

Partially Retargeted Balancing Weights for Causal Effect Estimation Under Positivity Violations

Martha Barnard*, Jared D. Huling, Julian Wolfson

Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota

Abstract

Positivity violations, which occur when some subgroups either always or never receive a treatment of interest, pose significant challenges for causal effect estimation with observational data. Recent balancing weight methods have proved to be highly effective in confounding control, however their utility is diminished in the presence of positivity violations, resulting in bias and excess variance. Approaches that deal with positivity violations, on the other hand, work by targeting a modified estimand that may be misaligned with the original research question. To address these challenges, we propose a novel balancing weights approach, which mitigates positivity violations while attempting to retain the original estimand by a targeted relaxation of the balancing constraints. Our proposed weighted estimator is consistent for the original estimand when either 1) the implied propensity score model is correct; or 2) all treatment effect modifiers are balanced to the target population. When these conditions do not hold, our estimator is consistent for a slightly modified treatment effect estimand. Furthermore, our proposed weighted estimator has reduced asymptotic variance when positivity does not hold. We evaluate our approach through applications to synthetic data, an observational study, and when transporting a treatment effect from a randomized trial.

Keywords: causal inference, direct balancing, overlap, transportability, observational studies

* barna126@umn.edu

1 Introduction

Confounding is a key challenge when estimating the causal effect of a binary treatment with observational data. While many methods exist for controlling confounding by balancing covariates between treatment groups, these methods do not fully control confounding when there are positivity violations. Positivity violations occur when there exists at least one combination of covariate values which perfectly predicts treatment assignment, a phenomenon which is especially likely to occur with high-dimensional covariates (D’Amour and Franks, 2021). When there are positivity violations, inverse probability weighting (IPW) estimators (Rosenbaum and Rubin, 1983; Hahn, 1998; Robins et al., 2000; Hirano and Imbens, 2001; Hirano et al., 2003; Imbens, 2004) have substantial statistical bias and inflated variance. A proposed alternative to IPW, direct balancing weights, (Hainmueller, 2012; Zubizarreta, 2015; Wang and Zubizarreta, 2020) tend to yield reduced estimator error and variance compared to IPW methods through achieving exact covariate balance for a prespecified set of covariates. However, when there are positivity violations there is often no solution to the direct balancing optimization problem prohibiting their use altogether.

To mitigate the impacts of positivity violations on causal effect estimation, a variety of methods, such as propensity score trimming, matching weights, and overlap weights, have been developed that modify the estimand and corresponding target population to achieve estimators with reduced variance (Crump et al., 2009; Li and Greene, 2013; Yang and Ding, 2018; Li et al., 2018, 2019). In fact, overlap weights (Li et al., 2018) do not require positivity by targeting the so-called “overlap population”; this results in an estimator with minimum asymptotic variance among weighted treatment effects under homoscedasticity. While these methods achieve reduced estimator variance, they target a population that is modified from the original population of interest such that the corresponding treatment effect estimate may be misaligned with the original research question. Yet, a key benefit of causal effect estimation with observational data is the ability to evaluate causal effects in a meaningful population of interest such that modifying the estimand is often either undesirable or challenging to interpret. Furthermore, these estimators are inconsistent for the original estimand when there is treatment effect heterogeneity. Modified treatment policy estimands result in improved estimator performance by defining an entirely different estimand arrived at by imagining modifications to the natural value of treatment (Kennedy, 2019). While the corresponding estimators perform well, these modified treatment policies do not always answer the original research question.

Thus, under positivity violations, existing methods either 1) target an estimand of interest but result in poor estimator performance or technical issues (e.g., direct balancing weights yielding no solution); or 2) alleviate these technical issues by targeting a modified estimand that may

be misaligned with the original estimand of interest (e.g., overlap weights). To address these challenges within estimator statistical performance and interpretation, we propose a novel balancing approach whose optimization problem yields a solution under positivity violations through a targeted relaxation of the balancing constraints. Our corresponding weighted estimator is consistent for the original estimand when either the implied propensity score model is correct or set of treatment effect modifiers is properly specified and smaller than the set of confounders. When these conditions do not hold, our weighted estimator is consistent for a slightly modified estimand and target population. Furthermore, we demonstrate that overlap weights are a special case of our proposed weights such that our method offers an intermediate between direct balancing and overlap weights. We derive asymptotic results which show that 1) our weights can be derived under a relaxed positivity assumption; and 2) our proposed estimator has reduced asymptotic variance when there are positivity violations. We propose a design- and model-based procedure for selecting the balancing constraints to relax and we evaluate the corresponding estimators through applications to synthetic data, an electronic health record (EHR) study, and when transporting a treatment effect from a randomized controlled trial (RCT) to a new population.

The remainder of the paper is organized as follows. In Section 2, we introduce notation and methodological background on balancing weights. We introduce and discuss the implications of our proposed partially retargeted balancing weights in terms of positivity, estimator interpretation, and extensions to additional estimation problems in Section 3. In Section 4, we discuss the asymptotic properties of our proposed weighted estimator. We propose design- and model-based procedures for identifying which balancing constraints to relax in Section 5. In Sections 6 and 7, we apply our methods to synthetic data, an EHR study, and when transporting a treatment effect from a RCT. We discuss results and future work in Section 8.

2 Methodological background

2.1 Notation, assumptions, and covariate balance

Consider an independent sample $\{(Z_i, Y_i, \mathbf{X}_i)\}_{i=1}^n$ of size n from a population where $Z_i = z$, indicates belonging to the treatment ($z = 1$) or control group ($z = 0$), Y_i is the outcome of interest, and $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$ is a p length vector of pre-treatment covariates. Let $Y(z)$ be the potential outcome that would be observed if assigned to treatment group z (Neyman (1990); Rubin (1974, 1978); Hernan and Robins (2024)). We assume the stable unit treatment value assumption (SUTVA) which implies that $Y_i = Y_i(Z_i)$ such that we only observe one potential outcome for each individual. We focus on the average treatment effect (ATE), $\tau = E[Y(1) - Y(0)]$ as the estimand of interest, although this work can be readily extended to other estimands of interest which we discuss in Section 3.3. Let $e(\mathbf{x}) = \Pr(Z_i = 1 | \mathbf{X}_i = \mathbf{x})$

be the propensity score. For the identification of τ the following assumptions are typically made:

Assumption 1: Strong ignorability. $\{Y(0), Y(1)\} \perp\!\!\!\perp Z | \mathbf{X}$ and

Assumption 2: Positivity. $0 < e(\mathbf{x}) < 1$ for all \mathbf{x} .

The focus of our work surrounds the positivity assumption, specifically, scenarios where this assumption does not hold. Let $\mu_z(\mathbf{x}) = E[Y(z) | \mathbf{X} = \mathbf{x}]$ and $\tau(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$. Then, $\tau = E \left\{ \frac{\mu_1(\mathbf{X})Z}{e(\mathbf{X})} - \frac{\mu_0(\mathbf{X})(1-Z)}{1-e(\mathbf{X})} \right\} = E \left\{ \frac{YZ}{e(\mathbf{X})} - \frac{Y(1-Z)}{1-e(\mathbf{X})} \right\}$ holds with the above assumptions, where the sample version of the last equation corresponds to the IPW estimator of the ATE. For a treatment effect estimator to be consistent, the expectations of $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ need to be balanced between the treated and control populations and to the target population. IPW (i.e., $w_1(\mathbf{x}) = \{e(\mathbf{x})\}^{-1}$ and $w_0(\mathbf{x}) = \{1 - e(\mathbf{x})\}^{-1}$) balance all functions of \mathbf{x} in this way such that the following equations hold,

$$E \{w_1(\mathbf{X})Zb(\mathbf{X})\} = E\{b(\mathbf{X})\} \text{ and } E \{w_0(\mathbf{X})(1 - Z)b(\mathbf{X})\} = E\{b(\mathbf{X})\} \quad (1)$$

$$E \{w_1(\mathbf{X})Zb(\mathbf{X})\} = E \{w_0(\mathbf{X})(1 - Z)b(\mathbf{X})\} \quad (2)$$

for any function $b(\mathbf{x})$, which we reference as three-way balance (Chan et al., 2016). Thus, IPW satisfy three-way balance of both $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$, regardless of their true forms. Specifically, Equation (1) balances the weighted treated and control group expectations of $b(\mathbf{x})$ to the population expectation of $b(\mathbf{x})$ and Equation (2) balances the weighted treated and control group expectations of $b(\mathbf{x})$ to each other. Thus, Equation (1) and (2) serve different purposes; while Equation (1) implies Equation (2), Equation (1) primarily serves to balance the treated and control groups to the estimand target population. When only Equation (2) holds, $b(\mathbf{x})$ may be balanced to modified target population that corresponds not to the ATE, but to some other causal estimand.

However, when there is a lack of overlap, even approximate balance of the empirical counterparts of Equations (1) and (2) may not be achieved after IPW, resulting in estimators with substantial error in addition to inflated variance. To achieve balance when there is a lack of overlap, some methods target a different estimand, $E[h(\mathbf{X})\tau(\mathbf{X})]$, which corresponds to a modified target population with improved overlap. Li et al. (2018) propose one such method, overlap weights ($h(\mathbf{x}) = e(\mathbf{x})\{1 - e(\mathbf{x})\}$), which are designed to ensure that the weights exist even when positivity does not hold. By modifying the target population, overlap weights can more easily satisfy Equation (1) yielding an estimator with reduced bias and variance with respect to the modified estimand, the ATO. However, the overlap weights estimator tends to be biased for the ATE when there is treatment effect heterogeneity. While the weights corresponding to these modified estimands tend to achieve improved finite-sample

balance in scenarios with a lack of overlap, they rarely achieve the empirical versions of Equations (1) and (2) resulting in estimator error. In fact, even when positivity does hold, exact finite-sample balance is rarely achieved by IPW-style estimators. While exact balance of covariate sample means is achieved by overlap weights when the propensity score is estimated with a logistic regression (Li et al., 2018), most IPW-style weights do not achieve exact balance of the sample means, especially under positivity violations. This has motivated the development of sample weights that explicitly achieve three-way finite sample balance.

2.2 Review of direct balancing weights

Direct balancing weights for the ATE achieve exact three-way balance for some pre-specified set of covariate functions; for the direct balancing weights estimator to be consistent for the ATE, $\mu_z(\mathbf{x})$ must be a linear combination of this set of covariate functions. Consider the case where we want to balance the covariate functions $b(\mathbf{x}) = \{b_j(\mathbf{x})\}_{j=1}^J$. Then, direct balancing weights for the ATE take the following form,

$$\{w_i^{bw}\}_{i=1}^n = \operatorname{argmin}_{w_i} \sum_{i=1}^n D(w_i) \quad \text{subject to} \quad (3a)$$

$$\frac{1}{n} \sum_{i=1}^n w_i I[Z_i = z] b_j(\mathbf{X}_i) - \frac{1}{n} \sum_{i=1}^n b_j(\mathbf{X}_i) = 0 \quad j = 1, \dots, J, \quad z = 0, 1, \quad (3b)$$

where $D(\cdot)$ is a convex measure of dispersion. The constraints $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ are often added to ensure that the weights do not extrapolate the observed outcomes. The resulting direct balancing weights estimator is,

$$\hat{\tau}_{w^{bw}} = \frac{1}{n} \sum_{i=1}^n w_i^{bw} Z_i Y_i - \frac{1}{n} \sum_{i=1}^n w_i^{bw} (1 - Z_i) Y_i. \quad (4)$$

In this constrained optimization problem, constraint (3b) ensures that the treated and control covariate functions are balanced to the sample population. These constraints also imply that the treated and control covariate functions are balanced to each other (i.e., $\frac{1}{n} \sum_{i=1}^n w_i Z_i b_j(\mathbf{X}_i) = \frac{1}{n} \sum_{i=1}^n w_i (1 - Z_i) b_j(\mathbf{X}_i)$) such that three-way balance is satisfied by $\{w_i^{bw}\}_{i=1}^n$ for $b(\mathbf{x})$. Since balance between treated and control groups is implicitly satisfied, Problem (3) is separable by treatment group such that two separate constrained optimization problems can be solved to derive $\{w_i^{bw}\}_{i=1}^n$ (Chan et al., 2016); we discuss implications of this separation in Section 3.1. When there is a lack of overlap between $\{b(\mathbf{x})\}_{z=1}$ and $\{b(\mathbf{x})\}_{z=0}$, often there is no solution to Problem (3) (i.e., Equations (3a)-(3b)) with the addition of the constraints $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ such that weights which achieve exact three-way balance without extrapolating cannot be derived. In these scenarios, the

exact balance constraint is relaxed and approximate balance is achieved by constraining the absolute difference between sample means to be less than or equal to some parameter δ ; weights derived through this process are called minimal weights (Wang and Zubizarreta, 2020). While the minimal weights estimator does have reduced variance compared to IPW and direct balancing estimators, minimal weights achieve none of the three balancing equations such that the corresponding estimator is often biased due to covariate imbalance. Thus, there is a need for methods that 1) achieve exact finite-sample covariate balance between treated and control groups to reduce estimator error; and 2) modify the target population minimally to yield an estimator with reduced variance when there is a lack of overlap.

3 A partially retargeted balancing approach

To address these challenges, we propose a novel constrained optimization problem for deriving weights, primarily for scenarios where there is no solution to direct balancing weights due to positivity violations. Consider the case where Assumption 2 does not hold, but $0 < \Pr\{Z_i = 1 | c(\mathbf{X}) = c(\mathbf{x})\} < 1$ for some $c(\mathbf{x}) \subset b(\mathbf{x})$ such that positivity holds when conditioning on this restricted subset. Let $g(\mathbf{x})$ be the remaining covariate functions in $b(\mathbf{x})$ such that $b(\mathbf{x}) = [c(\mathbf{x}) = \{c_k(\mathbf{x})\}_{k=1}^K, g(\mathbf{x}) = \{g_l(\mathbf{x})\}_{l=1}^L]$. The key idea of our method is to relax the balancing constraints by balancing all covariate functions, $b(\mathbf{x}) = \{c(\mathbf{x}), g(\mathbf{x})\}$, between treated and control groups, but only $c(\mathbf{x})$ to the sample population, yielding the following set of balancing equations:

$$\{w_i^*\}_{i=1}^n = \operatorname{argmin}_{w_i} \sum_{i=1}^n D(w_i) \quad \text{subject to} \quad (5a)$$

$$\frac{1}{n} \sum_{i=1}^n w_i I[Z_i = z] c_k(\mathbf{X}_i) - \frac{1}{n} \sum_{i=1}^n c_k(\mathbf{X}_i) = 0, \quad k = 1, \dots, K, \quad z = 0, 1 \quad (5b)$$

$$\sum_{i=1}^n w_i Z_i g_l(\mathbf{X}_i) - \sum_{i=1}^n w_i (1 - Z_i) g_l(\mathbf{X}_i) = 0, \quad l = 1, \dots, L. \quad (5c)$$

The primary difference between this constrained optimization problem and the direct balancing weights constrained optimization problem is Equation (5c); in this equation, the functions, $g(\mathbf{x}) = \{g_l(\mathbf{x})\}_{l=1}^L$ are only balanced between treated and control groups and not to the sample population. Thus, the derived weights satisfy exact three-way balance for functions $c(\mathbf{x}) = \{c_k(\mathbf{x})\}_{k=1}^K$ but only balance between treated and control groups for $g(\mathbf{x})$. In doing so, the derived weights do not balance $g(\mathbf{x})$ to a specific modified population (such as with overlap weights). However, the weighted treated and control group sample means of $g(\mathbf{x})$ are implicitly modified when minimizing weight dispersion subject to only the treated and control balance constraint. We call these weights, $\{w_i^*\}_{i=1}^n$, partially retargeted balancing

weights (PRTBW). In Section 3.1, we formally show that this approach is justified under a relaxed positivity assumption. In addition, the corresponding estimator to Problem (5) (i.e., Equations (5a) - (5c)) is consistent for the ATE when $c(\mathbf{x})$ contains all treatment effect modifiers, which we discuss further in Sections 3.2 and 4.

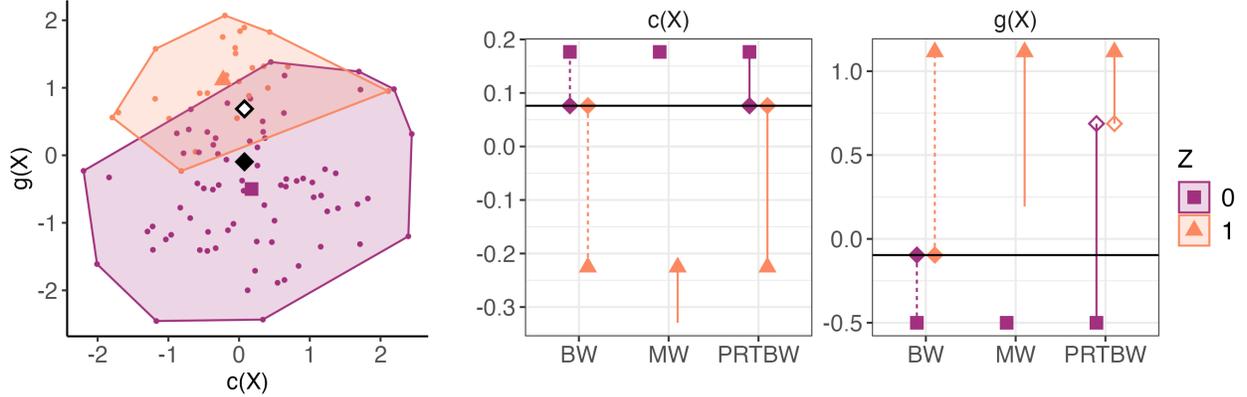


Figure 1: Simulated data illustrating the differences between direct balancing weights (BW), minimal weights (MW), and the proposed partially retargeted balancing weights (PRTBW). Squares (purple) and triangles (orange) indicate the unweighted control and treated sample means, respectively. The filled diamond indicates the sample population mean while the outlined diamond indicates the weighted sample mean given $\{w_i^*\}_{i=1}^n$ derived through Problem (5). In the right-hand figures, the horizontal line indicates the sample mean while the vertical lines end at the weighted treated and control means for each weighting method. For this simulated data, there is no solution to Problem (3), exhibited by the dashed vertical line.

Figure 1 presents a toy example of the differences in covariate balance and target population between balancing weights (Problem (3)), minimal weights, and our proposed partially retargeted balancing weights (Problem (5)); for this example, the constraints of the $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ are added to the optimization problems. For this simulated data, there is no solution to Problem (3), such that weights that exactly balance the treated and control sample means of $b(\mathbf{x}) = \{c(\mathbf{x}), g(\mathbf{x})\}$ cannot be derived. Our proposed weights balance the treated and control sample means of $c(\mathbf{x})$ to the population sample mean, but the treated and control sample means of $g(\mathbf{x})$ are balanced to a different value. In contrast, minimal weights do not achieve covariate balance for either $c(\mathbf{x})$ or $g(\mathbf{x})$. In fact, the minimally weighted treated sample means are unchanged and the minimally weighted control sample mean of $c(\mathbf{x})$ is further away from the population sample mean of $c(\mathbf{x})$ than the unweighted control sample mean. Thus, while both the minimal weights optimization problem and our proposed Problem (5) yield solutions for this data, they do this through different mechanisms such that each method yields a solution with a different interpretation.

To further explore the theoretical properties of our proposed weights, we identify the dual formulation of Problem (5). The addition of constraint (5c) ensures that the optimization problem is not separable across the treatment groups such that there is correspondingly a

single dual problem.

Theorem 1. *The dual of of Problem (5) is equivalent to the optimization problem*

$$\begin{aligned} \underset{\alpha_0, \alpha_1, \gamma}{\text{minimize}} \quad & \frac{1}{n} \sum_{i=1}^n -Z_i \rho\{\alpha_1^T c(\mathbf{X}_i) + \gamma^T g(\mathbf{X}_i)\} - (1 - Z_i) \rho\{\alpha_0^T c(\mathbf{X}_i) - \gamma^T g(\mathbf{X}_i)\} \\ & + \alpha_1^T c(\mathbf{X}_i) + \alpha_0^T c(\mathbf{X}_i) \end{aligned} \quad (6)$$

where $\alpha_{0_{K \times 1}}$, $\alpha_{1_{K \times 1}}$, and $\gamma_{L \times 1}$ are the dual variables associated with the balancing constraints in Equations (5b)-(5c), $\rho(t) = t - t(h')^{-1}(t) + h[(h')^{-1}(t)] - h(1)$ is strictly concave, and $h(t) = D(1 - t)$. Furthermore, the primal solution $w_i^* = Z_i w_{1i}^* + (1 - Z_i) w_{0i}^*$ satisfies $w_{1i}^* = \rho'\{\hat{\alpha}_1^T c(\mathbf{X}_i) + \hat{\gamma}^T g(\mathbf{X}_i)\}$ and $w_{0i}^* = \rho'\{\hat{\alpha}_0^T c(\mathbf{X}_i) - \hat{\gamma}^T g(\mathbf{X}_i)\}$ where $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\gamma}$ are the solution to the dual problem.

The proof of this theorem is in supplementary Section 1. In the following sections, we discuss the implications of Problem (5) in terms of positivity, the interpretation of the resulting estimator, and related estimation problems while drawing comparisons to direct balancing weights, minimal weights, and overlap weights.

3.1 Positivity considerations

Problem (5) is primarily motivated by the challenges posed by lack of overlap on causal effect estimation, generally, and direct balancing weights, specifically. In the following section, we show that a solution to Problem (5), with the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$, exists under a relaxed positivity assumption. Furthermore, we determine the propensity score model implied by Problem (5) and describe how this model connects to the relaxed positivity assumption and other methods. While positivity is required for the existence of a solution to Problem (3) with the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ (Zhao and Percival, 2017), by not balancing $g(\mathbf{x})$ to the target population, the weights derived through Problem (5) exist under a relaxed positivity assumption:

Assumption 2*: Relaxed positivity. $0 < Pr\{Z_i | c(\mathbf{X}_i) = c(\mathbf{x})\} < 1$ for all \mathbf{x} and there exists an \mathbf{x}^* such that $0 < Pr\{Z_i | c(\mathbf{X}_i) = c(\mathbf{x}^*), g(\mathbf{X}_i) = g(\mathbf{x}^*)\} < 1$.

The following proposition states that this assumption is sufficient for the existence of the weights in Problem (5) with the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ as $n \rightarrow \infty$.

Proposition 2. *Suppose Assumption 2* is satisfied and the expectation of $c(\mathbf{X})$ exist, then $P(w^* \text{ exists}) \rightarrow 1$ as $n \rightarrow \infty$.*

We leave the proof of this proposition to supplementary Section 1. When considering a finite

sample, the weights in Problem (3) exist if the convex hulls generated by $\{c(\mathbf{X}_i), g(\mathbf{X}_i)\}_{Z_i=1}$ and $\{c(\mathbf{X}_i), g(\mathbf{X}_i)\}_{Z_i=0}$ contain $\frac{1}{n} \sum_{i=1}^n \{c(\mathbf{X}_i), g(\mathbf{X}_i)\}$ (Zhao and Percival, 2017). In contrast, the weights in Problem (5) exist if 1) the convex hulls of $\{c(\mathbf{X}_i)\}_{Z_i=1}$ and $\{c(\mathbf{X}_i)\}_{Z_i=0}$ contain $\frac{1}{n} \sum_{i=1}^n c(\mathbf{X}_i)$; and 2) $\{c(\mathbf{X}_i), g(\mathbf{X}_i)\}_{Z_i=1} \cap \{c(\mathbf{X}_i), g(\mathbf{X}_i)\}_{Z_i=0} \neq \emptyset$. This is a substantial relaxation of the requirements for weight existence, especially for scenarios where one treatment group is much smaller than the other. Returning to Figure 1, the sample mean of $\{c(\mathbf{x}), g(\mathbf{x})\}$ is not in the intersection of the two convex hulls and thus there is no solution to Problem (3) as previously discussed. However, there are many observations in the intersection of the two convex hulls. Furthermore, when examining $c(\mathbf{x})$, its sample mean is within the observed range for both treatment groups. Thus, there is a solution to our proposed Problem (5), and in fact the corresponding weighted sample mean given by $\{w_i^*\}_{i=1}^n$ is within the intersection of the treated and control convex hulls.

The relaxed positivity assumption for Problem 5 is directly related to the propensity score model implied by Problem 5. Taking the expectation of population version of Equation (6) with respect to \mathbf{x} yields the following first order conditions,

$$\{e(\mathbf{x})\}^{-1} = \rho' \{ \boldsymbol{\alpha}_1^T c(\mathbf{x}) + \boldsymbol{\gamma}^T g(\mathbf{x}) \} = w_1^*, \quad (7a)$$

$$\{1 - e(\mathbf{x})\}^{-1} = \rho' \{ \boldsymbol{\alpha}_0^T c(\mathbf{x}) - \boldsymbol{\gamma}^T g(\mathbf{x}) \} = w_0^*, \quad \text{and} \quad (7b)$$

$$e(\mathbf{x}) = \frac{\rho' \{ \boldsymbol{\alpha}_0^T c(\mathbf{x}) - \boldsymbol{\gamma}^T g(\mathbf{x}) \}}{\rho' \{ \boldsymbol{\alpha}_0^T c(\mathbf{x}) - \boldsymbol{\gamma}^T g(\mathbf{x}) \} + \rho' \{ \boldsymbol{\alpha}_1^T c(\mathbf{x}) + \boldsymbol{\gamma}^T g(\mathbf{x}) \}} = \frac{w_0^*}{w_0^* + w_1^*}. \quad (7c)$$

Thus, the weights w_0^* and w_1^* correspond to IPW for a specific propensity score model, which we reference as the implied propensity score model of Problem (5). In this implied propensity score model, the dual variables $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}$ can be considered as the model coefficients. Furthermore, $\rho'(\cdot)$ may be regarded as the link function of the generalized linear propensity score model, where $D(\cdot)$ determines this link function (Wang and Zubizarreta, 2020).

Problem (5) restricts the coefficients corresponding to $g(\mathbf{x})$ in the models for $\{e(\mathbf{x})\}^{-1}$ and $\{1 - e(\mathbf{x})\}^{-1}$ to be opposites of each other. This ensures that $w_0^* = \{1 - w_1^{*-1}\}^{-1}$ such that only a single propensity score model is specified by Problem (5). However, there are substantial restrictions on the propensity score model that satisfies conditions (7a)-(7c). Specifically, $\text{Var}(Z|\mathbf{X} = \mathbf{x})$ is restricted to be a function of only $c(\mathbf{x})$ for specific dispersion measures (see supplementary Section 2 for further discussion). Since this generally does not hold when $E[Z|\mathbf{X} = \mathbf{x}]$ is a function of both $c(\mathbf{x})$ and $g(\mathbf{x})$, this model will only hold when $g(\mathbf{x})$ takes a specific functional form. However, it is explicitly through this restriction on the propensity score model that there is a solution to Problem (5) in scenarios with a lack of overlap, as overlap is improved through a *reduction* in the ability to predict treatment assignment (Clivio et al., 2024). Rather than attempt to specify a correct implied propensity

score model, Problem (5) imposes restrictions on the propensity score model to improve overlap and yield a solution to the optimization problem.

3.1.1 Connections to other methods

Problem (5) corresponds to a single propensity score model, while Problem (3) implies two incompatible propensity score models due to the separability of the optimization problem by treatment group. By implying a single propensity score model, the weights derived by our proposed method maintain the standard philosophical interpretation of IPW that direct balancing weights lack for the ATE. Furthermore, this allows for the doubly robust property (i.e., the estimator is consistent if either the implied propensity score model or outcome model is properly specified) and semiparametric efficiency of our method given the correct model specification (see Section 4 for further discussion of these properties). Minimal weights also correspond to two incompatible propensity score models but additionally place an L1 penalty on the dual variables (i.e., propensity score model coefficients). While the restrictions imposed on the implied propensity score model of Problem (5) are comparably less intuitive, by restricting only $\text{Var}(Z|\mathbf{X} = \mathbf{x})$ rather than both $E[Z|\mathbf{X} = \mathbf{x}]$ and $\text{Var}(Z|\mathbf{X} = \mathbf{x})$ our derived weights are able to achieve exact finite sample balance. Furthermore, in restricting $\text{Var}(Z|\mathbf{X} = \mathbf{x})$ our proposed method precisely restricts the aspect the propensity score model that primarily influences estimator variance.

