American Journal of Epidemiology
© The Author(s) 2018. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 188, No. 1
DOI: 10.1093/aje/kwy201
Advance Access publication:
September 5, 2018

## Practice of Epidemiology

# Addressing Extreme Propensity Scores via the Overlap Weights

**Fan Li*, Laine E. Thomas, and Fan Li****

* Correspondence to Fan (Frank) Li*, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, 2424 Erwin Road, Suite 1105, Durham, NC 27705 (e-mail: frank.li@duke.edu).

*Initially submitted June 1, 2018; accepted for publication August 28, 2018.*

The popular inverse probability weighting method in causal inference is often hampered by extreme propensity scores, resulting in biased estimates and excessive variance. A common remedy is to trim patients with extreme scores (i.e., remove them from the weighted analysis). However, such methods are often sensitive to the choice of cutoff points and discard a large proportion of the sample. The implications for bias and the precision of the treatment effect estimate are unclear. These problems are mitigated by a newly developed method, the overlap weighting method. Overlap weights emphasize the target population with the most overlap in observed characteristics between treatments, by continuously down-weighting the units in the tails of the propensity score distribution. Here we use simulations to compare overlap weights to standard inverse probability weighting with trimming, in terms of bias, variance, and 95% confidence interval coverage. A range of propensity score distributions are considered, including settings with substantial nonoverlap and extreme values. To facilitate practical implementation, we further provide a consistent estimator for the standard error of the treatment effect estimated using overlap weighting.

causal inference; clinical equipoise; epidemiologic methods; inverse probability weighting; overlap weighting; statistical efficiency; trimming

Abbreviation: IPW, inverse probability weighting; OW, overlap weighting; PS, propensity score.

Sources of data for observational treatment comparisons are expanding to include billing claims, large registries, and electronic health records. These resources have the potential to answer questions about the safety and effectiveness of treatments, although statistical methods must account for the lack of randomization. Inverse probability weighting (IPW) is a popular approach used to adjust for confounding due to differences between comparator groups that arise in observational data (1–4). Using a propensity score (PS) that summarizes differences in measured patient characteristics, IPW creates a weighted pseudopopulation in which both treatment groups resemble the total sample combined across treatment groups. Assuming that all of the important confounders are measured, this approach is appealing for its simplicity and alignment with a potential experiment: What if the entire sample had instead been randomized to the intervention of interest? In practice, IPW may perform poorly when the treatment groups are initially very different and when some patients have extreme propensity scores near 1 or 0—that is, almost always receive treatment and never receive treatment, respectively (5–7). Extreme propensities are particularly common in the setting of "big data," where inclusion criteria can be defined broadly. The increasing prevalence of large data sources precipitates the need to clarify best practice for handling extreme propensity scores.

When propensity scores approach 0 or 1, IPW has limitations that include: 1) large weights for individual patients, 2) bias, and 3) large variability in the estimated treatment effect (5–7). To address these problems, trimming methods have been proposed that exclude patients who have very high predicted probabilities of being in the treatment group (or in the control group) (8–10). Despite the potential gains from trimming, the decision regarding how many patients to exclude is ad hoc and can result in substantial loss of sample size. The relative performance of different approaches to trimming remains unclear, particularly in the setting of extreme propensity scores.

These problems are mitigated by a newly developed method, the overlap weighting (OW) method, in which each patient's weight is the probability of that patient being assigned to the opposite group (11). The properties of overlap weights have been demonstrated theoretically and include improvements in balance and precision relative to IPW. In addition, these weights are bounded and smoothly reduce the influence of patients at the tails of the PS distribution without making any exclusions.

Despite the theoretical advantages of OW, little is known about the relative performance of OW and IPW with trimming. In this article, we conduct a simulation study to compare OW to standard IPW with trimming, in terms of bias, variance, and 95% confidence interval coverage. To facilitate practical implementation, we further provide a consistent estimator for the standard error of the treatment effect estimated using the overlap weights.

