BMC Medical Research Methodology

RESEARCH

Open Access

Assessing racial disparities in healthcare expenditure using generalized propensity score weighting

Jiajun Liu[1*], Yi Liu[2,3], Yunji Zhou[4] and Roland A. Matsouaka[1,3]

Abstract

Purpose This paper extends current propensity score weighting methods for causal inference to better understand disparities in healthcare access across multiple racial groups. By treating each racial group as a distinct entity (or "treatment") in the causal inference framework, we can assess and evaluate heterogeneity in healthcare outcomes across various racial or ethnic categories. Furthermore, we leverage modern propensity score weighting techniques to address the challenges inherent to multiple group evaluations, such as violations of the positivity assumption, and compare the performance of different propensity score weights.

Methods We use generalized propensity score methods to assess racial disparities across 4 specific racial or ethnic groups: Whites, Hispanics, Asians, and Blacks. We first calculate weights that standardize the participants' characteristics and then compare their weighted outcomes. We consider four distinct measures (i.e., causal estimands) and estimation methods: the conventional average treatment effect on the treated (ATT), the ATT trimming, the ATT truncation, and the overlap weighted ATT (OWATT). These estimands are applied under a multi-valued "treatment" framework, where the said "treatment" is defined by non-manipulable racial or ethnic group memberships. Using data from the Medical Expenditure Panel Survey (MEPS), we assess disparities in healthcare expenditures across the 4 racial and ethnic groups.

Results We found significant disparities in healthcare expenditure between White participants and all the other racial or ethnic groups when using OWATT and ATT truncation. Conventional ATT and ATT trimming could indicate non-significant difference due to larger variance estimates. Moreover, the conventional ATT was found to be the least efficient estimation method, even when its variance was estimated via non-parametric bootstrapping. Overall, the OWATT emerges as a promising estimation method; it retains the available information from all samples, avoids subjectivity (inherent to choosing thresholds by its competitors) and mitigates judiciously pernicious inferential effects (such as the inflated variance estimates) by extreme propensity score weights.

Conclusion We found that generalized propensity score weighting (GPSW) methods are valuable quantitative tools to standardize and compare characteristics as well as outcomes of non-manipulable groups. This helps assess disparities across multiple racial and ethnic groups, as demonstrated in this study. These methods offer flexible and semi-parametric analysis on the primary causal parameters of interest (such as the racial disparities), with straightforward and intuitive interpretations. In addition, when there is violation of the positivity assumption, OWATT serves as an excellent alternative due to its greater efficiency, evidenced by relatively smaller variance. More importantly,

*Correspondence:
Jiajun Liu
jiajun.liu@duke.edu
Full list of author information is available at the end of the article

the OWATT uses the entire dataset by assigning weights to all participants, regardless of their propensity score values. This feature of OWATT circumvents the need to specify user-defined thresholds, as required in ATT trimming or truncation, and retains as much data information as possible, leading to more reliable estimation results.

Keywords Non-manipulable group membership, Survey sample, Multinomial regression, Overlap weights, Average treatment effect on the treated

Introduction

Despite advancements in healthcare delivery and utilization, disparities persist across racial and ethnic groups in the United States [1, 2]. Ethnic and racial minorities continue to experience lower rates of healthcare utilization and adherence to prescribed treatments [3] and often receive different treatment or recommendations from healthcare providers [4]. Consequently, these populations face worse health outcomes and lower life expectancy [3, 5, 6].

Statistical analyses aimed at estimating racial disparities often use standard methods but with a specific focus: comparing outcomes across racial and ethnic groups while adjusting for differences in a set of pre-specified baseline covariates. Such a comparison is often done by using propensity score weighting (PSW) methods [7], where the propensity score (PS) is the probability of being like a participant of a specific group, given one's baseline characteristics. If we only have two groups to compare, estimated propensity scores are used to balance the covariates between the groups and compare their PS-weighted outcomes. When data are collected across multiple racial or ethnic groups, much of the current literature still emphasizes the use of separate covariate-balancing and outcome comparisons two-by-two groups at a time [8–12]. Unfortunately, this analytic approach limits our ability to fully assess the data and may lead to misleading or erroneous results, since the propensity scores weighting are estimated based on different subsets of the data set. With multiple groups, the analysis should extend beyond the binary propensity score to a generalized propensity score, where the probabilities of being like a participant of each of the specific groups are estimated at once. Using the entire data set. This "multi-category" version of the propensity score allows us to estimate adequately average outcomes and evaluate the necessary contrasts across different groups [13, 14]. While this all-encompassing solution is laudable, it however comes with its own major drawback: estimations of the generalized propensity scores often exacerbate issues related to the positivity of the propensity scores.

The positivity assumption, which is crucial for a valid causal inference [15], requires that the propensity scores should be sufficiently bounded away from 0 and 1 [16–18]. It assumes that all the most important covariate patterns in the population of participants are also present across the different groups in our data set. This may not be the case if there is an imbalance in the group allocation of participants, especially if we have low proportions of certain racial or ethnic groups (underrepresented groups). In this case, certain participants' characteristics only apply to some of the groups. It may also be an issue when there a systematic covariate imbalance across groups [19]. Unfortunately, violations of the positivity assumption (also referred to as lack of positivity) often occur in practice. Examples of such violation include the analyses of North Carolina birth weight [19, 20] and racial disparities in health care expenditure [21]. Violations of the positivity assumption can lead to biased and inefficient (large standard error) estimations [22–25].

There are three mainstream methods for tackling the violations of positivity assumption: (i) propensity score trimming [26]; (ii) propensity score truncation [27–29]; and (iii) overlap weights (OW) [14, 22, 23]. These methods belong to a unified causal inference framework, the *weighted average treatment effect* (WATE), proposed by Li et al. [22]. The WATE centers around measuring the *average treatment effect* (ATE) under lack of positivity, which can be viewed as an "ATE-type" estimation problem. In addition to ATE, researchers might be interested in estimating *the average treatment effect on the treated* (ATT). In assessing racial disparities, the "treated" refers to a chosen reference racial or ethnic group, such as "White" among White, Black, Hispanic and non-Hispanic Asian populations. To address the lack of positivity in ATT estimation, Liu et al. [25] proposed the *weighted average treatment effect on the treated* (WATT) framework, which is centered around ATT estimation. Although the WATE framework has been applied to analyze racial disparities [7, 14, 21], there is limited work applying this recently proposed WATT framework. When we want to compare a group of participants to another specific one, which serves as a reference, such an assessment of disparities is often well-suited under the WATT framework. Therefore, in this paper, we consider the application of WATT to assess racial or ethnic disparities in healthcare expenditures using the 2003–2004 Medical Expenditure Panel Survey (MEPS), using generalized propensity score methods for multiple racial or ethnic groups.

The MEPS is a comprehensive set of large-scale surveys that collect detailed information on healthcare costs, utilization, and insurance coverage among families and individuals, their medical providers, and employers across the United States [30]. As the most complete source of data on healthcare expenditure and utilization, MEPS provides insights into the specific health services Americans use, their frequency, associated costs, and payment methods, as well as data on the scope and availability of health insurance for U.S. workers (https://meps.ahrq.gov//mepsweb/).

To investigate heterogeneity in healthcare spending across four racial groups (i.e., White, non-Hispanic Black, Hispanic and non-Hispanic Asian participants), we consider four estimands from the WATT framework: the conventional ATT, ATT trimming, ATT truncation, and overlap weighted ATT (OWATT). These methods offer flexible and semi-parametric estimations of racial disparities. By "flexible", we mean that these approaches differ from conventional outcome regression methods—which interpret the results based on the estimated regression parameters and may be sensitive to model misspecifications such as including incorrect predictors or choosing an inappropriate model. WATT framework uses background information to create pseudo subpopulations with balanced covariate distributions [31, 32]. By "semi-parametric", we refer to the models used to estimate the generalized propensity scores (and calculate the weights). These models can vary by their nature; they might include more straightforward, interpretable options like generalized linear regression (parametric) or alternative nonparametric machine learning models [33]. In addition, the interpretations of the methods we choose are straightforward and intuitive, as they directly define the quantities that we intend to measure, which facilitates a clearer insight of the racial disparities.

Our work also extends the findings of Cook et al. and Li and Li [7, 9], who used propensity scores under a binary treatment for racial disparities analysis. Our generalization to multi-valued treatment provides additional insights to practical investigators. This approach contrasts with the use of pairwise comparisons that only evaluate data two-by-two groups at a time, which can miss broader patterns and trends by not utilizing the groups are part of the full dataset.

The remainder of this paper is organized as follows. Section "Racial Disparity in Healthcare Expenditures" reviews the statistical framework for the WATT for multiple groups. In Section "Concluding Remarks", we apply these WATT to the MEPS data to examine racial disparities in healthcare expenditures. We especially compare the methods based on effective sample sizes, covariate balance performance, and both point and variance estimates. We conclude the paper, in Section 4, with some remarks, discussions, and future orientations.

Method

Average treatment effect on the treated in the context of racial disparities

We denote the outcome (healthcare expenditure) by Y and the racial or ethnic group categories by$Z$, with $Z \in \{1, \ldots, K\}$ and$K \geq 2$. We consider the generalized propensity scores defined by$e_i(X) = \Pr(Z = i|X)$, $i = 1, \ldots, K$. It can be modeled by a multinomial regression model, an extension of the binomial logistic regression, using an observed covariate vector $X = (X_1, X_2, \ldots, X_p)$, which contains information from multiple characteristics measured on the participants. The generalized propensity score $e_i(X)$ is the probability of being the race of group i $(i = 1, \ldots, K)$ given the covariates$X$, subject to$\sum_{i=1}^{K} e_i(X) = 1$.

