CVaR}_\alpha(\tau(X))=\mathbb{E}[\delta\mid\tau(X)\le F^{-1}_{\tau(X)}(\alpha)]$, and Equation (5) is intuitive: $\mathrm{CVaR}_\alpha(\delta)$ is worst average effect among *all* $(100\times\alpha)\%$-subpopulations, whereas $\mathrm{CVaR}_\alpha(\tau(X))$ only among X-defined subpopulations. This bound is also tight: given just $\tau(X)$, it cannot be improved.[6] Although it is tight, the bound may be *practically uninformative*, depending on how predictive X is and how the bound is used. Without covariates, the bound is trivial: $\mathrm{CVaR}_\alpha(\delta)\le\overline{\tau}$. Even if X is not very predictive but just enough to demonstrate $\mathrm{CVaR}_\alpha(\delta)\le\mathrm{CVaR}_\alpha(\tau(\mathrm{X}))<0$, the bound may still be very practically informative.

Remark 3 (CVaR as Summary of CATE). As discussed in Section 1, if we have protected groups (that we observe), we may consider CATE along that grouping. Otherwise, the significance of a CATE-learning prediction of $\tau(x)$ for some group $X=x$ is unclear. Theorem 1 shows that, regardless of substantive meanings of included covariates X, aggregate statistics of the $\tau(X)$ distribution provide insight into treatment effect risk, giving important meaning to the outputs of CATE learning with rich covariates. Nonetheless, it is insufficient to rely solely on CATE learning for estimation: As we explore in Section 4, averaging the 10% lowest predictions from a CATE learner suffers from both optimizer's curse and statistical instability, and we must develop special inferential procedures to target Equation (5).

Aside from being a bound, $\mathrm{CVaR}_\alpha(\tau(X))$ is also of interest as a summary of effect heterogeneity along meaningful covariates X. When X is more than a few discrete groups, understanding the many facets of CATE is challenging, both interpretationally and statistically. We could test for X heterogeneity (Gail and Simon 1985, Sawilowsky 1990, Davison 1992, Crump et al. 2008),[7] for example, omnibus test $H_0: 0\in\arg\min_\gamma \mathbb{E}(\tau(X)-\overline{\tau}-\gamma^\top(X-\mathbb{E}X))^2$ (Chernozhukov et al. 2018a). This, however, may detect minor heterogeneity in small subpopulations, may not assess magnitude or direction, and may be inappropriate if we expect heterogeneity. In contrast, $\mathrm{CVaR}_\alpha(\tau(X))$ is a simple, meaningful summary of $\tau(X)$. Inference, however, is a challenge. We tackle this in Section 4.

Remark 4 (Interquantile Averages of CATE). CVaR of CATE can in fact permit us to summarize average effects in the middle and not just the tails. Consider any $0<\alpha<\alpha'<1$. Provided that $F_{\tau(X)}(F^{-1}_{\tau(X)}(\alpha))=\alpha$, $F_{\tau(X)}(F^{-1}_{\tau(X)}(\alpha'))=\alpha'$ (e.g., $\tau(X)$ is continuous), we have that

$$\mathbb{E}[Y^*(1)-Y^*(0)\mid F^{-1}_{\tau(X)}(\alpha)<\tau(X)\le F^{-1}_{\tau(X)}(\alpha')] = \frac{\alpha'\mathrm{CVaR}_{\alpha'}(\tau(X))-\alpha\mathrm{CVaR}_\alpha(\tau(X))}{\alpha'-\alpha}. \tag{6}$$

Equation (6) is the average effect among individuals with CATE between the α- and α'-quantiles. A similar but different quantity is considered in Chernozhukov et al. (2018a): The average effect among individuals in interquantile ranges of an *estimate* of CATE fit on a split sample rather than the true CATE. They consider averaging this over splits, but that average still needs not correspond to Equation (6), and this approach is *not* robust to errors in the CATE estimate, meaning these errors will propagate to nonnegligible terms in the estimate and its variance. In contrast, by leveraging the unique optimization structure of CVaR, in Section 4, we provide an estimator that *is* robust to such errors, allowing us to estimate the CVaR of the true CATE rather than a split-sample-estimated CATE. By writing Equation (6) using CVaR, we can then leverage these results to get robust estimates for interquantile averages, as we will explain in Remark 6.

Remark 5 (Who Is Negatively Affected?). Suppose we find $\mathrm{CVaR}_\alpha(\tau(X)) < 0$, whereas $\overline{\tau} > 0$, where α is "substantial": The social-welfare benefit of the proposal is borne by some substantial negatively impacted subpopulation. Although that may already cool enthusiasm for the proposal, we may wonder *who* are the harmed individuals, for example, to help design a new, better treatment.

Assuming continuity, $\mathrm{CVaR}_\alpha(\tau(X))$ is the ATE among individuals with $\tau(X) \le F^{-1}_{\tau(X)}(\alpha)$: an *identifiable* group. A question is interpretation. This is easy if $\tau(X)$ is linear or tree (or estimated using such models, which still gives a bound per Theorem 7). We can also consider summaries of this group, for example, fraction belonging to sensitive groups, or learn simpler models to explain membership (Ribeiro et al. 2016, Lakkaraju et al. 2019). Alternatively, given we detect substantial inequities, we can *separately* investigate which variables negatively modulate treatment effect by, for example, studying $\arg\min_\gamma \mathbb{E}(\tau(X) - \overline{\tau} - \gamma^\top X)^2$ (Chernozhukov et al. 2018a, Kennedy 2020).

3.2. Lower Bounds Under Limited Residual Heterogeneity Range

Much as we try to best control for heterogeneity, disparate effect-predictiveness of covariates may mean some negative ITEs are averaged out and hidden while others are singled out. A remedy when concerned about disproportionate predictiveness among sensitive groups (e.g., race) would be to include these (or proxies) within X. However, we may always worry about missing something. A lower bound can provide assurances about what the upper bound may be missing.

This depends on how much residual heterogeneity remains. Our first set of lower bounds limit the range of residual heterogeneity, that is, almost-sure bounds on $\delta - \tau(X)$, whereas our second set of lower bounds limit its variance, that is, bounds on $\mathrm{Var}(\delta \mid X) = \mathbb{E}(\delta - \tau(X))^2$.

Theorem 2. *Suppose* $|\tau(X) - \delta| \le b$. *Then*

$$\mathrm{CVaR}_\alpha(\delta) \ge \sup_\beta \left(\beta + \frac{1}{2\alpha}\mathbb{E}[(\tau(X) - b - \beta)_-] + \frac{1}{2\alpha}\mathbb{E}[(\tau(X) + b - \beta)_-] \right). \tag{7}$$

Moreover, for any X distribution and integrable $\tau : \mathcal{X} \to \mathbb{R}$, there exists a (X, δ)-distribution with the given X distribution, $\tau(X) = \mathbb{E}[\delta \mid X]$, $|\tau(X) - \delta| \le b$, and Equation (7) holding with equality.

The right-hand side of Equation (7) is the α-CVaR of the equal mixture distribution of $\tau(X) - b$ and $\tau(X) + b$. It reduces to $\mathrm{CVaR}_\alpha(\tau(X))$ when $b = 0$ (equivalent to $\delta = \tau(X)$). When $\alpha = 1$, it becomes $\overline{\tau}$ for any $b \ge 0$ (as necessary for tightness). The lower bound is established via weak semi-infinite duality and its tightness by exhibiting the equal-mixture distribution.

Because $(\tau(X) \pm b - \beta)_- \ge (\tau(X) - \beta)_- - b$, Equation (7) upper bounds $\mathrm{CVaR}_\alpha(\tau(X)) - b$. This simpler bound is tight if we only assume a one-sided-bounded range.

