Propensity Score Weighting for Causal Inference with Multi-valued Treatments

Fan (Frank) Li[1] and Fan Li[2]

ABSTRACT

Causal or unconfounded descriptive comparisons between multiple groups are common in observational studies. Motivated from a racial disparity study in health services research, we propose a unified propensity score weighting framework, the balancing weights, for estimating causal effects with multiple treatments. These weights incorporate the generalized propensity scores to balance the weighted covariate distribution of each treatment group, all weighted toward a common pre-specified target population. The class of balancing weights include several existing approaches such as the inverse probability weights and trimming weights as special cases. Within this framework, we propose a set of target estimands based on linear contrasts. We further develop the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weighting scheme corresponds to the target population with the most overlap in covariates across the multiple treatments. These weights are bounded and thus bypass the problem of extreme propensities. We show that the generalized overlap weights minimize the total asymptotic variance of the moment weighting estimators for the pairwise contrasts within the class of balancing weights. We consider two balance check criteria and propose a new sandwich variance estimator for estimating the causal effects with generalized overlap weights. We apply these methods to study the racial disparities in medical expenditure between several racial groups using the 2009 Medical Expenditure Panel Survey (MEPS) data. Simulations were carried out to compare with existing methods.

KEY WORDS: balancing weights; generalized propensity score; generalized overlap weights; health services research; pairwise comparison; racial disparity

[1]Fan (Frank) Li is PhD student, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, 27710 (email: frank.li@duke.edu); [2]Fan Li is associate professor, Department of Statistical Science, Duke University, Durham, NC, 27705 (email: fli@stat.duke.edu).

1 Introduction

Propensity score weighting is a common method for balancing covariates and estimating treatment effects in causal inference. It is also applicable to unconfounded non-causal comparisons such as racial disparities studies (e.g. McGuire et al., 2006; Cook et al., 2009). There is a vast literature on propensity score weighting with binary treatments (for a review, see Ding and Li (2018)). This paper focuses on propensity score weighting strategies for multiple group comparisons, which have become increasingly common in practice. For example, in comparative effectiveness research, the interest often lies in comparing the effectiveness of several medical treatments; in health service research, the interest often lies in examining the disparities in health care utilization between more than two races or ethnicities (Zaslavsky and Ayanian, 2005).

For multiple group comparisons, Imbens (2000) has developed the generalized propensity score method; the key insight is that the scalar generalized propensity score of each treatment level can be exploited to separately estimate the average potential outcomes in that group. With the generalized propensity score device, matching and subclassification strategies have been discussed extensively; see, for instance, Lechner (2002); Zanutto et al. (2005); Rassen et al. (2013); Yang et al. (2016); Lopez and Gutman (2017). With the weighting strategy, the existing methods for multiple-group comparisons have largely focused on the pairwise average treatment effect (ATE), based on the inverse probability weighting (IPW) (Feng et al., 2012; McCaffrey et al., 2013). However, observational studies often rely on convenience samples, which does not necessarily represent a population of scientific meaning. In such cases, the automatic focus on ATE may be questionable because it is not clear what target population the causal conclusion is applicable to. Meanwhile, multiple treatments exacerbate the overlap issues as different treatments may be applicable only to certain subpopulations, and the ATE may correspond to an infeasible intervention. Regardless of the number of treatment levels, extreme propensity scores close to zero or one will likely result in bias and excessive variance of the IPW estimators (Li et al., 2018b). Crump et al. (2009) proposed an optimal trimming procedure that focuses on regions with good overlap and thus improves the efficiency of the IPW estimator for binary treatments; Yang et al. (2016) extended

the trimming rule to more than two treatments. Though easy to implement, propensity trimming often leads to an ambiguous target population and may discard a large number of units.

In this article, we propose a unified propensity score weighting framework for causal inference with multiple treatments. Specifically, we generalize the balancing weights framework for binary treatments (Li et al., 2018a) to balance the distribution of covariates from multiple treatment groups according to a pre-specified target population. Within this framework, we propose a set of target estimands based on linear contrasts. We further develop the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weights focus on the subpopulation with substantial probabilities to be assigned to all treatments. This target population aligns with the spirit of randomized clinical trials by emphasizing patients at clinical equipoise, and is thus of natural relevance to medical and policy studies. Under mild conditions, we show that the generalized overlap weights minimize the total asymptotic variance of the moment estimators for the pairwise contrasts within the class of balancing weights. These new weights are strictly bounded between zero and one, and thus automatically bypass the issue of extreme propensity scores.

Our methodological innovation is motivated by an application to racial disparities in medical expenditure. Identifying and tracking racial disparities in health care utilization represents a crucial step in developing health care policy and allocating health services resources. The *Unequal Treatment* report from the Institute of Medicine (IOM) defined health care disparity as the difference in treatment provided to social groups that is not justified by health status or treatment preference of the patient (IOM, 2003). Therefore, adjusting for the health status variables across different racial groups is necessary for producing interpretable disparity estimates concordant with the IOM definition. In this sense, these descriptive comparisons share the same nature with causal comparisons with respect to confounding control, and indeed propensity score methods have been widely used in health care disparity studies (Cook et al., 2012). One particular challenge is that the IOM definition of disparity includes racial differences in utilization mediated through factors other than health status and preference, such as many social factors (McGuire et al., 2006). Accordingly, a number of methods have been developed to account

for the social economic status variables in the propensity score analysis of racial disparities in health services (e.g. McGuire et al., 2006; Cook et al., 2009). In this paper, we combine one such method—the rank-and-replace adjustment—with the proposed generalized overlap weights to investigate racial disparities in medical expenditure between Whites, Blacks, Hispanics and Asians. This is in contrast to most existing racial disparity studies, which conducted separate comparisons of each White-minority pair (Cook et al., 2010).

The remainder of this article is organized as follows. Section 2 introduces the general framework of balancing weights. In Section 3, we propose the generalized overlap weights for pairwise comparisons with multiple treatments, discuss balance check criteria and variance estimation. In Section 4, we reanalyze the Medical Expenditure Panel Survey data and study the racial disparities in medical expenditure between several racial groups. Section 5 carries out simulations to examine the operating characteristics of the proposed method and compare with existing methods. Section 6 concludes.

2 Balancing Weights for Multiple Treatments

2.1 Basic Setup

We consider a sample of n units, each belonging to one of $J \geq 3$ groups for which covariate-balanced comparisons are of interest. Let $Z_i \in \mathbb{Z} = \{1, \ldots, J\}$ denote the treatment group membership, and $D_{ij} = \mathbb{1}\{Z_i = j\}$ the indicator of receiving treatment level j. For each unit, we observe an outcome Y_i and a set of p pre-treatment covariates $\boldsymbol{X}_i = (X_{i1}, ..., X_{ip})'$. For $J \geq 3$ treatments, Imbens (2000) defined the generalized propensity score, as follows.

DEFINITION **1** *(Generalized Propensity Scores) The generalized propensity score is the conditional probability of being assigned to each group given the covariates:*

$$e_j(\boldsymbol{X}) = \Pr(Z = j|\boldsymbol{X}), \quad j \in \mathbb{Z}.$$

By definition, the sum-to-unity restriction $\sum_{j=1}^{J} e_j(\boldsymbol{X}) = 1$ holds for all $\boldsymbol{X}$ in support $\mathbb{X}$, and hence each unit's propensity can be uniquely characterized by $J - 1$ scalar scores. Under the Stable

Unit Treatment Value Assumption (SUTVA), each unit has a potential outcome $Y_i(j)$ mapped to each treatment level $j \in \mathbb{Z}$, among which, only the one corresponding to the received treatment, $Y_i = Y_i(Z_i)$, is observed. To proceed, we make the following two standard assumptions.

ASSUMPTION **1** *(Weak Unconfoundedness) The assignment is weakly unconfounded if*

$$Y(j) \perp \mathbb{1}\{Z = j\}|\boldsymbol{X}, \quad \forall\, j \in \mathbb{Z}.$$

ASSUMPTION **2** *(Overlap) For all $\boldsymbol{X} \in \mathbb{X}$ and all group j, the probability of being assignment to any treatment group is bounded away from zero:*

$$e_j(\boldsymbol{X}) > 0, \quad \forall\, \boldsymbol{X} \in \mathbb{X},\ j \in \mathbb{Z}.$$

Assumption 1 imposes unconfoundedness separately for each level of the treatment, and is sufficient for identification of the population-level estimand (Imbens, 2000). This assumption implies that the potential outcome $Y(j)$ is independent of the assignment indicator $\mathbb{1}\{Z = j\}$, conditional on the scalar generalized propensity score $e_j(\boldsymbol{X})$. In other words, adjusting for the scalar score is sufficient to remove the bias in estimating the average value of $Y(j)$ over the target population. Assumption 2 restricts the study population to the covariate space where each unit has non-zero probability to receive any treatment.