We must point out that race and ethnicity are immutable in the sense that they cannot be experimented on, manipulated, or assigned to study participants as it is the case with treatment assignments [34, 35]. We do not view race or ethnicity, in this paper, as inherent biological factors that relate people to regions of ancestry, various genotypes and phenotypes [34, 35]. Although not manipulable, we view race and ethnicity as social constructs; hence we consider racial race or ethnic groups as entities from which we want to standardize characteristics across, following Li and Li [7]. This allows us to understand how race and ethnicity play out in people's lives. Therefore, we use the generalized propensity score as a quantitative tool to summarize the relationship of covariates and racial categories. In addition, the generalized propensity scores $e_i(X)$ here can be well-defined and modelled statistically, serving as a balancing score for balancing covariate distributions [14, 22] across multiple race groups, ensuring fairer comparisons on healthcare expenditures through weighting.

We work within the potential outcomes framework of Neyman and Rubin [15, 36, 37]. For the sake of comparison, we use the White race as the "treated" group of interest, i.e., the main group of interest (or the reference group) from which we will compare all the other groups. The goal is to create pseudo or counterfactual minority population, based on specific weights, which we will then compare to the White group, following Li and Li [7]. By comparing the White group to a counterfactual minority population whose covariate distributions are similar to that of the White population, we are interested in how disparities impact other races' health expenditure compared to the White group. The difference is estimated between any two groups, with the understanding that each participant has multiple *potential outcomes* [15, 37],

one for each racial or ethnic group (possibly contrary to the fact)–-under the stable unit treatment value assumption (SUTVA) [16, 37]. Let $Y(i)$ be the potential outcome if the participant is in group i $(i = 1, \ldots, K)$, while the observable outcome is such that $Y = Y(Z)$, by the consistency assumption, where $Z(Z \in \{1, \ldots, K\})$ indicates which group the participant belongs to.

Let $Z = j$ be the reference group and $Z = i(i \neq j)$ denotes one of the other groups. The measure (i.e., causal estimand) of interest is the average treatment effect on the treated (ATT):

$$\tau_{i,j}^{att} = E\big[Y(j) - Y(i)|Z = j\big] = E[Y|Z = j] - E[Y(i)|Z = j].$$

While the first term $E[Y|Z = j]$ can be determined from data, the second term is not as it involves counterfactual outcomes $Y(i)$ for participants who are in group i, but being evaluated at the reference group j. Therefore, we use the SUTVA and consistency assumption and leverage the data from the other group i to impute the potential outcome $Y(i)$ in the reference group j. This is done by weighting the observed outcomes of those in the group $i(i \neq j)$, where the weights ensure the balance of the covariate distributions between the other group i and the reference group j. Readers are referred to the end of Appendix A.1 for technical details and justification for this weight balancing process.

To estimate ATT, we need some additional statistical assumptions, which are listed in Appendix A.1 as technical details. Then, we can use the following *weighting estimator*:

$$\widehat{\tau}_{i,j}^{att} = \frac{\sum_{l=1}^{N} D_{lj} Y_l}{\sum_{l=1}^{N} D_{lj}} - \frac{\sum_{l=1}^{N} w_{i,j}(x_l) D_{li} Y_l}{\sum_{l=1}^{N} w_{i,j}(x_l) D_{li}}, \tag{1}$$

where $w_{i,j}(x_l) = \frac{e_j(x)}{e_i(x)}$ $(i \neq j)$ is the weight for a participant l in group i, $(i \neq j)$, $D_{lj} = \mathbb{I}\{Z_l = j\}$ indicates that the participant l is in group j, and $\mathbb{I}\{\cdot\}$ is the indicator function.

The estimator $\widehat{\tau}_{i,j}^{att}$ requires that we estimate the generalized propensity scores $e_i(x)$, for $\forall i \in \{1, \ldots, K\}$, by specifying a model. For this paper, we consider the multinomial regression model $\log(e_i(x)/e_1(x)) = x\prime\alpha_i$, where α_i is a regression coefficient vector, for $i = 2, \ldots, K$, to estimate the generalized propensity score $\widehat{e}_i(x) = \widehat{e}_1(x)\exp(x\prime\widehat{\alpha}_i)$, and plug it in the above weighting estimator, where $\widehat{e}_1(x) = \left(1 + \sum_{i=2}^{K} \exp(x\prime\widehat{\alpha}_i)\widehat{e}_i(x)\right)^{-1}$.

Based on statistical theory, the estimator $\widehat{\tau}_{i,j}^{att}$ is consistent to the true ATT under large sample if the generalized propensity scores are estimated consistently [14, 25]. Nevertheless, this estimator can be extremely unstable (i.e., with large variance) when the generalized propensity score estimates $\widehat{e}_i(x)$ are too small (close to 0) for some participants, which can lead to extreme weights $w_{i,j}(x_l)$. These small values of $\widehat{e}_i(x)$ can happen for various reasons, including by happenstance, model misspecification, random error, or intrinsic characteristics of the true generalized propensity score model. Participants who have such extreme generalized propensity score values are also referred to as violating the positivity assumption (as defined in Section "Introduction"). The statistical details about the positivity assumption can also be found in Appendix A.1.

In the following Section "Effective Sample Size", we outline some general strategies for addressing the violation of positivity based on the estimated generalized propensity scores from participants. We then specify how we especially address this issue in ATT estimation by introducing the weighted average treatment effect (WATT) framework [25]. All technical details can be found in Appendix A.2.

Addressing the Violation of Positivity in ATT Estimation

In literature of causal inference methodology, there are three mainstream methods for addressing the violation of positivity in our knowledge: (i) trimming [26]; (ii) truncation [27–29]; and (iii) overlap weights (OW) [14, 22, 23].

Trimming excludes participants whose generalized propensity scores $\widehat{e}_i(x_l)$ fall outside of a pre-specified range $[\alpha, 1]$, for some user-selected α, and $i = 1, \ldots, K$ with $i \neq j$ [14, 38]. Following Li et al. [23], after trimming, we re-run the same multinomial model, re-estimate the propensity scores $\widehat{e}_i(x_l)$ and update from the weights $w_{i,j}(x_l)$ using the remaining data. For binary treatment, we require the estimated propensity scores to be smaller than $1 - \alpha$ in ATT estimation. However, it is important that for multiple treatment groups, the $\widehat{e}_i(x_l)$ is on the denominator of the propensity score weights, and thus extremely small $\widehat{e}_i(x_l)$ tends to result in extreme large weights. Therefore, for multiple treatment groups, we only keep the participants have generalized propensity scores within $[\alpha, 1]$.

Truncation, also known as "weight-capping", assigns to participants whose generalized propensity scores $\widehat{e}_i(x_l)$ are below the threshold α a fixed weight $w_{i,j}(x_l) = \widehat{e}_j(x_l)/\alpha$, i.e., their generalized propensity scores are capped by the threshold value in the final estimation stage.

We need to point out that ATT estimation through both trimming and truncation is subjective since it involves ad hoc selections of a user-specified threshold parameter(s) α. this often leads to loss of information and sensitivity of the estimated racial disparities [23, 25, 39]. In contrast, the overlap weights (OW) avoid eliciting such a threshold by shifting the target of estimation to where the demand for positivity is essentially reduced (see details in Appendix A.2 for the definitions of OW).

By construction and by their nature, trimming, truncation, and OW change their target (i.e., the estimand and underlying population). For example, trimming by $\alpha = 0.1$ only keeps participants whose generalized propensity scores $\widehat{e}_i(x_l)$ are in [0.1, 1]. The shifted target is then based on those participants who cannot have their true generalized propensity scores outside of interval [0.1, 1]. Similarly, for truncation and OW, they have their own shifted target population. This is shifting phenomenon is often referred to as *moving the goalpost* [26]. This population shift has an important implication in practice, that is, the aforementioned methods always focus on participants with sufficient positivity, so that the treatment effects of interest (here the racial disparities) can always be estimated with good efficiency.

Weighted Average Treatment Effect on the Treated (WATT)

When the interested measure is ATT, we need to be more specific about the use of above methods in dealing with the lack of positivity in ATT estimation. In fact, the methods in Section "Effective Sample Size", are within the framework of WATT [25]. The WATT is a class of methods centering around ATT.

In this paper, building on the work of Liu et al. [25], we use the generalized propensity score designed for multiple race groups and considered four estimation methods from the WATT framework to assess the racial disparities in healthcare expenditure: (i) the conventional ATT; (ii) ATT trimming; (iii) ATT truncation; and (iv) overlap weighted ATT (OWATT). Methods in (ii)—(iv) especially target populations where the demand on the positivity is reduced.

The estimation of (ii)—(iv) follows similarly the weighting estimator (1) for ATT, with different respective weights; details are provided in Appendix A.2. For variance estimation, we use the non-parametric bootstrap method [40], which has shown valid for the WATT framework [25].