When $c(\mathbf{x}) = \emptyset$ and $g(\mathbf{x}) = b(\mathbf{x})$ in Problem (5), Assumption 2* becomes: “There exists an \mathbf{x}^* such that $0 < \Pr\{Z_i|c(\mathbf{X}_i) = c(\mathbf{x}^*), g(\mathbf{X}_i) = g(\mathbf{x}^*)\} < 1$ ” which is comparable to the lack of a positivity assumption for overlap weights. Thus, Assumption 2* is an intermediate between Assumption 2 and no positivity assumption; as more covariate functions are added to the $g(\mathbf{x})$ set, Assumption 2* becomes more relaxed. To further connect overlap weights and Problem (5), when $c(\mathbf{x}) = \emptyset$ and $g(\mathbf{x}) = b(\mathbf{x})$ the first order conditions in (7) simplify to $e(\mathbf{x})/\{1 - e(\mathbf{x})\} = w_0^*/w_1^*$. While the exact solution to the dual problem will depend on the specific dispersion measure, overlap weights satisfy this condition. Therefore, Problem (5) with $c(\mathbf{x}) = \emptyset$ and $g(\mathbf{x}) = b(\mathbf{x})$ can be considered the direct balancing equivalent of overlap weights; however, Problem (5) will always yield exact balance while overlap weights only yields exact balance when the propensity score is estimated with logistic regression.

3.2 Interpretation of the proposed estimator

While Problem (5) yields a solution when there is a lack of overlap, in doing so the population of $g(\mathbf{x})$ is modified away from the ATE target population, potentially resulting in estimator bias with respect to the ATE (termed bias due to estimand mismatch in [Barnard et al. \(2025\)](#)). Thus, we provide model- and design- based perspectives on the conditions under which our proposed estimator has minimal error with respect to the ATE. In addition, we provide a characterization of the modified population and estimand that $\hat{\tau}_{w^*}$ targets when

$\hat{\tau}_w^*$ does not have minimal error with respect to the ATE. We first note that for any weighted estimator for the ATE,

$$\hat{\tau}_w - \tau = \frac{1}{n} \sum_{i=1}^n \{w_i Z_i \mu_{Z_i}(\mathbf{X}_i) - \mu_1(\mathbf{X}_i)\} - \frac{1}{n} \sum_{i=1}^n \{w_i (1 - Z_i) \mu_{Z_i}(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)\} \quad (8a)$$

$$- \int \{\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\} d\{F - F_n\}(\mathbf{x}) \quad (8b)$$

$$+ \frac{1}{n} \sum_{i=1}^n w_i \epsilon_i Z_i - \frac{1}{n} \sum_{i=1}^n w_i \epsilon_i (1 - Z_i), \quad (8c)$$

where $\epsilon_i = Y_i(Z_i) - \mu_{Z_i}(\mathbf{X}_i)$ and $F_n(\mathbf{x}) = \sum_{i=1}^n I(\mathbf{X}_i \leq \mathbf{x})/n$ is the empirical CDF. In this expression, term (8b) goes to zero for a representative sample of the population and term (8c) has expectation zero. Thus, term (8a) is the primary contributor to estimator error; we focus on this term for the remainder of this section.

To better understand the conditions under which our balancing equations are justified, consider the case where $\mu_z(\mathbf{x})$ is a linear function of $c(\mathbf{x})$ and $g(\mathbf{x})$ such that $\mu_1(\mathbf{x}) = \beta_1^T c(\mathbf{x}) + \lambda_1^T g(\mathbf{x})$ and $\mu_0(\mathbf{x}) = \beta_0^T c(\mathbf{x}) + \lambda_0^T g(\mathbf{x})$ for some $\beta_z \in \mathbb{R}^K$ and $\lambda_z \in \mathbb{R}^L$. When weights are derived through Problem (3), term (8a) will be zero for the corresponding estimator. However when weights are derived with Problem (5), term (8a) simplifies to $\frac{1}{n} \sum_{i=1}^n \lambda_1^T g(\mathbf{X}_i) - \lambda_0^T g(\mathbf{X}_i)$ such that term (8a) is zero when $\lambda_1 = \lambda_0$. Thus, term (8a) is zero when $\mu_z(\mathbf{x}) = \beta_z^T c(\mathbf{x}) + \lambda^T g(\mathbf{x})$ and $\tau(\mathbf{x}) = (\beta_1^T - \beta_0^T)c(\mathbf{x})$. These models imply that our proposed estimator has minimal error with respect to the ATE when the $c(\mathbf{x})$ set contains all treatment effect modifiers. Thus, as the set $c(\mathbf{x})$ gets smaller and correspondingly the set $g(\mathbf{x})$ gets larger, our proposed estimator requires a stronger assumption about the number of effect modifiers to be justified for the ATE. However, this is a relaxation of the assumption $\tau(\mathbf{x}) = \tau$ required for the overlap weights estimator to be consistent for the ATE.

We can also explore term (8a) for our proposed estimator from a design-based, rather than modeling, prospective; an alternative simplification of term (8a) is $(\lambda_1^T - \lambda_0^T) \left\{ \frac{1}{n} \sum_{i=1}^n w_i I[Z_i = z] \times g(\mathbf{X}_i) - \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) \right\}$. From a design-based perspective, this term tends to be small when $1^T \left| \frac{1}{n} \sum_{i=1}^n w_i I[Z_i = z] g(\mathbf{X}_i) - \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) \right|$ is small. However, in Problem (5) we explicitly do not constrain $\frac{1}{n} \sum_{i=1}^n w_i I[Z_i = z] g(\mathbf{X}_i) - \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) = 0$ in order to yield a solution when there is a lack of overlap. Thus, $1^T \left| \frac{1}{n} \sum_{i=1}^n w_i I[Z_i = z] g(\mathbf{X}_i) - \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) \right|$ will be non-zero when weights are derived using Problem (5). Despite this, the error of our proposed estimator with respect to the ATE is due to differences between the weighted and sample population of $g(\mathbf{x})$, rather than the imbalance between the weighted treated and control populations. Thus, when $c(\mathbf{x})$ does not contain all treatment effect modifiers, our proposed estimator has minimal error with respect to the average treatment effect estimand over a modified population of $g(\mathbf{x})$, which can be characterized by $\frac{1}{n} \sum_{i=1}^n w_i^* Z_i g(\mathbf{X}_i) = \frac{1}{n} \sum_{i=1}^n w_i^* (1 - Z_i) g(\mathbf{X}_i)$.

While the ATO population can also be characterized through weighted sample means, it can be challenging to conceptualize this modified population especially with high-dimensional covariates. We provide further guidance and discussion on how to select $g(\mathbf{x})$ given these model- and design-based perspectives in Section 5 and supplementary Section 5.1.

3.3 Related estimation problems

Our proposed constrained optimization problem for deriving weights for ATE estimation can be readily extended to variety of additional estimation problems. We discuss extensions to the ATT and transporting effects below. Extensions to effect estimation with more than two treatment groups, weighted average treatment effects, and distributional balancing weights are discussed in supplementary Section 3.

3.3.1 ATT

When deriving either IPW or direct balancing weights for the ATT, only weights for the control group are derived such that the corresponding estimator is $\frac{1}{n} \sum_{i=1}^n Y_i - \frac{1}{n} \sum_{i=1}^n w_i(1 - Z_i)Y_i$. However, when adapting Problem (5) for ATT estimation, weights must be derived for both the treated and control group. To derive these weights, $\{w_i^{*ATT}\}$, we can simply replace the right-hand side of constraint (5b) with $\frac{1}{n} \sum_{i=1}^n Z_i c_k(\mathbf{X}_i)$ and compute Equation (4) with these weights. The design-based and model-based interpretations of the sets $c(\mathbf{x})$ and $g(\mathbf{x})$ remain the same as described in Section 3.2. However, the covariates that are effect modifiers may differ between the entire sample population and the treated population.

3.3.2 Transporting effects

While clinical trials yield “gold standard” causal effect estimates through randomization, trial inclusion/exclusion criteria is often restrictive such that the trial population may not reflect a clinically relevant population. In these cases, it is desirable to transport the effect to a more meaningful population. However, there is often a lack of overlap between the trial population and this more meaningful target population. Our procedure allows the analyst to improve overlap by modifying the target population for only a subset of covariates, rather than modifying all covariates or discarding observations, which is inconsistent with the goal of estimating the effect in a clinically relevant population. To adapt Problem 5 for transportability, we first introduce some additional notation. Let $R_i = r$ be an indicator of whether an observation belongs to the trial population ($r = 1$) or the target population ($r = 0$). We only observe Y_i and Z_i when $R_i = 1$ but observe \mathbf{X}_i for both populations. Let n_r indicate the sample size in population r . Given additional assumptions outlined in Dahabreh et al. (2020), the transported effect estimand is identifiable with observed data. We propose

the following modification of Problem (5) for transportability,

$$\{w_i^{*t}\}_{i=1}^n = \operatorname{argmin}_{w_i} \sum_{i=1}^n D(w_i) \quad \text{subject to} \quad (9a)$$

$$\frac{1}{n_1} \sum_{i=1}^n w_i I[Z_i = z] R_i c_k(\mathbf{X}_i) = \frac{1}{n_0} \sum_{i=1}^n (1 - R_i) c_k(\mathbf{X}_i), \quad k = 1, \dots, K, \quad z = 0, 1 \quad (9b)$$

$$\frac{1}{n_1} \sum_{i=1}^n w_i Z_i R_i g_l(\mathbf{X}_i) = \frac{1}{n_1} \sum_{i=1}^n w_i (1 - Z_i) R_i g_l(\mathbf{X}_i), \quad l = 1, \dots, L \quad (9c)$$

Then, our proposed estimator for the transported effect is $\tau_{w^{*t}} = \frac{1}{n} \sum_{i=1}^n w_i^{*t} Z_i R_i Y_i - \frac{1}{n} \sum_{i=1}^n w_i^{*t} (1 - Z_i) R_i Y_i$. The interpretation of $c(\mathbf{x})$ and $g(\mathbf{x})$ remains the same as described in Section 3.2, however the selection of $g(\mathbf{x})$ may be of even more importance in order to maintain a target population of interest.

4 Asymptotic properties of the proposed estimator

We derive asymptotic results for the solution to Equation (6) which we then use to develop asymptotic results for our proposed estimator, $\hat{\tau}_{w^*}$, under various conditions. All proofs of results in this section are in supplementary Section 1. We first require additional regularity conditions, similar to those in Hirano et al. (2003); Chan et al. (2016) and Källberg and Waernbaum (2023), on the set balancing functions, $b(\mathbf{x})$.

Assumption 3: Balance function conditions. *There exists a finite M , such that $|b_j(\mathbf{X})| < M$ for $j = 1, \dots, J$ almost surely and $B(\mathbf{X}) = [b_1(\mathbf{X}), \dots, b_J(\mathbf{X})]$ has full rank.*

For the remainder of this section, we notate w_i^* as $w^*(\mathbf{X}_i)$ to explicitly show the relationship to \mathbf{X}_i . Given Assumptions 1, 2*, and 3, $w_1^*(\mathbf{X}_i)$ and $w_0^*(\mathbf{X}_i)$ are consistent for their asymptotic counterparts, $\tilde{w}_1(\mathbf{X}_i)$ and $\tilde{w}_0(\mathbf{X}_i)$, which uniquely solve the population version of Equation (6) (supplementary Section 1, Proposition 5); a majority of the asymptotic results follow from this fact. Since Problem (5) implies a single propensity score model, our proposed estimator is doubly robust as stated below in Theorem 3.

Theorem 3. *Suppose Assumptions 1 and 3 hold and $\rho''\{\cdot\}$ is continuous over \mathbb{R} . Assume that the expectation of $c(\mathbf{X})$ exists and that $\operatorname{Var}\{Y(a)\} < \infty$ for $a = 0, 1$. Then, $\hat{\tau}_{w^*}$ is doubly robust in the sense that if either 1) Assumption 2 holds and $e(\mathbf{x})$ satisfies conditions (7a)-(7c) or; 2) Assumption 2* holds, $\mu_z(\mathbf{x})$ is linear in $c(\mathbf{x})$ and $g(\mathbf{x})$, and $\tau(\mathbf{x})$ is linear in $c(\mathbf{x})$, then $\hat{\tau}_{w^*}$ is consistent for the ATE.*

Note that $\rho''\{\cdot\}$ is continuous for common dispersion measures such as w^2 and $w \log(w)$. In contrast, $\hat{\tau}_{w^{bw}}$ is not doubly robust (Chan et al., 2016). Since our estimator is motivated by scenarios where positivity does not hold, the second condition that guarantees consistency

(i.e., $\tau(\mathbf{x})$ is linear in only $c(\mathbf{x})$) is of primary importance. Let $\tau_{g,\tilde{w}}(\mathbf{x}, z) = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T c(\mathbf{x}) + \boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{x}) z g(\mathbf{x}) - \boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{x}) (1 - z) g(\mathbf{x})$ for some $\boldsymbol{\beta}_z \in \mathbb{R}^K$ and $\boldsymbol{\lambda}_z \in \mathbb{R}^L$. When $\tau(\mathbf{x})$ is linear in both $c(\mathbf{x})$ and $g(\mathbf{x})$, our proposed estimator is consistent for the modified estimand $\tau_{g,\tilde{w}} = E[E[Y|Z = 1, c(\mathbf{X}), \tilde{w}_1(\mathbf{X})e(\mathbf{X})g(\mathbf{X})]] - E[E[Y|Z = 0, c(\mathbf{X}), \tilde{w}_0(\mathbf{X})\{1 - e(\mathbf{X})\}g(\mathbf{X})]] = E[\tau_{g,\tilde{w}}(\mathbf{X}, Z)]$ as stated in Corollary 3.1. Note that $\tau_{g,\tilde{w}}$ is indeed a causal estimand because $E[\tilde{w}_1(\mathbf{X})e(\mathbf{X})g(\mathbf{X})] = E[\tilde{w}_0(\mathbf{X})\{1 - e(\mathbf{X})\}g(\mathbf{X})]$ is a first order condition of the population version of Equation (6).

Corollary 3.1. *Suppose Assumptions 1 and 2* hold and $\rho''\{\cdot\}$ is continuous over \mathbb{R} . Assume that the expectation of $c(\mathbf{X})$ exists and that $\text{Var}\{Y(a)\} < \infty$ for $a = 0, 1$. Then, $\hat{\tau}_{w^*}$ is consistent for $\tau_{g,\tilde{w}}$ if $e(\mathbf{x})$ does not satisfy conditions (7a)-(7c) and $\mu_z(\mathbf{x})$ and $\tau(\mathbf{x})$ are both linear in $c(\mathbf{x})$ and $g(\mathbf{x})$.*

When positivity does hold and all models implied by Problem 5 are correctly specified, our proposed estimator is semiparametrically efficient as stated in Theorem 4.

Theorem 4. *Suppose Assumptions 1, 2, and 3 hold and $\text{Var}\{Y(a)|\mathbf{X} = \mathbf{x}\} < \infty$ for all \mathbf{x} and $a = 0, 1$. Suppose $e(\mathbf{x})$ satisfies conditions (7a)-(7c), $\mu_z(\mathbf{x})$ is linear in $c(\mathbf{x})$ and $g(\mathbf{x})$, and $\tau(\mathbf{x})$ is linear in $c(\mathbf{x})$. Then,*

$$\sqrt{n}(\hat{\tau}_{w^*} - \tau) \rightarrow N(0, V_{opt})$$

where $V_{opt} = E \left[\frac{\text{Var}(Y(1)|\mathbf{X})}{e(\mathbf{X})} + \frac{\text{Var}(Y(0)|\mathbf{X})}{1-e(\mathbf{X})} + \{\tau(\mathbf{X}) - \tau\}^2 \right]$ is the semiparametric efficiency bound (Hahn, 1998).

However, it is rare that the implied propensity score model is correctly specified and our proposed estimator is motivated by scenarios where Assumption 2 does not hold. Thus, we derive and present the asymptotic variance of our estimator when only the outcome and models are correctly specified in Corollary 4.1.

Corollary 4.1. *Suppose Assumptions 1, 2*, and 3 hold and $\text{Var}\{Y(a)|\mathbf{X} = \mathbf{x}\} < \infty$ for all \mathbf{x} and $a = 0, 1$. Suppose $e(\mathbf{x})$ does not satisfy conditions (7a)-(7c), $\mu_z(\mathbf{x})$ is linear in $c(\mathbf{x})$ and $g(\mathbf{x})$, and $\tau(\mathbf{x})$ is linear in $c(\mathbf{x})$. Then,*

$$\sqrt{n}(\hat{\tau}_{w^*} - \tau) \rightarrow N(0, V_{opt}^*)$$

where $V_{opt}^* = E[\tilde{w}_1(\mathbf{X})^2 e(\mathbf{X}) \text{Var}\{Y(1)|\mathbf{X}\} + \tilde{w}_0(\mathbf{X})^2 \{1 - e(\mathbf{X})\} \text{Var}\{Y(0)|\mathbf{X}\} + \{\tau(\mathbf{X}) - \tau\}^2]$.

Of key importance, this variance result only requires the relaxed positivity assumption, rather than the standard positivity assumption. However, to compare V_{opt} and V_{opt}^* , consider the

case where positivity does hold. Since the asymptotic weights tend to be less extreme (i.e., $\tilde{w}_1(\mathbf{X}_i) \leq \{e(\mathbf{X}_i)\}^{-1}$ and $\tilde{w}_0(\mathbf{X}_i) \leq \{1 - e(\mathbf{X}_i)\}^{-1}$) than IPW through the restriction on the implied propensity score model, $V_{opt}^* \leq V_{opt}$. Thus, our proposed estimator does indeed have reduced asymptotic variance, especially in scenarios where $e(\mathbf{x})$ is close to zero or positivity does not hold. Specifically, as the $g(\mathbf{x})$ set gets larger, $\tilde{w}_z(\mathbf{x})$ become less extreme and correspondingly V_{opt}^* becomes smaller. We further derive the asymptotic variance of our proposed estimator with respect to the modified estimand, $\tau_{g,\tilde{w}}$, when $\tau(\mathbf{x})$ is a linear function of both $c(\mathbf{x})$ and $g(\mathbf{x})$ in Corollary 4.2.

Corollary 4.2. *Suppose Assumptions 1, 2*, and 3 hold and $\text{Var}\{Y(a)|\mathbf{X} = \mathbf{x}\} < \infty$ for all \mathbf{x} and $a = 0, 1$. Assume that $\rho''\{\cdot\}$ and $\rho'''\{\cdot\}$ are continuous over \mathbb{R} . Suppose $e(\mathbf{x})$ does not satisfy conditions (7a)-(7c) and $\mu_z(\mathbf{x})$ and $\tau(\mathbf{x})$ are linear in both $c(\mathbf{x})$ and $g(\mathbf{x})$. Then,*

$$\sqrt{n}(\hat{\tau}_{w^*} - \tau_{g,\tilde{w}}) \rightarrow N(0, V_{g,\tilde{w}})$$

where $V_{g,\tilde{w}} = E[\tilde{w}_1(\mathbf{X})^2 e(\mathbf{X}) \text{Var}\{Y(1)|\mathbf{X}\} + \tilde{w}_0(\mathbf{X})^2 \{1 - e(\mathbf{X})\} \text{Var}\{Y(0)|\mathbf{X}\} + \{\tau_{g,\tilde{w}}(\mathbf{X}, Z) - \tau_{g,\tilde{w}} + r_1(\mathbf{X}, Z) - r_0(\mathbf{X}, Z)\}^2]$.

The functions $r_z(\mathbf{x}, z)$ account for the difference between the estimated and asymptotic weights. To avoid introducing substantial additional notation, we present the exact forms of these functions in the proof of Corollary 4.2 in supplementary Section 1. Note that the additional assumptions on $\rho\{\cdot\}$ hold for common dispersion measures such as w^2 and $w \log(w)$. To compare to Corollary 4.1, consider the case where the assumptions of Corollary 4.2 hold and the direct balancing weights estimator is computed. In this case, the asymptotic variance of the direct balancing estimator takes the same form as V_{opt}^* but with the asymptotic counterparts of direct balancing weights; we will reference this variance as V_{opt}^{bw} . Our proposed asymptotic weights are less extreme than the asymptotic direct balancing weights, however, $\text{Var}\{\tau_{g,\tilde{w}}(\mathbf{X}, Z)\} \gtrsim \text{Var}\{\tau(\mathbf{X})\}$ as $\text{Var}\{w_z(\mathbf{X})I[Z = z]g(\mathbf{X})\} \gtrsim \text{Var}\{g(\mathbf{X})\}$. Thus, there is an unclear relationship between $V_{g,\tilde{w}}$ and V_{opt}^{bw} ; in some scenarios $V_{g,\tilde{w}} > V_{opt}^{bw}$ while in others $V_{g,\tilde{w}} < V_{opt}^{bw}$. To further explore this relationship, we perform a simulation study to compare the asymptotic variance of these estimators (supplementary Section 4). When overlap is poor and there is low to moderate treatment effect heterogeneity, $V_{g,\tilde{w}}$ tends to be smaller than V_{opt}^{bw} (supplementary Figure 1). With the addition of the constraints $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ to both optimization problems, the direct balancing weights are even more extreme such that $\text{Var}(\hat{\tau}_{w^*}) \lesssim \text{Var}(\hat{\tau}_{w^{wb}})$ across a majority of data generating scenarios (supplementary Figure 3); in fact, when overlap is poor $\text{Var}(\hat{\tau}_{w^*}) \ll \text{Var}(\hat{\tau}_{w^{wb}})$.

5 Implementation details

To select the optimal $g(\mathbf{x})$ set, in terms of error with respect to the ATE, one approach is to examine all possible covariate function sets and select the set that 1) yields a solution to Problem (5); and 2) minimizes term (8a). However, examining all possible covariate subsets is computationally prohibitive, especially as covariate dimension increases. Therefore, we propose a greedy algorithm (Algorithm 1) for identifying an approximately optimal $g(\mathbf{x})$ subset according to a specific metric aligned with either the model- or design-based perspective on estimator error. We also propose an additional algorithm (supplementary Algorithm 1) that is less computationally intensive than Algorithm 1; we found these algorithms resulted in estimators with similar performance when there was no correlation between covariates (supplementary Figure 4). However, we generally recommend the use of Algorithm 1 if it is not computationally prohibitive.

Algorithm 1 Identifying an approximately optimal $g(\mathbf{X})$ set given metric m

- 1: **Inputs:** The matrix $\mathbf{X}_{n \times p}$, set of covariate functions $b_1(\mathbf{X}), \dots, b_J(\mathbf{X})$, and a metric m that can be computed for each $b_j(\mathbf{X})$
 - 2: **Initialize:** $b(\mathbf{X})^- = \emptyset$
 - 3: **while** $\{w_i^*\}_{i=1}^n$ given $g(\mathbf{X}) = b(\mathbf{X})^-$ does not exist **do**
 - 4: Compute m for $\{b_1(\mathbf{X}), \dots, b_j(\mathbf{X})\} \setminus b(\mathbf{X})^-$ and identify $b^*(\mathbf{X})$, the covariate with the minimum or maximum value of the metric m
 - 5: Let $b(\mathbf{X})^- = \{b(\mathbf{X})^-, b^*(\mathbf{X})\}$
 - 6: Solve Problem (5) with $g(\mathbf{X}) = b(\mathbf{X})^-$ and $c(\mathbf{X}) = \{b_1(\mathbf{X}), \dots, b_j(\mathbf{X})\} \setminus b(\mathbf{X})^-$
 - 7: **end while**
 - 8: **Outputs:**
 - 9: $b(\mathbf{X})^-$ \triangleright Smallest set that yields a solution to Problem (5) when $g(\mathbf{X}) = b(\mathbf{X})^-$
 - 10: $\{w_i^*\}_{i=1}^n$ \triangleright Resulting weights from Problem (5) with $g(\mathbf{X}) = b(\mathbf{X})^-$
-

From a design-based perspective, selecting the covariate functions that are each conditionally the most predictive of treatment will tend to result in the smallest $g(\mathbf{x})$ set that yields a solution to Problem (5). While selecting this set may not minimize the design-based simplification of estimator error, this set does tend to result in smaller design-based estimator error and simpler estimator interpretation. Thus, for the design-based metric, we propose and implement Spearman semipartial correlation; in step four, the maximum correlation is used to identify $b^*(\mathbf{x})$. We choose this metric because it is rank-based and reflects the impact of removing a covariate in a hypothetical propensity score model, however other metrics could work well here. Our design-based estimator then takes the same form as Equation (4) with weights computed with the $g(\mathbf{x})$ set selected by Algorithm 1.

From a model-based perspective, selecting the covariate functions that do not modify the

treatment effect as $g(\mathbf{x})$ will tend to minimize estimator error. Thus, we implement the absolute value of the one-step doubly robust estimator of treatment effect modification, proposed by [Boileau et al. \(2025\)](#), as the model-based metric to use within Algorithm 1; in step four, the minimum of this metric is used to identify $b^*(\mathbf{x})$. Our model-based estimator is a cross-fit style estimator to preserve downstream inference; we provide further discussion of this model-based metric and estimator in supplementary Section 5.2. For either metric/estimator, Algorithm 1 can be initiated with a non-empty $b(\mathbf{x})^-$ set (step 2) if there are covariates that one wants to include in the $g(\mathbf{x})$ set *a priori*. If minimizing error with respect to the ATE is not a priority, $g(\mathbf{x})$ could be chosen such that the modified $g(\mathbf{x})$ population is of scientific interest which we discuss further in supplementary Section 5.1.

6 Simulation experiments

We generate a variety of simulation scenarios that vary in the number of confounders (i.e., covariates), level of overlap, level of treatment effect heterogeneity, and proportion of covariates that are effect modifiers. Through these scenarios, we first aim to validate the estimator corresponding to the weights derived through Problem (5) when the true $\tau(\mathbf{x})$ function is known. In addition, we seek to determine in which data generating scenarios the proposed methods for a completely data-driven selection of $g(\mathbf{x})$ in Section 5 yield estimators with improved performance when compared to IPW, minimal weights, and overlap weights estimators when there is no solution to Problem (3).

6.1 Data generation

For clarity in presentation, we drop the observation index on the variables. For $p = 20$ and $p = 100$ we simulate $V_1 - V_p$ from a truncated normal distribution where $E[V_i] = 0$, $\text{Var}(V_i) = 1$ and $\text{Cov}(V_i, V_j) = 0$ for all $i = 1, \dots, p$ and $i \neq j$. To obtain a mix of continuous and categorical coefficients, we take $X_1 - X_{p/2} = V_1 - V_{p/2}$ and $X_j = I[V_j < 0]$ for $j = p/2 + 1, \dots, p$. We generate the true propensity scores with a logistic model, $e(\mathbf{X}) = \{1 - \exp(-\sum_{i=0}^p \alpha_i)\}^{-1}$. In this model, α_0 is selected to obtain the desired percent of treated observations (20%, 40%) and all other coefficients are determined as described in supplementary Section 6. We simulate treatment assignment as $Z \sim \text{Bernoulli}\{e(\mathbf{X})\}$ and the outcome as $Y = \mu_z(\mathbf{X}) + N(0, 1)$ for $Z = z$. Both $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ are modeled as linear combinations of X_1, \dots, X_p . For $\theta = 25\%$, 75% of covariates, the corresponding $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ coefficients are equal (i.e., these covariates are not effect modifiers). The remaining $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ coefficients are determined as described in supplementary Section 6.

We generate a total of 72 simulation scenarios of all combinations $p = 20, 100$, 20% and 40% treated, 25% and 75% of covariates being effect modifiers, three levels of overlap, and three levels of treatment effect heterogeneity. For each simulation scenario we generate 1,000

independent datasets with either $n = 1000$ ($p = 20$) or $n = 2000$ ($p = 100$) observations. We compute 1) IPW, direct balancing weights, and minimal weights estimators for the ATE; 2) our proposed estimator given the true $g(\mathbf{X})$ and the design- and model-based estimators proposed in Section 5; and 3) the IPW and balancing weights (i.e., all covariates in the set $g(\mathbf{x})$) estimators for the ATO. All balancing weights are computed with the stable balancing weights deviance measure (Zubizarreta, 2015) and with the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$. We compare the mean squared error (MSE) of these estimators with respect to the true ATE.