Theorem 3. *Suppose* $\tau(X) - \delta \le b$. *Then*

$$\mathrm{CVaR}_\alpha(\delta) \ge \mathrm{CVaR}_\alpha(\tau(X)) - b. \tag{8}$$

Moreover, for $\alpha < 1$, given any $\varepsilon > 0$, X distribution, and integrable $\tau : \mathcal{X} \to \mathbb{R}$, some (X, δ) distribution has the given X marginal, $\tau(X) = \mathbb{E}[\delta \mid X]$, $\tau(X) - \delta \le b$, and Equation (8) holding with equality up to ε error.

The lower bound is immediate, and its tightness is given by exhibiting a skewed two-point-mass distribution. For $\alpha = 1$, Equation (8) simply reads $\overline{\tau} \ge \overline{\tau} - b$, but for *any* $\alpha < 1$, Equation (8) is actually *tight*.

3.3. Lower Bounds Under Limited Residual Heterogeneity Variance

Limiting residual heterogeneity within a range may be implausible, or plausible only with large constants, yielding a weak bound. We next explore the implication of the residual ITE variance after controlling for X, which we can bound given observables.

Theorem 4. *Suppose* $\mathrm{Var}(\delta \mid X) \le \overline{\sigma}^2(X)$ *for some integrable* $\overline{\sigma}^2 : \mathcal{X} \to \mathbb{R}_+$. *Then,*

$$\mathrm{CVaR}_\alpha(\delta) \ge \sup_\beta \left(\beta + \frac{1}{2\alpha}\mathbb{E}\left[\tau(X) - \beta - \sqrt{(\tau(X) - \beta)^2 + \overline{\sigma}^2(X)} \right] \right). \tag{9}$$

Moreover, for any $\varepsilon > 0$, X distribution, and integrable $\tau : \mathcal{X} \to \mathbb{R}$, there exists a (X, δ) distribution with the given X distribution, $\tau(X) = \mathbb{E}[\delta \mid X]$, $\mathrm{Var}(\delta \mid X) \le \overline{\sigma}^2(X)$, and Equation (9) holding with equality up to ε error.

The proof of Theorem 4 leverages strong duality for convex semi-infinite optimization. Equation (9) equals

$\mathrm{CVaR}_\alpha(\tau(X))$ whenever $\overline{\sigma}^2(X)=0$ and $\overline{\tau}$ whenever $\alpha=1$. Because $|\delta-\tau(X)|\le b\Rightarrow \mathrm{Var}(\delta|X)\le b^2$, plugging $\overline{\sigma}^2(X)=b^2$ into Equation (9) must be looser than Equation (7) by tightness. Triangle inequality verifies this directly: $\sum_\pm(\tau(X)\pm b-\beta)_-=\tau(X)-\beta-\frac{1}{2}\sum_\pm|\tau(X)\pm b-\beta|\ge\tau(X)-\beta-\sqrt{(\tau(X)-\beta)^2+b^2}$.

A residual variance bound is both more plausible and easier to calibrate than an absolute bound. Letting $\rho(X)=\mathrm{Corr}(Y(0),Y(1)|X)\in[-1,1]$, we have

$$\mathrm{Var}(\delta|X)=\mathrm{Var}(Y|X,A=0)+\mathrm{Var}(Y|X,A=1)\\ -2\rho(X)\mathrm{Var}^{1/2}(Y|X,A=0)\mathrm{Var}^{1/2}(Y|X,A=1), \tag{10}$$

where all terms but $\rho(X)$ are identifiable. Thus, postulating different potential outcome correlations, we obtain different bounds. Equation (10) is maximized for $\rho(X)=-1$, which is tight, as all correlations are realizable. Thus, plugging $\overline{\sigma}^2(X)=(\mathrm{Var}^{1/2}(Y|X,A=0)+\mathrm{Var}^{1/2}(Y|X,A=1))^2$ into Equation (9) yields a tight lower bound on ITE-CVaR, given conditional expectations and variances. We may obtain better bounds if we postulate larger $\rho(X)$.

Theorem 4 also implies a simpler but looser bound.

Corollary 1. *We have*

$$0\le\mathrm{CVaR}_\alpha(\tau(X))-\mathrm{CVaR}_\alpha(\delta)\le\frac{1}{2\alpha}\mathbb{E}[\mathrm{Var}^{1/2}(\delta|X)] \tag{11}$$

$$\le\frac{1}{2\alpha}\mathbb{E}[\mathrm{Var}^{1/2}(Y|X,A=0)+\mathrm{Var}^{1/2}(Y|X,A=1)] \tag{12}$$

$$\le\frac{1}{2\alpha}\sqrt{\mathbb{E}[(Y-\mu(X,A))^2|A=0]}\\ +\frac{1}{2\alpha}\sqrt{\mathbb{E}[(Y-\mu(X,A))^2|A=1]}. \tag{13}$$

Equation (11) more transparently bounds the slack in Equation (5) in terms of residual effect variance. However, it is not tight, as can be seen for $\alpha=1$. Equation (12) is looser but appealing as it is identifiable. Equation (13) is even looser but depends only on the root-mean-squared error of regressing Y on X for each $A\in\{0,1\}$ (i.e., the numerator of nonparametric R^2).

4. Inference

We next turn to estimating the bounds developed in Section 3 and constructing CIs. Recall our data, $(X_i,A_i,\ Y_i)\sim(X,A,Y)$, $1\le i\le n$, may be experimental or observational. The only relevant technical difference between these two cases is whether propensity, $e(X)=\mathbb{P}(A=1|\ X)$, is known or not. Although it does not matter here, $e(X)$ is usually constant in experiments ($A\perp\!\!\!\perp X$). In observational settings, $e(X)$ may be estimated.

We focus here on inference on CATE-CVaR. We provide analogous procedures for the lower bounds of Theorems 2–4 and Corollary 1 in Online Appendix A. Fix α. Our inferential target is

$$\Psi=\mathrm{CVaR}_\alpha(\tau(X))=\beta^*+\frac{1}{\alpha}\mathbb{E}(\tau(X)-\beta^*)_-,\quad \text{where } \beta^*\\ =F^{-1}_{\tau(X)}(\alpha)=\inf\{\beta:\mathbb{P}\big(\tau(X)\le\beta\big)\ge\alpha\}.$$

Because $\tau(X)$ is not directly observed, the first step is fitting it. Fortunately, recent advances in causal machine learning provide excellent tools for this (Imai and Ratkovic 2013, Athey and Imbens 2016, Wager and Athey 2018, Künzel et al. 2019, Kennedy 2020, Nie and Wager 2021). Given an estimate $\hat{\tau}$, we might consider a plug-in approach: $\hat{\Psi}^{\text{plug-in}}=\sup_\beta\big(\beta+\frac{1}{n\alpha}\sum_{i=1}^n(\hat{\tau}(X_i)-\beta)_-\big)$. There are two challenges with this estimator. One is that the statistical behavior of $\hat{\Psi}^{\text{plug-in}}$ depends heavily on that of $\hat{\tau}$: If $\hat{\tau}$ converges slowly and/or has nonnegligible bias, as occurs when fit by flexible machine-learning methods, both estimation rates and valid inference may be imperiled for $\hat{\Psi}^{\text{plug-in}}$. Another is that it can be severely downward biased: it essentially averages the $(100\times\alpha)$% smallest CATE predictions, thus systematically picking out those with the most negative errors (optimizer's curse).