Racial disparity in healthcare expenditures

When assessing racial or ethnic disparities, the variable race or ethnicity (viewed as a "treatment group" variable) is mainly used to capture and adjust for differences in covariates across racial or ethnic groups and determine how they influence the outcome. Because race and ethnicity are social constructs that are not manipulable, they cannot be used to offer a causal interpretation of the difference in outcomes between groups of participants via the standard potential outcome framework. Nevertheless, according to the definition of disparity in healthcare provided by the Institute of Medicine (IOM), i.e., the difference in treatments assigned to society groups, such disparity can be well explained by health status and treatment preference [41]. Hence, controlling for health status covariates is expected to assist in the estimation and interpretation of disparities in healthcare. In a sense, comparisons of racial disparities in healthcare share a similar nature to the comparisons that control for confounders in a causal sense [11, 14]. Therefore, we can leverage the generalized propensity scores to balance the distributions of the covariates across racial or ethnic groups and apply generalized propensity score weighting methods on the Medical Expenditure Panel Survey (MEPS) data to examine the impact of racial or ethnic disparities on healthcare expenditure [8–11].

For this analysis, we consider the MEPS data from 2003 to 2004, where healthcare records and information from 20,446 participants were collected, including 9,830 White (48.1%), 5,280 Hispanic (25.8%), 1,431 Non-Hispanic Asian (7.0%), and 3,905 Non-Hispanic Black (19.1%) participants [9]. Thus, the data set has a sufficiently large sample size, and the consistency of ATT estimator is expected as long as the generalized propensity scores are estimated consistently, as we discussed in Section "Effective Sample Size". The baseline covariates were separated into Socioeconomic Status (SES) variables and health status variables. Following the IOM recommendations, estimates of healthcare disparities were adjusted for health status factors but not for SES variables [42]. Therefore, following Li and Li [16] and McGuire et al. [42], we only included health status variables to estimate the generalized propensity scores to assess covariate balance across the different racial or ethnic groups. The list of health status variables included gender, marital status, age, body mass index (BMI), MI, self-reported health status, SF-12 mental component summary, smoke, exercise, diabetes, asthma, stroke, cholesterol, cancer, and blood pressure. We use the total medical expenditure as the outcome and choose the White group as the treated group to which we compare all the other racial or ethnic groups.

To ensure that the analysis results accurately represent the underlying target population, we incorporate the survey weights of each participant into our analysis. Each participant's survey weight is available in the dataset as the "PERWT" variable. Incorporating these survey weights addresses potential biases and imbalances in the sampling process through a two-stage approach [43, 44]. First stage: Generalized propensity scores are estimated using a weighted multinomial regression that involve survey weights. In this model, the "PERWT" variable affects the estimation of model parameters by weighting the loss function or likelihood. As a result, the distribution of generalized propensity scores—and the effectiveness of the weighting approach in balancing covariates—is implicitly influenced by the survey weights. Second stage: During the estimation phase, we construct a combined weight for each

participant by multiplying the normalized generalized propensity score weight with the normalized survey weight. This step is equivalent to applying our generalized propensity score weighting framework to a rescaled outcome: the survey-weighted healthcare expenditure. Therefore, we incorporate the survey weight variable in two ways: first, by including it in the generalized propensity score model, and second, by using it to rescale the outcome.

By systematically integrating survey weights into both stages, our methodology aims to correct for sampling biases, ensuring more reliable and representative analysis for the underlying target population.

Overlap and covariates balance

To assess the overlap of generalized propensity scores for the four racial or ethnic groups and evaluate the balance of covariates, we first estimate the propensity scores using multinomial regression on these groups, adjusting for the health status variables as aforementioned. The survey weights are incorporated into the multinomial regression by setting the weight argument. In Fig. 1, each plot represents the distribution of generalized propensity scores for each racial or ethnic group and is categorized by the true racial group to which the participants belong.

These plots first evaluate the overlap of the generalized propensity scores across different racial groups. Overall, the estimated generalized propensity scores demonstrate a moderate degree of overlap across the groups, suggesting partial comparability of the distribution. Additionally, Fig. 1 reveals that some generalized propensity scores are near 0 and 1. This pattern indicates a potential violation of the positivity assumption, which can pose challenges when using propensity score weighting approaches, such as conventional ATT. Therefore, we shift from conventional ATT to other weighted ATT approaches to handle this problem, such as trimming, truncation, and overlap weighting, and further compare their performance in ATT estimation.

To explore how the four approaches perform differently in balancing covariates, we apply the absolute standardized mean differences to measure the overall difference in each covariate across multiple groups and show a love plot in Fig. 2 [25, 45, 46]. It is obvious that most of the covariates can achieve better balance across groups after imposing weights for Conventional ATT, Truncation ATT, and Overlap Weighted ATT (OWATT). However, trimming ATT weights can lead to larger imbalance for most of the covariates compared to unadjusted. In

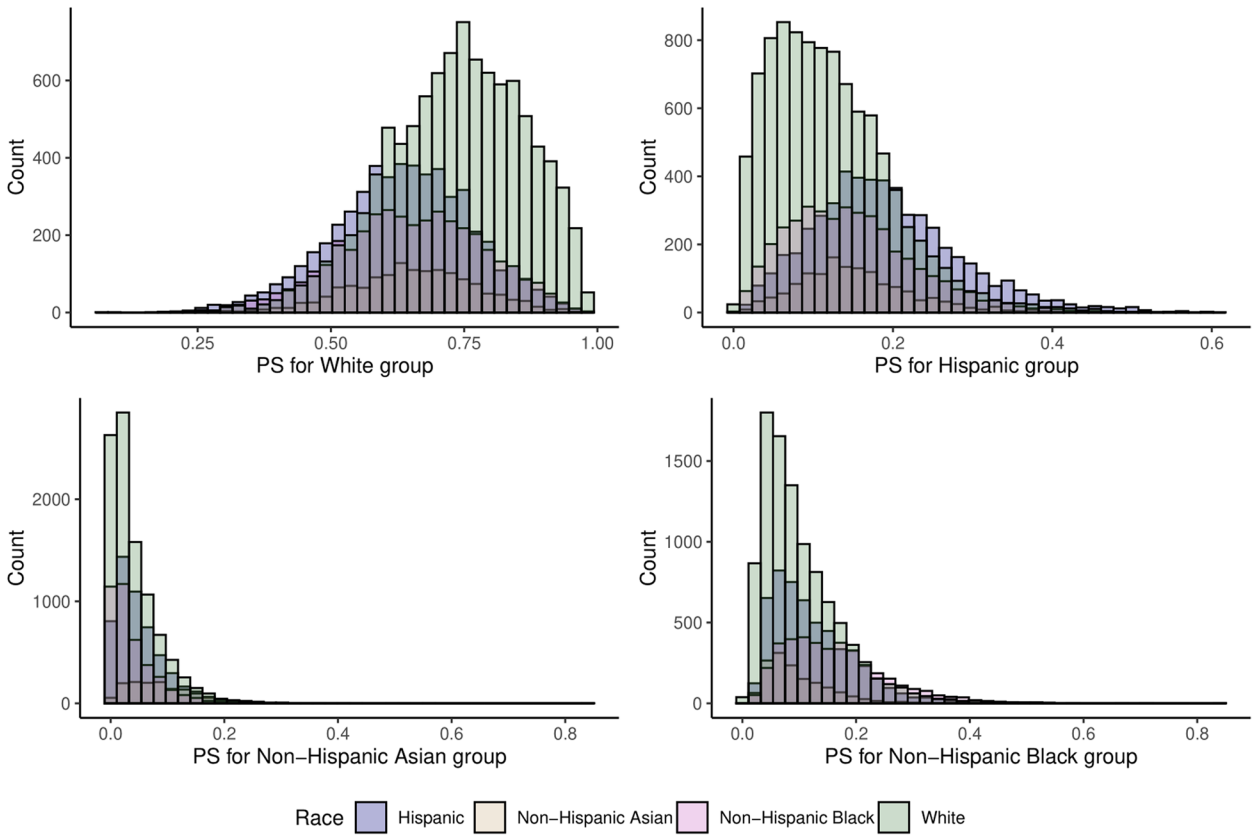

Fig. 1 Histogram of the estimated generalized propensity score for each racial group