## NOTATION AND METHODS

We denote the 2 potential outcomes for each unit by $Y_i(1)$ and $Y_i(0)$, corresponding to the outcome that would be ascertained under treatment ($Z_i = 1$) and control ($Z_i = 0$) status. Throughout, we maintain the standard assumptions in PS analysis: 1) consistency, 2) positivity, 3) conditional exchangeability, and 4) no interference (12). Consistency requires that the observed outcome, $Y_i$, for each unit is one of the 2 potential outcomes. Positivity requires that each unit has a nonzero probability of being assigned to either treatment. Conditional exchangeability requires that all confounders are observed and that randomization holds given the same values of all confounders.

### PS weighting

We assume that $\boldsymbol{X}_i$ is a vector of observed pretreatment covariates, which are sufficient to control for all sources of confounding. The PS, $e(\boldsymbol{X}_i) = P(Z_i = 1|\boldsymbol{X}_i)$, is the conditional probability of receiving treatment (13). This 1-dimensional summary score has a balancing property—that is, conditional on the value of the PS, the distribution of the multidimensional $\boldsymbol{X}$ is approximately the same between the treatment groups. Moreover, if the treatment assignment is unconfounded given $\boldsymbol{X}$, then it is also unconfounded given $e(\boldsymbol{X})$. In observational studies, the PS is unknown and must be estimated, most commonly through a logistic model: $e(\boldsymbol{X}_i; \boldsymbol{\beta}) = 1/(1 + \exp(-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}))$. The estimated PS, $\hat{e}_i = e(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}})$, is obtained by plugging in the estimated coefficients $\hat{\boldsymbol{\beta}}$. For notational simplicity, we assume that the first element of $\boldsymbol{X}_i$ is 1 so that the logistic model includes an intercept.

Once the propensity scores have been estimated, there are several ways of using them to estimate the treatment effects, including weighting, matching, and stratification. In this paper, we focus on weighting, where each unit is weighted by $w_i$—a function of its PS—to create a pseudopopulation where the covariate distributions of the treatment and control groups are balanced. There is a general class of such weights, called balancing weights (11), each of which corresponds to a specific target population and causal estimand—a weighted average treatment effect, denoted by $\Delta_w$ (see Web Appendix 1, available at https://academic.oup.com/aje, for its definition). An unbiased estimator of $\Delta_w$ is the weighted difference of the outcome between the groups:

$$\hat{\Delta}_w = \frac{\sum_{i=1}^{n} Z_i Y_i w_i}{\sum_{i=1}^{n} Z_i w_i} - \frac{\sum_{i=1}^{n} (1 - Z_i) Y_i w_i}{\sum_{i=1}^{n} (1 - Z_i) w_i}. \quad (1)$$

In practice, the weight $w_i$ is calculated on the basis of its estimated PS $\hat{e}_i$. Below we describe several balancing weights.

### Inverse probability weighting

IPW, also known as inverse probability of treatment weighting, is the most widely used balancing weighting scheme. IPW is defined as $w_i = 1/\hat{e}_i$ for treated units and $w_i = 1/(1 - \hat{e}_i)$ for control units. IPW assigns to each patient a weight proportional to the reciprocal of the probability of being assigned to the observed treatment group. The target population of IPW is the entire study cohort, and the causal estimand $\Delta_{\mathrm{IPW}}$ is the average treatment effect of the total sample combined over treatment.

### IPW with trimming

We examine 2 popular PS trimming methods. First, symmetric trimming excludes patients whose estimated PS is outside of the range $[\alpha, 1 - \alpha]$, where $\alpha$ is a threshold (9). When $\alpha = 0$, there is no trimming, and the analysis is equivalent to IPW. Crump et al. (9) suggested a rule of thumb of $\alpha = 0.1$. Second, we consider a version of asymmetric trimming proposed by Stürmer et al. (8). It involves multiple steps. First, we exclude patients with propensity scores outside of the common PS range formed by the treated and control patients. Then, among the treated units, we further exclude those patients whose PS is below the $q$ quantile of the treated units, and separately among control units, we exclude those whose PS is above the $(1 - q)$ quantile of the control units. When $q = 0$, only the first step is involved and the analysis is restricted to a common PS range. This form of asymmetric trimming was first proposed to handle unmeasured confounding, in the unique scenario where such confounding is exclusive to people with extreme propensity scores. However, asymmetric trimming has since been used as a generic trimming method, and our goal is to evaluate it for this purpose (10, 14). While ad hoc adaptations of trimming methods also exist, we focus on the published methods and discuss the implications subsequently.