To elaborate, we define the conditional expected potential outcomes in group j as $m_j(\boldsymbol{X}) = \mathbb{E}[Y(j)|\boldsymbol{X}]$. Under Assumption 1, we have $m_j(\boldsymbol{X}) = \mathbb{E}[Y|Z = j, \boldsymbol{X}]$, which is estimable from the observed data. As previously mentioned, the propensity score methods are also applicable to unconfounded descriptive (non-causal) comparisons where the group membership is a non-manipulable state, such as different races and different years. In these cases, a common objective is to compare the expected observed outcomes, $m_j(\boldsymbol{X}) = \mathbb{E}[Y|Z = j, \boldsymbol{X}]$; for example, when $J = 2$, Li et al. (2013) defined the contrast between $m_1(\boldsymbol{X})$ and $m_2(\boldsymbol{X})$ averaged over a population as the *average controlled difference* (ACD). For simplicity, henceforth we use the nomenclature of causal inference to generically refer to both causal and unconfounded descriptive settings, but stress that the methods developed here are applicable to both.

2.2 Balancing Weights

Assume the marginal density of the covariates, $f(\boldsymbol{X})$, exists, with respect to a base measure μ. In causal studies, the interest is on the average effects of units in a target population, whose density (up to a normalizing constant) we represent by $g(\boldsymbol{X}) = f(\boldsymbol{X})h(\boldsymbol{X})$, with $h(\boldsymbol{X})$ being a pre-specified function of covariates, which we refer to as a tilting function. We first define the expectation of the potential outcomes over the target population $g(\boldsymbol{X})$:

$$m_j^h \equiv \frac{\int_{\mathbb{X}} m_j(\boldsymbol{X}) f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}.$$

Then we characterize a class of additive estimands as a linear combination of the above expectations, with coefficients $\boldsymbol{a} = (a_1, \cdots, a_J)'$:

$$\tau^h(\boldsymbol{a}) \equiv \sum_{j=1}^{J} a_j m_j^h. \tag{1}$$

The causal estimand $\tau^h(\boldsymbol{a})$ generalizes the definition of weighted average treatment effect (WATE) in binary treatments (Hirano et al., 2003) where $J = 2$ and $\boldsymbol{a} = (1, -1)$. As will be seen in due course, $\tau^h(\boldsymbol{a})$ includes several existing causal estimands as special cases.

We next define the class of balancing weights. Let $f_j(\boldsymbol{X}) = f(\boldsymbol{X}|Z = j)$ be the density of $\boldsymbol{X}$ in the jth group over its support $\mathbb{X}_j$, we have $f_j(\boldsymbol{X}) \propto f(\boldsymbol{X})e_j(\boldsymbol{X})$. Given any pre-specified function h, we can weight the group-specific density $f_j(\boldsymbol{X})$ to the target population using the following weights, proportional up to a normalizing constant:

$$w_j(\boldsymbol{X}) \propto \frac{f(\boldsymbol{X})h(\boldsymbol{X})}{f(\boldsymbol{X})e_j(\boldsymbol{X})} = \frac{h(\boldsymbol{X})}{e_j(\boldsymbol{X})}, \quad \forall\, j \in \mathbb{Z}. \tag{2}$$

It is straightforward to show that the class of weights defined in (2) balance the weighted distributions of the covariates across J comparison groups:

$$f_j(\boldsymbol{X})w_j(\boldsymbol{X}) = f(\boldsymbol{X})h(\boldsymbol{X}), \quad \forall\, j \in \mathbb{Z}. \tag{3}$$

To apply the above framework, a key is to specify the coefficients $\boldsymbol{a}$ and the tilting function h, with the former defining the causal contrast and the latter representing the target population. We focus on the

case of multiple nominal treatments, where the scientific interest usually lies in pairwise comparisons. More specifically, the choice of $\boldsymbol{a}$ is contained in the finite set $\mathbb{S} = \{\boldsymbol{\lambda}_{j,j'} = \boldsymbol{\lambda}_j - \boldsymbol{\lambda}_{j'} : j < j'\}$, where $\boldsymbol{\lambda}_j$ is the $J \times 1$ unit vector with one at the jth position and zero everywhere else. In principle, the tilting function h can take any form, each leading to a unique type of balancing weights; statistical, scientific and policy considerations all play into the specification of h. We illustrate specifications of $\boldsymbol{a}$ and h by connecting the general definition (1) with existing estimands in the literature.

When $h(\boldsymbol{X}) = 1$, the target population $f(\boldsymbol{X})$ is the combined population from all groups and the weights become the standard inverse probability weights, $\{1/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$; the target estimand is the pairwise ATE as in Feng et al. (2012). When $h(\boldsymbol{X}) = e_{j'}(\boldsymbol{X})$, the target population is the subpopulation receiving treatment $Z = j'$, and the weights, $\{e_{j'}(\boldsymbol{X})/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$, are designed to estimate the average treatment effect for the treated (ATT). Define

$$\underline{e}_j = \max_{1\le l\le J}\{\min_{\boldsymbol{X}\in\mathbb{X}_l}\{e_j(\boldsymbol{X})\}\}, \quad \bar{e}_j = \min_{1\le l\le J}\{\max_{\boldsymbol{X}\in\mathbb{X}_l}\{e_j(\boldsymbol{X})\}\},$$

and an eligibility function $E_j(\boldsymbol{X}) = \mathbb{1}\{\underline{e}_j \le e_j(\boldsymbol{X}) \le \bar{e}_j\}$ for all $j \in \mathbb{Z}$. When $h(\boldsymbol{X}) = e_{j'}(\boldsymbol{X})\prod_{j=1}^{J} E_j(\boldsymbol{X})$, the target population is the subpopulation receiving treatment $Z = j'$ but remaining eligible for all other treatments (Lopez and Gutman, 2017). Similar eligibility functions were used earlier by van der Laan and Petersen (2007) and Moore et al. (2012) to develop improved causal models with time-varying treatments. Further, define a threshold α as the largest value such that

$$\alpha \le \frac{2\,\mathbb{E}\left[\sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) \middle| \sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) \le \alpha\right]}{\Pr\left(\sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) \le \alpha\right)}. \tag{4}$$

When $h(\boldsymbol{X}) = \mathbb{1}\{\boldsymbol{X} \in \mathbb{C}\}$ with $\mathbb{C} = \{\boldsymbol{X} \in \mathbb{X} | \sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) \le \alpha\}$, the target population is characterized by the subpopulation $\mathbb{C}$, and the inverse probability weights are formulated after applying the optimal trimming rule (Yang et al., 2016). Finally, when $h(\boldsymbol{X}) = \min_{1\le k\le J}\{e_k(\boldsymbol{X})\}$, one arrives at the generalized matching weights (Yoshida et al., 2017)—an extension of the matching weights of Li and Greene (2013) to multiple treatments. The matching weights is a weighting analogue to exact matching and the causal comparisons are made for the matched population. Moreover, one could choose

indicator functions for h that directly involves covariates of a subpopulation of interest, such as a specific gender or a range of age. Table 1 summarizes the above special cases.

Table 1: Examples of balancing weights and target populations for making pairwise comparisons with different tilting functions.

Target population	Tilting function $h(\boldsymbol{X})$	Weights $\{w_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Combined	1	$\{1/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Treated (j'th group)	$e_{j'}(\boldsymbol{X})$	$\{e_{j'}(\boldsymbol{X})/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Treated (restricted)	$e_{j'}(\boldsymbol{X})\prod_{j=1}^{J} E_j(\boldsymbol{X})$	$\{e_{j'}(\boldsymbol{X})\prod_{j=1}^{J} E_j(\boldsymbol{X})/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Trimming	$\mathbb{1}\{\boldsymbol{X} \in \mathbb{C}\}$	$\{\mathbb{1}\{\boldsymbol{X} \in \mathbb{C}\}/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Matching	$\min_{1\le k\le J}\{e_k(\boldsymbol{X})\}$	$\{\min_{1\le k\le J}\{e_k(\boldsymbol{X})\}/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Overlap	$(\sum_{k=1}^{J} 1/e_k(\boldsymbol{X}))^{-1}$	$\{(\sum_{k=1}^{J} 1/e_k(\boldsymbol{X}))^{-1}/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$

When the treatment levels are ordered categories, target estimands may differ from the pairwise comparisons and require different choice of $\boldsymbol{a}$. For instance, one may be interested in the quadratic contrasts between unit increases in the treatment level, namely $(m^h_{j+1} - m^h_j) - (m^h_j - m^h_{j-1})$. In other cases, one may estimate the weighted average of unit increase in the treatment level, $\sum_{j=1}^{J-1} \pi_j(m^h_{j+1} - m^h_j)$, or the accumulative effect of the maximum treatment, $m^h_J - m^h_1$. For the disparity study in Section 4, the multiple racial groups are unordered categories. For this reason, we mainly focus on multiple nominal groups, but note that the general framework of balancing weights remains applicable to multiple ordinal groups.