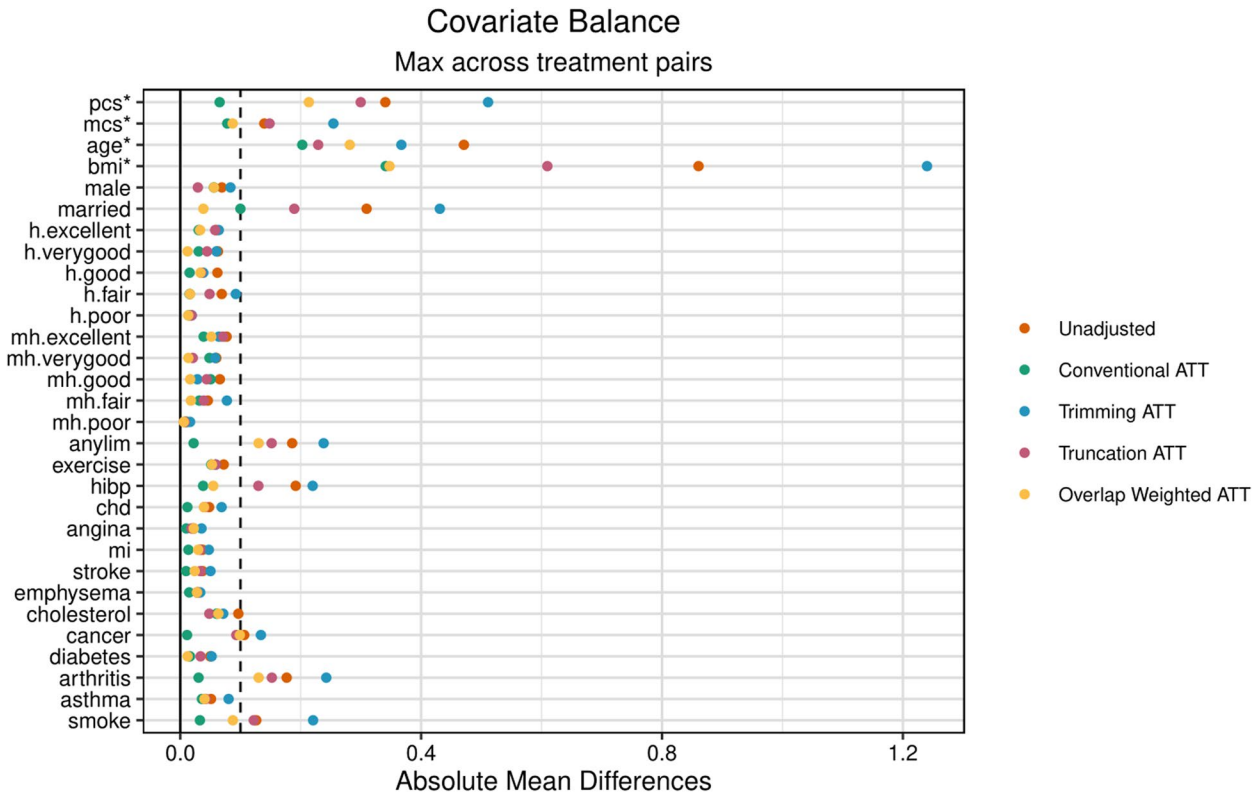

Fig. 2 Covariate balance under different ATT methods. Unadjusted: the simple difference in means of the two comparison groups without weighting; Conventional ATT: the original ATT estimation based on propensity score weighting but without handling positivity assumption violation; Trimming ATT: ATT estimation based on the subsample with propensity scores in $[\alpha, 1], \alpha = 0.10$; Truncation ATT: ATT estimation based on the entire sample but with propensity scores capped at $\alpha = 0.10$; Overlap Weighted ATT: ATT estimation based on the entire sample with individual-level overlap weights

comparison, OWATT leads most of the covariates have absolute standardized mean differences closer to 0 and bounded by 0.1, which can be considered a better covariate balance overall.

Effective sample size

To evaluate how well the information contained in the different weights we considered helps capture the essence of the data at hand and measures the efficiency of the weighting schemes, we estimated the effective sample size (ESS) across the different treatment groups [19, 20]. For a tilting function $h(X)$, the ESS for group i is defined as

$$ESS_i^h = \frac{\left(\sum_{l=1}^{N}\sum_{i=1}^{K} D_{li} w_i(X_l)\right)^2}{\sum_{l=1}^{N}\sum_{i=1}^{K} D_{li} w_i^2(X_l)}.$$

The ESS is a measure of the efficiency of the weighting scheme considered. It provides an approximate sample size of a simple random sample that is required to obtain an estimate with a similar level of precision than a weighted sample [19]. The closer the ESS to the original sample size the better. Since we used the White group as the pre-specified "treated" group, the ESS for the White group is the whole sample of 9830 White participants. With trimming and truncation, we illustrate the result in terms of different threshold, including $\alpha \in \{0.05, 0.1, 0.15\}$. As indicated in Table 1, the ESS decreases in all three racial groups other than White with conventional ATT, ATT trimming, ATT truncation, and OWATT compared to the original sample size (the "Unadjusted" row). In an overall sense, the conventional ATT and ATT trimming have smaller ESS compared to OWATT and ATT truncation. As trimming more participants, the ESS for that racial group tends to decrease.

Specifically, the ESS for the non-Hispanic Asian group is equal to 155.01 for the conventional ATT, which is small compared to the original sample size of 1,431 participants. This implies some extreme weights may exist in non-Hispanic Asians, which dilutes the contributions of some of the participants in the estimation. This result is consistent with the observations made in Fig. 1 and the findings of Li and Li [14]. The impact of these

Table 1 Effective sample size of each racial or ethnic group

	White	Hispanic	Non-Hispanic Asian	Non-Hispanic Black	Total
Unadjusted[1]	9830	5280	1431	3905	20,446
Conventional ATT[2]	9830	1837.92	155.01	1792.98	13,615.91
OWATT[3]	9830	3677.27	1105.31	2128.12	16,740.70
ATT Trimming (0.05)	9830	2728.68	495.78	2242.28	15,296.74
ATT Trimming (0.1)	9830	2645.63	260.80	1617.05	14,353.48
ATT Trimming (0.15)	9830	1559.88	106.04	1087.12	12,583.04
ATT Truncation (0.05)	9830	3218.58	1118.54	2543.93	16,711.05
ATT Truncation (0.1)	9830	4040.73	1302.78	3130.08	18,303.59
ATT Truncation (0.15)	9830	4521.17	1359.23	3452.28	19,162.68

[1] Unadjusted means the simple difference in means of the two comparison groups without weighting

[2] The original ATT estimation based on propensity score weighting but without handling positivity assumption violation

[3] Overlap weighted ATT

extreme weights was not adequately mitigated by trimming at $\alpha = 0.1$ as the ESS is equal to 260.80, which did not improve much compared to the original sample size. However, the OWATT and ATT truncation can increase the ESS for Non-Hispanic Asian to 1105.31 and 1118.54 ($\alpha = 0.05$) respectively, which are noticeable improvements. This outperformance of OWATT and truncation ATT is also reflected in Hispanic and Non-Hispanic Black group, where the ESS of trimming ATT and conventional ATT are smaller. Although truncation ATT has ESS larger than OWATT as the increase of threshold α, the larger α implies more original information is to be overwritten.

Estimated causal effects of racial disparity

To be more comprehensive, we considered three thresholds in estimating ATT with trimming or truncation: $\alpha = 0.05$, 0.10, and 0.15. The choice of a specific threshold has an impact on the results we will obtain since the choice influences how many participants' generalized propensity scores will be capped with α in truncation and how many participants will be dropped via trimming. Setting a higher threshold means more participants will be trimmed and have more information loss. To adequately compare the efficiency of the different methods, the variance (and thus the standard error [SE]) was estimated using the non-parametric bootstrap method with 1000 replicates.

The estimated health expenditure disparities across the different racial or ethnic groups under different methods and varying thresholds for trimming and truncation are presented in Table 2. Overall, all the methods agree on that there are some significant disparities in healthcare expenditure between White and Non-Hispanic Asian group. This result can be interpreted as, on average, White participants should have paid more in healthcare expenditure compared to Non-Hispanic Asian participants had they had similar distributions of the covariates as the White participants. Almost all methods consider the magnitude of the disparities between these two groups to be more than $2,500. Specifically, when using OWATT, the healthcare expenditure for White population is expected to be $2436.67 more than that for Non-Hispanic Asian population who shared similar covariate characteristics distribution.

For White and Hispanic, the conventional ATT and ATT trimming ($\alpha = 0.15$) reach a different conclusion compared to the other methods; they found no significant disparity in healthcare expenditure. As trimming too much (such as at$\alpha = 0.15$) leads to substantial loss of information, the bootstrap variance for ATT trimming is also very large and leads to a p-value that is larger than 0.05. Among the other methods that led to significant differences, the estimated differences in healthcare expenditures between the White group and the Hispanic group range from $1383.05 (ATT truncation ($\alpha = 0.05$)) to $2386.84 (OWATT). In the comparison, for White and Non-Hispanic Black participants, all the methods other than ATT trimming ($\alpha = 0.10$&$\alpha = 0.15$) indicate a significant disparity. Overall, when compared with White group, the estimated differences in healthcare expenditure for Non-Hispanic Black are smaller than that of the other two minority groups, ranging from $ 611.14 (ATT truncation($\alpha = 0.15$)) to $2077.52 (OWATT).

The estimation of the differences in healthcare expenditure varies substantially across different methods. As discussed at the end of Section "Effective Sample Size", OWATT, ATT trimming, and ATT truncation are shifting the target population from that is targeted by the conventional ATT, in order to handle the lack of positivity (under which the conventional ATT cannot often be well estimated). Therefore, the estimation results from these approaches are not expected to be the same as those from the conventional ATT. Case in point, compared to the conventional ATT, the OWATT, the ATT trimming, and the ATT truncation all reveal larger racial disparities between White and Non-Hispanic Asian groups and between White and Hispanic groups. In addition, the conventional ATT indicates that Hispanic group and Non-Hispanic Black group have the smallest disparity from White group. However, ATT trimming, ATT truncation,