### Overlap weighting

OW is a recently developed balancing weighting scheme that addresses some of the issues of IPW and trimming. The overlap weight is defined as $w_i = 1 - \hat{e}_i$ for a treated unit and $w_i = \hat{e}_i$ for a control unit. By construction, OW up-weights patients who have a substantial probability of receiving either treatment and smoothly down-weights the patients in the tails of the PS distribution (Web Appendix 1). Specifically, patients with propensity scores of 0.5 make the largest contribution to the effect estimate and patients with propensity scores close to 0 and 1 make the smallest contribution. The target population of OW emphasizes patients with the most overlap in their observed characteristic, and its corresponding estimand is the average treatment effect in the overlap population. Such an estimand is often of natural relevance to scientific investigation because it emphasizes the portion of the population where the most treatment equipoise exists in clinical practice.

OW has properties that are likely to be beneficial in the presence of extreme tails. By definition, the overlap weights are bounded between 0 and 1, and thus automatically overcome the large uncertainty issue caused by extreme propensity scores when using IPW. Indeed, Li et al. (11) theoretically proved that among all balancing weights, OW minimizes the large-sample variance of the estimator (1) of the treatment effect. Further, OW based on the PS estimated from a logistic model leads to exact balance between treatment groups for all covariates—including the original covariates and any derived covariates such as high-order terms and interactions—entering that logistic model (Web Appendix 2) (11).

In principle, one could use the bootstrap to obtain the variance of $\hat{\Delta}_{\mathrm{OW}}$ as in the article by Li et al. (11). To reduce the computational demand associated with bootstrapping large data sets, in this paper we provide a closed-form variance estimator using the empirical sandwich method (see Web Appendix 3 for a derivation) (15). Specifically, with the logistic PS model $e(\boldsymbol{X}_i) = 1/(1 + \exp(-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}))$, we show that a consistent estimator of the variance of $\hat{\Delta}_{\mathrm{OW}}$ is $(n\hat{\theta})^{-2}\sum_{i=1}^{n}\hat{I}_i^2$, where $\hat{\theta} = n^{-1}\sum_{i=1}^{n}\hat{e}_i(1-\hat{e}_i)$:

$$\hat{I}_i = Z_i(Y_i - \hat{\tau}_1)(1 - \hat{e}_i) - (1 - Z_i)(Y_i - \hat{\tau}_0)\hat{e}_i - (Z_i - \hat{e}_i)\hat{\boldsymbol{H}}_{\boldsymbol{\beta}}^{T}\hat{\boldsymbol{E}}_{\boldsymbol{\beta\beta}}^{-1}\boldsymbol{X}_i. \quad (2)$$

$$\hat{\boldsymbol{H}}_{\boldsymbol{\beta}} = n^{-1}\sum_{i=1}^{n}[Z_i(Y_i - \hat{\tau}_1) + (1 - Z_i)(Y_i - \hat{\tau}_0)] \times \hat{e}_i(1 - \hat{e}_i)\boldsymbol{X}_i. \quad (3)$$

$$\hat{\boldsymbol{E}}_{\boldsymbol{\beta\beta}} = n^{-1}\sum_{i=1}^{n}\hat{e}_i(1 - \hat{e}_i)\boldsymbol{X}_i\boldsymbol{X}_i^{T}. \quad (4)$$

Our empirical sandwich variance estimator is based on the theory of M-estimation, and it will adequately quantify the true variability even if the logistic PS model is misspecified. However, misspecification is likely to result in biased point estimates for the average treatment effect in the overlap population and compromise the validity of causal inference. We will examine the performance of the proposed variance formula in the subsequent simulation studies assuming that the logistic PS model is correctly specified.