2.3 Large-sample Properties of Moment Estimators

For any pre-specified vector $\boldsymbol{a}$ and tilting function h, we could first use the plug-in sample moment estimator to obtain the expectation of the potential outcomes among the target population

$$\hat{m}^h_j = \frac{\sum_{i=1}^{n} D_{ij} Y_i w_j(\boldsymbol{X}_i)}{\sum_{i=1}^{n} D_{ij} w_j(\boldsymbol{X}_i)}, \tag{5}$$

and then estimate $\tau^h(\boldsymbol{a})$ by a linear combination, $\hat{\tau}^h(\boldsymbol{a}) = \sum_{j=1}^{J} a_j \hat{m}_j^h$, where the sum is over a sample drawn from density $f(\boldsymbol{X})$. Below we establish three large-sample results of $\hat{\tau}^h(\boldsymbol{a})$; the proofs are given in the Web Supplementary.

PROPOSITION **1** *Given any h and $\boldsymbol{a}$, $\hat{\tau}^h(\boldsymbol{a})$ is a consistent estimator of $\tau^h(\boldsymbol{a})$.*

Denote the collection of treatment assignment $\underline{\boldsymbol{Z}} = \{Z_1, \ldots, Z_n\}$ and covariate design points $\underline{\boldsymbol{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. The next two results concern the variance of the sample estimator, which is decomposed as

$$\mathbb{V}[\hat{\tau}^h(\boldsymbol{a})] = \mathbb{E}_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\, \mathbb{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}] + \mathbb{V}_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\, \mathbb{E}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}].$$

The first term is the variation due to residual variance in $\hat{\tau}^h(\boldsymbol{a})$ conditional on the design points. The second term arises from the dependence of the expectation of the plug-in estimator on the sample, and estimating it involves the outcome model (associations between $Y(j)$ and $\boldsymbol{X}$). As individual variation is typically much larger than conditional mean variation, the benefit of further optimizing the weights by a preliminary look at the outcomes, which mixes the design and analysis, would usually not justify the risk of biasing model specification to attain desired results (Imbens, 2004). Hence, we focus on the first term.

PROPOSITION **2** *Given $\boldsymbol{a}$, suppose the family of residual variances $\{\mathbb{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}], n \geq 1\}$ is uniformly integrable. Then the expectation of the conditional variance converges*

$$n \cdot \mathbb{E}_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\, \mathbb{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}] \to Q(\boldsymbol{a}, h) \equiv \int_{\mathbb{X}} \Big(\sum_{j=1}^{J} a_j^2 v_j(\boldsymbol{X})/e_j(\boldsymbol{X}) \Big) h^2(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X}) / C_h^2,$$

where $v_j(\boldsymbol{X}) = \mathbb{V}[Y(j)|\boldsymbol{X}]$ and $C_h \equiv \int_{\mathbb{X}} h(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X})$ is a constant.

When the residual variance of the potential outcome is homoscedastic across all groups such that $v_j(\boldsymbol{X}) = v$, then the limit $Q(\boldsymbol{a}, h)$ can further simplify and the following result holds.

PROPOSITION **3** *Under homoscedasticity, the function*

$$\tilde{h}(\boldsymbol{X}) \propto \frac{1}{\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X})}$$

gives the smallest asymptotic variance for the moment estimator $\hat{\tau}^h(\boldsymbol{a})$ *among all* h*'s, and* $\min_h Q(\boldsymbol{a}, h) = v/C_{\tilde{h}}$.

A more general result of Proposition 3 can be obtained under heteroscedasticity. In that case, the optimal tilting function,

$$\tilde{h}(\boldsymbol{X}) \propto \frac{1}{\sum_{j=1}^{J} a_j^2 v_j(\boldsymbol{X})/e_j(\boldsymbol{X})},$$

explicitly depends on the residual variances of the potential outcomes. Although estimates of $v_j(\boldsymbol{X})$ can be obtained by outcome regression modeling in the analysis stage, it is rarely the case that accurate prior information is available in the design stage. Therefore, such a tilting function is difficult to specify for design purposes and may find limited use without peeking at the outcomes. For such considerations, we motivate the generalized overlap weights in Section 3 under homoscedasticity. These asymptotic results generalize those for binary treatments in Li et al. (2018a); they also extend the asymptotic results on propensity score trimming in Crump et al. (2009) and Yang et al. (2016), who have similarly assumed homoscedasticity but restricted the class of tilting functions to indicator functions.

3 Generalized Overlap Weighting for Pairwise Comparisons

3.1 The Generalized Overlap Weights

For nominal treatments, scientific interest often lies in comparing outcomes between each pair of treatment groups in a common target population. In this case, as $\boldsymbol{a} \in \mathbb{S}$, we propose to choose the tilting function h that minimizes the total asymptotic variance of the sample estimators for all pairwise comparisons; in other words, the objective function is

$$\sum_{j<j'} Q(\boldsymbol{\lambda}_{j,j'}, h) \propto Q(\mathbf{1}_J, h),$$

where $\mathbf{1}_J$ is the $J\times 1$ vector of ones. According to Proposition 3, the function $h(\boldsymbol{X}) = (\sum_{j=1}^{J} 1/e_j(\boldsymbol{X}))^{-1}$—the harmonic mean of the generalized propensity scores—minimizes $Q(\mathbf{1}_J, h)$ among all h. Based on this optimal tilting function h, we define the generalized overlap weights for $j = 1, ..., J$:

$$w_j(\boldsymbol{X}) \propto \frac{1/e_j(\boldsymbol{X})}{\sum_{k=1}^{J} 1/e_k(\boldsymbol{X})}.$$

For binary treatments ($J = 2$), the generalized overlap weights reduce to the overlap weights in Li et al. (2018a), namely the propensity of assignment to the other group: $w_1(\boldsymbol{X}) \propto 1 - e_1(\boldsymbol{X}) = e_2(\boldsymbol{X})$, $w_2(\boldsymbol{X}) \propto 1 - e_2(\boldsymbol{X}) = e_1(\boldsymbol{X})$.

The maximum of the harmonic mean function h is attained when $e_j(\boldsymbol{X}) = 1/J$ for all j, that is, when the units have the same propensity to each of the treatments. Heuristically, the tilting function h gives the most relative weight to the covariate regions in which none of the propensities are close to zero. While it is generally difficult to visualize the optimal h in higher dimensions, we could do so with $J = 3$ treatments. Figure 1 provides a ternary plot of h when $J = 3$. It is clear that the optimal tilting function gives the most relative weight to the covariate regions in which none of the propensities are close to zero, and down-weights the region where there is lack of overlap in at least one dimension. Therefore, we can interpret the corresponding target population to be the subpopulation with the most overlap in covariates among all groups, and term the target estimand as the pairwise average treatment effect among the overlap population (ATO). As the overlap population tilts $f(\boldsymbol{X})$ most heavily towards equipoise, it is naturally of policy and clinical relevance. Especially for clinical practice, this target population aligns with the spirit of randomized studies and emphasizes patients with clinical equipoise, whose treatment decisions remain unclear and thus for whom comparative information is most needed. Analogously, in descriptive studies for racial disparities, the overlap population represents individuals with most similarity in observed health-related characteristics, based on whom subsequent policy interventions on health care utilization become most meaningful.

Besides asymptotic efficiency, the generalized overlap weights have several attractive features. First, the harmonic mean function h is strictly bounded

$$0 < \min_{1\le k\le J}\{e_k(\boldsymbol{X})\}/J \le h(\boldsymbol{X}) \le \min_{1\le k\le J}\{e_k(\boldsymbol{X})\} < 1,$$

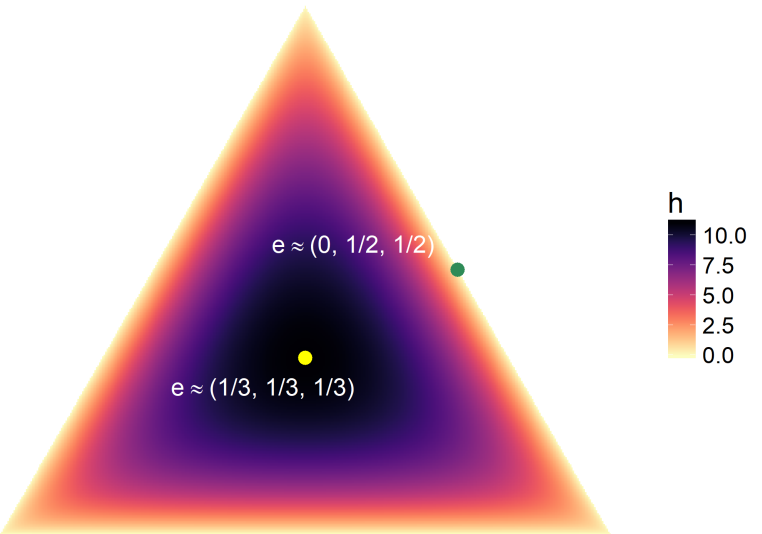

Figure 1: Ternary plot of optimal h (up to a proportionality constant) as a function of the generalized propensity score vector with $J = 3$ treatments. Each point in the triangular plane represents a unit with certain values of the generalized propensity scores. The value of each generalized propensity score is proportional to the orthogonal distance from that point to each edge. It is evident that the new weighting scheme emphasizes the centroid region with good overlap, e.g., units with $e(\boldsymbol{X}) \approx (1/3, 1/3, 1/3)$, and smoothly down-weights the edges, e.g., units with $e(\boldsymbol{X}) \approx (0, 1/2, 1/2)$.