Instead, we develop a debiasing approach that is *insensitive* to CATE estimation, accommodating both misspecified parametric models and flexible-but-imprecise machine-learning CATE estimators. The main challenge is estimating β^*, which cannot be expressed by an estimating equation in $X,Y(0),Y(1)$, so its efficient/orthogonal estimation is unclear, unlike the case of quantile/CVaR treatment effects (Firpo 2007, Belloni et al. 2017, Kallus et al. 2019). Fortunately, we care only about Ψ, not β^*, and special optimization structure in Ψ gives robustness to perturbations. so even rough estimates suffice. We therefore treat both τ and β^* as nuisance parameters, together with e,μ, and ensure simultaneous orthogonality to all four nuisances.

Algorithm 1 (Point Estimate and CI for $\mathrm{CVaR}_\alpha(\tau(X))$)

```
Input: Level $\alpha\in(0,1)$, data $\{(X_i,A_i,Y_i):i=1,\dots,n\}$, number of folds $K$, and $e,\mu,\tau$-estimators
for $k=1,\dots,K$ do estimate $\hat{e}^{(k)},\hat{\mu}^{(k)},\hat{\tau}^{(k)}$ using data $\{(X_i,A_i,Y_i):i\not\equiv k-1\ (\mathrm{mod}\ K)\}$
Set $\hat{\beta}=\frac{1}{K}\sum_{k=1}^K\inf\{\beta:\sum_{i\equiv k-1\ (\mathrm{mod\ K})}(\mathbb{I}[\hat{\tau}^{(k)}(X_i)\le\beta]-\alpha)\ge0\}$
for $i=1,\dots,n$ do set $\phi_i=\phi(X_i,A_i,Y_i;\hat{e}^{(i\,\mathrm{mod}\,K)},\hat{\mu}^{(i\,\mathrm{mod}\,K)},\hat{\tau}^{(i\,\mathrm{mod}\,K)},\hat{\beta})$
Set $\hat{\Psi}=\frac{1}{n}\sum_{i=1}^n\phi_i$, $\hat{\mathrm{se}}=\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^n(\phi_i-\hat{\Psi})^2}$
Return $\hat{\Psi}$ as point estimate and $[\hat{\Psi}\pm\Phi^{-1}((1+\gamma)/2)\hat{\mathrm{se}}]$ as $\gamma$ CIs
```

Algorithm 1 summarizes our procedure, where for Line 3 define

$$\phi(X,A,Y;\check{e},\check{\mu},\check{\tau},\check{\beta}) = \check{\beta} + \frac{1}{\alpha}\mathbb{I}[\check{\tau}(X) \le \check{\beta}]\Bigg(\check{\mu}(X,1) - \check{\mu}(X,0) + \frac{A-\check{e}(X)}{\check{e}(X)(1-\check{e}(X))}(Y-\check{\mu}(X,A)) - \check{\beta}\Bigg). \tag{14}$$

Note that $\Psi = \mathbb{E}\phi(X,A,Y,e,\mu,\tau,\beta^*)$, which Algorithm 1 essentially approximates by using estimates for the unknown e,μ,τ,β^* and replacing the mean over (X, A, Y) by a sample average over the data (X_i, A_i, Y_i). In particular, we use a "cross-fitting" scheme so that nuisance estimates are independent of the sample being averaged (Schick 1986, Zheng and van der Laan 2011, Chernozhukov et al. 2018b).[8]

Our specific construction in Equation (14) is crucial for the success of Algorithm 1. The plug-in approach is essentially the same approach applied instead to $\tilde{\phi}(X, A, Y;\check{\tau},\check{\beta}) = \check{\beta} + \frac{1}{\alpha}(\check{\tau}(X) - \check{\beta})_-$. The problem with it arises from the fact that $\mathbb{E}\tilde{\phi}(X,A,Y;\check{\tau},\check{\beta})$ is sensitive to changes to $\check{\tau},\check{\beta}$ near τ,β^* (nonzero derivative) so that small errors propagate. In contrast, we show $\mathbb{E}\phi(X,A,Y;\check{e},\check{\mu},\check{\tau},\check{\beta})$ has *zero* derivative in $\check{e},\check{\mu},\check{\tau},\check{\beta}$ at e,μ,τ,β^* (Lemma EC.1, which also bounds the second derivative and second moments), a condition known as Neyman orthogonality that ensures that small errors are insignificant (Chernozhukov et al. 2018b). This leads to local robustness (Theorem 5) and double robustness (Theorem 6) guarantees. The orthogonality in e,μ is the same as in ATE estimation (Robins et al. 1994, Chernozhukov et al. 2022) because, for every $\check{\tau},\check{\beta}$, $\mathbb{E}\phi(X,A,Y;e,\mu,\check{\tau},\check{\beta})$ is just a subgroup ATE (rescaled and recentered). The orthogonality in τ,β^* is a consequence of a saddle-point formulation

$$\Psi = \sup_{\check{\beta}\in\mathbb{R}}\inf_{\check{\tau}:\mathcal{X}\to\mathbb{R}}\mathbb{E}\phi(X,A,Y;e,\mu,\check{\tau},\check{\beta}) = \inf_{\check{\tau}:\mathcal{X}\to\mathbb{R}}\sup_{\check{\beta}\in\mathbb{R}}\mathbb{E}\phi(X,A,Y;e,\mu,\check{\tau},\check{\beta})$$

and first-order optimality conditions. One complication is making differentiability in $\check{\tau}$ at τ formal, which we tackle using Assumption 2. This saddle-point formulation also shows that if we get τ wrong we still obtain an upper bound on Ψ (Lemma 2), yielding a double validity guarantee (Theorem 7).

As we discuss in detail in Section 4.3, we treat τ as a separate nuisance from μ even though $\tau(x) = \mu(x,1) - \mu(x,0)$. This enables the use of specialized CATE learners. We also treat β^* as a separate nuisance (not as a parameter as in Kallus et al. 2019) and fit it as the quantile of $\hat{\tau}(X)$. As simple regressions, e and μ can be fit by parametric regression or standard machine-learning methods such as random forests, gradient boosting, neural networks, and so on.

Remark 6 (Comparing Different Levels and Interquantile Averages). To assess disparities, we may want to compare $\mathrm{CVaR}_\alpha(\tau(X))$ to ATE (equivalently, $\mathrm{CVaR}_1(\tau(X))$). To get good CIs on $\mathrm{CVaR}_{\alpha'}(\tau(X)) - \mathrm{CVaR}_\alpha(\tau(X))$, we can replace ϕ_i in Line 3 of Algorithm 1 with the difference of ϕ_is for α' and α (using the same nuisances except $\hat{\beta}^{(k)}$). Setting $\alpha' = 1$, this will, in particular, correctly yield smaller CIs on $\overline{\tau} - \mathrm{CVaR}_\alpha(\tau(X))$ for α near one. Similarly, if we want CIs on interquantile average effects as in Remark 4, then per Equation (6), we may simply replace ϕ_i in Line 3 of Algorithm 1 with the difference of ϕ_is for α' and α, weighted by $\frac{\alpha'}{\alpha'-\alpha}$ and $\frac{\alpha}{\alpha'-\alpha}$, respectively. We may also consider covariances of ϕ_i corresponding to many α levels for constructing simultaneous intervals.

Remark 7 (Partial-Identification Intervals). Although Algorithm 1 focuses on CATE-CVaR, which upper bounds ITE-CVaR, in Online Appendix A, we provide inference procedures for lower bounds on ITE-CVaR. These can be combined to construct intervals containing ITE-CVaR with probability γ. By union bound, we can simply combine the one-sided $(1+\gamma)/2$ CIs for the lower and upper bounds. However, coverage may be conservative $(>\gamma)$ for the partial identification interval given by the bounds. For calibrated γ-coverage (asymptotically), we must account for correlation between lower- and upper-bound estimates, given by the correlation between ϕ_is for each procedure. Then, we can construct calibrated intervals following appendix A.4 of Kallus et al. (2022).