## SIMULATION DESIGN

We carry out a series of simulation studies to compare IPW, PS trimming, and OW, where the prevalence of extreme weights is varied. For each data-generating process, we first generate 6 variables $V_1$–$V_6$ from a multivariate normal distribution with zero mean and unit marginal variance. We assume that the correlation between each pair of covariates is 0.5 and so the covariance structure for $V_1$–$V_6$ is compound symmetric. We retain $V_1$–$V_3$ as the continuous covariates, denoted by $X_1$–$X_3$, and dichotomize $V_4$–$V_6$ at zero to create the binary covariates $X_4$–$X_6$ (i.e., $V = I_{\{X<0\}}$), so that the marginal prevalence of each binary covariate is approximately 0.5. We then calculate the true PS using a logistic model,

$$e(\boldsymbol{X}) = \{1 + \exp[-(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6)]\}^{-1},$$

and simulate the observed treatment independently from a Bernoulli model. The continuous outcome variable $Y$ is assumed to satisfy

$$E(Y|Z, \boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \Delta Z.$$

This outcome model assumes a homogenous and additive treatment effect $\Delta$ for all units, and thus the magnitudes of the true causal estimands corresponding to all weighting schemes are identical: $\Delta_{\mathrm{IPW}} = \Delta_{\mathrm{OW}} = \Delta$.

We consider a range of scenarios with increasingly strong tails in the PS distributions. We choose the parameters in the PS model $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.15\gamma, 0.3\gamma, 0.3\gamma, -0.2\gamma, -0.25\gamma, -0.25\gamma)$ and vary $\gamma$ from 1 to 4 to reflect increasingly sparse overlap between groups. The intercept $\alpha_0$ in the PS model is chosen so that the overall treatment prevalences (proportions of receipt of treatment) are approximately 0.4 and 0.1. Figure 1 and Web Figure 1 summarize the PS distributions for each PS model indexed by $\gamma$, at each level of treatment prevalence. When $\gamma = 1$, there is substantial overlap between treatment groups and no tails; as the value of $\gamma$ increases, the PS distributions exhibit increasingly strong tails with less overlap. We also report the proportion of propensity scores outside of the range [0.1, 0.9] for each PS model in Table 1. For the outcome model, we fix the intercept $\beta_0 = 0$ and choose $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) = (-0.5, -0.5, -1.5, 0.8, 0.8, 1.0)$. We generate the observed outcome $Y$ from a normal distribution with mean $E(Y|Z, \boldsymbol{X})$ and standard deviation 1.5. We fix the treatment effect as $\Delta = 0.75$ so that higher values of $Y$ reflect the beneficial effect of the treatment. Overall, the simulation parameters are selected so that persons with a covariate profile indicating worse outcomes are those who are more likely to be treated. For each scenario, we simulate 1,000 data sets with 2 sample sizes, $n = 500$ and $n = 2{,}000$. For each sample size, our simulation design corresponds to a 2 × 2 factorial design with 2 levels of overall treatment prevalence and 4 levels of PS tails indexed by $\gamma$.

For each simulation scenario, the treatment effect is estimated using IPW, OW, and 2 IPW trimming approaches: For symmetric trimming, $\alpha = 0.05$, 0.10, or 0.15, and for asymmetric trimming, $q = 0$, 0.01, or 0.05. These values correspond to those in the prior publications (8, 9). To quantify the degree of confounding bias, we also include the crude estimator that is the mean difference of the outcomes between groups, which is expected to be biased. We obtain the variance of the IPW and OW estimates using the formula of Lunceford and Davidian (16) and the proposed formulas shown in equations 2–4, respectively. Whenever trimming is performed, we reestimate the PS on the basis of the trimmed sample. We calculate the bias, relative efficiency, and 95% confidence interval coverage for each estimator.