Table 2 Point Estimates and SEs for racial disparity among three comparisons

Comparison	Method	Point Estimate	Standard Error	*P*-value
White vs. Hispanic	Conventional ATT[1]	626.79	405.55	0.122
	OWATT[2]	2386.84	225.31	< 0.001
	ATT Trimming ($\alpha = 0.05$)	1627.18	269.84	< 0.001
	ATT Trimming ($\alpha = 0.10$)	2324.48	393.45	< 0.001
	ATT Trimming ($\alpha = 0.15$)	3047.6	5126.87	0.552
	ATT Truncation ($\alpha = 0.05$)	1383.05	258.14	< 0.001
	ATT Truncation ($\alpha = 0.10$)	1898.53	207.72	< 0.001
	ATT Truncation ($\alpha = 0.15$)	2191.3	192.84	< 0.001
White vs Non-Hispanic Asian	Conventional ATT	1597.95	579.48	0.006
	OWATT	2436.67	304.96	< 0.001
	ATT Trimming ($\alpha = 0.05$)	2679.41	421.02	< 0.001
	ATT Trimming ($\alpha = 0.10$)	3307.70	416.10	< 0.001
	ATT Trimming ($\alpha = 0.15$)	2634.83	768.38	< 0.001
	ATT Truncation ($\alpha = 0.05$)	2336.69	302.04	< 0.001
	ATT Truncation ($\alpha = 0.10$	2552.4	267.32	< 0.001
	ATT Truncation ($\alpha = 0.15$)	2634.19	257.70	< 0.001
White vs Non-Hispanic Black	Conventional ATT	875.51	279.11	0.002
	OWATT	2077.52	222.06	< 0.001
	ATT Trimming ($\alpha = 0.05$)	873.23	323.08	0.007
	ATT Trimming ($\alpha = 0.10$)	663.21	413.79	0.109
	ATT Trimming ($\alpha = 0.15$)	611.14	1506.99	0.685
	ATT Truncation ($\alpha = 0.05$)	942.46	267.42	< 0.001
	ATT Truncation ($\alpha = 0.10$)	912.97	255.73	< 0.001
	ATT Truncation ($\alpha = 0.15$)	918.9	247.73	< 0.001

[1] The original ATT estimation based on propensity score weighting but without handling positivity assumption violation

[2] Overlap weighted ATT

and OWATT show an opposite pattern, i.e., the disparities are more substantial for White vs. Hispanic and White vs. Non-Hispanic Asian, with the difference being at least greater than $1383.05. Under the same threshold, the estimated disparities between Non-Hispanic Asian and White are similar for ATT trimming and ATT truncation. Nevertheless, compared to ATT truncation, the estimated expenditure difference is larger for ATT trimming when the Hispanic group is the comparator and smaller when Non-Hispanic Black is the comparator.

Furthermore, the estimated racial disparity in health expenditure increases with more participants being trimmed or capped when comparing Hispanic and Non-Hispanic Asian participants to White but decreases when comparing Non-Hispanic Black to White participants. Regardless of the groups being compared, the difference in healthcare expenditure using OWATT seems to be consistent and remains around $2,000, with the minimum difference being $2,077.52 (White vs. Non-Hispanic Black) and the maximum equal to $2,436.67 (White vs. Non-Hispanic Asian).

As seen in Table 2, compared to the conventional ATT and the ATT trimming, the OWATT standard error estimates are always smaller among all three comparisons, regardless of which threshold being selected. This indicates that OWATT achieves higher efficiency than these two other methods in ATT estimation. Indeed, when comparing White and Non-Hispanic Black groups, OWATT demonstrates the highest efficiency in ATT estimation, achieving the smallest bootstrap variance—smaller even than that of truncation ATT. Although ATT truncation results in slightly smaller standard errors when more participants' generalized propensity scores are truncated ($\alpha = 0.15$), OWATT remains a promising alternative, because it preserves participants' information and avoids relying on the subjective choice of the truncation threshold.

Concluding remarks

Summary

In this paper, we apply the WATT framework of Liu et al. [25], combined with the generalized propensity score methods [14] to analyze the MEPS data to assess

racial or ethnic disparities in healthcare expenditures. We specifically highlight OWATT as a useful method within the class of WATT. OWATT mitigates the subjectivity inherent to trimming and truncation methods and has smaller estimated standard errors (higher efficiency) than both the conventional ATT method and the ATT trimming.

Furthermore, by modelling the generalized propensity scores and using weights that standardize group-specific covariates, we compare each minoritized racial or ethnic group to the White group (as reference). The conventional ATT shows that the difference between the White and Hispanic participants in health expenditure is non-significant, while with ATT trimming (except for threshold of 0.15), ATT truncation, and OWATT, we demonstrate noticeable disparities in healthcare expenditure between White participants and the Hispanic population at the 0.05 significance level. For Non-Hispanic Black and Non-Hispanic Asian, most of the methods agree on the existence of healthcare spending disparities. Even though the racial disparity estimation differs by approach and choice of threshold, all the methods addressing the lack of positivity agree that such disparities between White population and the three minorities cannot be ignored. ATT trimming sometimes disagree with the common conclusions reached by other approaches by yielding a large variance estimation [47], which is probably due to loss of information.

Our paper also indicates that while trimming or truncating participants can reduce variance in some scenarios, we risk losing valuable information, as shown by a reduction in effective sample size (ESS) (Table 1). Finally, when extreme propensity scores are present, the ATT trimming yields smaller variance than the conventional ATT, but the OWATT typically outperforms the ATT trimming, with performance similar to ATT truncation, while circumventing the issues of overweighting individual participant's contribution or overall information unnecessarily and subjectively thresholds. We have also seen that, in fact, excessive trimming may even reverse the conclusions reached by other methods. By preserving more data, OWATT offers a greater accuracy, making it a more promising choice for researchers and clinicians seeking minimal subjectivity and maximum data retention.

Discussion

In addition to our findings, we acknowledge some limitations related to our method and propose potential extensions for future research.

First, similar to ATT trimming and truncation, OWATT shifts the goalpost (i.e., the underlying population of interest) to the overlap population [26]. While the conventional ATT estimates the healthcare expenditure disparities of the reference race group by leveraging outcomes from all participants of other groups, the OWATT achieves the same objective. It does not trim or truncate any observations but uses smooth weighting to select suitable comparators in the other groups (see details in Appendix A.2 and Fig. 3 in Appendix 1). However, as opposed to the conventional ATT, the OWATT does not induce any extreme weights when the generalized propensity scores of participants in the non-reference groups are close to 0. As such, the OWATT targets a slightly different subpopulation of participants among those who are not in the reference group (compared to the conventional ATT). Hence, the OWATT offers a better trade-off between efficiency and bias reduction than the ATT trimming and the truncation.

Similar to the OWATT, both the ATT trimming and the ATT truncation also move the goalpost of inference but require an additional tuning parameter: the trimming or truncation threshold. Though data-driven strategies for choosing optimal thresholds were described by Crump et al. [26], unfortunately, they did not clearly demonstrate how a consistent and adequate threshold can be selected and implemented for practical applications. The choice of $\alpha = 0.10$ as the threshold is just a rule-of thumb, which does not always work in practice [39]. Therefore, such a choice for an appropriate threshold can be driven by subjective considerations, which may leave open the possibility for a fishing expedition to reach a desired significant p-value. Like many statistical methodologies, the OWATT is also involves inherent trade-offs. It makes a minor change in the target population and results interpretation (compared to the conventional ATT) but gains significantly in estimation and efficiency without the burden to select arbitrarily tuning parameters. In addition, it includes all data information available and produces reasonable propensity score weights [19, 48, 49].

As a side, since we used overlap weights, we must clarify that the interpretation of these weights in the WATT framework differs from the WATE framework of Li et al. [22]. Specifically, our interpretation cannot be equated to targeting the "clinical equipoise" often referenced by researchers in related fields [50, 51].

Furthermore, the current paper opens several possible extensions. First, we can easily apply the method to the average treatment effect on the control (ATC) when the effect on the controls (i.e., some non-reference groups) is of interest. We just need to pre-specify which group should be the control group and follow in a similar fashion the work we have presented. The racial disparity results can be useful in closing the healthcare gap by designing interventions that mitigate the disparities.

To guide policy decision-making and propose wiser policies that promote racial equality in health care, the results presented here can be framed in the context of specific minoritized racial or ethnic groups. Second, while we have used bootstrap for variance estimation, it can be important to investigate other methods, such as the model-based sandwich variance method or the wild bootstrap [21]. Third, our results are adjusted for health status variables, which may be different when considering SES variables. A future study can investigate racial disparities in healthcare expenditure while considering both health status and SES covariates [8–12]. Furthermore, it is also of interest to consider augmented or doubly robust estimators of the class of weighted average treatment effect on the treatment (WATT), by combining inverse probability weighting and additional outcome modeling to possibly achieve the semiparametric efficiency bound and mitigate the effect of extreme propensity scores [49, 52–55]. To our knowledge, there is no publication for such an augmented estimator for the general WATT. Other weighting strategies may also apply, including calibration weighting for alternative covariate balance estimators [34–37], data-driven-based trimming, or weight modification [24, 25, 39, 40].

In addition, in our analysis, we leverage survey weights to provide a more valid estimation and interpretation for target population compared to previous studies analyzing the same data set. Although we handle the extreme weights derived from generalize propensity scores (implicitly mitigate possible extreme survey weights), we did not develop the methods to handle extreme survey weights specifically. In survey statistics, Gelman's methods [56, 57] are adopted by sample survey researchers when dealing with extreme survey weights, presenting a promising future direction for combining them to our WATT framework. Finally, we can also consider applying the estimator proposed in this paper to survival outcomes [58, 59], multi-source data [60, 61], and conformal inference [62].