Remark 8 (Monotonicity). Although $\mathrm{CVaR}_\alpha(\tau(X))$ is monotone in α, Algorithm 1's output for different α may not be due to estimation errors. We can postprocess to ensure monotonicity using rearrangement (Hardy et al. 1952), which only improves estimation and does not affect inference (Chernozhukov et al. 2010). We use this in Section 5.

4.1. Local Robustness and CIs

We now establish favorable guarantees for Algorithm 1. First, we show it is insensitive to slow but consistent estimation of nuisances, having first-order behavior as if we used true values.

We will need some minimal regularity.

Assumption 1 (Boundedness). *Suppose $\overline{e} \le e \le 1-\overline{e}$ and $|Y| \le B$ for positive constants $\overline{e}, B > 0$.*

The first condition of Assumption 1 ensures that the X distributions of experimental groups *overlap*. It is usually guaranteed in randomized experiments by setting $e(X)$ constant $(A \perp\!\!\!\perp X)$. In unconfounded observational studies, it is a standard assumption. The second condition requires bounded outcomes and is largely technical to make analysis tractable.

Assumption 2 (Margin). *Suppose $F_{\tau(X)}$ is continuously differentiable at $F^{-1}_{\tau(X)}(\alpha)$ with a positive derivative.*

Assumption 2 prohibits degeneracy of the quantile and essentially ensures two things at once: limited sensitivity to errors in the quantile and the learnability of the quantile itself.

On the one hand, Assumption 2 implies that the probability of $\tau(X)$ being within δ of its α-quantile scales at most linearly with δ, known as a margin condition with exponent 1 (Mammen and Tsybakov 1999, Audibert and Tsybakov 2007). Margin conditions have been used in causal inference to ensure a smooth dependence on $\tau(X)$ and robustness to errors therein (equation (16) in Luedtke and van der Laan 2016b; assumption 2.2 in Kitagawa and Tetenov 2018; and assumption 6 in Kennedy et al. 2020). Assumption 2 also implies the α-quantile is unique, ensuring regularity; nonuniqueness may require additional stabilization as in Luedtke and van der Laan (2016b).

Assumption 2 simultaneously implies that the probability near the α-quantile scales at *least* linearly, which ensures rates of estimation for the quantile. The same is needed for asymptotic normality of sample quantiles of *observed* variables (corollary 21.5 in Van der Vaart 1998). Compared with standard analysis of quantile estimation, an added complexity here is that we only have an estimate of the variable, $\tau(X)$, whose quantile we wish to estimate. We next deal with this complexity, showing how rates for $\hat{\tau}^{(k)}$ translate to rates for $\hat{\beta}$ under Assumption 2.

Lemma 1. *Suppose Assumption 2 holds. Then, $\hat{\beta}$ in Line 2 of Algorithm 1 satisfies*

$$|\hat{\beta}-\beta^*| = O_p\left(\|\hat{\tau}^{(k)}-\tau\|_q^{\frac{q}{q+1}} + n^{-1/2}\right) \quad \forall q \in [1,\infty],$$

where throughout we interpret $\frac{aq+b}{cq+d} = \frac{a}{c}$ for $q=\infty$.

We now show that $\hat{\Psi}$ enjoys *local robustness* (Chernozhukov et al. 2022 and references therein). Aside from a now-standard analysis of cross-fitting (Chernozhukov et al. 2018b), the crucial steps are characterizing the first and second functional derivatives of our special ϕ construction (Lemma EC.1) and leveraging Lemma 1. A slight deviation from the usual cross-fitting analysis is carefully handling the fact that all but one nuisance ($\hat{\beta}$) are cross-fit.

Theorem 5. *Suppose Assumptions 1 and 2 hold and that for $k=1,\ldots,K$, $\|\hat{e}^{(k)}-e\|_2 = o_p(1)$, $\|\hat{\mu}^{(k)}-\mu\|_2 = o_p(1)$, $\|\hat{e}^{(k)}-e\|_2\|\hat{\mu}^{(k)}-\mu\|_2 = o_p\left(n^{-\frac{1}{2}}\right)$, $\|\hat{\tau}^{(k)}-\tau\|_q = o_p\left(n^{-\frac{q+1}{4q}}\right)$, $\mathbb{P}(\|\hat{\mu}^{(k)}\|_\infty \le B) \to 1$, and $\mathbb{P}(\overline{e} \le \hat{e}^{(k)} \le 1-\overline{e}) \to 1$. Then $\hat{\Psi}$, $\hat{\mathrm{se}}$ in Line 4 of Algorithm 1 satisfy*

$$\hat{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\phi(X,A,Y;e,\mu,\tau,\beta^*) + o_p(n^{-1/2}) = \Psi + O_p(n^{-1/2}),$$

$$\mathbb{P}(\Psi \in [\hat{\Psi} \pm \Phi^{-1}((1+\gamma)/2)\hat{\mathrm{se}}]) \to \gamma \quad \forall\gamma.$$

The rate assumptions on $\hat{e}^{(k)}$ and $\hat{\mu}^{(k)}$ are lax: It suffices to have $o_p(n^{-1/4})$ rates on both or to have no rate on $\hat{\mu}^{(k)}$ at all if e is known. This parallels standard conditions in double machine-learning ATE estimation, achievable by a variety of machine-learning methods (Chernozhukov et al. 2018b). We explore the condition on $\hat{\tau}^{(k)}$ in Section 4.3.

4.2. Double Robustness and Double Validity

Theorem 5 guarantees good performance if all nuisances are estimated slowly but still consistently. However, even if nuisances are inconsistent, we perform well.

First, we establish a property mirroring doubly robust ATE estimation (Robins et al. 1994): Even if e or μ is inconsistent, we remain consistent, provided τ is consistently estimated, albeit slowly.

Theorem 6 (Double Robustness). *Fix any $\tilde{e},\tilde{\mu}$ with $\overline{e} \le \tilde{e} \le 1-\overline{e}$, $\|\tilde{\mu}\|_\infty \le B$. Let $r_n \to 0$ be a deterministic sequence. Suppose Assumptions 1 and 2 hold and that for $k=1,\ldots,K$, $\|\hat{e}^{(k)}-\tilde{e}\|_2 = o_p(1)$, $\|\hat{\mu}^{(k)}-\tilde{\mu}\|_2 = o_p(1)$, $\|\hat{\tau}^{(k)}-\tau\|_q = O_p\left(r_n^{\frac{q+1}{2q}}\right)$, $\mathbb{P}(\|\hat{\mu}^{(k)}\|_\infty \le B) \to 1$, $\mathbb{P}(\overline{e} \le \hat{e}^{(k)} \le 1-\overline{e}) \to 1$, and*

$$\textit{either } \|\hat{e}^{(k)}-e\|_2 = O_p(r_n) \quad \textit{or} \quad \|\hat{\mu}^{(k)}-\mu\|_2 = O_p(r_n).$$

Then $\hat{\Psi}$ in Line 4 of Algorithm 1 satisfies

$$\hat{\Psi} = \Psi + O_p(r_n \vee n^{-1/2}).$$

Theorem 6 is particularly strong in experiments (e known), so we can get away with $\hat{\mu}^{(k)} = 0$.