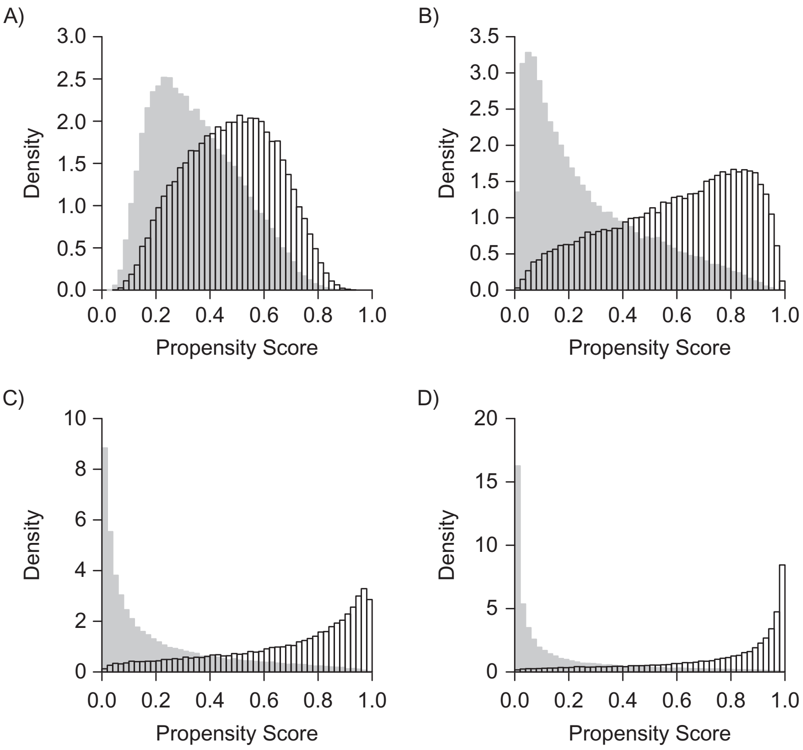

**Figure 1.** Distributions of propensity scores associated with 4 different data-generating processes with an overall treatment prevalence equal to 0.4. The unshaded bars indicate the treated group; the gray shaded bars indicate the control group. Parts A–D show the distributions of propensity scores when $\gamma = 1$, $\gamma = 2$, $\gamma = 3$, and $\gamma = 4$, respectively.

## RESULTS

### Numerical comparisons

Table 2 presents the bias for each estimator in the presence of increasingly strong tails in the PS distributions with $n = 2{,}000$. As $\gamma$ increases, the degree of confounding bias also increases, as quantified by the bias of the crude estimator. The standard IPW estimator is unbiased in the absence of tails ($\gamma = 1$) but quickly exhibits bias with an increasing level of tails. Symmetric trimming excludes the units with extreme propensity scores and removes the bias from IPW for all 3 choices of $\alpha$. By contrast, asymmetric trimming provides biased estimates; the bias becomes more pronounced with increasing levels of tail $\gamma$ and threshold $q$.

**Table 1.** Proportion of Propensity Scores Outside of the Range [0.1, 0.9] According to Different Propensity Score Models Used in Simulation Models[a]

| Treatment Prevalence | Data-Generating Model | | | |
|---|---|---|---|---|
| | $\gamma = 1$ | $\gamma = 2$ | $\gamma = 3$ | $\gamma = 4$ |
| 0.4 | 1 | 21 | 41 | 54 |
| 0.1 | 61 | 69 | 74 | 78 |

[a] Values were evaluated using a large sample with 100,000 observations.

Table 3 shows the relative efficiency with $n = 2{,}000$. We define the relative efficiency as the ratio between the Monte Carlo variance for the crude estimator and that for the weighted estimator under consideration; a more efficient estimator corresponds to a larger value of relative efficiency. Consistently across all of the simulation scenarios in Table 3, OW is much more efficient than IPW, especially with strong PS tails. Symmetric trimming substantially improves the efficiency of IPW but remains less efficient than OW. Finally, the improvement in efficiency due to asymmetric trimming is much less substantial than that due to symmetric trimming.

Table 4 presents the 95% normality-based 95% confidence interval coverage with $n = 2{,}000$. In general, the results of coverage align with those for bias in that biased estimates tend to be associated with poor coverage. Notably, the confidence interval based on the proposed variance estimator for OW provides nominal coverage, for both levels of treatment prevalence and across all data-generating processes with different tail properties. Finally, the result patterns for bias, efficiency, and coverage are similar for a smaller sample of 500 and are presented in Web Tables 1–3.