6.2 Results

Our proposed estimator implemented with the true $\tau(\mathbf{x})$ function has the minimum estimator MSE with respect to the ATE across almost all simulation scenarios where there is a solution to Problem (5) for the true $\tau(\mathbf{x})$ function (Figure 2 and supplementary Figure 7). Specifically, this estimator tends to be minimally biased and have comparable variance to the minimal weights and ATO estimators (supplementary Figures 8-11). Furthermore, when there is a lack of overlap and no solution to Problem (3), our proposed estimator is the only estimator that has minimal bias for the ATE. However, it is rare that the true $\tau(\mathbf{x})$ is known, so we focus on the performance of our proposed design- and model-based estimators that use a data-driven procedure to select $g(\mathbf{x})$.

Across almost all simulation scenarios, either our proposed design- or model-based estimator has minimum estimator MSE with respect to the ATE compared to either the minimal weights or ATO estimators (Figure 2 and supplementary Figure 7). Furthermore, the design-based estimator is the only estimator that has 1) a solution to Problem (5); and 2) lower MSE than the IPW ATE estimator across all scenarios. The minimal weights estimator tends to have the largest MSE, even larger than the ATE and ATO estimators, when there is moderate-high treatment effect heterogeneity and low overlap. When there is low treatment effect heterogeneity, the minimal weights estimator tends to have the smallest MSE. However, this is due to the estimator achieving minimal variance but possessing substantial statistical bias; thus, our proposed estimators may still be preferable in these scenarios (supplementary Figure 8). Our model-based estimator tends to perform better than the design-based estimator for $p = 20$, while the reverse is true when $p = 100$. We likely see this trend because there may not be sufficient sample size for estimating treatment effect modification and computing the weights with sample splitting for $p = 100$. We see similar trends in estimator MSE between the 20% treated (Figure 2) and 40% treated scenarios (supplementary Figure 7), though for the 40% treated scenarios all estimators substantially outperform the IPW estimator.

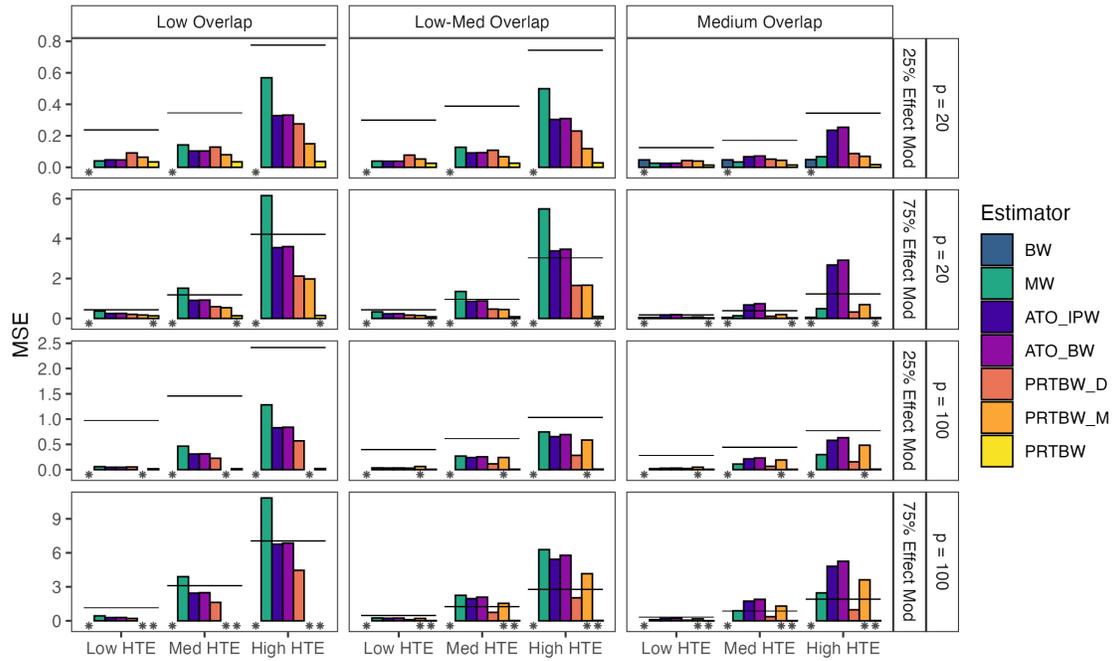


Figure 2: Estimator mean squared error (MSE) for all 20% treated simulation scenarios. The line indicates the MSE for the IPW ATE estimator for each scenario. The asterisk, *, indicates estimators where a solution to the balancing optimization problem did not exist for all datasets. The horizontal axis indicates levels of treatment effect heterogeneity while each panel is labeled by different overlap levels, number of predictors, and percent of covariates that are treatment effect modifiers.

7 Two illustrative applications

We apply our proposed weighted estimators for ATE estimation to an EHR study and adapt Problem (5) for transportability to transport a treatment effect from an RCT to a Midwestern academic health center population. We compare our proposed estimators to IPW, direct balancing, and minimal weights estimators for the ATE and ATO and demonstrate their superior performance in terms of covariate balance and estimator variance for these two applications.

7.1 EHR study on the effect of indwelling arterial catheters on mortality

We perform a reanalysis of [Hsu et al. \(2015\)](#) which estimates the ATE of indwelling arterial catheters (IACs) on mortality in patients with respiratory failure using observational data from the MIMIC-III v1.4 critical care database ([Johnson et al., 2016](#)). The data has 2,522 observations of mechanically ventilated patients, where 51.5% of patients received IAC. The outcome of interest is an indicator of mortality 28 days within hospital admission. Missing data is imputed with single imputation. Pre-treatment covariates include demographic information, baseline measurements (e.g., blood pressure, lab values), risk scores, and missing data indicators. In total, we identify 72 pre-treatment covariates of interest. There is a lack of overlap in the estimated propensity scores in this data with many propensity scores close to 1 (supplementary Figure 12). Due to this lack of overlap, there is no solution to Problem (3) with the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$. We estimate the ATE with IPW, minimal weights, and our proposed design- and model-based estimators that use Algorithm (1) to select $g(\mathbf{x})$. Since rare binary covariates tend to be challenging to exactly balance, we also compute our proposed estimators when initializing Algorithm (1) with the set of binary covariates whose outcomes occur for $< 5\%$ of observations. We also estimate the ATO using IPW and balancing weights (i.e., all covariates are in the $g(\mathbf{x})$ set). All balancing weights are computed with the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$. We compare the covariate balance, treatment effect estimates, and Wald-type bootstrapped 95% confidence intervals (CIs) between all methods.

The treatment effect estimates and 95% CIs are similar for all estimators except the IPW estimator, which corresponds to a more negative estimate and a larger 95% CI (Figure 3). However, all estimates and 95% CIs indicate that IAC non-significantly reduces 28 day mortality by a small percentage. The fact that the minimal weights, overlap weights, and PRTBW estimates are similar likely indicates that there is limited treatment effect heterogeneity. Thus, these estimators primarily differ in their target populations and the covariate balance achieved. Figure 3 shows the weighted treated and control sample means for three covariates in comparison to the sample mean (the black horizontal line). IPW and

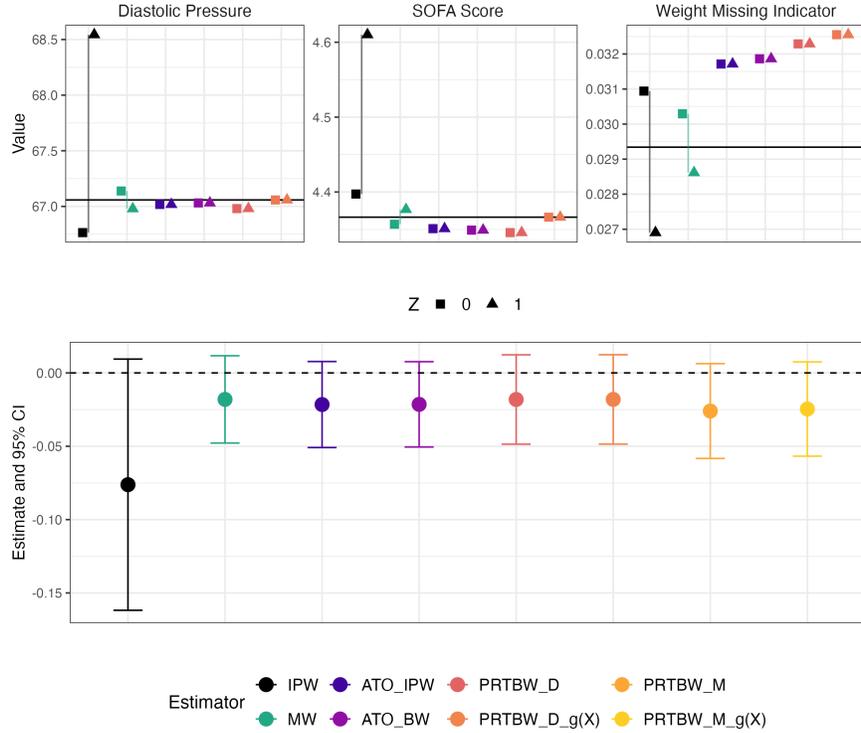


Figure 3: ATE estimates and 95% confidence intervals of the effect of IAC on 28-day mortality (bottom figure) and weighted covariate balance of three covariates (top figures) for eight different weighting/estimation methods. For the top figures, the horizontal line indicates the sample mean of the covariates in the target population and vertical lines indicate the extent of covariate imbalance for each method.

minimal weights do not exactly balance any of the covariates, with corresponding average standardized mean differences (SMD) of 0.161 and 0.008, respectively. In contrast, all other weights exactly balance the treated and control distributions. The smallest $g(\mathbf{x})$ set is selected for our method when Algorithm 1 is initiated with rare binary covariates. In Figure 3, the corresponding design-based weights balance both diastolic pressure and the SOFA score to the target population, which is not true of any of the other weights. Supplementary Figures 13-15 show the SMD for all covariates 1) between treated and control groups; and 2) between treated/control groups and the target population. While these estimators yield similar estimates and 95% CIs for this application, our proposed estimators tend to have improved covariate balance and correspond to a minimally modified target population.

7.2 Transporting a “health care hotspotting” RCT to a Midwestern U.S. academic health center population

A recent intervention of popular interest, termed “health care hotspotting”, aimed to improve healthcare delivery for individuals who interact substantially with the healthcare system through intensive, targeted follow-up after admission. Despite the promising nature of this

intervention, an RCT in Camden, NJ found a null effect of this intervention on readmission rates (Finkelstein et al., 2020). However, there is evidence that the treatment effect varies by population (Yang et al., 2023) such that we are interested in determining the effect of this intervention in a Midwestern U.S. academic health center population. We perform a complete case analysis (< 30 missing observations) with a total of 781 trial population observations and 1305 target population observations. The outcome of interest is a readmission within 30 days. We include all common pre-treatment covariates between both datasets, which includes demographics and hospital stay characteristics, for a total of 21 covariates. There is minimal overlap in the estimated probability of being in the target population primarily due to race; the trial population is 84.9% non-white while the target population is 13.6% non-white (supplementary Figure 16). Thus, we implement our proposed design-based estimator and our weighted estimator when $g(\mathbf{x})$ *a priori* only includes race. We compare this estimator to the normalized inverse odds weighted (IOW) estimator (Dahabreh et al., 2020), the IOW estimator with population probabilities outside of [0.1, 0.9] trimmed (Crump et al., 2009), and the direct balancing weights estimator. All balancing weights are computed with the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$. We report covariate balance, transported treatment effect estimates, and Wald-type bootstrapped 95% CIs for these methods.

The IOW and direct balancing weights estimates and 95% CIs indicate a non-significant reduction in 30-day readmission (Figure 4); however, there is no solution to the direct balancing weights optimization problem for many of the bootstrapped datasets such that a true 95% bootstrapped CI cannot be derived. The trimmed estimator has the largest treatment effect estimate, but also the largest 95% CI, likely due to the small sample size after trimming. While the 95% CIs for both of our proposed estimators include zero, the intervals are substantially smaller and the estimates indicate that the intervention may actually increase hospital readmission rates in the target population. Both the IOW and trimmed estimators achieve poor covariate balance with average SMDs of 0.140 and 0.382. Supplementary Figures 17-19 show the SMD for all covariates 1) between treated and control groups; and 2) between treated/control groups and the target population. Direct balancing weights and our proposed weights achieve exact covariate balance; however, our proposed weights do modify the sample mean of the race covariate such that the estimate applies to a population that is approximately 80% non-white. Therefore, the PRTBW estimates indicate that the intervention may in fact increase 30-day readmission rates for either 1) this slightly modified target population; or 2) for the original target population when race is not a treatment effect modifier.

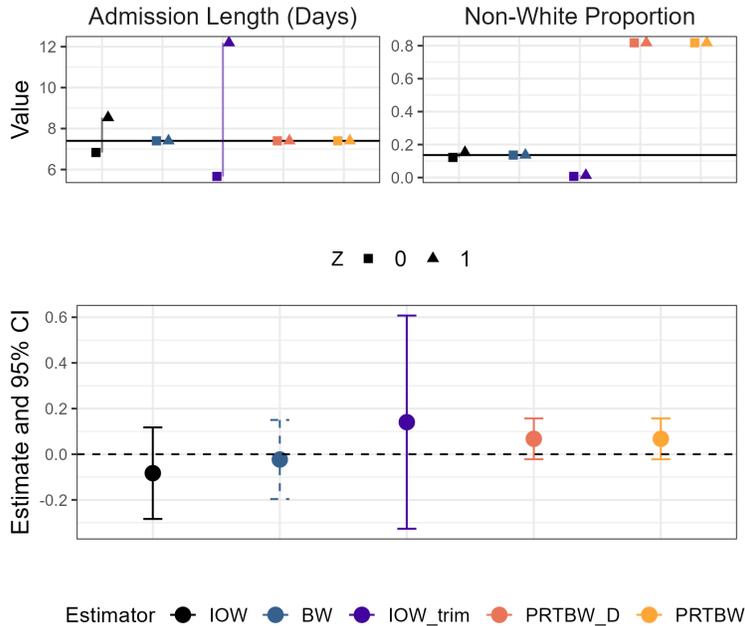


Figure 4: Transported treatment effect estimates and 95% confidence intervals of the effect of “health care hotspotting” on 30-day readmission rates for a Midwestern healthcare center population (bottom figure) and weighted covariate balance of two covariates (top figures) for five different weighting/estimation methods. For the top figures, the horizontal line indicates the sample mean of the covariates in the target population and vertical lines indicate the extent of covariate imbalance for each method.

8 Discussion

Positivity violations create substantial challenges when estimating causal effects with observational data. These challenges necessitate researchers to confront the trade-offs between estimator bias, variance, and maintaining the original target population of interest. In this work, we proposed a novel weighting procedure that achieves reduced estimator bias and variance through relaxing the balancing constraints to the target population. In doing so, our proposed weights can be derived under an relaxed positivity assumption. Our approach offers a choice in between direct balancing and overlap weights (which are special cases of our method), allowing the analyst to navigate between preservation of the original estimand and mitigation of positivity violations. We have shown that our proposed estimator 1) is consistent for the original estimand when either the implied propensity score model is correct or the set of treatment effect modifiers is properly specified; and 2) is consistent for a slightly modified estimand that is simple to characterize when these conditions do not hold. Furthermore, we have shown that our proposed estimator achieves reduced asymptotic variance under the relaxed positivity assumption. To aid in the practical use of our method, we proposed an algorithm for identifying the set of covariates balanced to a modified population; the corresponding design- and model-based estimators perform well across applications to

synthetic data, EHR data, and when transporting RCT effects.

While we focus on estimating the ATE, our proposed weighting procedure can be readily extended to other causal effect estimation problems as discussed in Section 3.3 and supplementary Section 3. Our proposed extension to distributional balancing weights may be especially promising as these weights do not require strong parametric assumptions on the outcome models. Future work includes exploring the theoretical properties of these extensions. There is also potential for combining the intuition behind our proposed method and modified treatment policies to obtain estimators with reduced bias and variance that correspond to both a treatment and population of interest. While our weighting procedure does perform well when there are positivity violations and higher dimensional covariates (e.g, $p = 100$), there is a need for additional methods that address the combined challenges of variable selection (in the context of unmeasured confounding) and positivity violations in $p \gg n$ scenarios. Further, while the infeasibility of Problem (3) is a clear indication that our proposed method or minimal weights need to be used to derive weights, it is generally unclear when estimator performance may be impacted by overlap levels for a given research problem. Thus, future work includes developing measures of overlap that are directly related the impact of overlap on estimator MSE.

Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2237827.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability Statement

The data that supports Section 7 are 1) MIMIC-III data: available with queries provided at <https://github.com/MIT-LCP/mimic-code>; 2) RCT data: available at <https://doi.org/10.7910/DVN/ZJVVQZ>; and 3) Health system target population data: not publicly available.

References

- Barnard, M., Huling, J. D., and Wolfson, J. (2025). A Unified Framework for Causal Estimand Selection. arXiv:2410.12093 [stat].
- Boileau, P., Leng, N., Hejazi, N. S., van der Laan, M., and Dudoit, S. (2025). A nonparametric framework for treatment effect modifier discovery in high dimensions. Journal of the Royal

Statistical Society Series B: Statistical Methodology, 87(1):157–185.

- Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally Efficient Non-Parametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(3):673–700.
- Clivio, O., Bruns-Smith, D., Feller, A., and Holmes, C. C. (2024). Towards Principled Representation Learning to Improve Overlap in Treatment Effect Estimation.
- Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96(1):187–199.
- Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. Statistics in Medicine, 39(14):1999–2014. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8426>.
- D’Amour, A. and Franks, A. (2021). Deconfounding Scores: Feature Representations for Causal Effect Estimation with Weak Overlap. arXiv:2104.05762 [stat].
- Finkelstein, A., Zhou, A., Taubman, S., and Doyle, J. (2020). Health Care Hotspotting - A Randomized, Controlled Trial. The New England Journal of Medicine, 382(2):152–162.
- Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. Econometrica, 66(2):315–332.
- Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. Political Analysis, 20(1):25–46.
- Hernan, M. A. and Robins, J. M. (2024). Causal Inference: What If. CRC Press.
- Hirano, K. and Imbens, G. W. (2001). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services and Outcomes Research Methodology, 2(3):259–278.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica, 71(4):1161–1189.
- Hsu, D. J., Feng, M., Kothari, R., Zhou, H., Chen, K. P., and Celi, L. A. (2015). The Association Between Indwelling Arterial Catheters and Mortality in Hemodynamically Stable Patients With Respiratory Failure: A Propensity Score Analysis. Chest, 148(6):1470–1476.

- Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. The Review of Economics and Statistics, 86(1):4–29.
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3:160035.
- Kennedy, E. H. (2019). Nonparametric Causal Effects Based on Incremental Propensity Score Interventions. Journal of the American Statistical Association, 114(526):645–656. Publisher: ASA Website _eprint: <https://doi.org/10.1080/01621459.2017.1422737>.
- Källberg, D. and Waernbaum, I. (2023). Large Sample Properties of Entropy Balancing Estimators of Average Causal Effects. Econometrics and Statistics.
- Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. Journal of the American Statistical Association.
- Li, F., Thomas, L. E., and Li, F. (2019). Addressing Extreme Propensity Scores via the Overlap Weights. American Journal of Epidemiology, 188(1):250–257.
- Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. The International Journal of Biostatistics, 9(2):215–234.
- Neyman, J. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Statistical Science, 5(4):465–472.
- Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass.), 11(5):550–560.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701.
- Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics, 6(1):34–58.
- Wang, Y. and Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. Biometrika, 107(1):93–105.
- Yang, Q., Wiest, D., Davis, A. C., Truchil, A., and Adams, J. L. (2023). Hospital Readmissions by Variation in Engagement in the Health Care Hotspotting Trial: A Secondary Analysis of a Randomized Clinical Trial. JAMA Network Open, 6(9):e2332715.

- Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. Biometrika, 105(2):487–493.
- Zhao, Q. and Percival, D. (2017). Entropy Balancing is Doubly Robust. Journal of Causal Inference, 5(1). Publisher: De Gruyter.
- Zubizarreta, J. R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. Journal of the American Statistical Association, 110(511):910–922.

Supplementary Materials for Partially Retargeted Balancing Weights for Causal Effect Estimation Under Positivity Violations

1 Technical results and proofs

For all proofs, we notate $\|\cdot\|$ as the L2 norm.

1.1 Theorem 1

Proof. We rewrite the problem in matrix notation:

$$\begin{aligned} & \operatorname{argmin}_{w_i} \sum_{i=1}^n h(s_i) \\ & \text{subject to } Q_{2K+L \times 2n} s_{2n \times 1} = 0_{2K+L \times 1} \end{aligned}$$

where

$$s_{2n \times 1} = \begin{bmatrix} (1 - Z_i w_i)_{n \times 1} \\ (1 - (1 - Z_i) w_i)_{n \times 1} \end{bmatrix}$$

$$Q_{2K+L \times 2n} = \begin{bmatrix} c_1(\mathbf{X}_1) & c_1(\mathbf{X}_2) & \cdots & c_1(\mathbf{X}_n) & 0 & 0 & \cdots & 0 \\ c_K(\mathbf{X}_1) & c_K(\mathbf{X}_2) & \cdots & c_K(\mathbf{X}_n) & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & c_1(\mathbf{X}_1) & c_1(\mathbf{X}_2) & \cdots & c_1(\mathbf{X}_n) \\ 0 & 0 & \cdots & 0 & c_K(\mathbf{X}_1) & c_K(\mathbf{X}_2) & \cdots & c_K(\mathbf{X}_n) \\ g_1(\mathbf{X}_1) & g_1(\mathbf{X}_2) & \cdots & g_1(\mathbf{X}_n) & -g_1(\mathbf{X}_1) & -g_1(\mathbf{X}_2) & \cdots & -g_1(\mathbf{X}_n) \\ g_L(\mathbf{X}_1) & g_L(\mathbf{X}_2) & \cdots & g_L(\mathbf{X}_n) & -g_L(\mathbf{X}_1) & -g_L(\mathbf{X}_2) & \cdots & -g_L(\mathbf{X}_n) \end{bmatrix}.$$

This problem is now in the form of [Tseng and Bertsekas \(1991\)](#). The dual of this problem is

$$\begin{aligned} & \operatorname{minimize}_{\lambda} q(\lambda) \\ & \text{subject to no constraint on } \lambda, \end{aligned}$$

where $q(\lambda) = \sum_{j=1}^{2n} h_j^*(Q_j^T \lambda)$ and $h_j^*(t) = \sup_{s_j} \{ts_j - h(s_j)\}$. Consider $j < n + 1$. Then we have that

$$\begin{aligned}
h_j^*(t) &= \sup_{s_j} \{ts_j - h(s_j)\} \\
&= \sup_{w_j} \{t - tZ_jw_j - h(1 - Z_jw_j)\} \\
&= \sup_{w_j} \{t - tZ_jw_j - Z_jh(1 - w_j) - (1 - Z_j)h(1)\} \\
&= t - tZ_jw_j^* - Z_jh(1 - w_j^*) - (1 - Z_j)h(1).
\end{aligned}$$

Next we have that

$$\begin{aligned}
h_{j+n}^*(t) &= \sup_{s_{j+n}} \{ts_{j+n} - h(s_{j+n})\} \\
&= \sup_{w_j} \{t - t(1 - Z_j)w_j - h[1 - (1 - Z_j)w_j]\} \\
&= t - t(1 - Z_j)w_j^* - (1 - Z_j)h(1 - w_j^*) - Z_jh(1),
\end{aligned}$$

where w_j^* satisfies the first order conditions:

$$\begin{aligned}
-tZ_j + Z_jh'(1 - w_j^*) &= 0 \\
-t(1 - Z_j) + (1 - Z_j)h(1 - w_j^*) &= 0 \\
\implies 1 - (h')^{-1}(t) &= w_j^*.
\end{aligned}$$

Then we have that

$$\begin{aligned}
h_j^*(t) &= t - tZ_j(1 - (h')^{-1}(t)) - Z_jh(1 - 1 + (h')^{-1}(t)) - (1 - Z_j)h(1) \\
&= t - tZ_j\{1 - (h')^{-1}(t)\} - Z_jh\{(h')^{-1}(t)\} - (1 - Z_j)h(1) \\
&= -Z_j\{t - t(h')^{-1}(t) + h[(h')^{-1}(t)] - h(-1)\} - h(1) + t
\end{aligned}$$

and

$$\begin{aligned}
h_{j+n}^*(t) &= t - t(1 - Z_j)\{1 - (h')^{-1}(t)\} - (1 - Z_j)h\{1 - 1 + (h')^{-1}(t)\} - Z_jh(1) \\
&= t - t(1 - Z_j)\{1 - (h')^{-1}(t)\} - (1 - Z_j)h\{(h')^{-1}(t)\} - Z_jh(1) \\
&= -(1 - Z_j)\{t - t(h')^{-1}(t) + h[(h')^{-1}(t)] - h(1)\} - h(1) + t.
\end{aligned}$$

Let $\rho(t) = t - t(h')^{-1}(t) + h[(h')^{-1}(t)] - h(1)$. This gives

$$\begin{aligned}
h_j^*(t) &= -Z_j\rho(t) - h(1) + t \\
h_{j+n}^*(t) &= -(1 - Z_j)\rho(t) - h(1) + t.
\end{aligned}$$

Now, note that

$$\begin{aligned}
\rho'(t) &= 1 - (h')^{-1}(t) - t\{(h')^{-1}(t)\}' + h'[(h')^{-1}(t)] \times \{(h')^{-1}(t)\}' \\
&= 1 - (h')^{-1}(t) - t\{(h')^{-1}(t)\}' + t\{(h')^{-1}(t)\}' \\
&= 1 - (h')^{-1}(t) \\
\implies \rho'(t) &= w_j^*.
\end{aligned}$$

Thus, the dual formulation becomes

$$\begin{aligned}
&\text{minimize}_{\lambda} \quad q(\lambda) \\
&\text{subject to} \quad \text{no constraint on } \lambda,
\end{aligned}$$

where

$$\begin{aligned}
q(\lambda) &= \frac{1}{n} \sum_{j=1}^n -Z_j \rho(Q_j^T \lambda) - (1 - Z_j) \rho(Q_{j+n}^T \lambda) + Q_j^T \lambda + Q_{j+n}^T \lambda - 2nh(1) \\
&= \frac{1}{n} \sum_{j=1}^n -Z_j \rho(Q_j^T \lambda) - (1 - Z_j) \rho(Q_{j+n}^T \lambda) + Q_j^T \lambda + Q_{j+n}^T \lambda.
\end{aligned}$$