and thus the weighting scheme is robust to extreme weights, in contrast to IPW. Second, the target population defined by the generalized overlap weights is adaptive to the covariate distributions among the J comparison groups. For example, when the propensity of assignment to treatment j is small

compared to others so that $e_j(\boldsymbol{X}) \approx 0$, the tilting function

$$h(\boldsymbol{X}) \propto \prod_{l=1}^{J} e_l(\boldsymbol{X}) / \sum_{k=1}^{J} \prod_{l \neq k} e_l(\boldsymbol{X}) \approx \prod_{l=1}^{J} e_l(\boldsymbol{X}) / \prod_{l \neq j} e_l(\boldsymbol{X}) = e_j(\boldsymbol{X}),$$

suggesting that the target population is similar to the jth treatment group and the associated estimand approximates the ATT. On the other hand, if the treatment groups are almost balanced in size and covariate distribution so that $e_j(\boldsymbol{X}) \approx 1/J$ for all j, we have $h(\boldsymbol{X}) \propto 1$ and the target estimand approximates the pairwise ATE. Arguably this adaptiveness enables the generalized overlap weighting scheme to define a scientific question that may be best answered nonparametrically by the available data at hand. Finally, the generalized matching weights (Yoshida et al., 2017)—defined by $h(\boldsymbol{X}) = \min_{1 \leq j \leq J}\{e_j(\boldsymbol{X})\}$—share some of the above advantages, but these weights are not asymptotically efficient and are non-smooth, which renders the variance calculation more complex.

3.2 Estimate Generalized Propensity Scores and Balance Check

In practice, usually the propensity scores are not known and must be estimated from the data. For multiple nominal treatments, the generalized propensity scores are frequently modeled by a multinomial logistic regression,

$$\begin{aligned} e_1(\boldsymbol{X}_i) &= \frac{1}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \\ e_j(\boldsymbol{X}_i) &= \frac{\exp(\alpha_j + \boldsymbol{X}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \quad j = 2, \ldots, J, \end{aligned} \tag{6}$$

where the covariate vector $\boldsymbol{X}$ are allowed to contain higher-order moments, splines and interactions. Model parameters $\boldsymbol{\theta} = (\alpha_2, \ldots, \alpha_J, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_J^T)^T$ can be estimated by standard maximum likelihood, from which we obtain the estimated propensity scores. To assess the fit of the propensity score model, we check the weighted covariate balance in the target population. We consider two ways for balance check motivated by the population balancing constraint (3). First, constraint (3) implies the weighted covariate balance between each group and the target population. Therefore, we inspect, for each treatment level, the weighted covariate mean deviation from that of the target population. Specifically, we define $\bar{X}_j =$

$\sum_{i=1}^{n} D_{ij} X_i w_j(\boldsymbol{X}_i) \Big/ \sum_{i=1}^{n} D_{ij} w_j(\boldsymbol{X}_i)$ as the weighted mean of covariate X from the jth group and $S^2_{X,j}$ as the unweighted variance. Further, we define $\bar{X}_p = \sum_{i=1}^{n} X_i h(\boldsymbol{X}_i) \Big/ \sum_{i=1}^{n} h(\boldsymbol{X}_i)$ as the average value of covariate X in the target population and $S^2_X = J^{-1} \sum_{j=1}^{J} S^2_{X,j}$ as the averaged unweighted variance. The population standardized difference (PSD) is then defined for each covariate and each treatment level as $\text{PSD}_j = |\bar{X}_j - \bar{X}_p|/S_X$. Similar to McCaffrey et al. (2013), we then use $\max_j |\text{PSD}_j|$ as the balance metric for each covariate X and inspect the adequacy of the propensity score model. If a covariate is not well balanced in one group, interaction terms of that variable with other variables can be added to the model, and the new model is re-fit and re-evaluated until balance is deemed satisfactory. On the other hand, the population balance constraint (3) also implies pairwise balance $f_j(\boldsymbol{X}) w_j(\boldsymbol{X}) = f_{j'}(\boldsymbol{X}) w_{j'}(\boldsymbol{X})$ for all $j \neq j'$, and so we could alternatively assess balance by checking the pairwise absolute standardized differences (ASD), $\text{ASD}_{j,j'} = |\bar{X}_j - \bar{X}_{j'}|/S_X$. The balance metric for each covariate can then be similarly specified as $\max_{j<j'} |\text{ASD}_{j,j'}|$.

Finally, a special property of the overlap weights with binary treatments is exact balance, that is, when the propensities scores are estimated from a logistic model, the standardized difference of all the covariates entering the propensity model is zero, i.e, $\text{ASD}_{1,2} = 0$ for $J = 2$ (Li et al., 2018a, Theorem 3). However, this exact balance property is due to the happenstance that the logistic score equations exploit the covariate-balancing moment conditions, and does not directly extend to the generalized overlap weights with $J \geq 3$ when the propensity score is estimated by a multinomial logistic model. Therefore, we still recommend the conventional iterative fitting-checking procedure to improve the propensity model.

3.3 Variance Estimation

The asymptotic variance results in Section 2.3 are not directly useful for calculating the sample variance of $\hat{\tau}^h(\boldsymbol{\lambda}_{j,j'})$ in practice because the $v_j(\boldsymbol{X})$'s are not known. Moreover, one has to account for the additional uncertainty in estimating the propensities in the variance estimation. Here we derive an empirical sandwich variance estimator (Stefanski and Boos, 2002) that accounts for the uncertainty in estimating the generalized overlap weights from the multinomial logistic model (6). We provide the

following theorem to motivate the closed-variance calculation for the pairwise ATO estimates. The proof is given in the Web Supplementary.

THEOREM **1** *Under standard regularity conditions, when the generalized propensity scores are estimated by multinomial logistic regression (6), the resulting ATO estimator between groups j and j' is asymptotically normal*

$$\sqrt{n}\{\hat{\tau}^h(\boldsymbol{\lambda}_{j,j'}) - \tau^h(\boldsymbol{\lambda}_{j,j'})\} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left\{\psi_{ij} - \psi_{ij'}\right\}^2 / [\mathbb{E}\{h(\boldsymbol{X})\}]^2\right),$$

where

$$\psi_{ij} = D_{ij}(Y_i - m_j^h)w_j(\boldsymbol{X}_i) + \mathbb{E}\left\{D_{ij}(Y_i - m_j^h)\frac{\partial}{\partial \boldsymbol{\theta}^T}w_j(\boldsymbol{X}_i)\right\}\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{S}_{\boldsymbol{\theta},i},$$

and $\boldsymbol{S}_{\boldsymbol{\theta},i}$, $\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ are the individual score and information matrix of $\boldsymbol{\theta}$, respectively.

Theorem 1 suggests the following consistent variance estimator. Denote $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i}$, $\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ as the maximum likelihood estimator of $\boldsymbol{\theta}$, the plug-in consistent estimators for the individual score and information matrix, the variance estimator for the estimated ATO is expressed by

$$\hat{\mathbb{V}}[\hat{\tau}^h(\boldsymbol{\lambda}_{j,j'})] = \frac{\sum_{i=1}^n \left(\hat{\psi}_{ij} - \hat{\psi}_{ij'}\right)^2}{\left[\sum_{i=1}^n \{\sum_{k=1}^J 1/\hat{e}_k(\boldsymbol{X}_i)\}^{-1}\right]^2}, \tag{7}$$

where

$$\hat{\psi}_{ij} = D_{ij}(Y_i - \hat{m}_j^h)w_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}) + \left\{\frac{1}{n}\sum_{i=1}^n D_{ij}(Y_i - \hat{m}_j^h)\frac{\partial}{\partial \boldsymbol{\theta}^T}w_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})\right\}\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i}.$$

The true generalized propensity score is generally unknown in applications and will be substituted by its sample analogue. Hirano et al. (2003) suggested that a consistent estimator of the propensity score leads to more efficient estimation of the WATE with binary treatments than the true propensity score. Our derivation of the variance estimator re-interprets their findings in the context of multiple treatments. Specifically, with a consistent estimator for the generalized propensity score, the influence function for estimating m_j^h, $\psi_{ij}/\,\mathbb{E}\{h(\boldsymbol{X})\}$, can be viewed as the residual of $D_{ij}(Y_i - m_j^h)w_j(\boldsymbol{X}_i)/\,\mathbb{E}\{h(\boldsymbol{X})\}$—the influence function for estimating m_j^h using the true propensity score—after projecting it onto the nuisance tangent space of $\boldsymbol{\theta}$. Therefore, the efficiency implications from Hirano et al. (2003) carry over to our pairwise comparisons emphasizing the overlap population.