To demonstrate the exact balance of OW, we show the absolute standardized difference in Figure 2 and Web Figure 2. The absolute standardized difference, defined in Web Appendix 2, measures covariate balance in the weighted target population. OW consistently demonstrates its superb balance property, as the absolute standardized difference is identically zero regardless of PS tails, while IPW is subject to large imbalance under strong tails. Although symmetric trimming improves

**Table 2.** Bias of the Estimators in the Presence of Increasingly Strong Tails in the Propensity Score Distributions ($n$ = 2,000)

| Estimator | Data-Generating Model | | | |
|---|---|---|---|---|
| | γ = 1 | γ = 2 | γ = 3 | γ = 4 |
| *Treatment Prevalence = 0.4* | | | | |
| Crude | −2.01 | −3.18 | −3.76 | −4.06 |
| Overlap weighting | 0.00 | 0.00 | 0.00 | 0.00 |
| IPW | | | | |
| No trimming | 0.00 | 0.00 | −0.06 | −0.22 |
| Symmetric trimming | | | | |
| α = 0.05 | 0.00 | 0.00 | −0.01 | −0.01 |
| α = 0.1 | 0.00 | 0.00 | 0.00 | 0.00 |
| α = 0.15 | 0.00 | 0.00 | 0.00 | 0.00 |
| Asymmetric trimming | | | | |
| $q$ = 0 | 0.03 | 0.15 | 0.37 | 0.58 |
| $q$ = 0.01 | −0.12 | −0.43 | −0.75 | −0.93 |
| $q$ = 0.05 | −0.46 | −1.20 | −1.75 | −1.98 |
| *Treatment Prevalence = 0.1* | | | | |
| Crude | −2.08 | −3.46 | −4.19 | −4.58 |
| Overlap weighting | 0.00 | 0.00 | 0.00 | 0.00 |
| IPW | | | | |
| No trimming | 0.00 | −0.06 | −0.48 | −1.17 |
| Symmetric trimming | | | | |
| α = 0.05 | 0.00 | 0.00 | 0.00 | 0.00 |
| α = 0.1 | 0.00 | 0.00 | 0.00 | −0.01 |
| α = 0.15 | 0.00 | 0.00 | 0.00 | 0.00 |
| Asymmetric trimming | | | | |
| $q$ = 0 | 0.13 | 0.42 | 0.63 | 0.60 |
| $q$ = 0.01 | −0.01 | −0.19 | −0.35 | −0.44 |
| $q$ = 0.05 | −0.27 | −0.86 | −1.21 | −1.31 |

Abbreviation: IPW, inverse probability weighting.

**Table 3.** Relative Efficiency of the Estimators Relative to the Crude Estimator in the Presence of Increasingly Strong Tails in the Propensity Score Distributions ($n$ = 2,000)

| Estimator | Data-Generating Model | | | |
|---|---|---|---|---|
| | γ = 1 | γ = 2 | γ = 3 | γ = 4 |
| *Treatment Prevalence = 0.4* | | | | |
| Crude | 1.00 | 1.00 | 1.00 | 1.00 |
| Overlap weighting | 3.63 | 2.48 | 1.65 | 1.16 |
| IPW | | | | |
| No trimming | 2.35 | 0.31 | 0.05 | 0.02 |
| Symmetric trimming | | | | |
| α = 0.05 | 2.42 | 1.24 | 0.88 | 0.68 |
| α = 0.1 | 2.77 | 1.80 | 1.21 | 0.85 |
| α = 0.15 | 3.09 | 1.95 | 1.17 | 0.81 |
| Asymmetric trimming | | | | |
| $q$ = 0 | 2.68 | 0.49 | 0.08 | 0.03 |
| $q$ = 0.01 | 2.44 | 0.78 | 0.30 | 0.16 |
| $q$ = 0.05 | 1.66 | 0.61 | 0.25 | 0.13 |
| *Treatment Prevalence = 0.1* | | | | |
| Crude | 1.00 | 1.00 | 1.00 | 1.00 |
| Overlap weighting | 3.84 | 2.38 | 1.68 | 1.18 |
| IPW | | | | |
| No trimming | 0.68 | 0.07 | 0.02 | 0.02 |
| Symmetric trimming | | | | |
| α = 0.05 | 2.14 | 1.34 | 0.95 | 0.69 |
| α = 0.1 | 2.06 | 1.39 | 1.10 | 0.83 |
| α = 0.15 | 1.33 | 1.17 | 1.04 | 0.77 |
| Asymmetric trimming | | | | |
| $q$ = 0 | 0.94 | 0.14 | 0.04 | 0.04 |
| $q$ = 0.01 | 0.82 | 0.20 | 0.11 | 0.08 |
| $q$ = 0.05 | 0.61 | 0.21 | 0.12 | 0.08 |

Abbreviation: IPW, inverse probability weighting.

covariate balance, asymmetric trimming exacerbates imbalance with a larger value of $q$, because it actually causes lack of overlap (see Web Appendix 4 for further explanation).