Let $B^+(\mathbf{X}_j) = \{c_1(\mathbf{X}_j), \dots, c_K(\mathbf{X}_j), 0_K, g_1(\mathbf{X}_j), \dots, g_L(\mathbf{X}_j)\}$

and $B^-(\mathbf{X}_j) = \{0_K, c_1(\mathbf{X}_j), \dots, c_K(\mathbf{X}_j), -g_1(\mathbf{X}_j), \dots, -g_L(\mathbf{X}_j)\}$. Then, the primal solution w_j^* satisfies

$$\begin{aligned}
w_j^* &= \rho'\{B^+(\mathbf{X}_j)^T \hat{\lambda}\} \text{ for } Z_j = 1 ; j = 1, \dots, n \\
w_j^* &= \rho'\{B^-(\mathbf{X}_j)^T \hat{\lambda}\} \text{ for } Z_j = 0 ; j = 1, \dots, n,
\end{aligned}$$

where $\hat{\lambda}$ is the solution to the dual problem. Let $\lambda = \begin{bmatrix} \boldsymbol{\alpha}_{0_{K \times 1}} \\ \boldsymbol{\alpha}_{1_{K \times 1}} \\ \boldsymbol{\gamma}_{L \times 1} \end{bmatrix}$ and let $c(\mathbf{X}_j) = \{c_1(\mathbf{X}_j), \dots, c_K(\mathbf{X}_j)\}$ and $g(\mathbf{X}_j) = \{g_1(\mathbf{X}_j), \dots, g_L(\mathbf{X}_j)\}$. Then the dual formulation simplifies to

$$\begin{aligned} & \underset{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}}{\text{minimize}} && q(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}) \\ & \text{subject to} && \text{no constraint on } \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}, \end{aligned}$$

$$\begin{aligned} \text{where } q(\alpha_0, \alpha_1, \gamma) &= \frac{1}{n} \sum_{i=1}^n -Z_i \rho \{ \boldsymbol{\alpha}_1^T c(\mathbf{X}_i) + \boldsymbol{\gamma}^T g(\mathbf{X}_i) \} - (1 - Z_i) \rho \{ \boldsymbol{\alpha}_0^T c(\mathbf{X}_i) - \boldsymbol{\gamma}^T g(\mathbf{X}_i) \} \\ &\quad + \boldsymbol{\alpha}_1^T c(\mathbf{X}_i) + \boldsymbol{\gamma}^T g(\mathbf{X}_i) + \boldsymbol{\alpha}_0^T c(\mathbf{X}_i) - \boldsymbol{\gamma}^T g(\mathbf{X}_i) \\ &= \frac{1}{n} \sum_{i=1}^n -Z_i \rho \{ \boldsymbol{\alpha}_1^T c(\mathbf{X}_i) + \boldsymbol{\gamma}^T g(\mathbf{X}_i) \} - (1 - Z_i) \rho \{ \boldsymbol{\alpha}_0^T c(\mathbf{X}_i) - \boldsymbol{\gamma}^T g(\mathbf{X}_i) \} \\ &\quad + \boldsymbol{\alpha}_1^T c(\mathbf{X}_i) + \boldsymbol{\alpha}_0^T c(\mathbf{X}_i) \end{aligned}$$

and the primal solution $w_i^* = Z_i w_{1i}^* + (1 - Z_i) w_{0i}^*$ satisfies $w_{1i}^* = \rho' \{ \hat{\boldsymbol{\alpha}}_1^T c(\mathbf{X}_i) + \hat{\boldsymbol{\gamma}}^T g(\mathbf{X}_i) \}$ and $w_{0i}^* = \rho' \{ \hat{\boldsymbol{\alpha}}_0^T c(\mathbf{X}_i) - \hat{\boldsymbol{\gamma}}^T g(\mathbf{X}_i) \}$ where $\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\gamma}}$ is the solution to the dual problem. \square

1.2 Proposition 2

Proof. We use similar proof techniques to Appendix A of [Zhao and Percival \(2017\)](#). Let $B_\epsilon(\mathbf{x}^*) = \{ \mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|_\infty \leq \epsilon \}$ for some $\epsilon > 0$ and let $\text{CH}\{\mathbf{X}_i\}_{Z_i=z}$ be the convex hull generated by $\{\mathbf{X}_i\}_{Z_i=z}$. For the weights to exist, we need 1) $0 < \Pr\{Z_i | c(\mathbf{X}_i) = c(\mathbf{x})\} < 1$ for all \mathbf{x} as shown by ([Zhao and Percival, 2017](#)); and 2) $\text{CH}\{\mathbf{X}_i\}_{Z_i=1} \cap \text{CH}\{\mathbf{X}_i\}_{Z_i=0} \neq \emptyset$. It remains to determine when $\text{CH}\{\mathbf{X}_i\}_{Z_i=1} \cap \text{CH}\{\mathbf{X}_i\}_{Z_i=0} \neq \emptyset$ holds. We show that if there exists \mathbf{x}^* such that that $0 < P(Z_i | \mathbf{X}_i = \mathbf{x}^*) < 1$, a slightly stronger claim holds: $P\{\text{CH}\{\mathbf{X}_i\}_{Z_i=1} \cap \text{CH}\{\mathbf{X}_i\}_{Z_i=0} = B_\epsilon(\mathbf{x}^*)\} = P\{B_\epsilon(\mathbf{x}^*) \in \text{CH}\{\mathbf{X}_i\}_{Z_i=1}, B_\epsilon(\mathbf{x}^*) \in \text{CH}\{\mathbf{X}_i\}_{Z_i=0}\} \rightarrow 1$ as $n \rightarrow \infty$.

Let $R_i(\mathbf{x}^*) = i = 1, \dots, 3^p$ be the 3^p boxes centered at $\mathbf{x}^* + \frac{3}{2}\epsilon b$ where $b \in \mathbb{R}^p$ contains entries of 0, -1, 1. Then, $P(B_\epsilon(\mathbf{x}^*) \in \text{CH}\{\mathbf{X}_i\}_{Z_i=1}, B_\epsilon(\mathbf{x}^*) \in \text{CH}\{\mathbf{X}_i\}_{Z_i=0}) \geq P(\exists \mathbf{X}_i \in R_i(\mathbf{x}^*), Z_i = 0, \exists \mathbf{X}_j \in R_j(\mathbf{x}^*), Z_j = 1, i, j = 1, \dots, 3^p)$.

Assume that for some \mathbf{x}^* , for all $\mathbf{x} \in B_{2\epsilon}(\mathbf{x}^*)$, $0 < P(Z_i | \mathbf{X}_i = \mathbf{x}) < 1$. This assumption implies that $\rho_1 = \min_i P(\mathbf{X} \in R_i(\mathbf{x}^*) | Z = 1) > 0$ and $\rho_0 = \min_i P(\mathbf{X} \in R_i(\mathbf{x}^*) | Z = 0) > 0$.

Then we have the following,

$$\begin{aligned}
& P(\exists \mathbf{X}_i \in R_i(\mathbf{x}^*), Z_i = 0, \exists \mathbf{X}_j \in R_j(\mathbf{x}^*), Z_j = 1, i, j = 1, \dots, 3^p) \\
&= P(\exists \mathbf{X}_i \in R_i(\mathbf{x}^*), Z_i = 0, i = 1, \dots, 3^p) \\
&\quad \times P(\exists \mathbf{X}_j \in R_j(\mathbf{x}^*), Z_j = 1, j = 1, \dots, 3^p) \\
&\geq \{1 - \sum_{i=1}^{3^p} P(X \notin R_i(\mathbf{x}^*) | Z = 0)^n\} \\
&\quad \times \{1 - \sum_{i=1}^{3^p} P(X \notin R_j(\mathbf{x}^*) | Z = 1)^n\} \\
&= \{1 - 3^p(1 - \rho_0)^n\} \{1 - 3^p(1 - \rho_1)^n\} \\
&\quad \lim_{n \rightarrow \infty} \rightarrow 1,
\end{aligned}$$

where the first equality holds because each (\mathbf{X}_i, Z_i) is an iid sample. We see that letting $\epsilon \rightarrow 0$, the only assumption required is the existence of some \mathbf{x}^* such that $0 < P(Z_i | \mathbf{X}_i = \mathbf{x}^*) < 1$.

□

1.3 Proposition 5

Proposition 5. *Suppose Assumptions 1, either 2 or 2*, and 3 hold and ρ has a continuous first derivative. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\gamma})$. Then,*

- i) $\hat{\boldsymbol{\theta}} \rightarrow_p \tilde{\boldsymbol{\theta}}$ where $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\gamma}})$. is the unique minimizer of $E[-e(\mathbf{X})\rho\{\boldsymbol{\alpha}_1^T c(\mathbf{X}) + \boldsymbol{\gamma}^T g(\mathbf{X})\} - \{1 - e(\mathbf{X})\}\rho\{\boldsymbol{\alpha}_0^T c(\mathbf{X}) - \boldsymbol{\gamma}^T g(\mathbf{X})\} + \boldsymbol{\alpha}_1^T c(\mathbf{X}) + \boldsymbol{\alpha}_0^T c(\mathbf{X})]$ and
- ii) $w_1^*(\mathbf{X}) = \rho'\{\hat{\boldsymbol{\alpha}}_1^T c(\mathbf{X}) + \hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} \rightarrow_p \tilde{w}_1(\mathbf{X}) = \rho'\{\tilde{\boldsymbol{\alpha}}_1^T c(\mathbf{X}) + \tilde{\boldsymbol{\gamma}}^T g(\mathbf{X})\}$ and $w_0^*(\mathbf{X}) = \rho'\{\hat{\boldsymbol{\alpha}}_0^T c(\mathbf{X}) + -\hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} \rightarrow_p \tilde{w}_0(\mathbf{X}) = \rho'\{\tilde{\boldsymbol{\alpha}}_0^T c(\mathbf{X}) - \tilde{\boldsymbol{\gamma}}^T g(\mathbf{X})\}$.

We use similar proof techniques to the proof of proposition 1 in [Källberg and Waernbaum \(2023\)](#). First, we prove the following lemma.

Lemma 6. *Given Assumption 3, $l(\boldsymbol{\theta}) = E[-e(\mathbf{X})\rho\{\boldsymbol{\alpha}_1^T c(\mathbf{X}_i) + \boldsymbol{\gamma}^T g(\mathbf{X})\} - \{1 - e(\mathbf{X})\}\rho\{\boldsymbol{\alpha}_0^T c(\mathbf{X}) - \boldsymbol{\gamma}^T g(\mathbf{X})\} + \boldsymbol{\alpha}_1^T c(\mathbf{X}) + \boldsymbol{\alpha}_0^T c(\mathbf{X})]$ has a unique minimizer $\tilde{\boldsymbol{\theta}}$ that satisfies $\inf_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = l(\tilde{\boldsymbol{\theta}})$.*

Proof. Let $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(t)$, $t \geq 1$, be a sequence of random vectors where $\lim_{t \rightarrow \infty} l(\boldsymbol{\theta}^*) = l(\tilde{\boldsymbol{\theta}}) < \infty$. Without loss of generality assume $\boldsymbol{\theta}^*/\|\boldsymbol{\theta}^*\| \rightarrow_p \mathbf{a}$ where \mathbf{a} is a constant vector. If $\|\boldsymbol{\theta}^*\|$ is bounded, $\boldsymbol{\theta}^* \rightarrow_p \tilde{\boldsymbol{\theta}}$ for some $\tilde{\boldsymbol{\theta}}$ where $\inf_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = l(\tilde{\boldsymbol{\theta}})$ by the continuity of ρ such that $\tilde{\boldsymbol{\theta}}$ is a minimum of $l(\boldsymbol{\theta})$. The solution $\tilde{\boldsymbol{\theta}}$ is unique if $l(\boldsymbol{\theta})$ is strictly convex. Thus, it is sufficient to show that $\|\boldsymbol{\theta}^*\|$ is bounded and $l(\boldsymbol{\theta})$ is strictly convex.

We must first show that $\|\boldsymbol{\theta}^*\|$ is bounded. We will use proof by contradiction and assume that $\|\boldsymbol{\theta}^*\| \rightarrow \infty$. Since the covariance matrix of $b(\mathbf{X}) = \{c(\mathbf{X}), g(\mathbf{X})\}$ is non-singular by

Assumption 3, $\alpha_1^T c(\mathbf{X}_i) + \gamma^T g(\mathbf{X})\} \rightarrow \pm\infty$ and $\alpha_0^T c(\mathbf{X}_i) - \gamma^T g(\mathbf{X})\} \rightarrow \pm\infty$ with non-zero probability. Then since ρ is strictly concave, this implies that $-\rho\{\alpha_1^T c(\mathbf{X}_i) + \gamma^T g(\mathbf{X})\} \rightarrow \infty$ and $-\rho\{\alpha_0^T c(\mathbf{X}_i) - \gamma^T g(\mathbf{X})\} \rightarrow \infty$. However, this implies that $\lim_{t \rightarrow \infty} l(\boldsymbol{\theta}^*) \rightarrow \infty$ with non-zero probability, a contradiction to $\lim_{t \rightarrow \infty} l(\boldsymbol{\theta}^*) = l(\tilde{\boldsymbol{\theta}})$. Thus, $\|\boldsymbol{\theta}^*\|$ is bounded and $\boldsymbol{\theta}^* \rightarrow_p \tilde{\boldsymbol{\theta}}$ for some $\tilde{\boldsymbol{\theta}}$ such that $\inf_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = l(\tilde{\boldsymbol{\theta}})$. It remains to show that this solution is unique by proving the strict convexity of $l(\boldsymbol{\theta})$. We will show that the Hessian of $l(\boldsymbol{\theta})$, $\mathbf{H}(\boldsymbol{\theta})$, is positive-definite. Now $\mathbf{H}(\boldsymbol{\theta})$ is given by the following (row, column),

$$\begin{aligned}\mathbf{H}_{k_1, k_2}(\boldsymbol{\theta}) &= E[-e(\mathbf{X})\rho''\{\alpha_1^T c(\mathbf{X}) + \gamma^T g(\mathbf{X})\}c_{k_1}(\mathbf{X})c_{k_2}(\mathbf{X}) \\ &\quad - \{1 - e(\mathbf{X})\}\rho''\{\alpha_0^T c(\mathbf{X}) - \gamma^T g(\mathbf{X})\}c_{k_1}(\mathbf{X})c_{k_2}(\mathbf{X})], \\ \mathbf{H}_{l_1, l_2}(\boldsymbol{\theta}) &= E[-e(\mathbf{X})\rho''\{\alpha_1^T c(\mathbf{X}) + \gamma^T g(\mathbf{X})\}g_{l_1}(\mathbf{X})g_{l_2}(\mathbf{X}) \\ &\quad - \{1 - e(\mathbf{X})\}\rho''\{\alpha_0^T c(\mathbf{X}) - \gamma^T g(\mathbf{X})\}g_{l_1}(\mathbf{X})g_{l_2}(\mathbf{X})], \\ \mathbf{H}_{k_1, l_1}(\boldsymbol{\theta}) &= E[-e(\mathbf{X})\rho''\{\alpha_1^T c(\mathbf{X}_i) + \gamma^T g(\mathbf{X})\}c_{k_1}(\mathbf{X})g_{l_1}(\mathbf{X}) \\ &\quad - \{1 - e(\mathbf{X})\}\rho''\{\alpha_0^T c(\mathbf{X}_i) - \gamma^T g(\mathbf{X})\}c_{k_1}(\mathbf{X})g_{l_1}(\mathbf{X})].\end{aligned}$$

We can construct an estimator for $\mathbf{H}(\boldsymbol{\theta})$, $\hat{\mathbf{H}}(\boldsymbol{\theta})$ using the sample equivalents of $\mathbf{H}_{k_1, k_2}(\boldsymbol{\theta})$, $\mathbf{H}_{l_1, l_2}(\boldsymbol{\theta})$, and $\mathbf{H}_{k_1, l_1}(\boldsymbol{\theta})$. Let $c_i = -e(\mathbf{X}_i)\rho''\{\alpha_1^T c(\mathbf{X}_i) + \gamma^T g(\mathbf{X}_i)\} + -\{1 - e(\mathbf{X}_i)\}\rho''\{\alpha_0^T c(\mathbf{X}_i) - \gamma^T g(\mathbf{X}_i)\}$. By the strict concavity of ρ , $c_i > 0$. Let $v^* = [\sqrt{c_1}, \dots, \sqrt{c_n}]$. Let v be any vector of length n . Then $v^T \hat{\mathbf{H}}(\boldsymbol{\theta})v = v^T \odot v^{*T} B(\mathbf{X})^T B(\mathbf{X})v \odot v^* > 0$ since $B(\mathbf{X})$ is full rank by Assumption 3. Then $\hat{\mathbf{H}}$ is positive definite and since $E[\hat{\mathbf{H}}(\boldsymbol{\theta})] = \mathbf{H}(\boldsymbol{\theta})$, $\hat{\mathbf{H}}(\boldsymbol{\theta}) \rightarrow_p \mathbf{H}(\boldsymbol{\theta})$, $\mathbf{H}(\boldsymbol{\theta})$ is positive definite and $l(\boldsymbol{\theta})$ is strictly convex as desired. \square

Now we will prove Proposition 5.

Proof. By Theorem 6.3 in Boos and Stefanski (2013) and the boundedness of $b_j(\mathbf{X})$ (Assumption 3), $\hat{l}(\boldsymbol{\theta}) \rightarrow l(\boldsymbol{\theta})$ uniformly almost surely for all $\boldsymbol{\theta}$ in every compact subset of \mathbb{R}^{2K+L} where $\hat{l}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n -Z_i \rho\{\alpha_1^T c(\mathbf{X}_i) + \gamma^T g(\mathbf{X}_i)\} - (1 - Z_i) \rho\{\alpha_0^T c(\mathbf{X}_i) - \gamma^T g(\mathbf{X}_i)\} + \alpha_1^T c(\mathbf{X}_i) + \alpha_0^T c(\mathbf{X}_i)$. By Lemma 6, we know the Hessian of $\hat{l}(\boldsymbol{\theta})$ is positive definite and thus that there is a unique solution to $\hat{l}(\boldsymbol{\theta})$. Then, let $R_{m,r}$ be the indicator that the solution is within distance $r > 0$ of $\tilde{\boldsymbol{\theta}}$ for sample size $n = m$. We want to show that $\{R_{m,r}\}_{m \geq 1}$ has finitely many zeros such that $\sum_{m=1}^{\infty} (1 - R_{m,r}) < \infty$ for an arbitrary r ; this is equivalent to $\lim_{n \rightarrow \infty} P(|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}| > \epsilon) = 0$ for any $\epsilon > 0$ as desired. Let $\Delta = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\| \leq r\}$ and $\Delta^* = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\| = r\}$. Then, $\hat{l}(\boldsymbol{\theta}) \rightarrow l(\boldsymbol{\theta})$ implies that there exists a number n^* such that $|\hat{l}(\boldsymbol{\theta}) - l(\boldsymbol{\theta})| < r/2$ and $|\hat{l}(\tilde{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})| < r/2$ almost surely for $n \geq n^*$. Since $\tilde{\boldsymbol{\theta}}$ minimizes $l(\boldsymbol{\theta})$, $\hat{l}(\boldsymbol{\theta})$ has a minima in Δ and a solution for $n \geq n^*$. This implies that $R_{m,r} = 1$ for $m \geq n^*$ and thus that $\sum_{m=1}^{\infty} (1 - R_{m,r}) < \infty$. Then, $\hat{\boldsymbol{\theta}} \rightarrow_p \tilde{\boldsymbol{\theta}}$ where $\tilde{\boldsymbol{\theta}}$ is the unique minimizer of $l(\boldsymbol{\theta})$

by Lemma 6.

Then since ρ' is continuous, by the continuous mapping theorem we have that $w_1^*(\mathbf{X}) = \rho'\{\hat{\boldsymbol{\alpha}}_1^T c(\mathbf{X}) + \hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} \rightarrow_p \tilde{w}_1(\mathbf{X}) = \rho'\{\tilde{\boldsymbol{\alpha}}_1^T c(\mathbf{X}) + \tilde{\boldsymbol{\gamma}}^T g(\mathbf{X})\}$ and $w_0^*(\mathbf{X}) = \rho'\{\hat{\boldsymbol{\alpha}}_0^T c(\mathbf{X}) - \hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} \rightarrow_p \tilde{w}_0(\mathbf{X}) = \rho'\{\tilde{\boldsymbol{\alpha}}_0^T c(\mathbf{X}) - \tilde{\boldsymbol{\gamma}}^T g(\mathbf{X})\}$. \square

1.4 Theorem 3

Proof. (1)

Assume that the true models are $\{e(\mathbf{x})\}^{-1} = \rho'\{\tilde{\boldsymbol{\alpha}}_1^T c(\mathbf{x}) + \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x})\}$ and $\{1 - e(\mathbf{x})\}^{-1} = \rho'\{\tilde{\boldsymbol{\alpha}}_0^T c(\mathbf{x}) - \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x})\}$. Then given Assumptions 1 and 2,

$$\begin{aligned} \tau &= E \left[\frac{Z_i Y_i}{e(\mathbf{X}_i)} \right] - E \left[\frac{(1 - Z_i) Y_i}{1 - e(\mathbf{X}_i)} \right] \\ &= E[\rho'\{\tilde{\boldsymbol{\alpha}}_1^T c(\mathbf{x}) + \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x})\} Z_i Y_i] - E[\rho'\{\tilde{\boldsymbol{\alpha}}_0^T c(\mathbf{x}) - \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x})\} (1 - Z_i) Y_i]. \end{aligned}$$

Further,

$$\begin{aligned} \tau_{w^*} &= \sum_{i=1}^n w^*(\mathbf{X}) Z_i Y_i - \sum_{i=1}^n w^*(\mathbf{X}) (1 - Z_i) Y_i \\ &= \sum_{i=1}^n \rho'\{\hat{\boldsymbol{\alpha}}_1^T c(\mathbf{X}) + \hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} Z_i Y_i - \sum_{i=1}^n \rho'\{\hat{\boldsymbol{\alpha}}_0^T c(\mathbf{X}) - \hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} (1 - Z_i) Y_i \end{aligned}$$

Then, given Proposition 5, Assumption 3, and the assumption that $\rho''\{\cdot\}$ is continuous over \mathbb{R} , the law of large numbers for averages with estimated parameters applies (Boos and Stefanski (2013) Theorem 7.3) and

$$\begin{aligned} \sum_{i=1}^n \rho'\{\hat{\boldsymbol{\alpha}}_1^T c(\mathbf{X}) + \hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} Z_i Y_i &\rightarrow_p E[\rho'\{\tilde{\boldsymbol{\alpha}}_1^T c(\mathbf{x}) + \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x})\} Z_i Y_i], \\ \sum_{i=1}^n \rho'\{\hat{\boldsymbol{\alpha}}_0^T c(\mathbf{X}) - \hat{\boldsymbol{\gamma}}^T g(\mathbf{X})\} (1 - Z_i) Y_i &\rightarrow_p E[\rho'\{\tilde{\boldsymbol{\alpha}}_0^T c(\mathbf{x}) - \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x})\} (1 - Z_i) Y_i]. \end{aligned}$$

It follows directly that $\tau_{w^*} \rightarrow_p \tau$. \square

Proof. (2)

Assume $\mu_1(\mathbf{x}) = \boldsymbol{\beta}_1^T c(\mathbf{x}) + \boldsymbol{\lambda}^T g(\mathbf{x})$ and $\mu_0(\mathbf{x}) = \boldsymbol{\beta}_0^T c(\mathbf{x}) + \boldsymbol{\lambda}^T g(\mathbf{x})$ for some $\boldsymbol{\beta}_z \in \mathbb{R}^K$ and $\boldsymbol{\lambda} \in \mathbb{R}^L$. We will use a similar proof technique by Zhao and Percival (2017) to show that augmenting the our proposed estimator with estimated outcome regressions does not change the estimator. Let $\hat{\mu}_0(\mathbf{x})$, $\hat{\mu}_1(\mathbf{x})$, and $\hat{e}(\mathbf{x})$ be outcome regression and propensity score

estimates. Then the standard doubly robust estimator for the ATE is,

$$\hat{\tau}_{DR} = \frac{1}{n_1} \sum_{Z_i=1} \left[\frac{Y_i - \hat{\mu}_1(\mathbf{X}_i)}{\hat{e}(\mathbf{X}_i)} - \hat{\mu}_1(\mathbf{X}_i) \right] - \frac{1}{n_0} \sum_{Z_i=1} \left[\frac{Y_i - \hat{\mu}_0(\mathbf{X}_i)}{1 - \hat{e}(\mathbf{X}_i)} - \hat{\mu}_0(\mathbf{X}_i) \right]$$

Now consider using w^* to estimate the propensity score. Then,

$$\begin{aligned} \hat{\tau}_{DR} - \hat{\tau}_{w^*} &= - \sum_{Z_i=1} w^*(\mathbf{X}_i) \hat{\mu}_1(\mathbf{X}_i) + \frac{1}{n} \sum_i \hat{\mu}_1(\mathbf{X}_i) + \sum_{Z_i=0} w(\mathbf{X}_i)^* \hat{\mu}_0(\mathbf{X}_i) - \frac{1}{n} \sum_i \hat{\mu}_0(\mathbf{X}_i) \\ &= - \sum_{Z_i=1} w^*(\mathbf{X}_i) [\hat{\beta}_1^T c(\mathbf{X}_i) + \hat{\lambda}^T g(\mathbf{X}_i)] + \frac{1}{n} \sum_i \hat{\beta}_1^T c(\mathbf{X}_i) + \hat{\lambda}^T g(\mathbf{X}_i) \\ &\quad + \sum_{Z_i=0} w^*(\mathbf{X}_i) [\hat{\beta}_0^T c(\mathbf{X}_i) + \hat{\lambda}^T g(\mathbf{X}_i)] - \frac{1}{n} \sum_i \hat{\beta}_0^T c(\mathbf{X}_i) + \hat{\lambda}^T g(\mathbf{X}_i) \\ &= \hat{\beta}_1^T \left[\sum_i c(\mathbf{X}_i) - \sum_i w^*(\mathbf{X}_i) Z_i c(\mathbf{X}_i) \right] \\ &\quad - \hat{\beta}_0^T \left[\sum_i c(\mathbf{X}_i) - \sum_i w^*(\mathbf{X}_i) (1 - Z_i) c(\mathbf{X}_i) \right] \\ &\quad + \hat{\lambda}^T \left[\sum_i w^*(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i) - \sum_i w^*(\mathbf{X}_i) Z_i g(\mathbf{X}_i) \right] \\ &= 0, \end{aligned}$$

and $\hat{\tau}_{DR} = \hat{\tau}_{w^*}$. Since $\hat{\tau}_{DR} \rightarrow_p \tau$ (Bang and Robins, 2005) we also have $\hat{\tau}_{w^*} \rightarrow_p \tau$. \square

1.5 Corollary 3.1

Proof. Assume $\mu_1(\mathbf{x}) = \beta_1^T c(\mathbf{x}) + \lambda_1^T g(\mathbf{x})$ and $\mu_0(\mathbf{x}) = \beta_0^T c(\mathbf{x}) + \lambda_0^T g(\mathbf{x})$ for some $\beta_z \in \mathbb{R}^K$ and $\lambda_z \in \mathbb{R}^L$. Then, given Proposition 5, Assumption 3, and the assumption that $\rho''\{\cdot\}$ is continuous over \mathbb{R} , the law of large numbers for averages with estimated parameters applies (Boos and Stefanski (2013) Theorem 7.3) and

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n w^*(\mathbf{X}_i) Z_i Y_i &\rightarrow_p E[\tilde{w}(\mathbf{X}_i) Z_i Y_i] \\ &= E[\tilde{w}(\mathbf{X}_i) Z_i \{\beta_1^T c(\mathbf{X}_i) + \lambda_1^T g(\mathbf{X}_i)\}] \\ &= E[\beta_1^T c(\mathbf{X}_i) + \tilde{w}(\mathbf{X}_i) Z_i \lambda_1^T g(\mathbf{X}_i)] \\ &= E[\beta_1^T c(\mathbf{X}_i) + \lambda_1^T \tilde{w}(\mathbf{X}_i) e(\mathbf{X}_i) g(\mathbf{X}_i)] \\ &= E[E[Y_i(1) | c(\mathbf{X}_i), \tilde{w}(\mathbf{X}_i) e(\mathbf{X}_i) g(\mathbf{X}_i)]] \\ &= E[E[Y_i | Z_i = 1, c(\mathbf{X}_i), \tilde{w}(\mathbf{X}_i) e(\mathbf{X}_i) g(\mathbf{X}_i)]]. \end{aligned}$$

The second equality holds by \tilde{w} satisfying the first order conditions of the population version of Equation (6). Similarly, $\frac{1}{n} \sum_{i=1}^n w^*(\mathbf{X}_i) (1 - Z_i) Y_i \rightarrow E[E[Y_i | Z_i = 0, c(\mathbf{X}_i), \tilde{w}(\mathbf{X}_i) \{1 -$

$e(\mathbf{X}_i)\}g(\mathbf{X}_i)]$. Then $\hat{\tau}_{w^*} \rightarrow \tau_{g, \bar{w}}$ as desired. \square

1.6 Theorem 4

Proof. We can decompose $\hat{\tau}_{w^*} - \tau$ into the following:

$$\begin{aligned}
\hat{\tau}_{w^*} - \tau &= \frac{1}{n} \sum_{i=1}^n w^*(\mathbf{X}_i)Z_i Y_i - w^*(\mathbf{X}_i)(1 - Z_i)Y_i - \tau \\
&= \frac{1}{n} \sum_{i=1}^n w_1^*(\mathbf{X}_i)Z_i \{Y_i - \mu_1(\mathbf{X}_i)\} + \frac{1}{n} \sum_{i=1}^n (w_1^*(\mathbf{X}_i)Z_i - 1)\mu_1(\mathbf{X}_i) \\
&\quad - \frac{1}{n} \sum_{i=1}^n w_0^*(\mathbf{X}_i)(1 - Z_i)\{Y_i - \mu_0(\mathbf{X}_i)\} + \frac{1}{n} \sum_{i=1}^n \{w_0^*(\mathbf{X}_i)(1 - Z_i) - 1\}\mu_0(\mathbf{X}_i) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \tau(\mathbf{X}_i) - \tau \\
&= \frac{1}{n} \sum_{i=1}^n S_i + R_1 + R_2,
\end{aligned}$$

where

$$\begin{aligned}
S_i &= \frac{Z_i}{e(\mathbf{X}_i)} \{Y_i - \mu_1(\mathbf{X}_i)\} - \frac{1 - Z_i}{1 - e(\mathbf{X}_i)} \{Y_i - \mu_0(\mathbf{X}_i)\} + \{\tau(\mathbf{X}_i) - \tau\} \\
R_1 &= \frac{1}{n} \sum_{i=1}^n \left(w_1^*(\mathbf{X}_i) - \frac{1}{e(\mathbf{X}_i)} \right) Z_i \{Y_i - \mu_1(\mathbf{X}_i)\} - \left(w_0^*(\mathbf{X}_i) - \frac{1}{1 - e(\mathbf{X}_i)} \right) (1 - Z_i) \{Y_i - \mu_0(\mathbf{X}_i)\} \\
R_2 &= \frac{1}{n} \sum_{i=1}^n \left(w_1^*(\mathbf{X}_i)Z_i - 1 \right) \mu_1(\mathbf{X}_i) - \left(w_0^*(\mathbf{X}_i)(1 - Z_i) - 1 \right) \mu_0(\mathbf{X}_i) \\
&= \frac{1}{n} \sum_{i=1}^n (w_1^*(\mathbf{X}_i)Z_i - 1) \{ \beta_1^T c(\mathbf{X}_i) + \boldsymbol{\lambda}^T g(\mathbf{X}_i) \} - \{ w_0^*(\mathbf{X}_i)(1 - Z_i) - 1 \} \{ \beta_1^T c(\mathbf{X}_i) + \boldsymbol{\lambda}^T g(\mathbf{X}_i) \} \\
&= \frac{1}{n} \sum_{i=1}^n (w_1^*(\mathbf{X}_i)Z_i - 1) \beta_1^T c(\mathbf{X}_i) - \sum_{i=1}^n \{ w_0^*(\mathbf{X}_i)(1 - Z_i) - 1 \} \beta_0^T c(\mathbf{X}_i) \\
&\quad + \sum_{i=1}^n \{ w_1^*(\mathbf{X}_i)Z_i - w_0^*(\mathbf{X}_i)(1 - Z_i) \} \boldsymbol{\lambda}^T g(\mathbf{X}_i).
\end{aligned}$$

Note that $R_2 = 0$ by the constraints in the optimization problem. Then, since S_i takes the same form as the efficient score for the ATE in [Hahn \(1998\)](#), $\hat{\tau}_{w^*}$ is asymptotically normal and semiparametrically efficient as long as R_1 is $o_p(n^{-1/2})$.