It would appear we must consistently estimate CATE to have hope of estimating its CVaR. Although true, we next show that *even* if we mis-estimate CATE *and also* one of e,μ, we *still* get an upper bound on CATE-CVaR (hence on ITE-CVaR). This appears to be the second finding of a double-validity property since being first documented in sensitivity analysis (Dorn et al. 2021).

We first establish the population-level bound behavior and then state the implication for estimation.

Lemma 2. *Fix any $\tilde{\tau}:\mathcal{X}\to\mathbb{R}$. Let $\tilde{\beta} = F^{-1}_{\tilde{\tau}(X)}(\alpha)$. Suppose Assumption 2 holds with τ replaced with $\tilde{\tau}$. Then,*

$$\mathrm{CVaR}_\alpha(\tau(X)) \le \tilde{\beta} + \frac{1}{\alpha}\mathbb{E}[\mathbb{I}[\tilde{\tau}(X) \le \tilde{\beta}](\tau(X)-\tilde{\beta})]. \tag{15}$$

Theorem 7 (Double Validity). *Fix any $\tilde{e},\tilde{\mu},\tilde{\tau}$ with $\overline{e} \le \tilde{e} \le 1-\overline{e}$, $\|\tilde{\mu}\|_\infty \le B$, $\|\tilde{\tau}\|_\infty \le 2B$. Let $r_n \to 0$ be a deterministic sequence. Suppose Assumption 1 holds, Assumption 2 holds with τ replaced with $\tilde{\tau}$, and that for $k=1,\ldots,K$, $\|\hat{e}^{(k)}-\tilde{e}\|_2 = o_p(1)$, $\|\hat{\mu}^{(k)}-\tilde{\mu}\|_2 = o_p(1)$, $\|\hat{\tau}^{(k)}-\tilde{\tau}\|_q = O_p\left(r_n^{\frac{q+1}{q}}\right)$, $\mathbb{P}(\|\hat{\mu}^{(k)}\|_\infty \le B) \to 1$, $\mathbb{P}(\overline{e} \le \hat{e}^{(k)} \le 1-\overline{e}) \to 1$, and*

$$\textit{either } \|\hat{e}^{(k)}-e\|_2 = O_p(r_n) \quad \textit{or } \|\hat{\mu}^{(k)}-\mu\|_2 = O_p(r_n).$$

Then $\hat{\Psi}$ *in Line* 4 *of Algorithm* 1 *satisfies*

$$\hat{\Psi} \geq \Psi - O_p(r_n \vee n^{-1/2}).$$

Theorem 7 guarantees extensive robustness and suggests a practical, black box–free approach in experimental settings: set $\hat{\mu}^{(k)} = 0$ and use simple *misspecified* parametric models (e.g., linear) for CATE estimation, and we still estimate a valid ITE-CVaR bound at fast $O_p(n^{-1/2})$ rates.

4.3. CATE Estimation and Rates

Algorithm 1 accepts separate learners for *both* μ *and* τ. Therefore, although $\tau(x) = \mu(x,1) - \mu(x,0)$, we need *not* have $\hat{\tau}^{(k)}(X) = \hat{\mu}^{(k)}(x,1) - \hat{\mu}^{(k)}(x,0)$, and in fact we should not. Recent work advocates and provides specialized methods for *directly* estimating CATE (Imai and Ratkovic 2013, Athey and Imbens 2016, Wager and Athey 2018, Künzel et al. 2019, Kennedy 2020, Nie and Wager 2021).

This is important because Algorithm 1 uses the μ and τ estimates differently and, correspondingly, our theoretical results impose different assumptions on each. The τ estimate is used for approximating the event $\mathbb{I}[\tau(X) \leq \beta^*]$, which is crucial for targeting CVaR correctly. In contrast, the μ estimate is just used to estimate a weighted average treatment effect, given the weights $\mathbb{I}[\tau(X) \leq \beta^*]$, and is therefore interchangeable with propensity.

We next review different options for CATE estimation and how these ensure the conditions of Theorems 5–7. We emphasize that these need not be understood as exhaustive list of which learners to use: Practically, the nuisance estimation rates are high-level assumptions that generally say one may safely plug-in black box machine-learning estimators to Algorithm 1: No restrictions are made but rates (no metric-entropy conditions), estimators can be flexible/nonparametric in that rates can be much slower than "parametric" $O_p(n^{-1/2})$ rates, and results are exceedingly robust to inconsistent estimation.

4.3.1. Experimental Settings. A major issue with CATE estimation by differencing outcome regressions is that effect signals are easily lost. CATE is generally simpler and less variable than baseline mean outcomes, $\mu(X, 0), \mu(X,1)$. For example, many variables often help predict outcomes, but few modulate the treatment effect. It is therefore imperative to learn CATE directly.

In experimental settings (e known) we can construct a pseudo-outcome $\Delta = \frac{A - e(X)}{e(X)(1-e(X))} Y$ and, because $\tau(X) = \mathbb{E}[\Delta | X]$, learn CATE by regressing Δ on X, using any supervised-learning method. Setting either $q = \infty$ or $q = 2$, either $\|\hat{\tau}^{(k)} - \tau\|_\infty = o_p(n^{-1/4})$ or $\|\hat{\tau}^{(k)} - \tau\|_2 = o_p(n^{-3/8})$ suffices to satisfy the rate condition in Theorem 5. One case that theoretically ensures $\|\hat{\tau}^{(k)} - \tau\|_\infty = o_p(n^{-1/4})$ is when $\tau(x)$ is more-than-$d/2$-smooth in $x \in \mathbb{R}^d$ (Stone 1982, theorem 1). Another option is $\tau(x)$ linear with $o(\sqrt{n}/\log d)$ nonzero coefficients (Belloni et al. 2017). Alternatively, to theoretically ensure $\|\hat{\tau}^{(k)} - \tau\|_2 = o_p(n^{-3/8})$, we may use nonparametric least squares assuming τ belongs to any function class with log covering number at radius ϵ at most ϵ^{-p} with $p < 2/3$ (Wainwright 2019). This works *regardless* of μ being nice.

We may avoid black box models (and cross-fitting) altogether by using simple linear regression of Δ on X to obtain a valid bound per Theorem 7.

To satisfy the other conditions, for Theorems 6 and 7, we can set $\mu = 0$, and for Theorem 5, we need only estimate μ consistently without rate. We can either estimate μ directly or only estimate $\overline{\mu}(X) = \mathbb{E}[Y|X]$ and set $\hat{\mu}^{(k)}(X,A) = \hat{\overline{\mu}}^{(k)}(X) + (A - e(X))\hat{\tau}^{(k)}(X)$. Consistency for either is immediate from $\mathbb{E}Y^2 < \infty$ (Györfi et al. 2002).