Appendix 1

Details for statistical methodology

A.1 Statistical assumptions for consistent ATT estimation

We follow all notations made in Section "Racial Disparity in Healthcare Expenditures". To identify and estimate ATT from observational data, there are two crucial assumptions:

Assumption 1 (Unconfoundness) $Y(i) \perp \mathbb{I}\{Z = i\}|X$, for $\forall i \in \{1,2,\ldots,K\}$;

Assumption 2 (Positivity/Overlap) $P(e_i(X) > 0) = 1$ for $\forall i \in \{1,2,\ldots,K\}\backslash j$,

where the symbol $\perp$ indicates independence of random variables

Assumption 1 implies that all the important confounders (variables associated with both $Y(i)$ and $Z$for $\forall i \in \{1,2,\ldots,K\}$) are available in our data and are included in the covariate vector X. This ensures that the propensity score $e_i(X)$ can be a balancing score, i.e., $Y(i) \perp \mathbb{I}\{Z = i\}|e_i(X)$. While the unconfoundedness assumption is not verifiable (for instance, through a statistical test), it can hopefully be assessed and evaluated via domain knowledge, in practice, by field experts. The use of directed acyclic graphs (DAGs) has played an important role in this regard [63, 64]

Assumption 2 is the positivity assumption we alluded to previously. It means that for all participants in the race groups other than the reference group j, they should have positive generalized propensity scores that are strictly greater than zero. In other words, they must have some similar characteristics to the participants in the reference race group j for allowing causal comparison

Under the above two assumptions, the ATT can be rewritten using the following formula:

$$\tau_{i,j}^{att} = E[Y|Z = j] - E[Y(i)|Z = j]$$

$$= \frac{E\left[Y\mathbb{I}\{Z = j\}\right]}{E\left[\mathbb{I}\{Z = j\}\right]} - \frac{E\left[\mathbb{I}\{Z = j\}E\{Y|Z = i, X\}\right]}{E\left[\mathbb{I}\{Z = j\}\right]}$$

$$= \frac{E\left[Y\mathbb{I}\{Z = j\}\right]}{P(Z = j)} - \frac{E\left[E\left[\mathbb{I}\{Z = j\}|X\right]E\{Y|Z = i, X\}\right]}{P(Z = j)}$$

$$= \frac{E\left[Y\mathbb{I}\{Z = j\}\right]}{P(Z = j)} - \frac{E\left[P(Z = j|X)P(Z = i|X)^{-1}E\{Y\mathbb{I}\{Z = i\}|X\}\right]}{P(Z = j)}$$

$$= \frac{1}{P(Z = j)}E\left[Y\mathbb{I}\{Z = j\} - \frac{P(Z = j|X)}{P(Z = i|X)}Y\mathbb{I}\{Z = i\}\right]$$

$$= \frac{1}{P(Z = j)}E\left[Y\mathbb{I}\{Z = j\} - \frac{e_j(X)}{e_i(X)}Y\mathbb{I}\{Z = i\}\right].$$

It can also be verified that, $P(Z = j)$ can also be written as $E\left[\frac{e_j(X)}{e_i(X)}\mathbb{I}\{Z = i\}\right]$. Therefore, this formula illustrates that why we can use estimator (1) for estimating ATT consistently, and why we can refer the weight $\frac{e_j(X)}{e_i(X)}$ as the balancing weight, since it re-weights the outcome of group i to be alike that of group j.

A.2 Details of methods dealing with lack of positivity in ATT estimation

As mentioned in Section "Effective Sample Size" and Section "Estimated Causal Effects of Racial Disparity", to circumvent the lack of positivity for participants whose generalized propensity scores are extreme, there are three main methodologies: (i) trimming; (ii) truncation; and (ii) overlap weights (OW).

When looking at the reference race group to estimate ATT, trimming is executed only in other race groups, where whose generalized propensity scores fall outside of $[\alpha, 1]$ are to be excluded [14, 38]. As a remark, in the introduction of Section "Effective Sample Size", this range is $[\alpha, 1]$ and the trimming is conducted for participants from all groups. This difference is because the original literature by Crump et al. [26] proposing this trimming method targets the ATE, while in our paper, we target ATT or ATT-like measures. It can be verified that once those who are in other race groups with generalized propensity scores fall outside of $[\alpha, 1]$ are excluded, there are no longer extreme weights in ATT estimation. Thus, trimming requirements for ATT are less strict than those for ATE.

In our study, except for the pre-determined reference group j, participants whose generalized propensity scores fall outside of the range $[\alpha, 1]$ in the comparison group $i (i \neq j)$ are trimmed, where i is in $\{1, \ldots, K\} \backslash j$. Then, the generalized propensity scores for all remaining participants should be re-estimated and updated and by re-performing the multinomial regression from the remaining data [23].

As opposed to trimming, truncation does not exclude participants but requires a threshold to determine which participant's generalized propensity score to amend. A participant in group whose generalized propensity score is less than α will be set to α otherwise, the generalized propensity scores remain unchanged.

Once the generalized propensity scores for all participants are either trimmed or truncated, the estimation of ATT follows the same framework as discussed in Section "Overlap and Covariates Balance", including the derivation of individual weights and ATT estimates. Truncation has been shown to be efficient in handling extreme weights and is considered to have a good bias-variance trade-off [65]. The estimands for trimming and truncation generally deviate from the conventional ATT since the weights are altered and depend on the choice of the threshold. As indicated in Section "Estimated Causal Effects of Racial Disparity", both ATT trimming and ATT truncation fall in the general class of estimands, the WATT.

Moreover, instead of relying on a user-specified trimming or truncation threshold, Liu et al. [25] proposed the overlap weighted average treatment effect on treated (OWATT) by using the OW in the WATT framework. To introduce OWATT, we first review the set-up of a general WATT.

For simplicity, we first consider the general WATT under a binary treatment framework, defined by

$$\tau_{0,1}^{watt} = \mathbb{E}[Y|Z=1] - \frac{\mathbb{E}[\omega_{0h}(X)(1-Z)Y]}{\mathbb{E}[\omega_{0h}(X)(1-Z)]},$$

where $\omega_{0h}(X) = \omega_0(X)h(X)$. The weights $\omega_0(X) = \frac{e(X)}{1-e(X)}$ with $e(X) = P(A=1|X)$ shape the covariates distribution of non-reference race group to be more alike the reference race group by adjusting the contribution of non-reference race group participants. The function $h(X)$ is the "tilting function", which further weights the covariate distribution of the non-reference race group participants whenever needed. For example, $h(X) = \mathbb{I}\{e(X) \leq 1-\alpha\}$ defines the ATT trimming, where we only select participants with propensity score below the $1-\alpha$ threshold in the controls to estimate the treatment effect. This can be useful when some non-reference race group participants violate the positivity assumption, but as discussed in Section "Effective Sample Size", the selection of the α parameter needs careful consideration and can often be ad-hoc and subjective. In addition to trimming, the WATT also includes the (i) conventional ATT, with $h(X) = 1$; (ii) ATT truncation, where $h(X) = \mathbb{I}\{e(X) < 1-\alpha\} + (1-\alpha)\alpha^{-1}\omega_0(X)^{-1}\mathbb{I}\{e(X) \geq 1-\alpha\}$; and (iii) OWATT for which $h(X) = e(X)(1-e(X))$ [25].

The sample weighting estimator for WATT under a binary treatment is:

$$\widehat{\tau}_{0,1}^{watt} = \frac{\sum_{l=1}^{N} Z_l Y_l}{\sum_{l=1}^{N} Z_l} - \frac{\sum_{l=1}^{N} (1-Z_l)\widehat{\omega}_{0h}(X_l) Y_l}{\sum_{l=1}^{N} (1-Z_l)\widehat{\omega}_{0h}(X_l)},$$

where $\widehat{\omega}_{0h}(X_l) = \widehat{\omega}_0(X_l)\widehat{h}(X_l)$, $\widehat{\omega}_0(X_l) = \frac{\widehat{e}(X_l)}{1-\widehat{e}(X_l)}$, and $\widehat{e}(X_l)$ the estimated propensity scores.

What distinguishes OWATT from the other estimands is its distinctive feature when it comes to the violation of the positivity assumption: it smoothly reduces the influence of extreme propensity scores as $e(X)$ moves away from 0.5 and gets closer to 1. This impacts the contributions of the participants' outcomes to the overall treatment effect estimation; participants with extreme propensity scores are allocated smaller weights $\omega_{0h}(X)$. Thus, the contribution of participants from the non-reference race group i $(i \neq j)$ to the weighted mean outcome is adjusted accordingly, which reduces their unduly influence on the overall assessment of the treatment effect.

To better understand the OWATT, we illustrate the trends of different propensity score weights for non-reference race group participants under a binary treatment in Fig. 3 in Appendix 1. In this figure, we compare

the propensity score weights defined by the conventional ATT, ATT trimming (using $\alpha = 0.1$), ATT truncation (using $\alpha = 0.1$), and OWATT. As shown in Fig. 3 in Appendix 1, ATT weights increase to infinity as the propensity score approaches to one, while ATT trimming sets the weights to zero for participants with propensity scores below the trimming threshold ($\alpha = 0.1$) and ATT truncation caps the weights at this threshold. In contrast, only OWATT assigns non-zero, bounded, and different weights to all individual participants, including those with extreme propensity scores. Under sufficient positivity (e.g., when all propensity scores are below 0.75), OWATT produces a weight curve similar to the other methods, maintaining alignment with other approaches. However, in scenarios lacking sufficient positivity, OWATT displays a gently increasing curve, effectively up-weighting non-reference race group participants with higher propensity scores while ensuring weights remain bounded to avoid extreme values. This allows OWATT to remain robust by providing stable weight distributions and avoiding extreme weights. Therefore, OWATT maintains reliable estimates by appropriately weighting non-reference race group participants based on their propensity scores without discarding any observations.