We will focus on the first component of R_1 , which we decompose as $R_1 = R_1^1 + R_1^2$. We can

bound the first term as

$$\begin{aligned}
|R_1^1| &= \left| \frac{1}{n} \sum_{i=1}^n \left(w_1^*(\mathbf{X}_i) - \frac{1}{e(\mathbf{X}_i)} \right) Z_i \{Y_i - \mu_1(\mathbf{X}_i)\} \right| \\
&= \left| \int \left(w_1^*(\mathbf{X}_i) - \frac{1}{e(\mathbf{X}_i)} \right) \{Z_i Y_i - Z_i \mu_1(\mathbf{X}_i)\} dF_n(\mathbf{x}) \right| \\
&\leq \left\| w_1^*(\mathbf{X}_i) - \frac{1}{e(\mathbf{X}_i)} \right\| \cdot \|\{Z_i Y_i - Z_i \mu_1(\mathbf{X}_i)\}\| \\
&= o_p(1) O_p(n^{-1/2}) = o_p(n^{-1/2}),
\end{aligned}$$

where the third inequality follows from Cauchy-Schwartz and the fourth equality follows from Theorem 1 and the central limit theorem. Similar arguments hold for R_1^2 and we get that

$$\begin{aligned}
R_1 &\leq |R_1^1| + |R_1^2| \\
&= o_p(n^{-1/2}) + o_p(n^{-1/2}) = o_p(n^{-1/2})
\end{aligned}$$

as desired. □

1.7 Corollary 4.1

Proof. When the propensity score model is not correctly specified, R_1 will no longer go to zero. Then, the variance of $\hat{\tau}_{w^*}$ will be $\text{Var}(S_i + R_{1i})$ where

$$\begin{aligned}
S_i + R_{1i} &= w_1^*(\mathbf{X}_i) Z_i \{Y_i - \mu_1(\mathbf{X}_i)\} - w_0^*(\mathbf{X}_i) (1 - Z_i) \{Y_i - \mu_0(\mathbf{X}_i)\} + \{\tau(\mathbf{X}_i) - \tau\} \\
&= w_1^*(\mathbf{X}_i) Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - w_0^*(\mathbf{X}_i) (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} + \{\tau(\mathbf{X}_i) - \tau\} \\
&= \{w_1^*(\mathbf{X}_i) - \tilde{w}_1(\mathbf{X}_i)\} Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - \{w_0^*(\mathbf{X}_i) - \tilde{w}_0(\mathbf{X}_i)\} (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} \\
&\quad + \tilde{w}_1(\mathbf{X}_i) Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} + \{\tau(\mathbf{X}_i) - \tau\}.
\end{aligned}$$

Using similar arguments used for R_1 above and Proposition 3,

$$\{w_1^*(\mathbf{X}_i) - \tilde{w}_1(\mathbf{X}_i)\} Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - \{w_0^*(\mathbf{X}_i) - \tilde{w}_0(\mathbf{X}_i)\} (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} = o_p(n^{-1/2})$$

such that

$$\text{Var}(S_i + R_{1i}) = \text{Var}[\tilde{w}_1(\mathbf{X}_i) Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} + \{\tau(\mathbf{X}_i) - \tau\}].$$

Then,

$$\begin{aligned}
\text{Var}(S_i + R_{1_i}) &= \text{Var}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}] + \text{Var}[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}] \\
&\quad + \text{Var}\{\tau(\mathbf{X}_i) - \tau\} \\
&\quad - 2\text{Cov}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}, \tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}] \\
&\quad + 2\text{Cov}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}, \{\tau(\mathbf{X}_i) - \tau\}] \\
&\quad - 2\text{Cov}[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}, \{\tau(\mathbf{X}_i) - \tau\}].
\end{aligned}$$

We have that

$$\begin{aligned}
&-2\text{Cov}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}, \tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}] \\
&\quad = -2E[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}] \\
&\quad\quad \times E[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}] \\
&\quad = -2E(E[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i]) \\
&\quad\quad \times E(E[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}|\mathbf{X}_i]) \\
&\quad = -2E(\tilde{w}_1(\mathbf{X}_i)E[Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i]) \\
&\quad\quad \times E(\tilde{w}_0(\mathbf{X}_i)E[(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}|\mathbf{X}_i]) \\
&\quad \text{(by ignorability)} = -2E[\tilde{w}_1(\mathbf{X}_i)e(\mathbf{X}_i)\{\mu_1(\mathbf{X}_i) - \mu_1(\mathbf{X}_i)\}] \\
&\quad\quad \times E[\tilde{w}_0(\mathbf{X}_i)\{1 - e(\mathbf{X}_i)\}\{\mu_0(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)\}] \\
&\quad = 0 \times 0 = 0.
\end{aligned}$$

By a similar argument, $E[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}]E[\{\tau(\mathbf{X}_i) - \tau\}] = E[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}]E[\{\tau(\mathbf{X}_i) - \tau\}] = 0$ so it remains to show that $E[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}\{\tau(\mathbf{X}_i) - \tau\}] = E[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}\{\tau(\mathbf{X}_i) - \tau\}] = 0$. Now,

$$\begin{aligned}
E[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}\{\tau(\mathbf{X}_i) - \tau\}] &= E(E[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}\{\tau(\mathbf{X}_i) - \tau\}|\mathbf{X}_i]) \\
&= E(\tilde{w}_1(\mathbf{X}_i)\{\tau(\mathbf{X}_i) - \tau\}E[Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i]) \\
&\quad \text{(by ignorability)} = E[\tilde{w}_1(\mathbf{X}_i)\{\tau(\mathbf{X}_i) - \tau\}e(\mathbf{X}_i)\{\mu_1(\mathbf{X}_i) - \mu_1(\mathbf{X}_i)\}] \\
&= 0
\end{aligned}$$

and by a similar argument $E[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}\{\tau(\mathbf{X}_i) - \tau\}] = 0$ such that all

covariance terms are zero. Next,

$$\begin{aligned}
\text{Var}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}] &= E(\text{Var}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i]) \\
&\quad + \text{Var}(E[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i]) \\
(\text{by similar argument to above}) &= E(\text{Var}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i]) + \text{Var}(0) \\
&= E(\tilde{w}_1(\mathbf{X}_i)^2\text{Var}[Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i]).
\end{aligned}$$

Now, note that

$$\begin{aligned}
\text{Var}[Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i] &= \text{Var}(Z_i|\mathbf{X}_i)\text{Var}\{\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i\} \\
&\quad + \text{Var}(Z_i|\mathbf{X}_i)E\{\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i\}^2 \\
&\quad + \text{Var}\{\{Y_i(1) - \mu_1(\mathbf{X}_i)\}|\mathbf{X}_i\}E[Z_i|\mathbf{X}_i]^2 \\
&= e(\mathbf{X}_i)\{1 - e(\mathbf{X}_i)\}\text{Var}\{Y_i(1)|\mathbf{X}_i\} \\
&\quad + e(\mathbf{X}_i)^2\{1 - e(\mathbf{X}_i)\}\text{Var}\{Y_i(1)|\mathbf{X}_i\} \\
&= e(\mathbf{X}_i)(1 - e(\mathbf{X}_i) + e(\mathbf{X}_i))\text{Var}\{Y_i(1)|\mathbf{X}_i\} \\
&= e(\mathbf{X}_i)\text{Var}\{Y_i(1)|\mathbf{X}_i\}.
\end{aligned}$$

Then,

$$\text{Var}[\tilde{w}_1(\mathbf{X}_i)Z_i\{Y_i(1) - \mu_1(\mathbf{X}_i)\}] = E[\tilde{w}_1(\mathbf{X}_i)^2e(\mathbf{X}_i)\text{Var}\{Y_i(1)|\mathbf{X}_i\}]$$

and by a similar argument,

$$\text{Var}[\tilde{w}_0(\mathbf{X}_i)(1 - Z_i)\{Y_i(0) - \mu_0(\mathbf{X}_i)\}] = E[\tilde{w}_0(\mathbf{X}_i)^2\{1 - e(\mathbf{X}_i)\}\text{Var}\{Y_i(0)|\mathbf{X}_i\}].$$

Finally, we have that

$$\begin{aligned}
\text{Var}\{\tau(\mathbf{X}_i) - \tau\} &= E[\{\tau(\mathbf{X}_i) - \tau\}^2] - E[\tau(\mathbf{X}_i) - \tau]^2 \\
&= E[\{\tau(\mathbf{X}_i) - \tau\}^2] - (E[\tau(\mathbf{X}_i)] - \tau)^2 \\
&= E[\{\tau(\mathbf{X}_i) - \tau\}^2].
\end{aligned}$$

Then, we have that the asymptotic variance of the estimator when the propensity score model is misspecified is

$$E[\tilde{w}_1(\mathbf{X}_i)^2e(\mathbf{X}_i)\text{Var}\{Y_i(1)|\mathbf{X}_i\} + \tilde{w}_0(\mathbf{X}_i)^2\{1 - e(\mathbf{X}_i)\}\text{Var}\{Y_i(0)|\mathbf{X}_i\} + \{\tau(\mathbf{X}_i) - \tau\}^2].$$

□

1.8 Corollary 4.2

Proof. Assume that $\mu_z(\mathbf{x})$ and $\tau(\mathbf{x})$ are linear in both $c(\mathbf{x})$ and $g(\mathbf{x})$ such that $\mu_1(\mathbf{x}) = \beta_1^T c(\mathbf{x}) + \lambda_1^T g(\mathbf{x})$ and $\mu_0(\mathbf{x}) = \beta_0^T c(\mathbf{x}) + \lambda_0^T g(\mathbf{x})$ for some $\beta_z \in \mathbb{R}^K$ and $\lambda_z \in \mathbb{R}^L$ where $\beta_0 \neq \beta_1$ and $\lambda_0 \neq \lambda_1$. Instead of targeting the ATE, τ , $\hat{\tau}_{w^*}$ targets $\tau_{g, \tilde{w}} = E[E[Y|Z = 1, c(\mathbf{X}), \tilde{w}(\mathbf{X})e(\mathbf{X})g(\mathbf{X}_i)] - E[E[Y|Z = 0, c(\mathbf{X}), \tilde{w}(\mathbf{X})\{1 - e(\mathbf{X})\}g(\mathbf{X})]]] = E[\tau_{g, \tilde{w}}(\mathbf{X}, Z)]$ where $\tau_{g, \tilde{w}}(\mathbf{x}, z) = (\beta_1 - \beta_0)^T c(\mathbf{x}) + \lambda_1^T \tilde{w}_1(\mathbf{x})zg(\mathbf{x}) - \lambda_0^T \tilde{w}_0(\mathbf{x})(1 - z)g(\mathbf{x})$. We have the following decomposition of estimator error:

$$\begin{aligned}
\hat{\tau}_{w^*} - \tau_{g, \tilde{w}} &= \frac{1}{n} \sum_{i=1}^n w_1^*(\mathbf{X}_i) Z_i Y_i - \frac{1}{n} \sum_{i=1}^n w_0^*(\mathbf{X}_i) (1 - Z_i) Y_i - \tau_{g, \tilde{w}} \\
&= \frac{1}{n} \sum_{i=1}^n \tilde{w}_1(\mathbf{X}_i) Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} - \tau_{g, \tilde{w}} \\
&\quad + \frac{1}{n} \sum_{i=1}^n \{w_1^*(\mathbf{X}_i) - \tilde{w}_1(\mathbf{X}_i)\} Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} \\
&\quad - \frac{1}{n} \sum_{i=1}^n \{w_0^*(\mathbf{X}_i) - \tilde{w}_0(\mathbf{X}_i)\} (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} \\
&\quad + \frac{1}{n} \sum_{i=1}^n w_1^*(\mathbf{X}_i) Z_i \mu_1(\mathbf{X}_i) - \frac{1}{n} \sum_{i=1}^n w_0^*(\mathbf{X}_i) (1 - Z_i) \mu_0(\mathbf{X}_i) \\
&= \frac{1}{n} \sum_{i=1}^n \tilde{w}_1(\mathbf{X}_i) Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} \\
&\quad + \frac{1}{n} \sum_{i=1}^n \beta_1^T c(\mathbf{X}_i) - \beta_0^T c(\mathbf{X}_i) \\
&\quad + E[\lambda_1^T \tilde{w}_1(\mathbf{X}) Z g(\mathbf{X})] - E[\lambda_0^T \tilde{w}_0(\mathbf{X}) (1 - Z) g(\mathbf{X})] - \tau_{g, \tilde{w}} \\
&\quad + \frac{1}{n} \sum_{i=1}^n \{w_1^*(\mathbf{X}_i) - \tilde{w}_1(\mathbf{X}_i)\} Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} \\
&\quad - \frac{1}{n} \sum_{i=1}^n \{w_0^*(\mathbf{X}_i) - \tilde{w}_0(\mathbf{X}_i)\} (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} \\
&\quad + \frac{1}{n} \sum_{i=1}^n \lambda_1^T w_1^*(\mathbf{X}_i) Z_i g(\mathbf{X}_i) - E[\lambda_1^T \tilde{w}_1(\mathbf{X}) Z g(\mathbf{X})] \\
&\quad - \frac{1}{n} \sum_{i=1}^n \lambda_0^T w_0^*(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i) + E[\lambda_0^T \tilde{w}_0(\mathbf{X}) (1 - Z) g(\mathbf{X})].
\end{aligned}$$

By Assumption 3 and the assumption that $\rho''\{\cdot\}$ and $\rho'''\{\cdot\}$ are continuous over \mathbb{R} , the results of Theorems 5.28 and 7.2 in Boos and Stefanski (2013) apply. Thus, there exist some

$q_1(\mathbf{X}_i, Z_i)$ and $q_0(\mathbf{X}_i, Z_i)$ with $E[q_z(\mathbf{X}_i, Z_i)] = 0$, $z = 0, 1$ such that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \boldsymbol{\lambda}_1^T w_1^*(\mathbf{X}_i) Z_i g(\mathbf{X}_i) - E[\boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{X}_i) Z_i g(\mathbf{X}_i)] &= \frac{1}{n} \sum_{i=1}^n q_1(\mathbf{X}_i, Z_i) + R_2 \\ \frac{1}{n} \sum_{i=1}^n \boldsymbol{\lambda}_0^T w_0^*(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i) - E[\boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i)] &= \frac{1}{n} \sum_{i=1}^n q_0(\mathbf{X}_i, Z_i) + R_3 \end{aligned}$$

where $R_2, R_3 = o_p(n^{-1/2})$. Note that we know the exact form of these two functions, $q_1(\mathbf{X}_i, Z_i)$ and $q_0(\mathbf{X}_i, Z_i)$, as specified in Theorem 5.28 in [Boos and Stefanski \(2013\)](#). Then $\hat{\tau}_w - \tau_{g, \tilde{w}} = S_i + R_1 + R_2 + R_3$ where

$$\begin{aligned} S_i &= \tilde{w}_1(\mathbf{X}_i) Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} - \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} \\ &\quad + \boldsymbol{\beta}_1^T c(\mathbf{X}_i) - \boldsymbol{\beta}_0^T c(\mathbf{X}_i) - E[\boldsymbol{\beta}_1^T c(\mathbf{X}_i) - \boldsymbol{\beta}_0^T c(\mathbf{X}_i)] + q_1(\mathbf{X}_i, Z_i) - q_0(\mathbf{X}_i, Z_i) \\ R_1 &= \frac{1}{n} \sum_{i=1}^n \{w_1^*(\mathbf{X}_i) - \tilde{w}_1(\mathbf{X}_i)\} Z_i \{Y_i(1) - \mu_1(\mathbf{X}_i)\} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \{w_0^*(\mathbf{X}_i) - \tilde{w}_0(\mathbf{X}_i)\} (1 - Z_i) \{Y_i(0) - \mu_0(\mathbf{X}_i)\} \end{aligned}$$

where R_1 has previously been shown to be $o_p(n^{-1/2})$. Then,

$$\begin{aligned} \text{Var}(\hat{\tau}_w - \tau_{g, \tilde{w}}) &= \text{Var}(S_i) \\ &= E[\tilde{w}_1(\mathbf{X}_i)^2 e(\mathbf{X}_i) \text{Var}\{Y_i(1)|\mathbf{X}_i\} + \tilde{w}_0(\mathbf{X}_i)^2 \{1 - e(\mathbf{X}_i)\} \text{Var}\{Y_i(0)|\mathbf{X}_i\}] \\ &\quad + E \left[\left\{ \boldsymbol{\beta}_1^T c(\mathbf{X}_i) - \boldsymbol{\beta}_0^T c(\mathbf{X}_i) - E[\boldsymbol{\beta}_1^T c(\mathbf{X}_i) - \boldsymbol{\beta}_0^T c(\mathbf{X}_i)] \right. \right. \\ &\quad \left. \left. + q_1(\mathbf{X}_i, Z_i) - q_0(\mathbf{X}_i, Z_i) \right\}^2 \right] \end{aligned}$$

as all covariance terms cancel as shown in Corollary 4.1. It remains to determine $q_1(\mathbf{X}_i, Z_i)$ and $q_0(\mathbf{X}_i, Z_i)$. By Theorem 5.28 in [Boos and Stefanski \(2013\)](#),

$$\begin{aligned} q_1(\mathbf{X}_i, Z_i) &= \boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{X}_i) Z_i g(\mathbf{X}_i) - E[\boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{X}_i) Z_i g(\mathbf{X}_i)] + r_1(\mathbf{X}_i, Z_i) \quad \text{and} \\ q_0(\mathbf{X}_i, Z_i) &= \boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i) - E[\boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i)] + r_0(\mathbf{X}_i, Z_i), \end{aligned}$$

where $r_z(\mathbf{X}_i, Z_i)$ depend on the asymptotic distribution of $w_1^*(\mathbf{X}_i)$ and $w_0^*(\mathbf{X}_i)$. Therefore, to derive $r_z(\mathbf{X}_i, Z_i)$, we first derive the asymptotic distribution of $\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\gamma}}$ using the theory of M-estimation.

Let $m_k = E[c_k(\mathbf{X}_i)]$ and $\mathbf{m} = (m_1, \dots, m_K)$. Let $\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0, \boldsymbol{\gamma})$. Let

$$\begin{aligned}\phi_k(\mathbf{X}, \mathbf{m}) &= c_k(\mathbf{X}) - m_k \\ \varphi_k^{(1)}(\mathbf{X}, Z, \mathbf{m}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}) &= \rho' \{ \boldsymbol{\alpha}_1^T c(\mathbf{X}) + \boldsymbol{\gamma}^T g(\mathbf{X}) \} Z c_k(\mathbf{X}) - m_k \\ \varphi_k^{(0)}(\mathbf{X}, Z, \mathbf{m}, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}) &= \rho' \{ \boldsymbol{\alpha}_0^T c(\mathbf{X}) - \boldsymbol{\gamma}^T g(\mathbf{x}) \} (1 - Z) c_k(\mathbf{X}) - m_k \\ \varphi_l^{(2)}(\mathbf{X}, Z, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}) &= \rho' \{ \boldsymbol{\alpha}_1^T c(\mathbf{X}) + \boldsymbol{\gamma}^T g(\mathbf{X}) \} Z g_l(\mathbf{X}) - \rho' \{ \boldsymbol{\alpha}_0^T c(\mathbf{X}) - \boldsymbol{\gamma}^T g(\mathbf{X}) \} (1 - Z) g_l(\mathbf{X})\end{aligned}$$

Let $\psi(\mathbf{X}, Z, \boldsymbol{\theta}) = (\phi, \varphi^{(1)}, \varphi^{(0)}, \varphi^{(2)})$. Then $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\gamma}$ is the solution to

$$\frac{1}{n} \sum_{i=1}^n \psi(\mathbf{X}_i, Z_i, \boldsymbol{\theta}) = 0.$$

Given the assumptions on $\rho\{\cdot\}$, Assumption 3, and similar arguments to the proof of Lemma 6, by the theory of M-estimation and Theorem 7.2 in Boos and Stefanski (2013),

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = \sum_{i=1}^n \mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}_i, Z_i, \tilde{\boldsymbol{\theta}}) + R_4$$

where $R_4 = o_p(n^{-1/2})$. We will now determine $\mathbf{A}(\tilde{\boldsymbol{\theta}})$. For ease in notation, let $h'_1 = \rho'' \{ \tilde{\boldsymbol{\alpha}}_1^T c(\mathbf{x}) + \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x}) \} Z$ and $h'_0 = \rho'' \{ \tilde{\boldsymbol{\alpha}}_0^T c(\mathbf{x}) - \tilde{\boldsymbol{\gamma}}^T g(\mathbf{x}) \} (1 - Z)$ and $\mathbf{H}_{\mathbf{a}, \mathbf{b}} = E(\mathbf{a}\mathbf{b}^T) = \text{Cov}(\mathbf{a}, \mathbf{b}) + E[\mathbf{a}]E[\mathbf{b}]^T$. Then,

$$\begin{aligned}A(\tilde{\boldsymbol{\theta}}) &= E \begin{bmatrix} \mathbf{I}_K & \mathbf{0}_K & \mathbf{0}_K & \mathbf{0}_{K \times L} \\ \mathbf{I}_K & -h'_1 c(\mathbf{X}) c(\mathbf{X})^T & \mathbf{0}_K & -h'_1 c(\mathbf{X}) g(\mathbf{X})^T \\ \mathbf{I}_K & \mathbf{0}_K & -h'_0 c(\mathbf{X}) c(\mathbf{X})^T & -h'_0 c(\mathbf{X}) g(\mathbf{X})^T \\ \mathbf{0}_{L \times K} & -h'_1 g(\mathbf{X}) c(\mathbf{X})^T & h'_0 g(\mathbf{X}) c(\mathbf{X})^T & \{-h'_1 - h'_0\} g(\mathbf{X}) g(\mathbf{X})^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_K & \mathbf{0}_K & \mathbf{0}_K & \mathbf{0}_{K \times L} \\ \mathbf{I}_K & -\mathbf{H}_{h'_1 c(\mathbf{X}), c(\mathbf{X})} & \mathbf{0}_K & -\mathbf{H}_{h'_1 c(\mathbf{X}), g(\mathbf{X})} \\ \mathbf{I}_K & \mathbf{0}_K & -\mathbf{H}_{h'_0 c(\mathbf{X}), c(\mathbf{X})} & -\mathbf{H}_{h'_0 c(\mathbf{X}), g(\mathbf{X})} \\ \mathbf{0}_{L \times K} & -\mathbf{H}_{h'_1 g(\mathbf{X}), c(\mathbf{X})} & \mathbf{H}_{h'_0 g(\mathbf{X}), c(\mathbf{X})} & -\mathbf{H}_{(h'_1 + h'_0) g(\mathbf{X}), g(\mathbf{X})} \end{bmatrix}\end{aligned}$$

Symbolically inverting $\mathbf{A}(\tilde{\boldsymbol{\theta}})$ is non-trivial, so we proceed with $\mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}, Z, \tilde{\boldsymbol{\theta}})$ rather than the simplified form. Let $\mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}, Z, \tilde{\boldsymbol{\theta}})_{\tilde{\delta}}$ be the rows of $\mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}, Z, \tilde{\boldsymbol{\theta}})$ corresponding

to any $\tilde{\delta} \subset \tilde{\theta}$. Then, by Theorem 5.28 in [Boos and Stefanski \(2013\)](#),

$$\begin{aligned} q_1(\mathbf{X}_i, Z_i) &= \boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{X}_i) Z_i g(\mathbf{X}_i) - E[\boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{X}_i) Z_i g(\mathbf{X}_i)] \\ &\quad + E \left[\frac{\partial}{\partial(\tilde{\boldsymbol{\alpha}}_1, \tilde{\gamma})} \boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{X}_i) Z_i g(\mathbf{X}_i) \right] \mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}_i, Z_i, \tilde{\boldsymbol{\theta}})_{(\tilde{\boldsymbol{\alpha}}_1, \tilde{\gamma})} \\ q_0(\mathbf{X}_i, Z_i) &= \boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i) - E[\boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i)] \\ &\quad + E \left[\frac{\partial}{\partial(\tilde{\boldsymbol{\alpha}}_0, \tilde{\gamma})} \boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i) \right] \mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}_i, Z_i, \tilde{\boldsymbol{\theta}})_{(\tilde{\boldsymbol{\alpha}}_0, \tilde{\gamma})}, \end{aligned}$$

where

$$\begin{aligned} E \left[\frac{\partial}{\partial(\tilde{\boldsymbol{\alpha}}_1, \tilde{\gamma})} \boldsymbol{\lambda}_1^T \tilde{w}_1(\mathbf{X}_i) Z_i g(\mathbf{X}_i) \right] &= [\boldsymbol{\lambda}_1^T \mathbf{H}_{h'_1 g(\mathbf{X}), c(\mathbf{X})}, \boldsymbol{\lambda}_1^T \mathbf{H}_{h'_1 g(\mathbf{X}), g(\mathbf{X})}] \text{ and} \\ E \left[\frac{\partial}{\partial(\tilde{\boldsymbol{\alpha}}_0, \tilde{\gamma})} \boldsymbol{\lambda}_0^T \tilde{w}_0(\mathbf{X}_i) (1 - Z_i) g(\mathbf{X}_i) \right] &= [\boldsymbol{\lambda}_0^T \mathbf{H}_{h'_0 g(\mathbf{X}), c(\mathbf{X})}, -\boldsymbol{\lambda}_0^T \mathbf{H}_{h'_0 g(\mathbf{X}), g(\mathbf{X})}]. \end{aligned}$$