4 Application to Racial Disparities in Medical Expenditure

4.1 The Data

Our application is based on the 2009 Medical Expenditure Panel Survey (MEPS) data. The sample contains health information, social-economic status (SES) and total health care expenditure for four racial groups with adult aged at least 18 years: 9830 non-Hispanic Whites, 1446 Asians, 4020 Blacks, 5150 Hispanics. We are interested in estimating the health care disparity in the yearly total health care expenditure, after controlling for the differences due to patient health status, i.e., variables reflecting clinical appropriateness and need. Using the MEPS data, Cook et al. (2010) estimated the racial disparities between each White-minority pair. One potential limitation of such separate binary comparisons is the non-transitivity among the pairwise estimates, as each comparison may be made for a different target population (see supplementary material for detailed discussions on transitivity). Here we focus on the simultaneous multiple-group comparisons by defining a common target population.

The MEPS data is well-suited to study racial disparities since it records a wide range of patient-level health characteristics. As previously mentioned, the IOM definition of disparity excludes differences in health status and patient preferences, but includes differences in social-economic status and discrimination. For this reason, we follow McGuire et al. (2006) and distinguish between the set of health status variables ($\boldsymbol{X}_H$) and the set of SES variables ($\boldsymbol{X}_S$), with the former including body mass index, SF-12 physical and mental component summary, comprehensive measurements of health conditions, age, gender, marital status and the latter including poverty status, education, health insurance and geographical region. As there is no gold standard in measuring patient preferences (McGuire et al., 2006), we do not interpret any variables as preference measurements, but acknowledge that the lack of this information represents a limitation in implementing the IOM definition. From the first column of the two boxplots in Figure 2, we observe substantial differences in the health status distributions among the four racial groups, which indicate the necessity of adjustment.

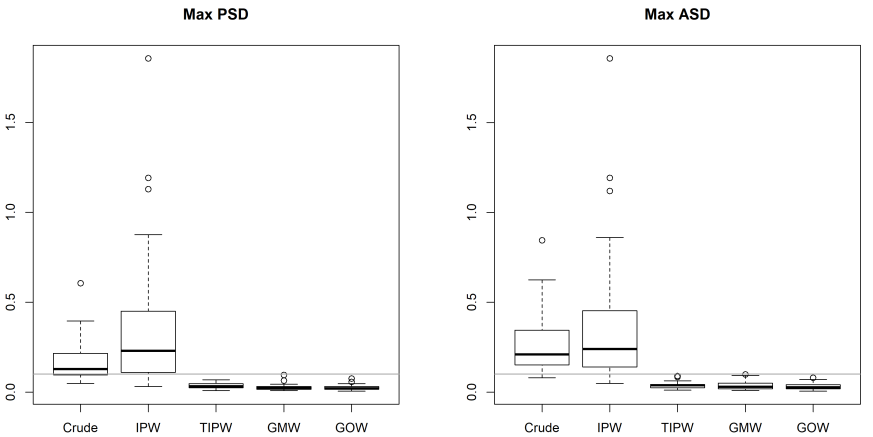

Figure 2: Boxplots for the maximum population standardized difference (PSD) and maximum absolute standardized difference (ASD) for all health status covariates corresponding to each adjustment method. The gray horizontal line indicates adequate balance at 0.1. Crude: unweighted; IPW: inverse probability weighting; TIPW: inverse probability weighting combined with optimal trimming; GMW: generalized matching weighting; GOW: generalized overlap weighting.

4.2 Balance Check and Effective Sample Size

We employ the generalized propensity scores to balance the health status variables among the four racial groups. If the generalized propensity scores are well estimated, then the propensity-score-weighted populations should be balanced with respect to the health status variables, thus removing the contribution of health status differences to the disparity estimates. This is the general idea behind the application of a health status propensity score to estimate White-minority disparity in the health services literature (Cook et al., 2012). We estimate the generalized propensity scores using a multinomial logistic regression including the main effects of all health status variables. The distributions of the estimated scores are presented in Figure 3. There is a moderate lack of overlap especially regarding the Asian group. As such, balancing the health status variables toward the combined population through IPW inevitably

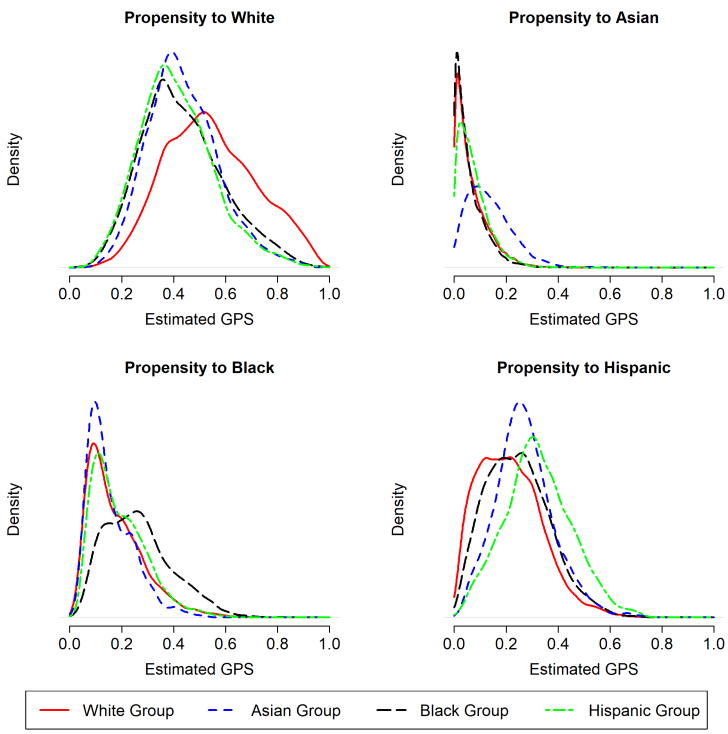

Figure 3: Marginal distributions of the estimated health status generalized propensity scores.

emphasizes the patients atypical for their own racial groups, producing disparity estimates lacking policy relevance. By contrast, balancing the health status variables toward the overlap population via the generalized overlap weighting (GOW) emphasizes a naturally comparable subpopulation that are most typical in each respective group, and leads to disparity estimates of greater policy interest. Based on the estimated propensity scores, we calculate for each health status variable the values of $\max_j |\text{PSD}_j|$ and $\max_{j<j'} |\text{ASD}_{j,j'}|$, which are defined in Section 3.2 to examine balance in the weighted populations. Due to the lack of overlap, IPW results in severe imbalances in more than a few health status variables, presenting worse results than no weighting at all. On the other hand, GOW provides the best balance among the overlap population. Two other competing methods, optimal trimming (TIPW) and generalized matching weighting (GMW) also perform adequately in balancing the health status variables in their respective target populations. The balance results are similar between the two balance criteria.

To quantify the amount of information in different target populations, we report the corresponding effective sample size (ESS). Following McCaffrey et al. (2013), we define the ESS for group j as

$$\text{ESS}_j^h = \frac{\left(\sum_{i=1}^n \sum_{j=1}^J D_{ij} w_j(\boldsymbol{X}_i)\right)^2}{\sum_{i=1}^n \sum_{j=1}^J D_{ij} w_j^2(\boldsymbol{X}_i)}.$$

As weighting generally increases the variance compared to the unweighted estimates based on the same sample, the ESS may serve as a conservative measure to characterize the variance inflation or precision loss due to weighting. It is evident from Table 2 that all weighting methods reduce ESS compared to the original sample. However, IPW results in a very small ESS for Asians relative to the original group size, signaling the presence of extreme weights and lack of overlap. By contrast, GOW corresponds to the largest total ESS, matching its theoretical efficiency optimality.

Table 2: Effective sample size of each (weighted) group. Crude: unweighted; IPW: inverse probability weighting; TIPW: inverse probability weighting combined with optimal trimming; GMW: generalized matching weighting; GOW: generalized overlap weighting.

	Whites	Asians	Blacks	Hispanics	Total
Crude	9830	1446	4020	5150	20446
IPW	8371	10	2549	2482	13412
TIPW	6524	695	2183	3071	12473
GMW	4937	1285	1875	3176	11273
GOW	6015	1166	2234	3756	13171

4.3 Analysis 1: Health Status Propensity Score Weighting

We calculate the pairwise racial disparities as the weighted average controlled difference in total health care expenditure using GOW, and report point estimates and 95% confidence intervals (based on the sandwich variance) in the last row of Table **??**. This weighting scheme emphasizes a naturally comparable subpopulation with similar health status, namely patients who, based on their health conditions and

Table 3: Racial disparity estimates in total health care expenditure (in dollars). The point estimates are obtained as average controlled differences by propensity score weighting. The associated 95% confidence intervals are obtained by the sandwich variance (IPW, TIPW and GOW) or bootstrap (GMW).