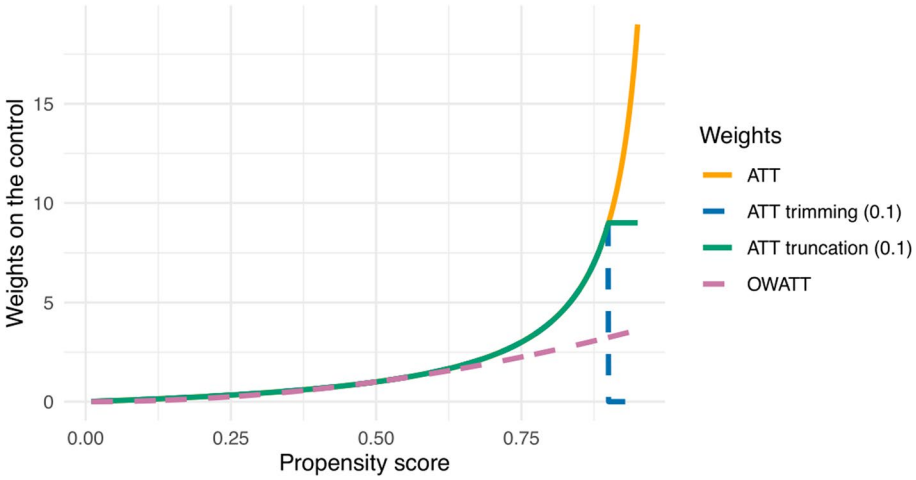

Fig. 3 Propensity score weights on the control participants defined by different estimands

For multiple treatments, where $Z \in \{1,2,\ldots,K\}$ and $K \geq 3$, we choose $j \in \{1,2,\ldots,K\}$ to indicate the treated (or reference) group. The generalized propensity score is given by

$$e_i(X) = Pr(Z = i|X), i \in \{1,2,\ldots,K\},$$

and indicates the probability of being assigned to race group i given the covariates X. The sum of generalized propensity scores across all the treatment groups $\sum_{i=1}^{K} e_i(X)$ is equal to 1.

The generalized WATT estimand for multiple treatment groups is defined by

$$\tau_{i,j}^{watt} = \mathbb{E}[Y|Z = j] - \frac{\mathbb{E}\left[\omega_{i,h}(X)\mathbb{I}\{Z = i\}Y\right]}{\mathbb{E}\left[\omega_{i,h}(X)\mathbb{I}\{Z = i\}\right]},$$

where $\omega_{i,h}(X) = \frac{e_j(X)h(X)}{e_i(X)}$ are the weights that allow the covariate distribution(s) of participants in group i to be similar to those from the reference race group j. Its corresponding estimator is given by

$$\widehat{\tau}_{i,j}^{watt} = \frac{\sum_{l=1}^{N} D_{lj}Y_l}{\sum_{l=1}^{N} D_{lj}} - \frac{\sum_{l=1}^{N} \widehat{e}_j(X_l)\widehat{e}_i(X_l)^{-1}\widehat{h}(X_l)D_{li}Y_l}{\sum_{l=1}^{N} \widehat{e}_j(X_l)\widehat{e}_i(X_l)^{-1}\widehat{h}(X_l)D_{li}},$$

where $\widehat{h}(X_l)$ is the estimated tilting function given the covariates for participant l is $X_l (l = 1,\ldots,N)$, and $D_{li} = \mathbb{I}\{Z_l = i\}$ is the indicator for the group membership.

To estimate the generalized OWATT, we use the tilting function

$$h(X) = \left(\sum_{i=1}^{K} \frac{1}{e_i(X)}\right)^{-1},$$

which is proportional to the harmonic mean of the generalized propensity scores. One can verify that, when

$K = 2$, the tilting function is the usual overlap tilting function $h(X) = e_1(X)e_2(X) = e_1(X)(1 - e_1(X))$.

Appendix 2

Additional statistical analyses

In this section, we additionally provide some statistical analysis as references.

B.1 Diagnostic analysis of multinomial regression for propensity Scores

To validate the multinomial regression model for estimating propensity scores, we provide some assessments of model diagnostic to reflect how the model fit the data.

First, Nagelkerke's R^2 is 0.243 and McFadden's R^2 is 0.104, which imply that the multinomial regression model for the generalized propensity score achieves a moderate and reasonable fit of the data [66–69]. Specifically, the generalized propensity score model captures a significant portion of the variability but also leaves some rooms for further improvement.

Second, we also explore the area under the curve (AUC) for each racial group as shown in Table 3 in Appendix 2. From this, we can see that the model has moderate classification ability to identify White participants, acceptable classification ability to identify Hispanic and Non-Hispanic Black participants and has good ability to discriminate Non-Hispanic Asian. Therefore, in an overall sense, the multinomial propensity score regression model fit the dataset moderately good, still have room to improve though.

Table 3 AUC of multinomial regression model for each racial group

	White	Hispanic	Non-Hispanic Asian	Non-Hispanic Black
AUC	0.70	0.70	0.76	0.70

B.2 Probability Density Plots of WATT Estimates with Different Methods

For the variance estimation in this paper, we used the bootstrap approach by repeatedly estimating ATT under different estimators. The bootstrap method approximates the sampling distribution for each estimator and captures the variability of estimates through repeated estimations. To present further insights into how these estimates be distributed, we provide density plots for the point estimates based on four different estimators, derived from 1000 bootstrap replicates.

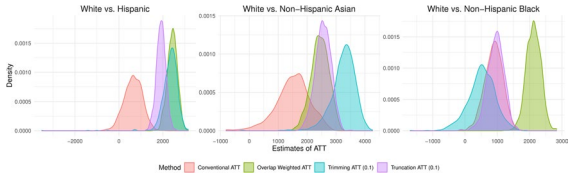

Fig. 4 Density distribution of ATT estimates by comparisons ($\alpha = 0.1$)

For here, we focus on $\alpha = 0.1$, which is one of the typical thresholds for trimming ATT and truncation ATT. As shown in Fig. 4 in Appendix 2, the ATT estimator without any adjustment for extreme weights exhibits the largest variance. When the truncation threshold is 0.1, the variance pattern of OWATT is similar to that of ATT truncation, particularly in the comparisons between White and Hispanic and between White and Non-Hispanic Asian. The OWATT achieves a reduction in variance compared to other estimators in the White vs. Non-Hispanic Black comparison, which is aligned with the results in Table 2.

B.3 G-computation approach for ATT estimation

Additionally, we apply G-computation approach to estimate the racial disparity in healthcare expenditure among multiple comparisons [70]. First, we fit the outcome model by letting healthcare expenditure as the outcome, including all the covariates that used in the multinomial regression for previous generalized PS estimation and race indicators as predictors. The survey weights are defined in the weight argument to account for the possible imbalance in the survey stage. Afterwards, we predict the potential outcomes or counterfactual health expenditures for the treated group, i.e., White, under all treatment levels. For example, for White vs. Hispanic, we take the difference of the protentional health expenditures for the White group and the protentional health expenditures for the White group if they are Hispanic. Since G-computation approach does not involve propensity score, we disregard the violation of positivity assumption here. The results of ATT estimation with G-computation approach are provided in Table 4 in Appendix 2.

Table 4 ATT estimation with G-computation approach

Comparison	Point Estimate	Standard Error	*P*-value
White vs. Hispanic	1394.81	168.63	< 0.001
White vs. Non-Hispanic Asian	1368.29	226.55	< 0.001
White vs. Non-Hispanic Black	889.74	226.72	< 0.001

When use G-computation approach, all the comparisons indicate noticeable racial disparities in healthcare expenditure between all the three other racial groups and White group. The conclusion for G-computation is aligned with the conclusions when using PS-based approaches that discussed in this paper. G-computation, consistent with OWATT, ATT trimming, and ATT truncation methods, presents a more substantial disparity between White and Hispanic groups as well as between White and Non-Hispanic Asian groups. The estimates of ATT between White and Non-Hispanic Asian and between White vs. Non-Hispanic Black are close to the estimates of the conventional ATT without handling extreme weights. In general, for each comparison, G-computation leads to smaller estimates of racial disparities in healthcare spending compared to the three methods mainly discussed in this paper. This approach could also be an alternative as ATT estimation, which is not based on generalized propensity scores but rely on the outcome model specification.

Acknowledgements
Not applicable.

Authors' contributions
All authors contributed to the study's conception and design. J.L. developed the methodology, conducted the case study, summarized the results, prepared all tables and figures, and wrote the majority of the manuscript. Y.L. contributed to the methodology and wrote parts of the introduction and discussion. Y.Z. prepared the data and wrote parts of the introduction and discussion. R.M. wrote the data background and parts of the introduction and discussion. All authors reviewed, provided feedback, revised, and approved the final manuscript.

Funding
Yi Liu is supported by the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (NIH) under Award Number T32HL079896. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data availability
The data analyzed during the current study are available on the website of the Medical Expenditure Panel Survey (MEPS) [https://www.meps.ahrq.gov/mepsweb].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA. [2]Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA. [3]Duke Clinical Research Institute, Durham, NC 27701, USA. [4]Department of Biostatistics, University of Washington, Seattle, WA 98105, USA.