Then, we have that

$$\begin{aligned} \text{Var}(\hat{\tau}_w - \tau_{g, \tilde{w}}) &= E[\tilde{w}_1(\mathbf{X}_i)^2 e(\mathbf{X}_i) \text{Var}\{Y_i(1)|\mathbf{X}_i\} + \tilde{w}_0(\mathbf{X}_i)^2 \{1 - e(\mathbf{X}_i)\} \text{Var}\{Y_i(0)|\mathbf{X}_i\}] \\ &\quad + E[\{\tau_{g, \tilde{w}}(\mathbf{X}_i, Z_i) - \tau_{g, \tilde{w}} + r_1(\mathbf{X}_i, Z_i) - r_0(\mathbf{X}_i, Z_i)\}^2] \\ &= V_{g, \tilde{w}} \end{aligned}$$

for $r_1(\mathbf{X}_i, Z_i) = [\boldsymbol{\lambda}_1^T \mathbf{H}_{h'_1 g(\mathbf{X}), c(\mathbf{X})}, \boldsymbol{\lambda}_1^T \mathbf{H}_{h'_1 g(\mathbf{X}), g(\mathbf{X})}] \mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}_i, Z_i, \tilde{\boldsymbol{\theta}})_{(\tilde{\boldsymbol{\alpha}}_1, \tilde{\gamma})}$ and $r_0(\mathbf{X}_i, Z_i) = [\boldsymbol{\lambda}_0^T \mathbf{H}_{h'_0 g(\mathbf{X}), c(\mathbf{X})}, -\boldsymbol{\lambda}_0^T \mathbf{H}_{h'_0 g(\mathbf{X}), g(\mathbf{X})}] \mathbf{A}^{-1}(\tilde{\boldsymbol{\theta}}) \psi(\mathbf{X}_i, Z_i, \tilde{\boldsymbol{\theta}})_{(\tilde{\boldsymbol{\alpha}}_0, \tilde{\gamma})}$. Thus,

$$\sqrt{n}(\hat{\tau}_w - \tau_{g, \tilde{w}}) \rightarrow_d N(0, V_{g, \tilde{w}})$$

as desired. □

2 Restrictions on the implied propensity score model

We explore the restrictions on the implied propensity score model for two commonly used measures of dispersion: 1) entropy weights ([Hainmueller, 2012](#)); and 2) stable balancing weights ([Zubizarreta, 2015](#)). The first order conditions in (7) imply the following condition:

$$(\rho')^{-1}(\{e(\mathbf{x})\}^{-1}) + (\rho')^{-1}(\{1 - e(\mathbf{x})\}^{-1}) = (\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1)^T c(\mathbf{x})$$

Without a defining a specific measure of dispersion $D(\cdot)$ (and therefore a specific $\rho(\cdot)$) it is hard to connect this to $\text{Var}(Z|\mathbf{X})$. However, we explore two common deviance measures

(Wang and Zubizarreta, 2020):

1. Entropy weights: $D(t) = t \log(t) \implies \rho'(t) = e^{-t-1} \implies (\rho')^{-1}(t) = -\log(t) - 1$

2. Stable balancing weights: $D(t) = t^2 \implies \rho'(x) = -\frac{t}{2} \implies (\rho')^{-1}(x) = -2t$

Consider the entropy measure of weight dispersion first. Applying this to the first order conditions in (7), we get the following

$$\begin{aligned}\log\{e(\mathbf{x})\} &= \boldsymbol{\alpha}_1^T c(\mathbf{x}) + \boldsymbol{\gamma}^T g(\mathbf{x}) + 1 \\ \log\{1 - e(\mathbf{x})\} &= \boldsymbol{\alpha}_1^T c(\mathbf{x}) - \boldsymbol{\gamma}^T g(\mathbf{x}) + 1\end{aligned}$$

$$\begin{aligned}\log\{e(\mathbf{x})\} + \log\{1 - e(\mathbf{x})\} &= (\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1)^T c(\mathbf{x}) + 2 \\ \log\{e(\mathbf{x})(1 - e(\mathbf{x}))\} &= (\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1)^T c(\mathbf{x}) + 2 \\ \text{Var}(Z|\mathbf{X} = \mathbf{x}) &= \exp\{(\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1)^T c(\mathbf{x}) + 2\}\end{aligned}$$

such that $\text{Var}(Z|\mathbf{X} = \mathbf{x})$ only varies by $c(\mathbf{x})$ in the implied propensity score model. Now, consider the dispersion measure used for stable balancing weights:

$$\begin{aligned}\frac{1}{e(\mathbf{x})} &= -2\boldsymbol{\alpha}_1^T c(\mathbf{x}) - 2\boldsymbol{\gamma}^T g(\mathbf{x}) \\ \frac{1}{1 - e(\mathbf{x})} &= -2\boldsymbol{\alpha}_0^T c(\mathbf{x}) + 2\boldsymbol{\gamma}^T g(\mathbf{x}) \\ \frac{1}{e(\mathbf{x})} + \frac{1}{1 - e(\mathbf{x})} &= -2\boldsymbol{\alpha}_1^T c(\mathbf{x}) - 2\boldsymbol{\alpha}_0^T c(\mathbf{x}) \\ \frac{1}{e(\mathbf{x})(1 - e(\mathbf{x}))} &= -2\boldsymbol{\alpha}_1^T c(\mathbf{x}) - 2\boldsymbol{\alpha}_0^T c(\mathbf{x}) \\ \text{Var}(Z|\mathbf{X} = \mathbf{x}) &= \frac{1}{-2\boldsymbol{\alpha}_1^T c(\mathbf{x}) - 2\boldsymbol{\alpha}_0^T c(\mathbf{x})}\end{aligned}$$

and again, $\text{Var}(Z|\mathbf{X} = \mathbf{x})$ only varies by $c(\mathbf{x})$ in the implied propensity score model. While this relationship may not hold for every possible dispersion measure, for these two dispersion measures the implied propensity score model of Problem (5) allows $E[Z|\mathbf{X} = \mathbf{x}]$ to vary with all functions of covariates but $\text{Var}(Z|\mathbf{X} = \mathbf{x})$ to vary only with the set of covariates that are balanced to the population. Since this relationship does not hold for any arbitrary $c(\mathbf{x})$ and $g(\mathbf{x})$ sets, this model places substantial restrictions on what forms these sets can take for the implied propensity score model to be true. Given a specific $c(\mathbf{x})$ set, the $g(\mathbf{x})$ set that satisfies the implied propensity score model can be identified for each $D(\cdot)$ by using the first order conditions (7a)-(7c) to solve for $\boldsymbol{\gamma}^T g(\mathbf{x})$ as a function of $\boldsymbol{\alpha}_z^T c(\mathbf{x})$.

3 Connections to other estimation problems

3.1 Multiple treatment groups

Consider the case where there are $M > 2$ treatment groups such that $Z_i = m$ for $m = 0, \dots, M - 1$ and where we want to estimate the average treatment effect between treatments $m_1, m_2 = 0, \dots, M - 1$, $\tau_{m_1, m_2} = E[\mu_{m_1}(\mathbf{x}) - \mu_{m_2}(\mathbf{x})]$. Adapting Problem (5) to derive weights for estimating τ_{m_1, m_2} simply requires replacing $z = 0, 1$ with $z = m_1, m_2$ in Problem (5). However, if weights are derived in this way for a series of average treatment effects (e.g., $\tau_{m_1, m_2}, \tau_{m_2, m_3}, \tau_{m_1, m_3}$ etc.), each average treatment effect will correspond to a different population for the covariate set $g(\mathbf{x})$ or even a different $g(\mathbf{x})$ set. To maintain the same $g(\mathbf{x})$ covariate population for average treatment effects corresponding all contrasts for $M_1 \leq M$ treatments, we would include M_1 constraints balancing $c(\mathbf{x})$ in treatment groups to the sample population (e.g., constraint (5b)) and $M_1 - 1$ constraints balancing $g(\mathbf{x})$ between treatment groups (e.g., constraint (5c)). As M_1 increases, generally the set $g(\mathbf{x})$ will need to be larger for the existence of a solution. However, for there to be no estimator error for every contrast, $g(\mathbf{x})$ must contain no effect modifiers for *all* treatment groups. Thus, deriving weights for all contrasts separately may allow for a smaller $g(\mathbf{x})$ set, but this set and corresponding population may vary for each contrast. In contrast, deriving weights for all contrasts together will result in the same target population for all average treatment effect estimands but the set $g(\mathbf{x})$ may need to be large to derive the weights.

3.2 Weighted average treatment effects

Consider the estimand $E[h(\mathbf{X})\tau(\mathbf{X})]$ for some function $h(\mathbf{x})$. This as a weighted average treatment effect (WATE) where the population covariate density $f(\mathbf{x})$ is modified to $h(\mathbf{x})f(\mathbf{x})$. The ATO is the WATE when $h(\mathbf{x}) = e(\mathbf{x})\{1 - e(\mathbf{x})\}$. When adapting our proposed constrained optimization problem to derive weights for estimating the WATE, $\{w_i^{WATE}\}$, we can simply replace the right-hand side of constraint (5b) with $\sum_{i=1}^n h(\mathbf{X}_i)c_k(\mathbf{X}_i) / \sum_{i=1}^n h(\mathbf{X}_i)$. The resulting weights will balance $c(\mathbf{x})$ to the weighted population with density $h(\mathbf{x})f(\mathbf{x})$ while $g(\mathbf{x})$ will only be balanced between the treated and control groups. Similar to the ATT, all discussion and methods proposed in Sections 3.2 and 5 can be used for the WATE. However, if the population defined by $h(\mathbf{x})f(\mathbf{x})$ has improved overlap compared to the sample population, the set $g(\mathbf{x})$ required to yield a solution will be smaller than the set required to yield a solution for the ATE. Thus, various $h(\mathbf{x})$ functions could be used in combination with our proposed constrained optimization problem to improve overlap and yield a solution. Yet, this will modify a large portion of the sample covariate population to the population described by $h(\mathbf{x})f(\mathbf{x})$ such that this modified population ideally is of specific interest to the given research question.

3.3 Distributional balancing weights

Direct balancing weights tend to correspond to estimators with reduced variance in part because they focus only on balancing linear combinations of the set of basis functions $b(\mathbf{x})$. However, if the CATE is not only linear in this set of basis functions, the direct balancing weights estimator is generally biased for the ATE. This has motivated the development of distributional balancing weights which are derived to minimize the distance between each of the weighted treatment group empirical distributions and the population empirical distribution (Huling and Mak, 2024). For distances that are integral probability metrics, the corresponding distributional balancing weights balance all functions in a given class (e.g, weights proposed by Huling and Mak (2024) balance all functions in a particular Sobolev space). Thus, these weights balance a much wider class of functions than direct balancing weights, which is desirable as $\mu_z(\mathbf{X})$ is often unknown. Let \mathcal{D} a distributional distance measure, $F_n(\mathbf{x})$ be the empirical covariate distribution, and $F_{n,z,w}(\mathbf{x})$ be the weighted empirical covariate distribution in treatment group z . Then distributional weights commonly take the follow form,

$$\mathbf{w}_n = \operatorname{argmin}_{\mathbf{w}} \mathcal{D}\{F_{n,0,w}(\mathbf{x}), F_n(\mathbf{x})\} + \mathcal{D}\{F_{n,1,w}(\mathbf{x}), F_n(\mathbf{x})\}$$

where $\mathcal{D}\{F_{n,0,w}(\mathbf{x}), F_{n,1,w}(\mathbf{x})\}$ can also be added to minimization objective. While there is always a solution to this optimization problem, when there is a lack of overlap, the corresponding estimator may have inflated variance. To address this, Chen et al. (2024) propose weighting $\mathcal{D}\{F_{n,0,w}(\mathbf{x}), F_n(\mathbf{x})\} + \mathcal{D}\{F_{n,1,w}(\mathbf{x}), F_n(\mathbf{x})\}$ by a hyperparameter $0 \geq \alpha \geq 1$ and $\mathcal{D}\{F_{n,0,w}(\mathbf{x}), F_{n,1,w}(\mathbf{x})\}$ by $1 - \alpha$; increasing the relative weight of $\mathcal{D}\{F_{n,0,w}(\mathbf{x}), F_{n,1,w}(\mathbf{x})\}$ tends to decrease variance but increase bias due to deviation from the target population. When $\alpha = 0$, the corresponding weights target the ATO. However, this results in the entire covariate population changing for any $\alpha > 0$. Thus, we propose the following distributional weights,

$$\begin{aligned} \mathbf{w}_n = \operatorname{argmin}_{\mathbf{w}} & \mathcal{D}[F_{n,0,w}\{c(\mathbf{x})\}, F_n\{c(\mathbf{x})\}] + \mathcal{D}[F_{n,1,w}\{c(\mathbf{x})\}, F_n\{c(\mathbf{x})\}] \\ & + \mathcal{D}\{F_{n,0,w}(\mathbf{x}), F_{n,1,w}(\mathbf{x})\}, \end{aligned}$$

where the distribution of only a subset of the covariates (i.e., $c(\mathbf{x})$) is balanced to the empirical distribution, but all covariates are balanced between treated and control groups. In addition, $\mathcal{D}\{F_{n,0,w}(\mathbf{x}), F_{n,1,w}(\mathbf{x})\}$ could be replaced by $\mathcal{D}[F_{n,0,w}\{g(\mathbf{x})\}, F_{n,1,w}\{g(\mathbf{x})\}]$ if desired. Alternatively, Mak and Joseph (2018) define distributional distances with a kernel that allows for the variable influence of each covariate on the distance. This could be used within a distributional balancing weights set-up to weaken the impact of the $g(\mathbf{x})$ on the distributional distance between the weighted treatment groups and the sample population.

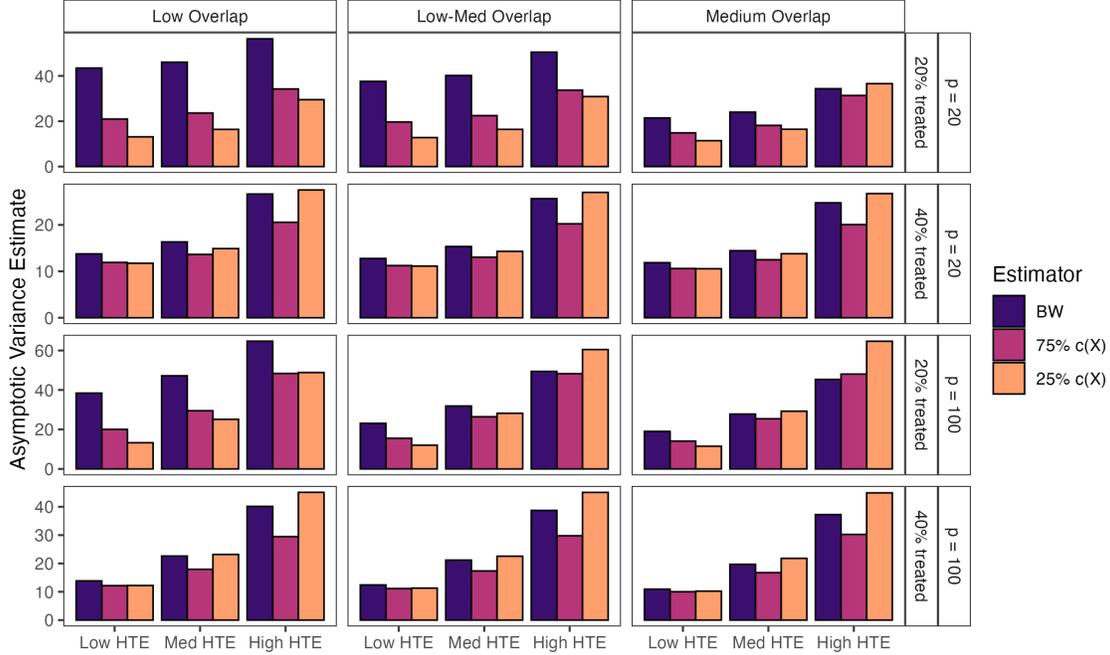


Figure 1: Estimated asymptotic variance estimates for all simulation scenarios. In the legend, “X% c(X)” indicates that X% of covariates are included in the $c(x)$ set for the PRTBW weighted estimators. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

4 Simulation study on estimator variance

We generate data using the same methods described in Section 6 and supplementary Section 6 with $\theta = 100\%$ of covariates being treatment effect modifiers. Thus, we generate a total of 36 simulation scenarios that are all combinations $p = 20, 100, 20\%$ and 40% treated, three levels of overlap, and three levels of treatment effect heterogeneity. We compute a Monte Carlo estimate of the asymptotic variance of the direct balancing weights estimator (Corollary 4.1) and the asymptotic variance of our proposed estimator when $c(\mathbf{x})$ includes 75% and 25% of all covariates (Corollary 4.2). We also compute the empirical variance of these estimators when computing the weights with and without the additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$.

Across a majority of scenarios, our proposed estimators have smaller estimated asymptotic variance than the direct balancing weights estimator (supplementary Figure 1). However, when there is high treatment effect heterogeneity, our proposed estimator with only 25% of covariates in the $c(\mathbf{x})$ set tends to have larger estimated asymptotic variance than the direct balancing weights estimator, likely due to the large variance of $\tau_{g, \tilde{w}}(\mathbf{X}, Z)$ compared to $\tau(\mathbf{X})$. Despite this, our proposed estimator with 75% of covariates in the $c(\mathbf{x})$ set has smaller

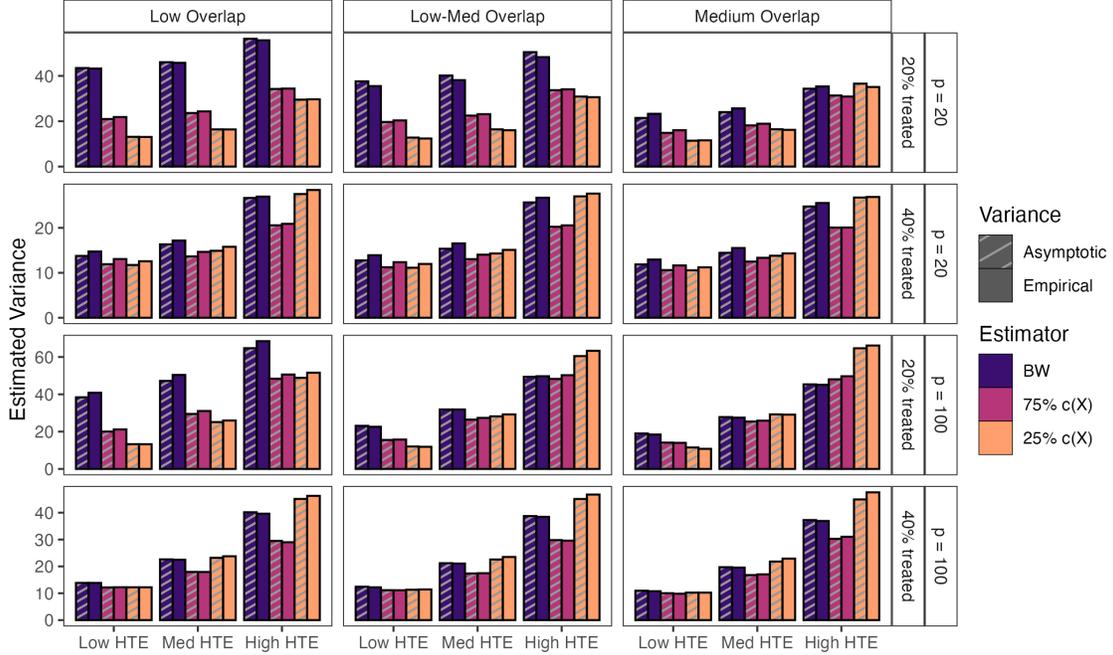


Figure 2: Estimated asymptotic variance estimates and empirical estimator variance for all simulation scenarios. In the legend, “X% c(X)” indicates that X% of covariates are included in the $c(x)$ set for the PRTBW weighted estimators. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

asymptotic variance than the direct balancing weights estimator across almost all scenarios. Thus, while our proposed estimator will not always have smaller asymptotic variance than the direct balancing estimator, our proposed estimator does have similar or lower variance across many scenarios. The empirical estimator variance is similar the estimated asymptotic variance for all estimators across all scenarios which further validates the estimated asymptotic variances (supplementary Figure 2).

While the asymptotic results do not directly apply with the additional constraints of $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ to Problem (5), we can explore the behavior of the estimators in this case with empirical estimator variance. In many scenarios, there is no solution to the direct balancing optimization problem with these additional constraints such that empirical estimator variance cannot be computed (supplementary Figure 3). Across all scenarios where direct balancing weights could be calculated, our proposed estimators have smaller or similar empirical variance to the direct balancing weights estimator; when overlap is poor our proposed estimators have substantially smaller empirical variance. Therefore, our proposed estimators tend to outperform the direct balancing weights estimator in terms of variance across almost all scenarios when the weights are constrained to not extrapolate.

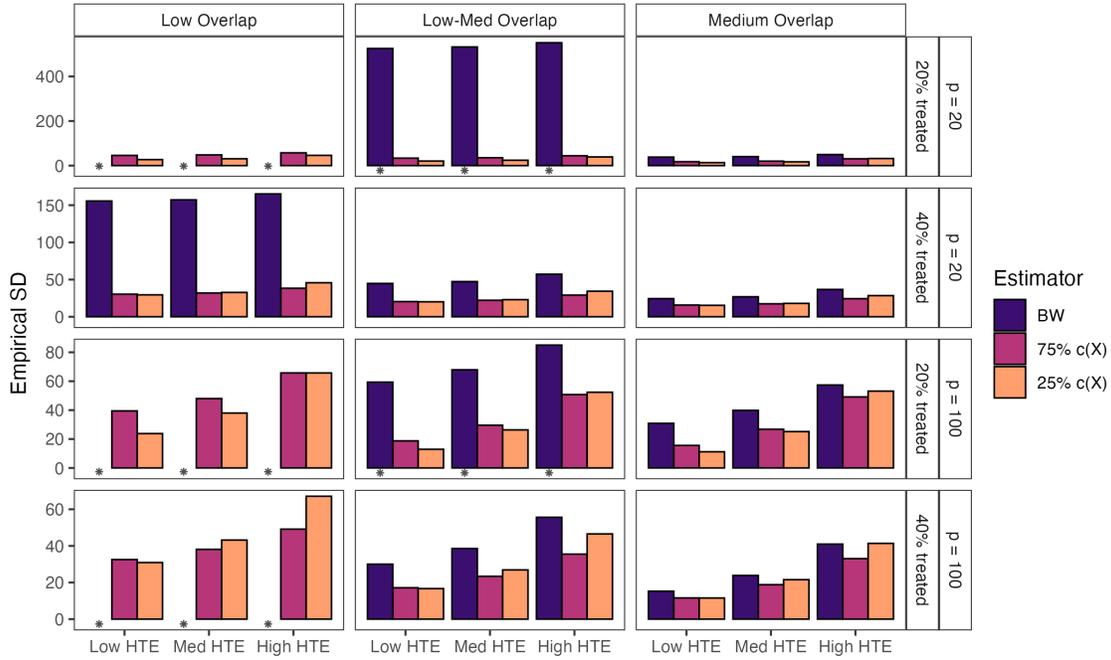


Figure 3: Empirical estimator variance for all simulation scenarios when including additional constraints of $w_i \geq 0$ and $\sum_{i=1}^n w_i Z_i = \sum_{i=1}^n w_i (1 - Z_i)$ in Problem (5). In the legend, “X% c(X)” indicates that X% of covariates are included in the $c(x)$ set for the weighted PRTBW estimators. The asterisk, *, indicates estimators where a solution to the balancing weights optimization problem did not exist for all 1,000 datasets. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

5 Additional guidance and implementation details

5.1 Guidance on selecting the $g(\mathbf{x})$ set

Selecting the $g(\mathbf{x})$ set requires researchers to evaluate the trade-offs between estimator variance and estimator bias due to the modified $g(\mathbf{x})$ population for their particular application. While these trade-offs are unavoidable when there are positivity violations, it is still desirable to select the $c(\mathbf{x})$ and $g(\mathbf{x})$ sets such that the corresponding estimator has minimal error with respect to an estimand of scientific interest, whether this the ATE or the average treatment effect for a modified population. Both subject matter knowledge as well as data driven techniques can be used to identify the $c(\mathbf{x})$ and $g(\mathbf{x})$ set; here, we provide further guidance on how a variety of factors and preferences may influence the selection of these two sets.

As mentioned in the previous section, the corresponding estimator for Problem (5) has minimal error with respect to the ATE when $c(\mathbf{x})$ contains all treatment effect modifiers effect modifiers. When the ATE is of scientific interest, it is therefore desirable to determine the set $g(\mathbf{x})$ that 1) yields a solution to Problem (5); and 2) minimally modifies the treatment effect. There are a variety of scenarios where set of effect modifiers is likely to be smaller than the set of confounders. Of particular interest are scenarios with high dimensional confounders, which tend to yield overlap issues (D’Amour et al., 2021), but where the set of effect modifiers potentially has moderate to low dimension. Subject matter knowledge may indicate which covariates are unlikely to be effect modifiers and thus which covariates should be in the $g(\mathbf{x})$. A data-driven procedure could also be used to select $g(\mathbf{x})$ by estimating how much each covariate influences treatment effect heterogeneity; in fact, we propose an algorithm and corresponding model-based estimator to do this Section 5.

In other scenarios, either 1) all covariates of interest may be effect modifiers or 2) the size of $g(\mathbf{x})$ required to ensure a solution to Problem (5) may be large such that some effect modifiers need to be included in the set. From a design-based perspective, $g(\mathbf{x})$ is the set of covariates that may differ from the original sample population; thus in these cases, it is desirable to choose $g(\mathbf{x})$ such that the resulting weighted population is still meaningful within the given scientific context. In these cases, $g(\mathbf{x})$ could be chosen as the set of covariates for which their population is of less interest or relevance to the original research question. However, when there are high dimensional covariates in may be challenging to determine which subset yields a solution to Problem (5) *and* whose corresponding population may be modified without substantially altering the original research question. As described in Section 3.1, the set $g(\mathbf{x})$ needs to be predictive of treatment assignment Z in order to substantially relax the positivity assumption. As such, there is some level of predictive capacity, η , that the covariate functions in $g(\mathbf{x})$ need to jointly satisfy in order for Problem (5) to yield a

solution; thus, there are many possible $g(\mathbf{x})$ sets from which to select. However, selecting the covariate functions that are individually the most predictive of treatment will tend to result in the smallest of such sets that have predictive capacity η . While selecting the smallest $g(\mathbf{x})$ set that yields a solution to Problem (5) may not minimize the design-based component of estimator error, it tends to both 1) simplify the interpretation of the resulting modified population; and 2) be more computationally efficient to operationalize. Thus, in Section 5, we proposed a design-based algorithm for identifying this set.

Furthermore, both model- and design-based perspectives can be used together to select a $g(\mathbf{x})$ set that best matches the original research objective (e.g., $g(\mathbf{x})$ contains a mix of covariates that are either unlikely to be effect modifiers or highly predictive of Z). Therefore, while a lack of overlap forces researchers to confront trade-offs between estimator variance, statistical bias, and bias due population modification, we propose a variety of intuitive arguments and tools such that researchers can incorporate their subject-matter knowledge and research priorities when navigating these tradeoffs within our proposed optimization procedure.