4.3.2. Observational Settings. When e is unknown, the pseudo-outcome construction needs refinement. One option is DR-leaner (Kennedy 2020): regress $\Delta = \hat{\mu}(X, 1) - \hat{\mu}(X,0) + \frac{A - \hat{e}(X)}{\hat{e}(X)(1-\hat{e}(X))}(Y - \hat{\mu}(X,A))$ on X, where $\hat{e}, \hat{\mu}$ are appropriately cross-fitted. Another is R-learner (Nie and Wager 2021): let $\hat{\tau}$ minimize the average of $(Y - \hat{\overline{\mu}}(X) - (A - \hat{e}(X))\hat{\tau}(X))^2$, where $\hat{e}, \hat{\overline{\mu}}$ are appropriately cross-fitted. Kennedy (2020, corollary 3) provides rates for local-polynomial R-learners: If $e(x)$ is s_e-smooth in $x \in \mathbb{R}^d$, $\overline{\mu}(x) s_\mu$-smooth, and $\tau(x)$ more-than-$d/2$-smooth, then we obtain $o_p(n^{-1/4})$ rate pointwise error, provided $s_e \geq s_\mu$, $\frac{s_e + s_\mu}{2} > \frac{d}{8}$. To convert pointwise error bounds to sup-norm error bounds, $\|\hat{\tau}^{(k)} - \tau\|_\infty = o_p(n^{-1/4})$, we may follow the discretization approach of Stone (1982), incurring only logarithms. Alternatively, we can implement a DR-learner using nonparametric least squares, and following Kennedy (2020, corollary 3) and Wainwright (2019), we will obtain $\|\hat{\tau}^{(k)} - \tau\|_2 = o_p(n^{-3/8})$ if τ belongs to any function class with log covering number at radius ϵ at most ϵ^{-p} with $p < 2/3$ and if $\|\hat{e}^{(k)} - e\|_2 \|\hat{\mu}^{(k)} - \mu\|_2 = o_p(n^{-3/8})$. Otherwise, we can simply use *misspecified* linear R- or DR-learners and still get a valid bound per Theorem 7.

5. Case Study

We now demonstrate our bounds and inference.[9] Although we consider a program evaluation example, we believe our results are also particularly relevant to A/B testing on online platforms, where, after testing, product innovations are usually either scrapped/reworked or broadly rolled out, and where ATEs are often small, creating an opportunity for many users to be negatively impacted despite positive average effects. Little data are public, however.

5.1. Background and Setup

Behaghel et al. (2014) analyze a large-scale randomized experiment comparing assistance programs offered to French unemployed individuals. They compare three arms: individuals in the "control" arm receive the standard services of the Public Employment Services, in "public" receive an intensive counseling program run by a public agency, and in "private" a similar program run by private agencies.

We consider a hypothetical scenario where the private-run counseling program ($A = 0$) is currently being offered to the unemployed and we consider the change to a public-run program ($A = 1$).[10] We take re-employment within six months as our (binary) outcome.

The ATE is 1.22 percentage points (90% CI, $[-0.35, 2.8]$), a 4.9% increase in re-employment. This suggests a positive/neutral effect, so a policymaker might hypothetically consider this an acceptable policy change, for example, if the public-run program provided cost savings.[11]

To apply our methodology, we consider all pretreatment covariates in table 2 of Behaghel et al. (2014), except we treat as numeric (rather than dichotomize) age, number children, years of experience, salary target, assignment timing, and number unemployment spells. Other variables quantify education, employment level and type, gender, martial status, national origin, region, unemployment reason, and long-term unemployment risk. The propensity is constant. As recommended in Section 4.3.1, we fit CATE using a pseudo-outcome linear regression. We estimate μ using cross-fitted gradient-boosting machines.

5.2. Upper Bounds

Figure 1 presents inference on CATE-CVaR using Algorithm 1 for $\alpha \in \{0.01, 0.02, \ldots, 1\}$. The line represents our point estimate, after rearrangement as recommended in Remark 8,[12] and the shaded region represents pointwise 90% CIs. Uncertainty grows for smaller α.

We see that the ATE estimate (right-most point) is positive with an interval containing zero. We find, however, that some 56% sized X-defined subpopulation has a negative effect at 90% confidence.[13] This strongly suggests that the change, if enacted could materially, negatively impact a large portion of the population, despite the positive/neutral ATE. Thus, considering treatment effect *risk* provides a crucial metric not reflected in the ATE. This risk is also *not* reflected in DTEs: The binary potential outcome distributions are *fully* specified by just $\mathbb{E}[Y(0)]$, $\mathbb{E}[Y(1)]$.[14]

Figure 1. Inference on $\text{CVaR}_\alpha(\tau(X))$

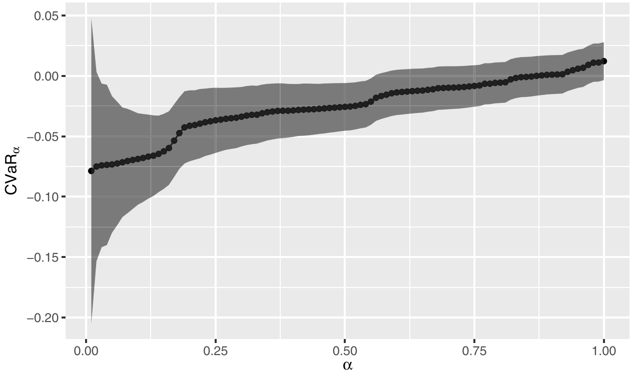
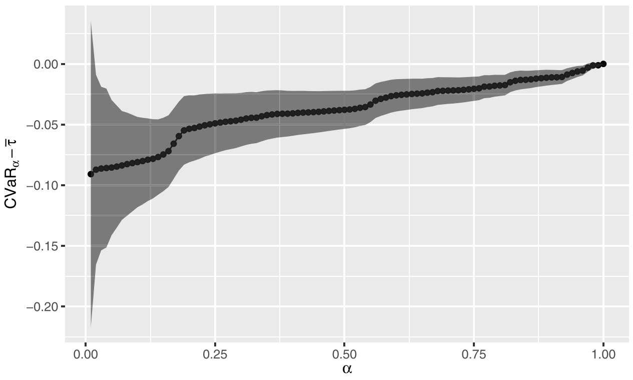


Figure 2. Inference on $\text{CVaR}_\alpha(\tau(X)) - \overline{\tau}$



In Figure 2 we focus on comparing CATE-CVaR to ATE following Remark 6. The only difference to Figure 1 is a slight vertical shift and that CIs (correctly) shrink to a point as $\alpha \to 1$, enabling more confident conclusions comparing subpopulations to the population.

In Figure 3, we consider estimating CATE-CVaR using the plug-in approach mentioned in Section 4, using standard errors of $\tilde{\phi}(X, A, Y; \hat{\tau}, \hat{\beta})$ to construct CIs. That is, we just compute the CVaR of CATE predictions on the data. The result is heavily downward biased with far-too-narrow CIs.

In Figure 4, we consider CATE-CVaR when we capture less heterogeneity, using only age, high school dropout, African national origin, and Paris region resident as covariates (X_1). This detects no significant risk. This illustrates that, although Theorem 1 is tight, the bound can be *practically uninformative* if covariates are

Figure 3. Plug-in Estimator

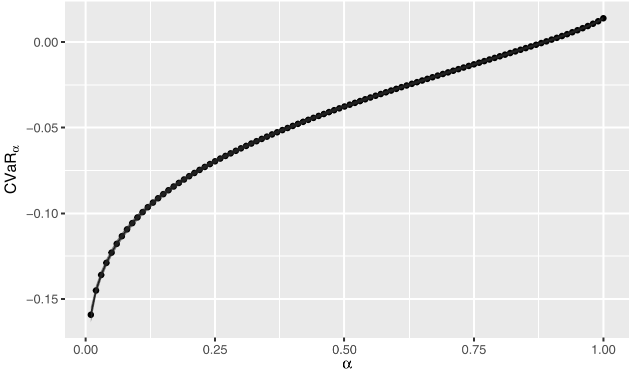
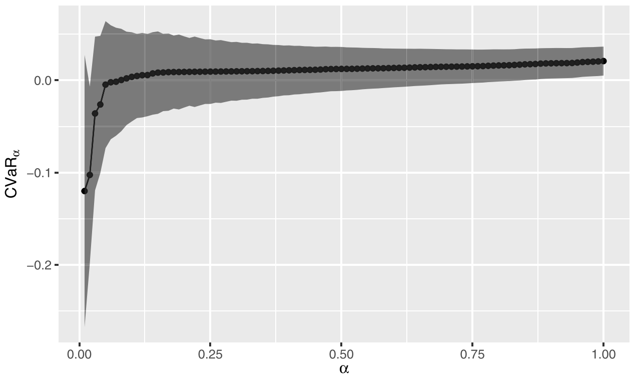


Figure 4. Restricted Covariate Set



not very predictive. Although using the full set of covariates X, we are able to detect that some 50% of the population has a negative effect, using only the subset X_1, we fail to detect this because we cannot informatively segment the population into heterogeneously impacted groups.