Received: 6 October 2024 Accepted: 13 February 2025
Published online: 07 March 2025

References

1. Macias-Konstantopoulos WL, Collins KA, Diaz R, Duber HC, Edwards CD, Hsu AP, et al. Race, Healthcare, and Health Disparities: A Critical Review and Recommendations for Advancing Health Equity. West J Emerg Med. 2023;24:906–18.
2. 2021 National Healthcare Quality and Disparities Report. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021.
3. Ma A, Sanchez A, Ma M. The Impact of Patient-Provider Race/Ethnicity Concordance on Provider Visits: Updated Evidence from the Medical Expenditure Panel Survey. J Racial Ethn Health Disparities. 2019;6:1011–20.
4. Bleich SN, Simon AE, Cooper LA. Impact of Patient-Doctor Race Concordance on Rates of Weight-Related Counseling in Visits by Black and White Obese Individuals. Obesity. 2012;20:562–70.
5. Traylor AH, Schmittdiel JA, Uratsu CS, Mangione CM, Subramanian U. Adherence to Cardiovascular Disease Medications: Does Patient-Provider Race/Ethnicity and Language Concordance Matter? J Gen Intern Med. 2010;25:1172–7.
6. Traylor AH, Subramanian U, Uratsu CS, Mangione CM, Selby JV, Schmittdiel JA. Patient Race/Ethnicity and Patient-Physician Race/Ethnicity Concordance in the Management of Cardiovascular Disease Risk Factors for Patients With Diabetes. Diabetes Care. 2010;33:520–5.
7. Li F, Li F. Using propensity scores for racial disparities analysis. Obs Stud. 2023;9:59–68.
8. Choi BY. Propensity score analysis for health care disparities: a deweighting approach. BMC Med Res Methodol. 2024;24:106.
9. Cook BL, McGuire TG, Meara E, Zaslavsky AM. Adjusting for Health Status in Non-Linear Models of Health Care Disparities. Health Serv Outcomes Res Methodol. 2009;9:1–21.
10. Cook BL, McGuire TG, Lock K, Zaslavsky AM. Comparing Methods of Racial and Ethnic Disparities Measurement across Different Settings of Mental Health Care. Health Serv Res. 2010;45:825–47.
11. Cook BL, McGuire TG, Zaslavsky AM. Measuring Racial/Ethnic Disparities in Health Care: Methods and Practical Issues. Health Serv Res. 2012;47:1232–54.
12. Zaslavsky AM, Ayanian JZ. Integrating Research on Racial and Ethnic Disparities in Health Care Over Place and Time. Med Care. 2005;43:303–7.
13. Imbens G. The role of the propensity score in estimating dose-response functions. Biometrika. 2000;87:706–10.
14. Li F, Li F. Propensity score weighting for causal inference with multiple treatments. Ann Appl Stat. 2019;13:2389–415.
15. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.
16. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
17. Imbens GW. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. Rev Econ Stat. 2004;86:4–29.
18. D'Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. J Econom. 2021;221:644–54.
19. Matsouaka RA, Zhou Y. Causal inference in the absence of positivity: The role of overlap weights. Biom J. 2024;66:2300156.
20. Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. Stat Methods Med Res. 2020;29:3721–56.
21. Matsouaka RA, Liu Y, Zhou Y. Overlap, matching, or entropy weights: what are we weighting for? Commun Stat - Simul Comput. 2024;0:1–20.
22. Li F, Morgan KL, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting. J Am Stat Assoc. 2018;113:390–400.
23. Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights. Am J Epidemiol. 2018. https://doi.org/10.1093/aje/kwy201.
24. Ma X, Wang J. Robust Inference Using Inverse Probability Weighting. J Am Stat Assoc. 2020;115:1851–60.
25. Liu Y, Li H, Zhou Y, Matsouaka RA. Average treatment effect on the treated, under lack of positivity. Stat Methods Med Res. 2024;33:1689–717.
26. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96:187–99.
27. Gruber S, Phillips RV, Lee H, van der Laan MJ. Data-Adaptive Selection of the Propensity Score Truncation Level for Inverse-Probability–Weighted and Targeted Maximum Likelihood Estimators of Marginal Point Treatment Effects. Am J Epidemiol. 2022;191:1640–51.

28. Ju C, Schwab J, van der Laan MJ. On adaptive propensity score truncation in causal inference. Stat Methods Med Res. 2019;28:1741–60.
29. Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. PLoS ONE. 2011;6:e18174.
30. Sullivan PW, Ghushchyan VH, Slejko JF, Belozeroff V, Globe DR, Lin S-L. The burden of adult asthma in the United States: Evidence from the Medical Expenditure Panel Survey. J Allergy Clin Immunol. 2011;127:363-369.e3.
31. Rubin DB. The design *versus* the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007;26:20–36.
32. Rubin DB. For objective causal inference, design trumps analysis. Ann Appl Stat. 2008;2:808–40.
33. Laan MJ van der, Polley EC, Hubbard AE. Super Learner. Stat Appl Genet Mol Biol. 2007;6:Article25.
34. Sen M, Wasow O. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. Annu Rev Polit Sci. 2016;19:499–522.
35. VanderWeele TJ, Hernán MA. Causal Effects and Natural Laws: Towards a Conceptualization of Causal Counterfactuals for Nonmanipulable Exposures, with Application to the Effects of Race and Sex. In: Berzuini C, Dawid P, Bernardinelli L, editors. Wiley Series in Probability and Statistics. 1st ed. Wiley; 2012. p. 101–13.
36. Splawa-Neyman J, Dabrowska DM, Speed TP. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Stat Sci. 1990;5.
37. Imbens GW, Rubin DB. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. 1st edition. Cambridge University Press; 2015.
38. Yang S, Imbens GW, Cui Z, Faries DE, Kadziola Z. Propensity Score Matching and Subclassification in Observational Studies with Multi-Level Treatments. Biometrics. 2016;72:1055–65.
39. Yang S, Ding P. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. Biometrika. 2018;105:487–93.
40. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall/CRC; 1994.
41. IOM. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care (with CD). Washington, D.C.: National Academies Press; 2003.
42. McGuire TG, Alegria M, Cook BL, Wells KB, Zaslavsky AM. Implementing the Institute of Medicine Definition of Disparities: An Application to Mental Health Care. Health Serv Res. 2006;41:1979–2005.
43. Ridgeway G, Kovalchik SA, Griffin BA, Kabeto MU. Propensity Score Analysis with Survey Weighted Data. J Causal Inference. 2015;3:237–49.
44. DuGoff EH, Schuler M, Stuart EA. Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys. Health Serv Res. 2014;49:284–303.
45. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28:3083–107.
46. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34:3661–79.
47. Sasaki Y, Ura T. Estimation and Inference for Moments of Ratios with Robustness Against Large Trimming Bias. Econom Theory. 2022;38:66–112.
48. Li L, Greene T. A Weighting Analogue to Pair Matching in Propensity Score Analysis. Int J Biostat. 2013;9:215–34.
49. Mao H, Li L, Greene T. Propensity score weighting analysis and treatment effect discovery. Stat Methods Med Res. 2019;28:2439–54.
50. Thomas LE, Li F, Pencina MJ. Overlap Weighting: A Propensity Score Method That Mimics Attributes of a Randomized Clinical Trial. JAMA. 2020;323:2417–8.
51. Parikh H, Ross R, Stuart E, Rudolph K. Who Are We Missing? A Principled Approach to Characterizing the Underrepresented Population. 2024.
52. Hirano K, Imbens GW, Ridder G. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica. 2003;71:1161–89.
53. Kurz CF. Augmented Inverse Probability Weighting and the Double Robustness Property. Med Decis Making. 2022;42:156–67.
54. Theory S, Data M. New York. NY: Springer; 2006.
55. Hahn J. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. Econometrica. 1998;66:315–31.
56. Gelman A. Struggles with Survey Weighting and Regression Modeling. Stat Sci. 2007;22:153–64.
57. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press; 2007.
58. Lee D, Yang S, Wang X. Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. J Causal Inference. 2022;10:415–40.
59. Transporting survival of an HIV clinical trial to the external target populations: Journal of Biopharmaceutical Statistics: Vol 0, No 0 - Get Access. https://www.tandfonline.com/doi/full/https://doi.org/10.1080/10543406.2024.2330216. Accessed 5 Sep 2024.
60. Han L, Hou J, Cho K, Duan R, Cai T. Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects. 2023.
61. Yang S, Gao C, Zeng D, Wang X. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. J R Stat Soc Ser B Stat Methodol. 2023;85:575–96.
62. Liu Y, Levis A, Normand S-L, Han L. Multi-Source Conformal Inference Under Distribution Shift. In: Proceedings of the 41st International Conference on Machine Learning. PMLR; 2024. p. 31344–82.
63. Lipsky AM, Greenland S. Causal Directed Acyclic Graphs. JAMA. 2022;327:1083.
64. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol. 2022;142:264–7.
65. Cole SR, Hernan MA. Constructing Inverse Probability Weights for Marginal Structural Models. Am J Epidemiol. 2008;168:656–64.
66. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78:691–2.
67. Menard S. Coefficients of Determination for Multiple Logistic Regression Analysis. Am Stat. 2000;54:17–24.
68. Louviere JJ, Hensher DA, Swait JD, Adamowicz W. Stated Choice Methods: Analysis and Applications. 1st edition. Cambridge University Press; 2000.
69. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Frontiers in Econo-metrics. Academic Press; 1973. p. 105–42.
70. Wang A, Nianogo RA, Arah OA. G-computation of average treatment effects on the treated and the untreated. BMC Med Res Methodol. 2017;17:3.

Publisher's Note