5.2 More details on the proposed model-based estimator

In Section 5, we propose a model-based estimator. To preserve downstream inference we use the following cross-fit style estimator,

$$\hat{\tau}_{w^*}^{CF} = \frac{|\mathcal{I}_1|}{n} \hat{\tau}^{\mathcal{I}_1} + \frac{|\mathcal{I}_2|}{n} \hat{\tau}^{\mathcal{I}_2}, \quad \hat{\tau}^{\mathcal{I}_1} = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \{\hat{w}_i^{*\mathcal{I}_2} Z_i Y_i - \hat{w}_i^{*\mathcal{I}_2} (1 - Z_i) Y_i\} \quad (10)$$

where \mathcal{I}_1 and \mathcal{I}_2 each contain half of the data, randomly split, $\hat{w}_i^{*\mathcal{I}_2}$ are weights computed with \mathcal{I}_2 and Algorithm 1 with a metric of treatment effect modification, and $\hat{\tau}^{\mathcal{I}_2}$ is defined in the same manner by swapping \mathcal{I}_1 and \mathcal{I}_2 . We implement the absolute value of the one-step doubly robust estimator of treatment effect modification,

$$TEM_j = \left| \frac{1}{\sum_i X_{ij}^2} \sum_{i=1}^n X_{ij} \left\{ (Y_i - \hat{\mu}_{Z_i}(\mathbf{X}_i)) \frac{2Z_i - 1}{Z_i \hat{e}(\mathbf{X}_i) + (1 - Z_i)(1 - \hat{e}(\mathbf{X}_i))} + \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i) \right\} \right|,$$

proposed by Boileau et al. (2025), as this treatment effect modification measure resulted in estimators with the lowest bias across a majority of the simulation scenarios, compared to the Hines et al. (2023) metric and the coefficients from a linear CATE model (supplementary Figure 4). In this estimator, $\hat{e}(\mathbf{x})$, $\hat{\mu}_z(\mathbf{x})$ are propensity score and outcome model estimates. Since propensity score estimates are likely to be extreme in scenarios with a lack of overlap we also explored $TEM_j^* = \text{abs}[\frac{1}{\sum_i X_{ij}^2} \sum_{i=1}^n X_{ij} \{\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)\}]$ which only requires outcome model estimates. In simulations, estimators corresponding to TEM_j tended to perform

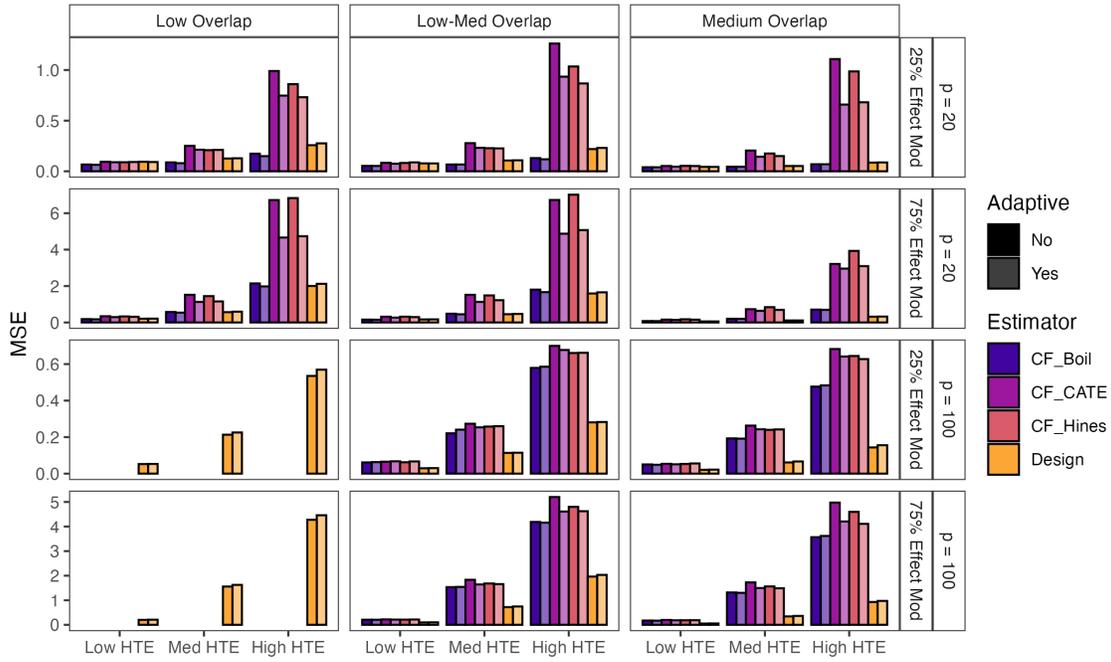


Figure 4: Comparison of estimator MSE with respect to the ATE for our proposed design-based estimator and model-based estimator (with the [Boileau et al. \(2025\)](#), [Hines et al. \(2023\)](#)), and linear CATE model coefficient metrics) for both supplementary Algorithm 1 (not-adaptive) and Algorithm 1 (adaptive). The scenarios are labeled on the top and right of the panels.

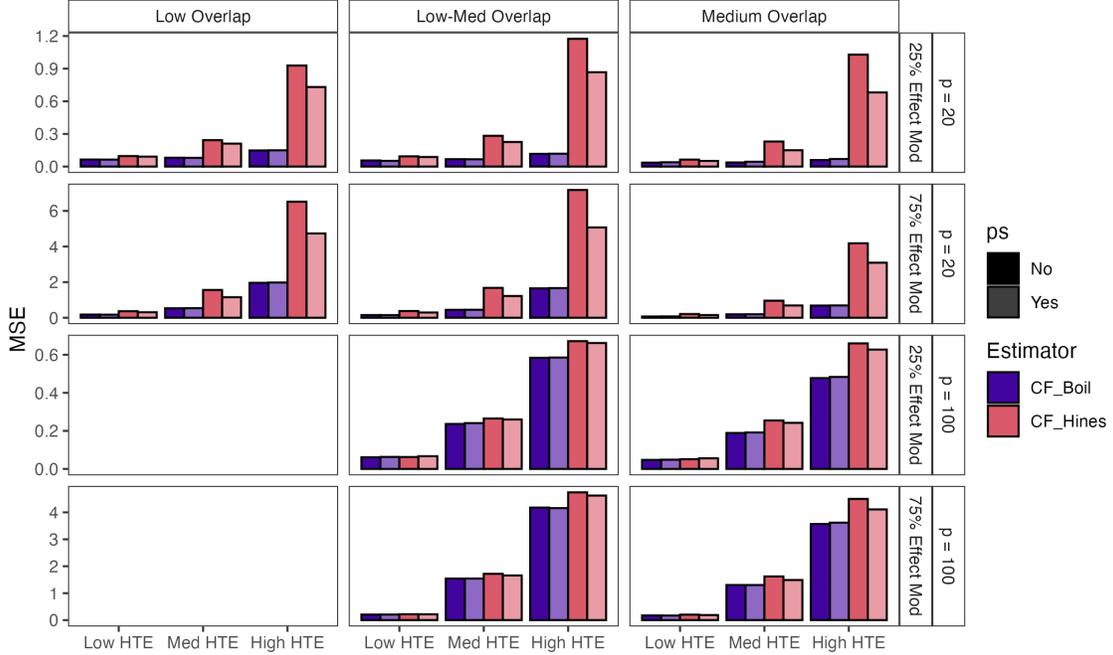


Figure 5: Comparison of estimator MSE for our model-based estimator with the [Boileau et al. \(2025\)](#) and [Hines et al. \(2023\)](#) metrics calculated with and without estimated propensity scores. The scenarios are labeled on the top and right of the panels.

similarly or better to estimators corresponding to TEM_j^* (supplementary Figure 5) such that we recommend and implement TEM_j in applications of our methods.

6 Additional simulation methods

For the data generation propensity score model, $\alpha_j = 0 \pm a$ for $a = 0.5/\gamma, 1.33/\gamma, 2.16/\gamma, 3/\gamma$ for $j > 0$. Here, a determines the strength of relationship between a given covariate and treatment assignment where there are an equal number of coefficients generated with each of $a = 0.5/\gamma, 1.33/\gamma, 2.16/\gamma, 3/\gamma$. As γ increases, propensity scores become more extreme and overlap decreases. We set $\mu_0(\mathbf{X}) = 0.25X_1 + \dots + 0.25X_p$ and $\mu_1(\mathbf{X}) = \sum_{i=0}^p \beta_i X_i$. We consider $\theta = 0.25, 0.75$ as the proportion of covariates are effect modifiers. We set $\beta_0 = 1$ and $\beta_j = 0.25 \pm c$ for $c = 0.75\delta, 0.6\delta, 0.45\delta, 0.3\delta, 0$. Here, c determines the strength of relationship between a given covariate and $\tau(\mathbf{X})$; for $(1 - \theta)p$ coefficients $\beta_j = 0$ and there are an approximately equal number of remaining coefficients set as each of $c = 0.75\delta, 0.6\delta, 0.45\delta, 0.3\delta$. Then, as δ increases, treatment effect heterogeneity increases for a given θ . Therefore, we generate 20 different potential types of covariates for each scenario (i.e., all combinations of α_j and β_j values). See supplementary Table 1 for the γ, δ values used for each scenario.

Table 1: Overlap and treatment heterogeneity hyperparameters used for data generation for each simulation scenario as described in Section 6 and supplementary Section 6.

	γ			δ		
	Low	Low-Med	Med	Low	Medium	High
20% Treated, p = 20	0.75	1	2	0.50	1	2
40% Treated, p = 20	0.50	0.75	1	0.50	1	2
20% Treated, p = 100	2	4	5	0.25	0.75	1.25
20% Treated, p = 100	1	2	3	0.25	0.75	1.25

Algorithm 1 Identifying an approximately optimal $g(\mathbf{X})$ set given metric m

- 1: **Inputs:** The matrix $\mathbf{X}_{n \times p}$ and set of covariate functions ordered according to some metric, $m, b_{(1)}(\mathbf{X}), \dots, b_{(J)}(\mathbf{X})$
 - 2: **Initialize:** $b(\mathbf{X})^- = \emptyset, j = 0$
 - 3: **while** $\{w_i^*\}_{i=1}^n$ given $g(\mathbf{X}) = b(\mathbf{X})^-$ does not exist **do**
 - 4: Let $j = j + 1$
 - 5: Let $b(\mathbf{X})^- = \{b(\mathbf{X})^-, b_{(j)}(\mathbf{X})\}$
 - 6: Solve Problem (5) for $g(\mathbf{X}) = b(\mathbf{X})^-$ and $c(\mathbf{X}) = \{b_{(1)}(\mathbf{X}), \dots, b_{(J)}(\mathbf{X})\} \setminus b(\mathbf{X})^-$
 - 7: **end while**
 - 8: **Outputs:**
 - 9: $b(\mathbf{X})^-$ \triangleright Smallest set that yields a solution to Problem (5) when $g(\mathbf{X}) = b(\mathbf{X})^-$
 - 10: $\{w_i^*\}_{i=1}^n$ \triangleright Resulting weights from (5) with $g(\mathbf{X}) = b(\mathbf{X})^-$
-

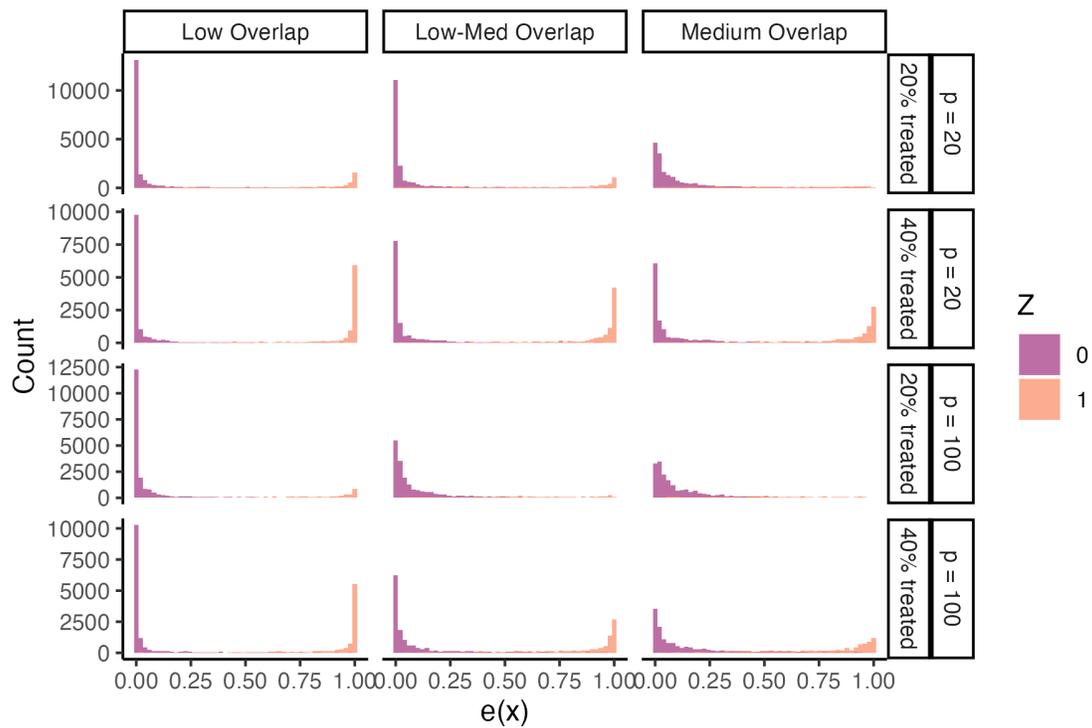


Figure 6: Propensity score distribution for the simulation scenarios described in Section 6. The scenarios are labeled on the top and right of the panels.

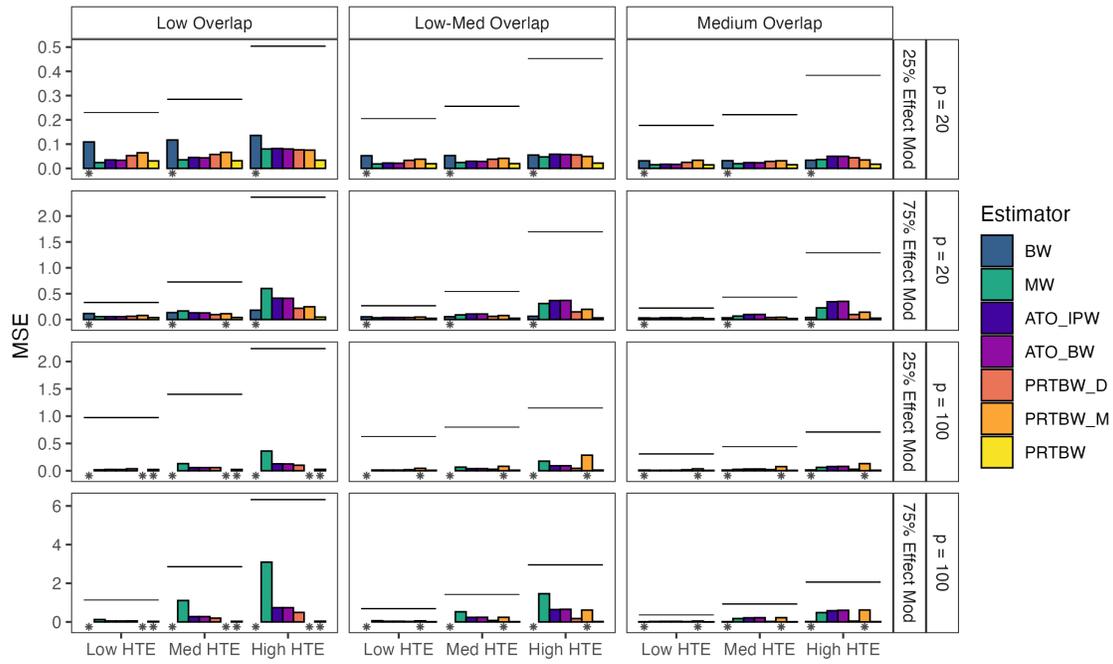


Figure 7: Estimator mean squared error (MSE) for all 40% treated simulation scenarios. The line indicates the MSE for the IPW ATE estimator for each scenario. The asterisk, *, indicates estimators where a solution to the balancing weights optimization problem did not exist for all datasets. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

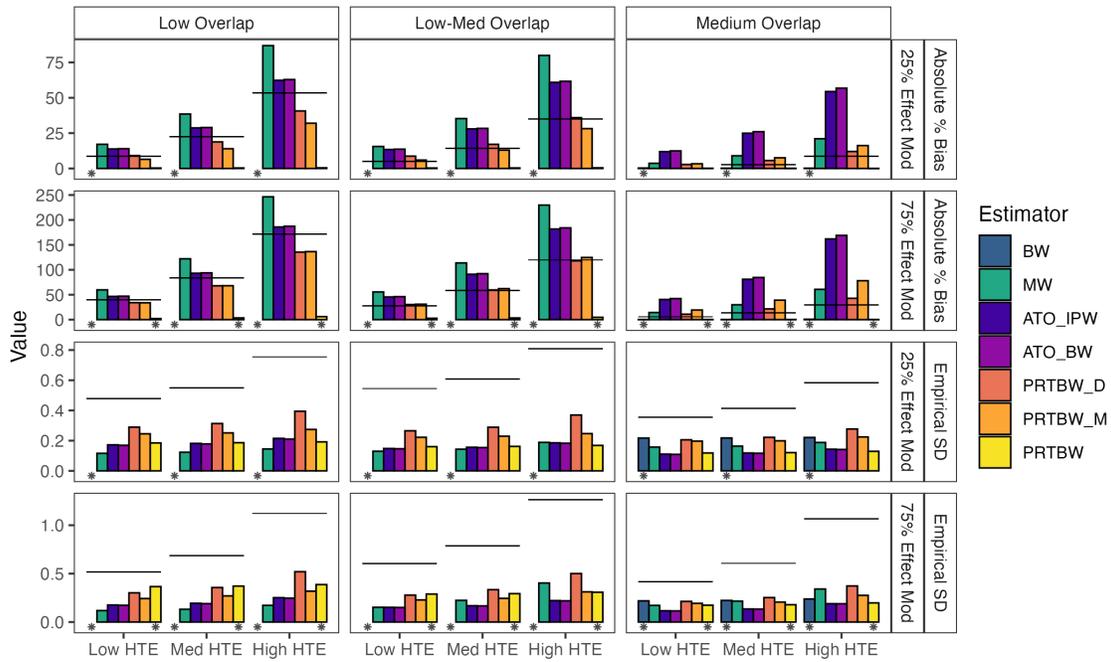


Figure 8: Estimator absolute percent bias with respect to the ATE and empirical standard deviation (SD) for 20% treated and $p = 20$ scenarios. The line indicates the MSE for the IPW ATE estimator for each scenario. The asterisk, *, indicates estimators where a solution to the balancing weights optimization problem did not exist for all datasets. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

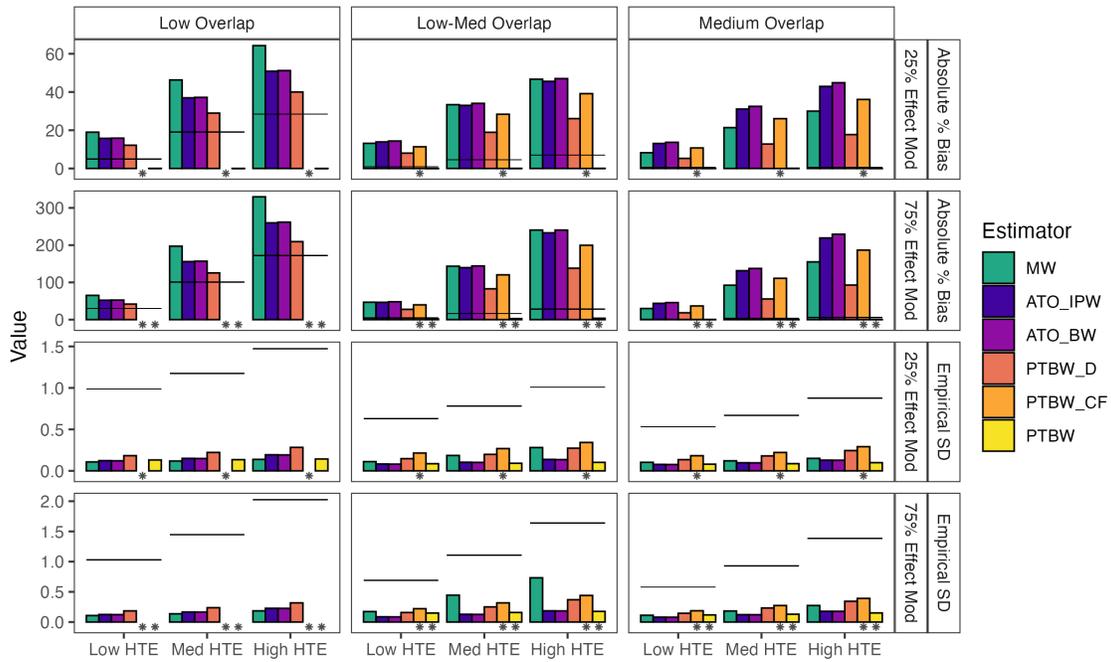


Figure 9: Estimator absolute percent bias with respect to the ATE and empirical standard deviation (SD) for 20% treated and $p = 100$ scenarios. The line indicates the MSE for the IPW ATE estimator for each scenario. The asterisk, *, indicates estimators where a solution to the balancing weights optimization problem did not exist for all datasets. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

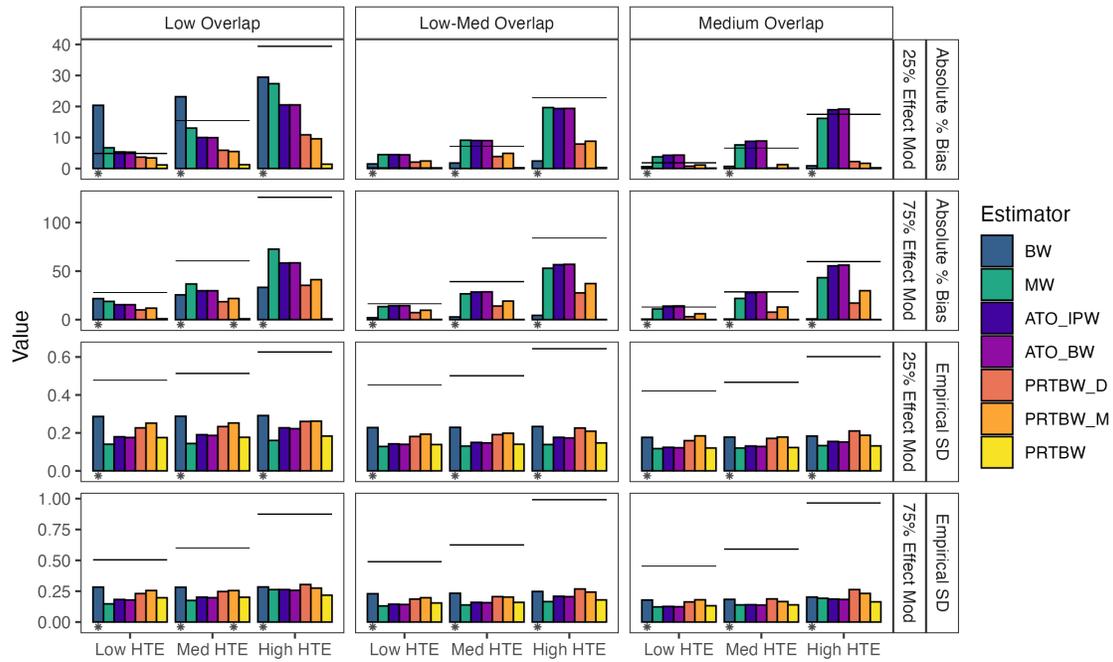


Figure 10: Estimator absolute percent bias with respect to the ATE and empirical standard deviation (SD) for 40% treated and $p = 20$ scenarios. The line indicates the MSE for the IPW ATE estimator for each scenario. The asterisk, *, indicates estimators where a solution to the balancing weights optimization problem did not exist for all datasets. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

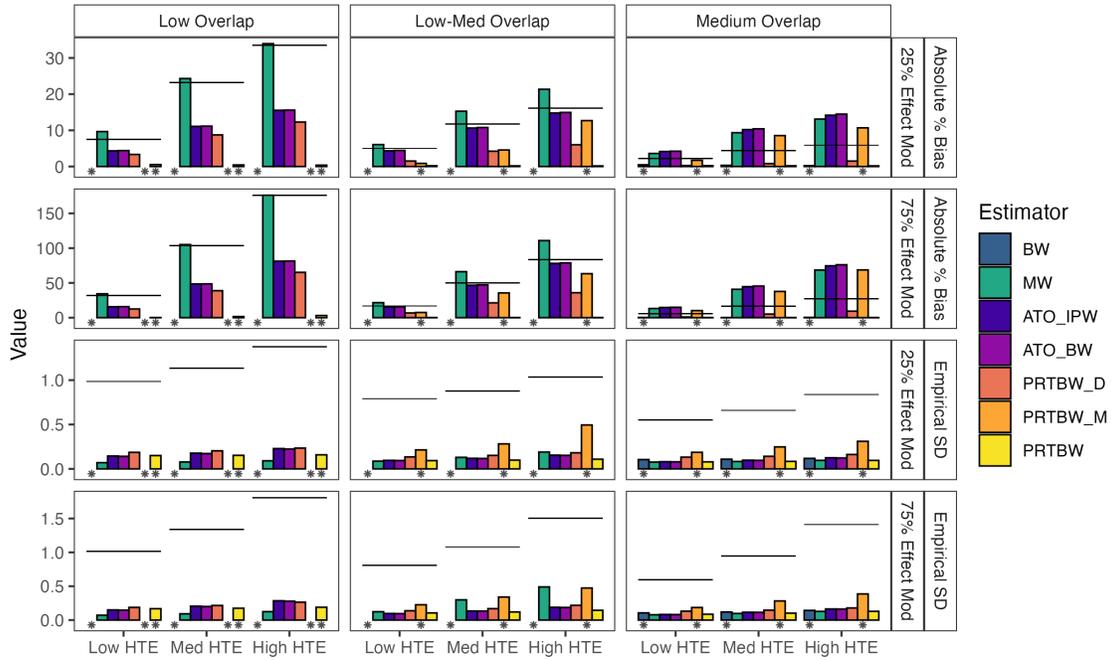


Figure 11: Estimator absolute percent bias with respect to the ATE and empirical standard deviation (SD) for 40% treated and $p = 100$ scenarios. The line indicates the MSE for the IPW ATE estimator for each scenario. The asterisk, *, indicates estimators where a solution to the balancing weights optimization problem did not exist for all datasets. The horizontal axis indicates levels of treatment effect heterogeneity while the simulation scenarios are labeled on the top and right of the panels.