5.3. Lower Bounds

Although the upper bounds show a significant subpopulation can be negatively harmed, being only bounds, it may be the subpopulation can be harmed even more or an even larger subpopulation can be harmed. Lower bounds help us understand how much greater the risk might be.

In Figure 5, we consider our lower bounds (versus ATE) when limiting the residual heterogeneity range given by Theorems 2 (two-sided range) and 3 (one-sided range).

Because it may be hard to justify and calibrate a limited range, in Figure 6, we consider lower bounds given by Theorem 4 and Corollary 1 by limiting residual heterogeneity variance. For the former, we fit $\mathrm{Var}(Y|A,X)$ using gradient-boosting machines and construct $\overline{\sigma}^2(X)$ per Equation (10) by varying constant values of $\rho(X) = \rho \in [-1,1]$. Recall $\rho = -1$ always yields an assumption-free bound. We use the same model to estimate the right-hand side of Equation (12). We compute the cross-validated root-mean-squared prediction error to estimate the right-hand side of Equation (13).

Figure 5. (Color online) Bounds Based on Residual-Heterogeneity Range

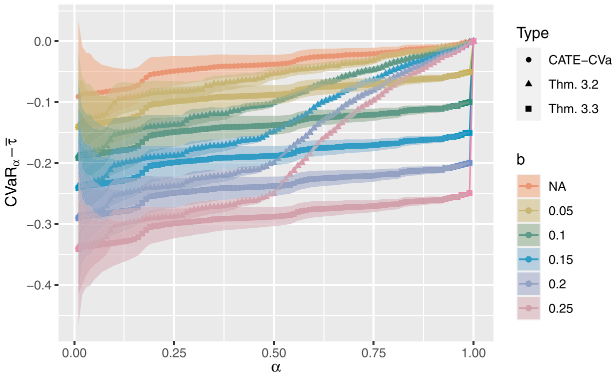
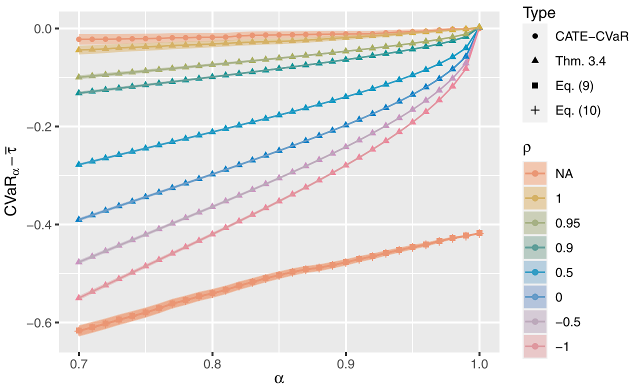


Figure 6. (Color online) Bounds Based on Residual-Heterogeneity Variance



We observe that assuming perfectly conditionally correlated potential outcomes yields a lower bound very close to the upper bound. The bounds of Corollary 1 appear loose; indeed, they are not tight. Nonetheless, the tightness of the other bounds does not mean they are practically informative, which depends on their use. The width of the bounds need not determine informativeness either. If the upper bound is negative, that is informative by itself of certain harm, regardless of how much more negative is the lower bound. If the upper bound is positive, however, a good lower bound may still help bound potential but uncertain harm.

6. Concluding Remarks

We study the average effect on those worst-affected by a proposed change as a measure of its *risk*, how to tightly bound it given covariates that explain some heterogeneity, and how to make robust inferences on these bounds even when this heterogeneity is roughly estimated. This provides very practical tools for assessing policy and product changes beyond their ATE and DTEs. We can safely use flexible yet biased/slow-to-converge machine learning, or we can avoid black box models and easily get good bounds by considering only linear projections of heterogeneity. In the hypothetical case study this detected that, what appeared to be a positive/neutral change could actually very negatively impact a substantial subpopulation.

We focused on experimental (or, unconfounded observational) settings without interference, where risk is already unidentifiable *despite* randomization. A future direction is to consider the impact of interference (Athey et al. 2018, Johari et al. 2022) or confounding (Tan 2006), where even ATEs are unidentifiable and fairness is harder to assess (Kallus and Zhou 2018, Jung et al. 2020, Kilbertus et al. 2020). Interestingly, for partial identification under Tan (2006)′ model, X-conditional outcome-CVaR plays a crucial role (Dorn et al. 2021). Another direction may be to consider other risk measures, such as given by Kullback-Leibler ambiguity sets (Ahmadi-Javid 2012). Per Endnote 5, the tight upper bound is still the risk measure applied to CATE, but it remains to compute lower bounds and design robust inference methods.

Endnotes

[1] We may still make some inferences on these even if we do not observe such attributes (Chen et al. 2019, Kallus et al. 2022).

[2] CVaR is sometimes defined for the right tail, corresponding to our $-\mathrm{CVaR}_{\alpha}(-Z)$.

[3] Also, $Y = Y^{*}(A)$ assumes noninterference (Rubin 1986).

[4] However, even this ideal can induce disparate impacts (Kallus and Zhou 2019).

[5] Equation (5) extends to any coherent risk by writing $\delta = \tau(X) + (\delta - \tau(X))$ and using subadditivity.

[6] The bound need not be tight given the (X, A, Y) distribution, which characterizes more than the mean of the $(\delta | X)$ distribution, as described by the Fréchet-Hoeffding bounds. We focus on best bounds given just by CATE, which is the common tool to understand effect heterogeneity in practice.

[7] There are also tests for heterogeneity *not* explained by X (Ding et al. 2016, 2019). These, like us, leverage bounds on unidentifiable quantities.

[8] We may avoid cross-fitting and fit nuisances once on the whole sample if we assume estimates belong to a Donsker class with probability tending to one; we omit this option for brevity.

[9] Replication code is available at https://github.com/CausalML/TreatmentEffectRisk.

[10] Some individuals assigned to the additional counseling refused it. We nonetheless restrict our attention to intent-to-treat interventions, considering hypothetically making available either the public-run or private-run counseling to unemployed individuals, who may decline it.

[11] Behaghel et al. (2014, section IV) discuss why public-run programs fare better.

[12] We present the figure without rearrangement in Online Appendix B.

[13] Becuase outcome is binary, the *largest* fraction that can have a negative effect is $(50 \times (1 - \overline{\tau}))\%$, so either $\overline{\tau} < 0$ or at most half may be negatively affected. The ATE interval indeed contains zero with confidence only 90%.

[14] In particular, the α-quantile DTE is uselessly *zero* for *all* $\alpha \in [0,1] \backslash \{1 - \mathbb{E}[Y(0)], 1 - \mathbb{E}[Y(1)]\}$ and the α-CVaR DTE is $\frac{1}{\alpha}(\mathbb{E}[Y(1)] - 1 + \alpha)_{+} - \frac{1}{\alpha}(\mathbb{E}[Y(0)] - 1 + \alpha)_{+}$, which is not even monotonic. For illustration we plot it in Online Appendix B.

References

Ahmadi-Javid A (2012) Entropic value-at-risk: A new coherent risk measure. *J. Optim. Theory Appl.* 155(3):1105–1123.