	IPW	TIPW	GMW	GOW
Whites-Asians	2402	1335	1112	1160
	(530, 4274)	(671, 1999)	(648, 1569)	(660, 1661)
Whites-Blacks	908	1148	839	886
	(505, 1311)	(781, 1515)	(455, 1239)	(518, 1253)
Whites-Hispanics	719	1257	1234	1221
	(129, 1309)	(804, 1711)	(813, 1623)	(849, 1593)
Asians-Blacks	-1494	-187	-273	-274
	(-3385, 397)	(-872, 499)	(-737, 281)	(-813, 264)
Asians-Hispanics	-1683	-77	122	61
	(-3621, 255)	(-812, 657)	(-385, 621)	(-479, 601)
Blacks-Hispanics	-189	109	395	335
	(-836, 459)	(-375, 594)	(-100, 820)	(-82, 752)

clinical need, could easily be either White or from each minority groups. Among this overlap subpopulation, Whites spent on average $ 1160, $886 and $1221 more than Asians, Blacks and Hispanics on health care, with directions and magnitudes comparable to earlier reports from 2003 and 2004 (Cook et al., 2009). All three 95% confidence intervals exclude zero, confirming that the disparity estimates are significantly different from the null. On the other hand, disparity estimates among the minority groups are not significantly different from zero among the overlap population.

Disparity estimates may be sensitive to the target population toward which the health status variables are balanced, and notably so with IPW. Here, IPW forces us to balance the health status toward a hypothetical combined population, which is an unrealistic target for policy intervention since it emphasizes patients atypical for their own racial group. The disparity estimates are also likely subject to bias since we found IPW fails to adequately balance the health status variables in Section 4.2. Besides, the lack of overlap leads to loss of efficiency. For example, the largest normalized inverse probability weight is 0.32, accounting for almost one third of the total weights out of 1446 Asians. As a consequence, it is not surprising to for IPW to report the Whites-Asians disparity that is more than twice the magnitude of the GOW estimate. The overlap issue is also apparent when we apply the optimal trimming (4), which excludes about 20% of the sample (2125 Whites, 44 Asians, 1001 Blacks and 603 Hispanics). Unlike IPW, both TIPW and GMW provide disparity estimates closer to GOW, although with wider confidence intervals.

4.4 Analysis 2: Health Status Propensity Score Weighting with Rank and Replace Adjustment

While the health propensity score weighting in Section 4.3 allows us to balance health status variables without peeking at the outcome distribution, it does not account for the contribution of SES variables. The IOM definition requires adjustment for $\boldsymbol{X}_H$ but includes justifiable differences in the distributions of SES variables $\boldsymbol{X}_S$; the latter reflect differential impact of operations of health care systems and regulatory climate (IOM, 2003). If variables in $\boldsymbol{X}_H$ are independent of variables in $\boldsymbol{X}_S$, then the analysis in Section 4.3 is IOM-concordant; if the variables in $\boldsymbol{X}_H$ are correlated with variables in $\boldsymbol{X}_S$, health

status propensity score weighting may inadvertently alter the distributions of $\boldsymbol{X}_S$ and only provides an approximation to the IOM-defined disparity (Balsa et al., 2007). To address such a concern, we apply the rank-and-replace adjustment method (McGuire et al., 2006) to undo the undesired weighting of $\boldsymbol{X}_S$ by the health status propensity score. Cook et al. (2010) combined binary overlap weights with rank-and-replace SES adjustment; here we extend the method to comparing multiple racial groups.

Table 4: Racial disparity estimates in total health care expenditure (in dollars). The point estimates are obtained as weighted average controlled differences by the combined propensity score and rank and replace method. The associated 95% confidence intervals are obtained by bootstrap.

	IPW	TIPW	GMW	GOW
Whites-Asians	-1194	1133	997	1023
	(-5307, 2534)	(258, 1877)	(486, 1530)	(464, 1584)
Whites-Blacks	1610	1610	1013	1069
	(1184, 1980)	(1248, 1942)	(668, 1299)	(728, 1357)
Whites-Hispanics	1899	1883	1374	1420
	(1381, 2352)	(1446, 2232)	(1082, 1673)	(1128, 1731)
Asians-Blacks	2804	476	16	46
	(-965, 6926)	(-367, 1323)	(-578, 551)	(-582, 594)
Asians-Hispanics	3093	749	377	397
	(-689, 7149)	(-83, 1565)	(-184, 902)	(-206, 967)
Blacks-Hispanics	289	273	361	351
	(-273, 805)	(-177, 629)	(41, 722)	(27, 721)

Following Cook et al. (2009), we perform the rank-and-replace adjustment based on a model-based SES index to equalize the weighted SES distributions and the unweighted marginals. We model the health care expenditure as a function of $\boldsymbol{X}_H$, $\boldsymbol{X}_S$ and racial group indicator: $g(\mathbb{E}[Y_i|\boldsymbol{X}_{H,i}, \boldsymbol{X}_{S,i}, Z_i]) = \gamma_0 + \boldsymbol{X}_{H,i}^T\boldsymbol{\gamma}_H + \boldsymbol{X}_{S,i}^T\boldsymbol{\gamma}_S + \sum_{j=1}^J \gamma_{1j} D_{ij}$, where the SES predictive index is denoted by $\boldsymbol{X}_{S,i}^T\boldsymbol{\gamma}_S$. We

choose g as the log link, and to allow for heteroscedastic variances (Buntin and Zaslavsky, 2004), apply the Park test to determine the variance power relative to the mean (Park, 1966; Manning and Mullahy, 2001). In other words, the model parameters are estimated by a Tweedie generalized linear model with data-driven specification of the power variance function (Jørgensen, 1997). The estimated coefficients provide the SES index value for each patient, and we obtain the weighted rank of $\boldsymbol{X}_{S,i}^T \boldsymbol{\gamma}_S$ within each racial group. The rank-and-replace method then restores the original group-specific SES distributions by replacing the propensity score weighted SES index values with the equivalently ranked unweighted SES index values. With this adjustment, the weighted distribution of the SES index values in each group is approximately the same as the original distribution of the index values in that group, and the resulting disparity estimates become IOM-concordant by recapturing the racial differences in SES.

We obtain the SES-adjusted expected expenditure for each patient through the generalized linear model, and calculate the weighted average controlled differences based on the adjusted expenditure. After balancing the health status variables toward the overlap population, factoring the SES differences into the calculation increases the Whites-Blacks, Whites-Hispanics disparity by \$183 and \$199 and decreases the Whites-Asians disparity by \$137, without modifying the direction and statistical significance. Such changes may be anticipated, for example, between Whites and Blacks in the following case. Given Whites have overall higher health status and SES and that $\boldsymbol{X}_H$, $\boldsymbol{X}_S$ are likely positively correlated, White patient with lower health status and lower SES will be weighted more heavily to balance $\boldsymbol{X}_H$. Assuming that White patients with lower SES have lower health care utilization, we would expect the slight increase in the Whites-Blacks disparity after restoring the original SES distributions. On the other hand, the SES adjustment had a larger effect on disparities among the minority groups, but the results remain statistically insignificant. Overall, the changes in the GOW estimates from Table 3 and Table 4 suggest that racial differences in health care utilization were slightly mediated through the SES variables. In contrast, the SES adjustment magnifies the undue influence of extreme propensities when IPW is used to balance $\boldsymbol{X}_H$, since for example, Whites are found to on average spend \$1194 less than Asians among the combined population. With IPW, not only the hypothetical combined population is of minimal policy relevance, but also the inherent bias due to extreme propensities complicates the

interpretation of the unusual direction in such a point estimate.

5 Simulations

To further shed light on the comparison between different weighting methods, we conduct simulations in the context of observational studies with multiple non-randomized treatments. Our data generating process is similar to Yang et al. (2016) except that we consider nonzero pairwise average treatment effect among the considered target populations. We generate covariates X_{i1}, X_{2i} and X_{3i} from a multivariate normal distribution with mean vector $(2, 1, 1)$ and covariances of $(1, -1, -0.5)$; $X_{4i} \sim$ Uniform$[-3, 3]$; $X_{5i} \sim \chi^2_1$ and $X_{6i} \sim$ Bernoulli(0.5), with the covariate vector $\boldsymbol{X}_i^T = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i})$. The assignment mechanism follows the multinomial logistic model

$$(D_{i1}, \ldots, D_{iJ})|\boldsymbol{X}_i \sim \text{Multinom}(e_1(\boldsymbol{X}_i), \ldots, e_J(\boldsymbol{X}_i)),$$

where D_{ij} is the treatment indicator defined in Section 2.1 and $e_j(\boldsymbol{X}_i) = \exp(\alpha_j + \boldsymbol{X}_i^T\boldsymbol{\beta}_j)/\sum_{k=1}^{J}\exp(\alpha_k + \boldsymbol{X}_i^T\boldsymbol{\beta}_k)$ is the true generalized propensity score with $\alpha_1 = 0$, $\boldsymbol{\beta}_1^T = (0, 0, 0, 0, 0, 0)$. In the first simulation with $J = 3$ treatment groups, $\boldsymbol{\beta}_2^T = \kappa_2 \times (1, 1, 1, -1, -1, 1)$ and $\boldsymbol{\beta}_3^T = \kappa_3 \times (1, 1, 1, 1, 1, 1)$. We set $(\kappa_2, \kappa_3) = (0.2, 0.1)$ to simulate a scenario with adequate covariate overlap and $(\kappa_2, \kappa_3) = (0.8, 0.4)$ to induce lack of overlap with strong propensity tails, i.e., the propensity to receive certain treatment is close to zero for specific design values. We further choose α_2 and α_3 so that the overall treatment proportions are fixed at $(0.3, 0.4, 0.3)$. The potential outcomes are generated from $Y_i(j) = \left(1, \boldsymbol{X}_i^T\right)\boldsymbol{\gamma}_j + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$, $\boldsymbol{\gamma}_1^T = (-1.5, 1, 1, 1, 1, 1, 1)$, $\boldsymbol{\gamma}_2^T = (-4, 2, 3, 1, 2, 2, 2)$ and $\boldsymbol{\gamma}_3^T = (3, 3, 1, 2, -1, -1, -1)$. In the second simulation with $J = 6$ groups, we similarly specify the parameters to simulate both adequate and lack of overlap. The detailed specification and visual inspection of the overlap in each simulation scenario can be found in the supplementary material. The total sample size is fixed at $n = 1500$ for $J = 3$ and $n = 6000$ for $J = 6$.