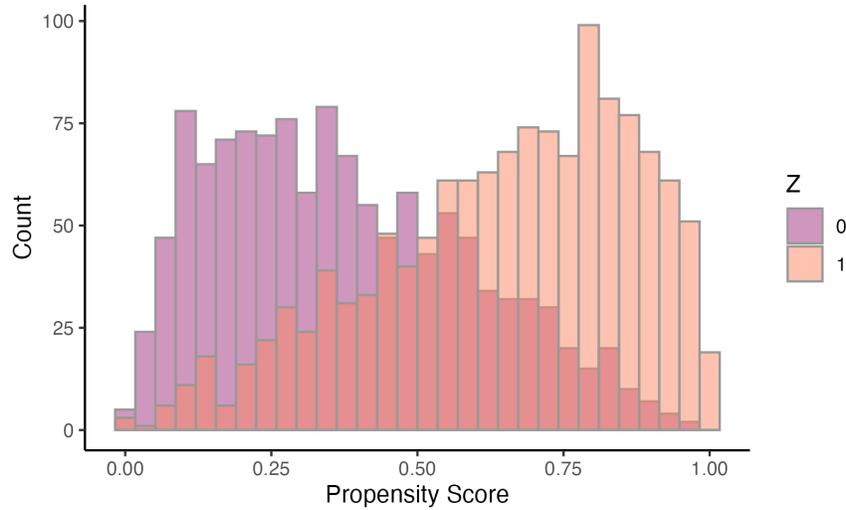


Figure 12: Distribution of estimated propensity scores for the study on indwelling arterial catheters (IAC) with data from the MIMIC-III database (Section 7.1)

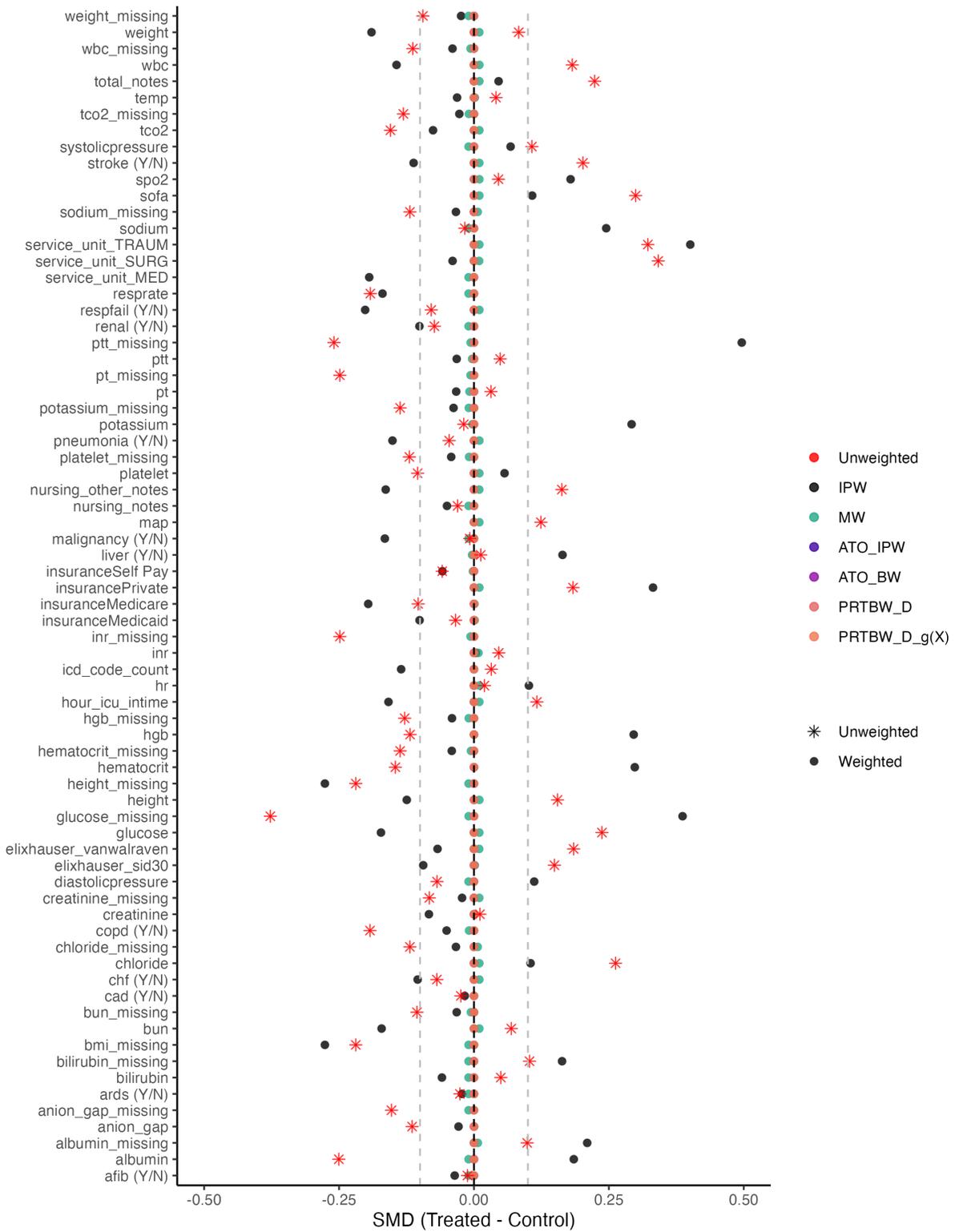


Figure 13: Treated and control group standardized mean differences (SMD) of all covariates for the study on indwelling arterial catheters (IAC) with data from the MIMIC-III database (Section 7.1).

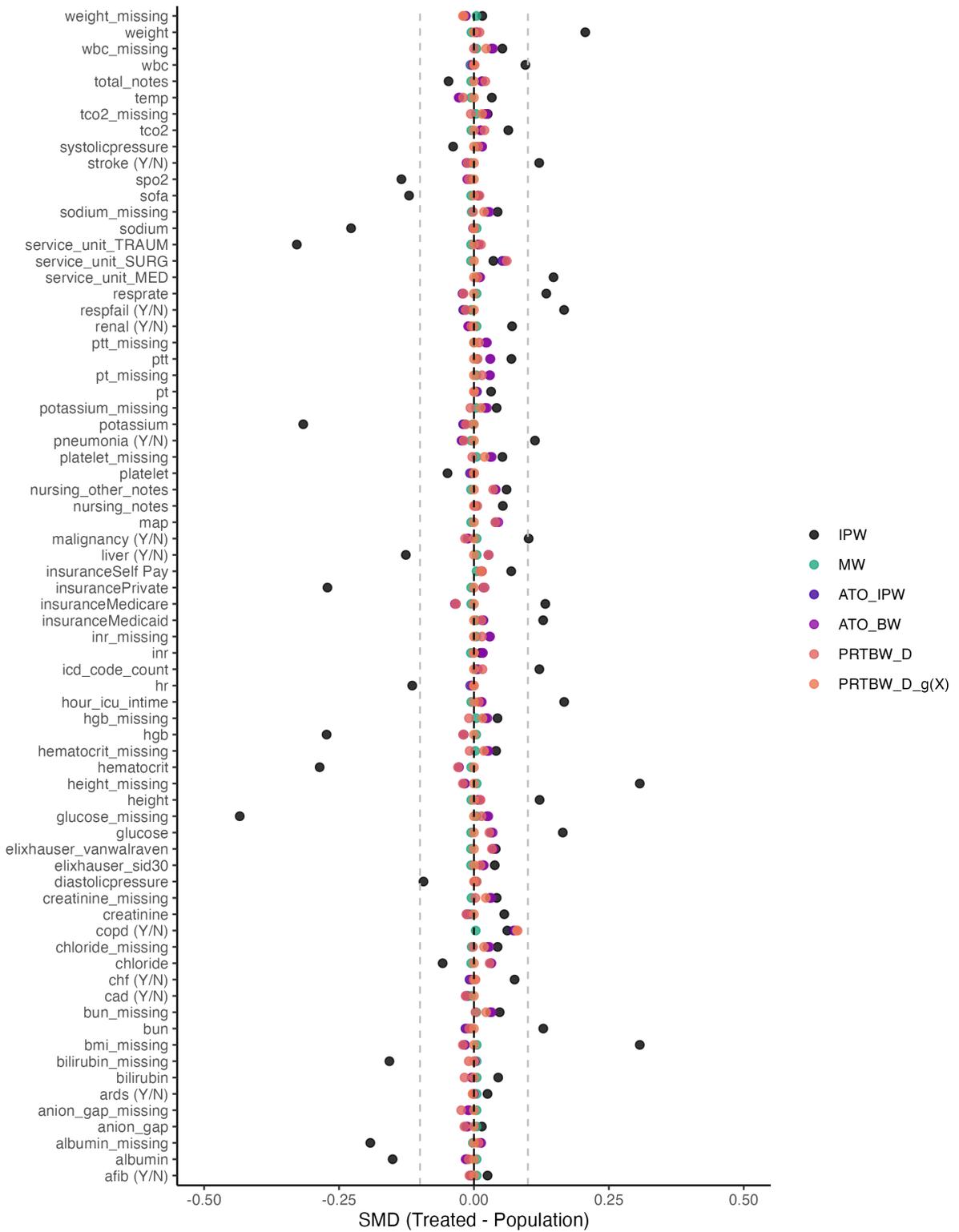


Figure 14: Treated and sample population standardized mean differences (SMD) of all covariates for the study on indwelling arterial catheters (IAC) with data from the MIMIC-III database (Section 7.1).

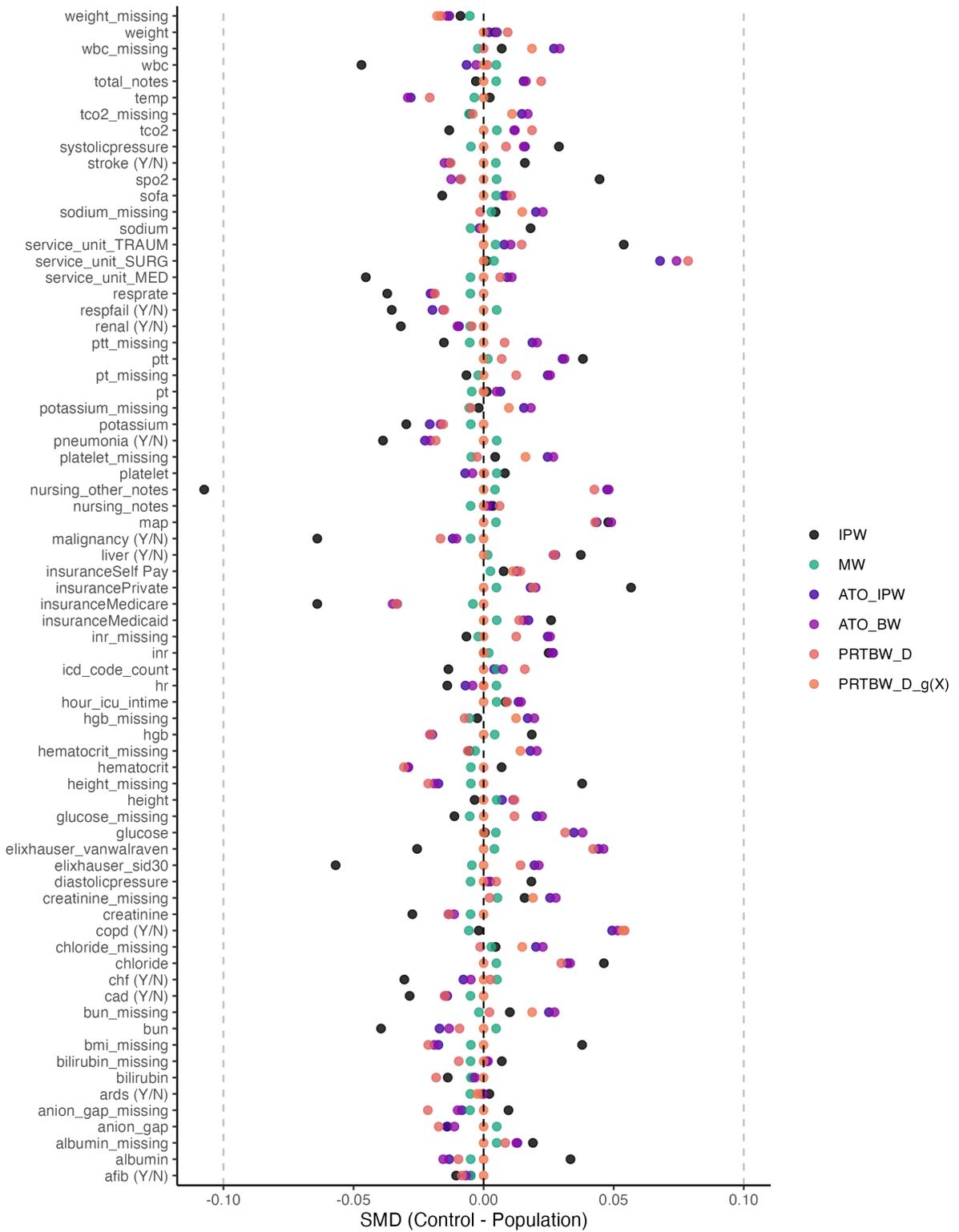


Figure 15: Control and sample population standardized mean differences (SMD) of all covariates for the study on indwelling arterial catheters (IAC) with data from the MIMIC-III database (Section 7.1).

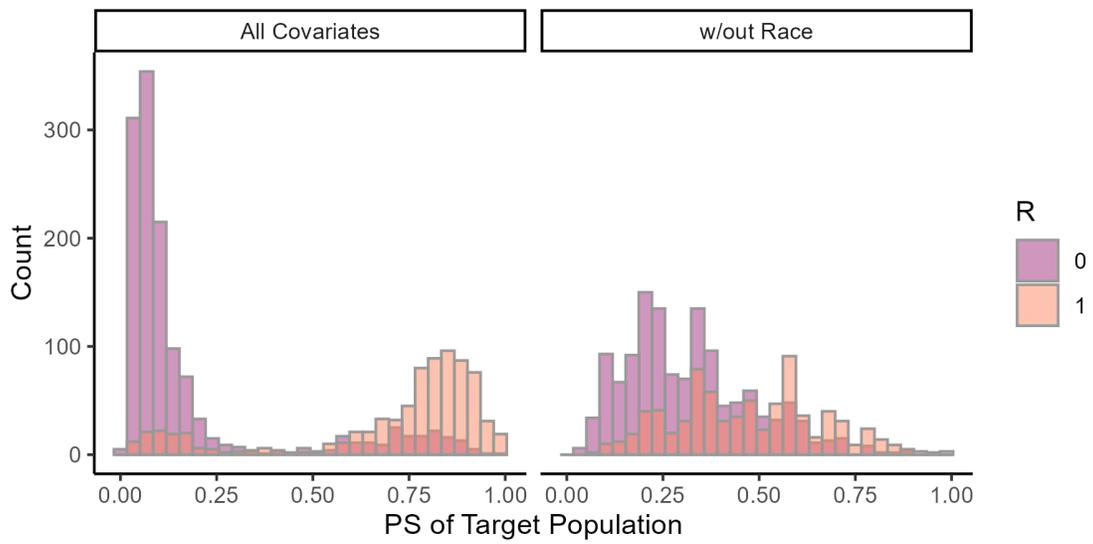


Figure 16: Distribution of the estimated probability of being in the target population when transporting the “health care hotspotting” RCT effect to a Midwestern academic health center population (Section 7.2)

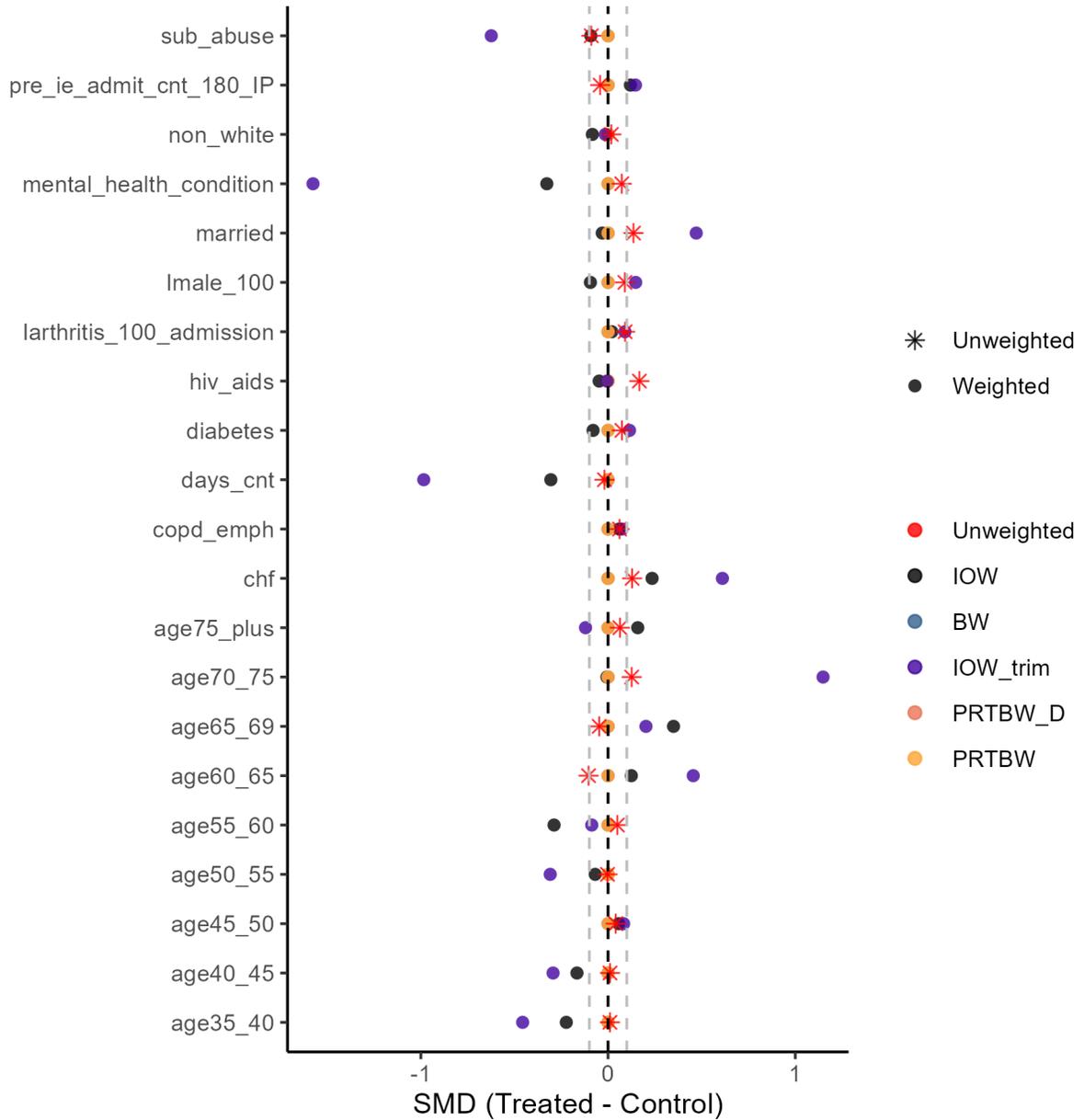


Figure 17: Treated and control group standardized mean differences (SMD) of all covariates when transporting the “health care hotspotting” RCT effect to a Midwestern academic health center population (Section 7.2).

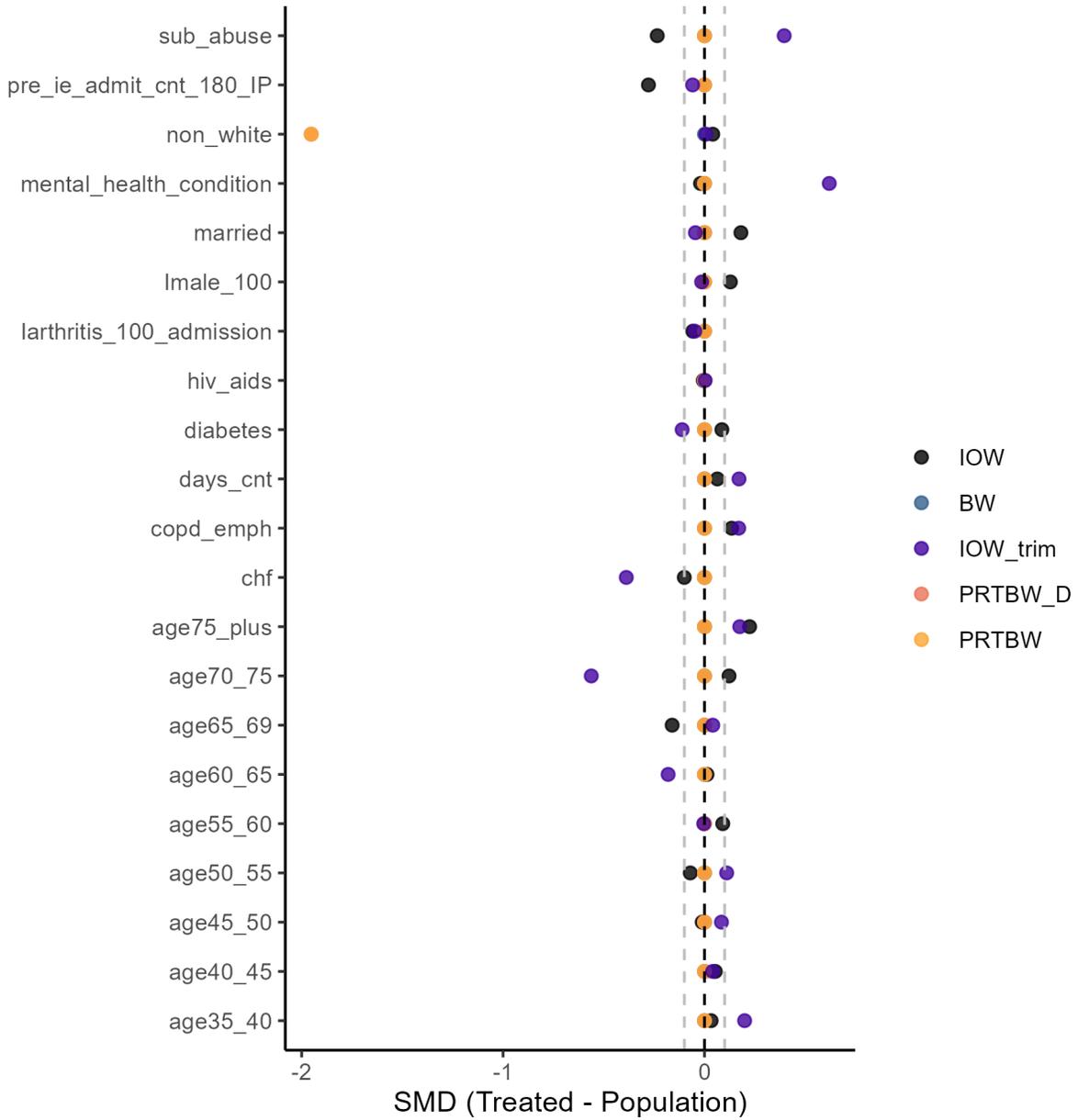


Figure 18: Treated and target population standardized mean differences (SMD) of all covariates when transporting the “health care hotspotting” RCT effect to a Midwestern academic health center population (Section 7.2).

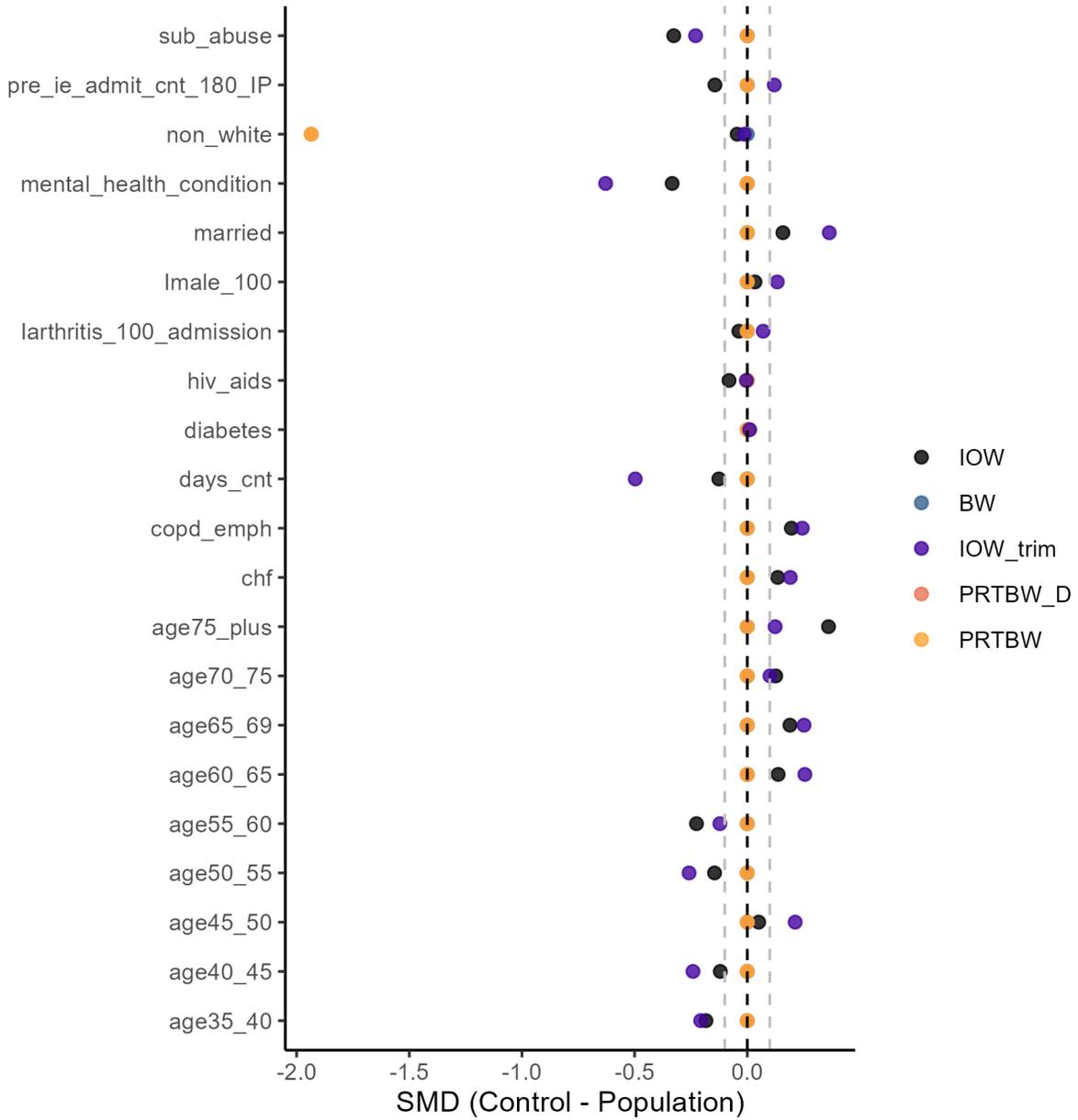


Figure 19: Control and target population standardized mean differences (SMD) of all covariates when transporting the “health care hotspotting” RCT effect to a Midwestern academic health center population (Section 7.2).

References

- Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973.
- Boileau, P., Leng, N., Hejazi, N. S., van der Laan, M., and Dudoit, S. (2025). A nonparametric framework for treatment effect modifier discovery in high dimensions. Journal of the Royal Statistical Society Series B: Statistical Methodology, 87(1):157–185.
- Boos, D. D. and Stefanski, L. A. (2013). Essential Statistical Inference, volume 120 of Springer Texts in Statistics. Springer, New York, NY.
- Chen, R., Huling, J. D., Chen, G., and Yu, M. (2024). Robust sample weighting to facilitate individualized treatment rule learning for a target population. Biometrika, 111(1):309–329.
- D’Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. Journal of Econometrics, 221(2):644–654.
- Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. Econometrica, 66(2):315–332.
- Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. Political Analysis, 20(1):25–46.
- Hines, O., Diaz-Ordaz, K., and Vansteelandt, S. (2023). Variable importance measures for heterogeneous causal effects. arXiv:2204.06030 [stat].
- Huling, J. D. and Mak, S. (2024). Energy balancing of covariate distributions. Journal of Causal Inference, 12(1).
- Källberg, D. and Waernbaum, I. (2023). Large Sample Properties of Entropy Balancing Estimators of Average Causal Effects. Econometrics and Statistics.
- Mak, S. and Joseph, V. R. (2018). Projected support points: a new method for high-dimensional data reduction. arXiv:1708.06897 [stat].
- Tseng, P. and Bertsekas, D. P. (1991). Relaxation Methods for Problems with Strictly Convex Costs and Linear Constraints. Mathematics of Operations Research, 16(3):462–481. Publisher: INFORMS.
- Wang, Y. and Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. Biometrika, 107(1):93–105.

Zhao, Q. and Percival, D. (2017). Entropy Balancing is Doubly Robust. Journal of Causal Inference, 5(1). Publisher: De Gruyter.

Zubizarreta, J. R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. Journal of the American Statistical Association, 110(511):910–922.