Artzner P, Delbaen F, Eber JM, Heath D (1999) Coherent measures of risk. *Math. Finance* 9(3):203–228.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc. National Acad. Sci. USA* 113(27):7353–7360.

Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89:133–161. https://doi.org/10.3982/ECTA15732.

Athey S, Eckles D, Imbens GW (2018) Exact p-values for network interference. *J. Amer. Statist. Assoc.* 113(521):230–240.

Audibert JY, Tsybakov AB (2007) Fast learning rates for plug-in classifiers. *Ann. Statist.* 35(2):608–633.

Bagnell JA (2005) *Robust Supervised Learning* (AAAI Press, Palo Alto, CA).

Behaghel L, Crépon B, Gurgand M (2014) Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *Amer. Econom. J. Appl. Econom.* 6(4):142–174.

Belloni A, Chernozhukov V, Fernández-Val I, Hansen C (2017) Program evaluation and causal inference with high-dimensional data. *Econometrica* 85(1):233–298.

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Sci.* 59(2):341–357.

Bertsimas D, Gupta V, Kallus N (2018) Robust sample average approximation. *Math. Programming* 171(1):217–282.

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019) Fairness under unawareness: Assessing disparity when protected class is unobserved. *Proc. Conf. Fairness, Accountability, and Transparency (FAT* '19)* (Association for Computing Machinery, New York), 339–348.

Chernozhukov V, Fernández-Val I, Galichon A (2010) Quantile and probability curves without crossing. *Econometrica* 78(3):1093–1125.

Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I (2018a) Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. Technical report, National Bureau of Economic Research, Cambridge, MA.

Chernozhukov V, Escanciano JC, Ichimura H, Newey WK, Robins JM (2022) Locally robust semiparametric estimation. *Econometrica* 90(4):1501–1535.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018b) Double/debiased machine learning for treatment and structural parameters. *Econom. J.* 21(1):C1–C68.

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2008) Nonparametric tests for treatment effect heterogeneity. *Rev. Econom. Statist.* 90(3):389–405.

Davison A (1992) Treatment effect heterogeneity in paired data. *Biometrika* 79(3):463–474.

Ding P, Feller A, Miratrix L (2016) Randomization inference for treatment effect variation. *J. Royal Statist. Soc. Ser. B Statist. Methodological* 78(3):655–671.

Ding P, Feller A, Miratrix L (2019) Decomposing treatment effect variation. *J. Amer. Statist. Assoc.* 114(525):304–317.

Dorn J, Guo K, Kallus N (2021) Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. Preprint, submitted December 21, https://doi.org/10.48550/arXiv.2112.11449.

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1):115–166.

Firpo S (2007) Efficient semiparametric estimation of quantile treatment effects. *Econometrica* 75(1):259–276.

Gail M, Simon R (1985) Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41(2):361–372.

Györfi L, Kohler M, Krzyżak A, Walk H (2002) *A Distribution-Free Theory of Nonparametric Regression* (Springer, Berlin).

Hardy GH, Littlewood JE, Pólya G, Pólya G (1952) *Inequalities* (Cambridge University Press, Cambridge, UK).

Heckman JJ, Smith J, Clements N (1997) Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Rev. Econom. Stud.* 64(4):487–535.

Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Statist.* 7(1):443–470.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, UK).

Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J. Econom. Literature* 47(1):5–86.

Johari R, Li H, Liskovich I, Weintraub GY (2022) Experimental design in two-sided platforms: An analysis of bias. *Management Sci.* 68(10): 7069–7089.

Jung J, Shroff R, Feller A, Goel S (2020) Bayesian sensitivity analysis for offline policy evaluation. *Proc. AAAI/ACM Conf. on AI, Ethics, and Society* (AAAI Press, Palo Alto, CA), 64–70.

Kallus N (2018) Balanced policy evaluation and learning. *Advan. Neural Inform. Processing Systems* 31.

Kallus N, Zhou A (2018) Residual unfairness in fair machine learning from prejudiced data. *Internat. Conf. Machine Learn.* 35:2439–2448.

Kallus N, Zhou A (2019) Assessing disparate impacts of personalized interventions: Identifiability and bounds. *Advan. Neural Inform. Processing Systems* 32.

Kallus N, Zhou A (2021) Minimax-optimal policy learning under unobserved confounding. *Management Sci.* 67(5):2870–2890.

Kallus N, Mao X, Uehara M (2019) Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond.

Kallus N, Mao X, Zhou A (2022) Assessing algorithmic fairness with unobserved protected class using data combination. *Management Sci.* 68(3):1959–1981.

Kearns M, Neel S, Roth A, Wu ZS (2018) *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness* (ICML).

Kennedy EH (2020) Towards optimal doubly robust estimation of heterogeneous causal effects. Preprint, submitted April 29, https://doi.org/10.48550/arXiv.2004.14497.

Kennedy EH, Balakrishnan S, G'Sell M (2020) Sharp instruments for classifying compliers and generalizing causal effects. *Ann. Statist.* 48(4):2008–2030.

Kilbertus N, Ball PJ, Kusner MJ, Weller A, Silva R (2020) *The Sensitivity of Counterfactual Fairness to Unmeasured Confounding* (UAI).

Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.

Krokhmal P, Palmquist J, Uryasev S (2002) Portfolio optimization with conditional value-at-risk objective and constraints. *J. Risk* 4:43–68.

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. National Acad. Sci. USA* 116(10):4156–4165.

Lahoti P, Beutel A, Chen J, Lee K, Prost F, Thain N, Wang X, Chi E (2020) Fairness without demographics through adversarially reweighted learning. *Advan. Neural Inform. Processing Systems*, vol. 33 (NeurIPS, San Diego).

Lakkaraju H, Kamar E, Caruana R, Leskovec J (2019) *Faithful and Customizable Explanations of Black Box Models* (AIES).

Luedtke AR, van der Laan MJ (2016a) Optimal individualized treatments in resource-limited settings. *Internat. J. Biostatist.* 12(1):283–303.

Luedtke AR, van der Laan MJ (2016b) Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* 44(2):713.

Luedtke AR, van der Laan MJ (2017) Evaluating the impact of treating the optimal subgroup. *Statist. Methods Medical Res.* 26(4):1630–1640.

Mammen E, Tsybakov AB (1999) Smooth discrimination analysis. *Ann. Statist.* 27(6):1808–1829.

Nie X, Wager S (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108(2):299–319.

Qian M, Murphy SA (2011) Performance guarantees for individualized treatment rules. *Ann. Statist.* 39(2):1180.

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" explaining the predictions of any classifier. *Proc. ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (ACM, New York) 22:1135–1144.

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression-coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* 89(427):846–866.

Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. *J. Risk* 2:21–42.

Rubin DB (1986) Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* 81(396):961–962.

Ruszczyński A, Shapiro A (2006) Optimization of convex risk functions. *Math. Oper. Res.* 31(3):433–452.

Sawilowsky SS (1990) Nonparametric tests of interaction in experimental design. *Rev. Ed. Res.* 60(1):91–126.

Schick A (1986) On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* 14(3):1139–1151.

Stone CJ (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* 10(4):1040–1053.

Tan Z (2006) A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* 101(476):1619–1637.

Van der Vaart A (1998) *Asymptotic Statistics* (Cambridge University Press, Cambridge, UK).

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113(523):1228–1242.

Wainwright MJ (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press, New York).

Zhao Y, Zeng D, Rush AJ, Kosorok MR (2012) Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* 107(499):1106–1118.

Zheng W, van der Laan MJ (2011) Cross-validated targeted minimum-loss-based estimation. *Targeted Learning* (Springer, Berlin), 459–474.