For each scenario, we simulate 1000 datasets and estimate the pairwise causal effects using alternative estimators. To quantify the confounding bias in each simulation scenario, we first report the raw difference in means (DIF). For comparison among weighting methods, we consider GOW, IPW,

Table 5: Simulation results with $J = 3$ treatment groups. With adequate overlap, the optimally trimming excludes at most 2% of the total sample. Under lack of overlap, the optimal trimming rule excludes 19% to 30% of the total sample.

Metric	Method	Adequate Overlap			Lack of Overlap		
		$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$	$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$
	DIF	0.46	0.60	0.14	0.43	0.64	0.21
	IPW	0.02	0.01	0.01	0.19	0.02	0.17
	TIPW	0.01	0.002	0.01	0.03	0.01	0.01
|Bias|	GPSM	0.02	0.01	0.01	0.25	0.10	0.15
	TGPSM	0.02	0.004	0.01	0.08	0.02	0.05
	GMW	0.02	0.01	0.02	0.001	0.01	0.01
	GOW	0.01	0.001	0.01	0.01	0.01	0.003
	DIF	0.55	0.65	0.37	0.50	0.68	0.38
	IPW	0.20	0.16	0.26	1.04	0.61	1.16
	TIPW	0.16	0.16	0.23	0.38	0.28	0.47
RMSE	GPSM	0.26	0.22	0.31	0.86	0.51	0.90
	TGPSM	0.25	0.23	0.31	0.53	0.37	0.60
	GMW	0.17	0.18	0.27	0.29	0.24	0.36
	GOW	0.15	0.15	0.22	0.28	0.23	0.35
	DIF	0.64	0.36	0.92	0.65	0.23	0.90
	IPW	0.92	0.95	0.95	0.79	0.88	0.91
	TIPW	0.94	0.94	0.94	0.93	0.90	0.91
Coverage	GPSM	0.99	0.97	0.97	0.88	0.91	0.91
	TGPSM	0.98	0.96	0.98	0.95	0.92	0.95
	GMW	0.95	0.96	0.94	0.95	0.95	0.95
	GOW	0.94	0.96	0.95	0.95	0.94	0.94

TIPW and GMW. We also examine a recent propensity score matching estimator proposed by Yang et al. (2016), both without and with the optimal trimming step (GPSM and TGPSM). GPSM separately exploits each scalar propensity score for estimating the average potential outcomes and thus resolves the issue of matching on high-dimensional propensity score vector. Because the target population may differ in different estimators, we assess the accuracy of estimators relative to their corresponding target estimands. Specifically, the target estimands of DIF, IPW and GPSM are pairwise ATE for the combined population and are analytically determined from the true potential outcome model, whereas the target estimands for GMW, GOW, TIPW and TGPSM are defined for subpopulations and evaluated numerically based on Monte Carlo integration. For each data replicate, we estimate the generalized propensity scores based on the correct multinomial logistic regression model including all covariates. The proposed sandwich variance (7) was used to obtain confidence intervals for GOW. The empirical sandwich variance (see supplementary material for details) and the Abadie and Imbens (2012) variance were used to obtain interval estimators for IPW and GPSM. Since the weight function $w_j(\boldsymbol{X})$ for GMW is not everywhere differentiable (with infinite-many non-differentiable points) and fails to satisfy the regularity conditions for deriving a sandwich variance, we use bootstrap for interval estimation. Finally, whenever trimming is used, the generalized propensity scores are re-estimated based on the trimmed sample as refitting improves the finite-sample performance of the resulting estimators (Li et al., 2018b); accordingly, variance calculation is carried out based on the trimmed sample.

Table 5 summarize the absolute bias, root mean squared error (RMSE) and coverage of each estimator with $J = 3$ groups. As expected, DIF shows substantial bias and under-coverage, indirectly characterizing the magnitude of confounding bias. All other approaches perform reasonably well when there is adequate overlap. With lack of overlap, IPW and GPSM are sensitive to extreme propensities and produce biased point estimates. The optimal trimming method excludes 19% to 30% of the total sample, reduces the bias and improves efficiency and coverage in estimating the subpopulation causal effects. By down-weighting extreme units, both GMW and GOW provide unbiased point estimates with nominal coverage. Overall, TIPW, GMW and GOW are associated with the smallest RMSE and are more efficient than the other methods. Among them, GOW has the smallest RMSE, matching the

theoretical predictions in Section 2.3.

The simulation results with $J = 6$ groups are presented in the supplementary material. With adequate overlap, all methods have good control of confounding bias, produce unbiased estimates and close to nominal coverage. GMW and GOW provide the lowest RMSE, with the latter demonstrating higher efficiency for estimating most of the causal contrasts (the ratio of total MSE is 1.18). With lack of overlap, the clear separation of covariate space makes it challenging to simultaneously remove all confounding for estimating the 15 pairwise contrasts. By discarding more than half of the sample, the optimal trimming method improves the bias, efficiency and coverage properties over IPW and GPSM, both of which are subject to bias and excessive variance with extreme propensities. GMW and GOW further improve the efficiency and coverage properties upon trimming by down-weighting the extreme units. Concordant with the large-sample theory, GOW produces more efficient estimates than GMW for 12 out of 15 causal contrasts (the ratio of total MSE is 1.17). In this challenging scenario, the bootstrap CI for GMW has slightly better finite-sample coverage than the closed-form CI for GOW based on the empirical sandwich variance, but the closed-form CI estimator for GOW demonstrates the best coverage among all the considered closed-form CI estimators. However, another substantial gain of GOW over GMW is the computational time: for each simulation, the bootstrap interval estimates for GMW with 1000 samples require more than 80 times longer running time than that of the closed-form GOW interval estimates, which can be very burdensome for large observational data sets.

6 Discussion

We proposed a unified propensity score weighting framework, the balancing weights, for causal inference with multiple treatments. Within this framework, we developed the generalized overlap weights for pairwise comparisons to emphasize the target population with the most covariate overlap. We applied these new weights to study health care disparities and found Whites had significantly more spending on health care than the minority groups in 2009, after controlling for differential distributions of health status. In contrast, the disparity estimates are not significantly different from zero between the minorities.

This patten persists regardless of considerations of the SES differences. These results could potentially help health policy decision makers direct more resources and infrastructures for the minority groups to improve their access to medical care as a means to minimize the White-minority disparities in utilization.

Although we do not intend to make a causal statement of the racial disparity in health care utilization, there may be a tendency to do so based on the parallel discussion on health disparity or inequality. While one should generally distinguish between *health care* disparity and *health* disparity/inequality as noted in McGuire et al. (2006), it is possible to borrow the weak causal perspective of VanderWeele and Robinson (2014a,b) to interpret the disparity estimates in Section 4. For instance, the estimates in Table 3 could be understood as the remaining differences in health care utilization if we were to, hypothetically, intervene on the differential health status across groups. Because such an interpretation is not typical in studying health care disparity, we resort to the non-causal descriptive interpretation as in Cook et al. (2010) and Li et al. (2013).

The proposed methods are highly relevant in comparative effectiveness research based on observational data. For example, the target estimand—the pairwise ATO—describes the causal comparison in the subpopulation with clinical equipoise, and may be preferred (Li et al., 2018b). With the increasing use of convenience samples in observational studies, the proposed generalized overlap weights represent a flexible adjustment method to regain a target population where current practice remains uncertain, rather than a target population dominated by extreme units for whom treatment decisions are already clear. Our presentation has focused exclusively on categorical treatments but the concept of target population remains relevant with a continuous treatment. In the latter setting, the weighted estimands (1) may also be cast as the average potential outcomes among the combined population under a stochastic intervention or modified treatment policy (Muñoz and van der Laan, 2012; Haneuse and Rotnitzky, 2013), which may provide an alternative interpretation.

There are several directions for extending the proposed method. First, as with all propensity score methods, a well-estimated propensity score is crucial to the analysis. To focus on the main message, this paper adopted a convenient parametric model to estimate the generalized propensity scores. A natural extension is to use flexible machine learning models to estimate the generalized propensity scores;

examples include the Generalized Boosting Model (McCaffrey et al., 2004, 2013), ensemble learning methods such as the Super Learner (Dudoit and van der Laan, 2005; Pirracchio et al., 2015), the debiased machine learning estimator (Chernozhukov et al., 2018), as well as Bayesian nonparametric models.

Second, the generalized overlap weights are obtained by setting the linear contrast coefficients $\boldsymbol{a}$ to allow for pairwise comparisons, which are of general scientific interest with multiple categorical treatments. When there is no strong *a priori* preference for $\boldsymbol{a}$, one possibility is to choose $\boldsymbol{a}$ based on minimizing a specific loss function (Hirshberg and Zubizarreta, 2017).

Third, this paper focused on the moment weighting estimators; these estimators are not semiparametric efficient even with a correct propensity score model (Hirano et al., 2003). An important avenue for improvement is to consider the class of augmented weighting estimators with balancing weights (Robins et al., 1994). One could construct, for each choice of the balancing weights, an augmented estimator as

$$\hat{m}_j^{h,\text{aug}} = \hat{m}_j^h - \frac{\sum_{i=1}^n (D_{ij} - e_j(\boldsymbol{X}_i)) w_j(\boldsymbol{X}_i) \hat{m}_j(\boldsymbol{X}_i)}{\sum_{i=1}^n h(\boldsymbol{X}_i)},$$

where $\hat{m}_j(\boldsymbol{X}_i) = \hat{\mathbb{E}}[Y(j)|\boldsymbol{X}]$ is the outcome regression function. It can be shown that $\hat{m}_j^{h,\text{aug}}$ is semiparametric efficient for estimating m_j^h when both the generalized propensity score model and the regression function are correctly specified. Of note, when the tilting function $h(\boldsymbol{X}_i) = 1$, $\hat{m}_j^{h,\text{aug}}$ has an additional doubly-robustness property such that it is consistent to $\mathbb{E}[Y(j)]$ when either the generalized propensity score model or the regression function is correctly specified, but not necessarily both. However, this robustness property does not hold for $\hat{m}_j^{h,\text{aug}}$ when h is a function of the propensity scores, such as the optimal tilting function considered in Section 3.1. Nevertheless, additional work is warranted to study the efficiency property of the augmented generalized overlap weighting estimator with multiple treatments.

Finally, the balancing weights framework pursues weighting by propensity scores to achieve balance, with different choices of weights targeting specific populations and causal estimands. An alternative strand of recent literature derives weights that directly balance the covariates, bypassing the estimation of propensity scores; examples include the entropy balancing (Hainmueller, 2012), the stabi-

lized balancing weights (Zubizarreta, 2015) and the approximate residual balancing (Athey et al., 2018). Those weights usually focus on the ATE or ATT estimand with binary treatments, and do not involve adaptively changing the target population as our general balancing weights framework. In practice, it is prudent for the analyst to choose a method according to the scientific question and settings of specific applications rather than fixating on one single method.

References

Abadie, A. and Imbens, G. W. (2012), "A martingale representation for matching estimators," *Journal of the American Statistical Association*, 107, 833–843.

Athey, S., Imbens, G. W., and Wager, S. (2018), "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80, 597–623.

Balsa, A., Cao, Z., and McGuire, T. (2007), "Does managed health care reduce health care disparities between minorities and Whites?" *Journal of Health Economics*, 27, 781–807.

Buntin, M. B. and Zaslavsky, A. M. (2004), "Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures," *Journal of Health Economics*, 23, 525–542.

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

Cook, B. L., Mcguire, T. G., Lock, K., and Zaslavsky, A. M. (2010), "Comparing methods of racial and ethnic disparities measurement across different settings of mental health care," *Health Services Research*, 45, 825–847.

Cook, B. L., Mcguire, T. G., Meara, E., and Zaslavsky, A. M. (2009), "Adjusting for health status in

non-linear models of health care disparities," *Health Services and Outcomes Research Methodology*, 9, 1–21.

Cook, B. L., McGuire, T. G., and Zaslavsky, A. M. (2012), "Measuring racial/ethnic disparities in health care: Methods and practical issues," *Health Services Research*, 47, 1232–1254.

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, 187–199.

Ding, P. and Li, F. (2018), "Causal inference: A missing data perspective," *Statistical Science*, 33, 214–237.

Dudoit, S. and van der Laan, M. J. (2005), "Asymptotics of cross-validated risk estimation in estimator selection and performance assessment," *Statistical Methodology*, 2, 131–154.

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., and Li, X. S. (2012), "Generalized propensity score for estimating the average treatment effect of multiple treatments," *Statistics in Medicine*, 31, 681–697.

Hainmueller, J. (2012), "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis*, 1, 25–46.

Haneuse, S. and Rotnitzky, A. (2013), "Estimation of the effect of interventions that modify the received treatment," *Statistics in Medicine*, 32, 5260–5277.

Hirano, K., Imbens, G., and Ridder, G. (2003), "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hirshberg, D. A. and Zubizarreta, J. R. (2017), "On two approaches to weighting in causal inference," *Epidemiology*, 28, 812–816.

Imbens, G. W. (2000), "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

— (2004), "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and Statistics*, 86, 4–29.

IOM (2003), *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*, Washington, DC: The National Academies Press.

Jørgensen, B. (1997), *Theory of Dispersion Models*, London, UK: Chapman and Hall.

Lechner, M. (2002), "Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies," *Review of Economics and Statistics*, 84, 205–220.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018a), "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, 113, 390–400.

Li, F., Thomas, L. E., and Li, F. (2018b), "Addressing extreme propensity scores via the overlap weights," *Forthcoming at American Journal of Epidemiology*.

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013), "Propensity score weighting with multilevel data," *Statistics in Medicine*, 32, 3373–3387.

Li, L. and Greene, T. (2013), "A weighting analogue to pair matching in propensity score analysis," *International Journal of Biostatistics*, 9, 1–20.

Lopez, M. J. and Gutman, R. (2017), "Estimation of causal effects with multiple treatments: A review and new ideas," *Statistical Science*, 32, 432–454.

Manning, W. and Mullahy, J. (2001), "Estimating log models: to transform or not to transform?" *Journal of Health Economics*, 20, 461–494.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013), "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in Medicine*, 32, 3388–3414.

McCaffrey, D. F., Ridgeway, G., , and Morral, A. (2004), "Propensity score estimation with boosted regression for evaluating causal effects in observational studies," *Psychological Methods*, 403–425.

McGuire, T. G., Alegria, M., Cook, B. L., Wells, K. B., and Zaslavsky, A. M. (2006), "Implementing the Institute of Medicine definition of disparities: An application to mental health care," *Health Services Research*, 41, 1979–2005.

Moore, K. L., Neugebauer, R., Van der Laan, M. J., and Tager, I. B. (2012), "Causal inference in epidemiological studies with strong confounding," *Statistics in Medicine*, 31, 1380–1404.

Muñoz, I. D. and van der Laan, M. (2012), "Population Intervention Causal Effects Based on Stochastic Interventions," *Biometrics*, 68, 541–549.

Park, R. (1966), "Estimation with heteroscedastic error terms," *Econometrica*, 34, 888.

Pirracchio, R., Petersen, M. L., and van der Laan, M. (2015), "Improving propensity score estimators' robustness to model misspecification using Super Learner," *American Journal of Epidemiology*, 181, 108–119.

Rassen, J. A., Shelat, A. A., Franklin, J. M., Glynn, R. J., Solomon, D. H., and Schneeweiss, S. (2013), "Matching by propensity score in cohort studies with three treatment groups," *Epidemiology*, 24, 401–409.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression-coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846–866.

Stefanski, L. A. and Boos, D. D. (2002), "The calculus of M-estimation," *American Statistician*, 56, 29–38.

van der Laan, M. J. and Petersen, M. L. (2007), "Causal effect models for realistic individualized treatment and intention to treat rules," *International Journal of Biostatistics*, 3, 1–51.

VanderWeele, T. J. and Robinson, W. R. (2014a), "On the causal interpretation of race in regressions adjusting for confounding and mediating variables," *Epidemiology*, 25, 473–484.

— (2014b), "Rejoinder: How to reduce racial disparities?: Upon what to intervene?" *Epidemiology*, 25, 491–493.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016), "Propensity score matching and subclassification in observational studies with multi-level treatments," *Biometrics*, 72, 1055–1065.

Yoshida, K., Hernández-Díaz, S., Solomon, D. H., Jackson, J. W., Gagne, J. J., Glynn, R. J., and Franklin, J. M. (2017), "Matching weights to simultaneously compare three treatment groups comparison to three-way matching," *Epidemiology*, 28, 387–395.

Zanutto, E., Lu, B., and Hornik, R. (2005), "Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign," *Journal of Educational and Behavioral Statistics*, 30, 59–73.

Zaslavsky, A. M. and Ayanian, J. Z. (2005), "Integrating research on racial and ethnic disparities in health care over place and time," *Medical Care*, 43, 303–307.

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