## Conclusions

Standard IPW is known to exhibit bias and excessive variance in the presence of PS tails. We confirm such limitations in our simulations with varying levels of PS tails. Symmetric trimming improves the standard IPW regarding bias, variance, and 95% confidence interval coverage. However, the trimming decisions may be arbitrary, and the (often unknown) optimal cutoff point depends on the tail property as well as the treatment prevalence. For instance, when the overall treatment prevalence is 0.4, symmetric trimming with $\alpha = 0.15$ provides a more efficient IPW estimator than $\alpha = 0.1$ under no PS tails or only mild PS tails ($\gamma \leq 2$), whereas symmetric trimming with $\alpha = 0.1$ has higher efficiency under stronger PS tails ($\gamma > 2$). When the treatment prevalence is low (0.1), symmetric trimming with $\alpha = 0.1$ corresponds to a more efficient IPW estimator than $\alpha = 0.15$ across all levels of $\gamma$. On the other hand, asymmetric trimming with $q = 0$ may not address extreme weights and tends to be biased (although coverage is adequate because of large estimated variance); asymmetric trimming with $q > 0$ intentionally creates PS nonoverlap, leading to imbalance, biased point estimates, and undercoverage. Overall, asymmetric trimming excludes fewer units than symmetric trimming (Web Figures 3 and 4) and becomes less attractive in our simulations with no unmeasured confounders.

Our simulations confirm the general theory that OW is more efficient than IPW and PS trimming, and they support the optimality of OW even under strong PS tails. Although the 95% confidence interval for OW with our proposed variance estimator had nominal coverage for most scenarios we examined, results in Web Table 3 indicate its slight undercoverage with a smaller sample size, when there are strong PS tails and a low treatment prevalence. This phenomenon is expected, since lower treatment prevalence corresponds to a further reduced effective sample size, which renders the asymptotic results less accurate. Nevertheless, in

**Table 4.** Coverage of the 95% Confidence Intervals for the Estimators in the Presence of Increasingly Strong Tails in the Propensity Score Distributions ($n$ = 2,000)

| Estimator | Data-Generating Model | | | |
|---|---|---|---|---|
| | γ = 1 | γ = 2 | γ = 3 | γ = 4 |
| *Treatment Prevalence = 0.4* | | | | |
| Crude | 0.0 | 0.0 | 0.0 | 0.0 |
| Overlap weighting | 94.7 | 95.4 | 94.8 | 95.0 |
| IPW | | | | |
| No trimming | 94.4 | 91.2 | 79.9 | 66.9 |
| Symmetric trimming | | | | |
| α = 0.05 | 94.4 | 94.8 | 94.8 | 94.0 |
| α = 0.1 | 94.5 | 94.7 | 94.6 | 94.8 |
| α = 0.15 | 94.0 | 94.9 | 95.1 | 94.3 |
| Asymmetric trimming | | | | |
| $q$ = 0 | 94.9 | 96.0 | 97.6 | 96.0 |
| $q$ = 0.01 | 70.1 | 9.1 | 2.1 | 1.0 |
| $q$ = 0.05 | 0.4 | 0.0 | 0.0 | 0.0 |
| *Treatment Prevalence = 0.1* | | | | |
| Crude | 0.0 | 0.0 | 0.0 | 0.0 |
| Overlap weighting | 96.7 | 94.4 | 95.4 | 94.7 |
| IPW | | | | |
| No trimming | 91.6 | 77.7 | 54.0 | 32.3 |
| Symmetric trimming | | | | |
| α = 0.05 | 95.0 | 94.0 | 93.1 | 93.3 |
| α = 0.1 | 94.9 | 94.1 | 94.2 | 94.7 |
| α = 0.15 | 93.5 | 93.1 | 93.8 | 94.3 |
| Asymmetric trimming | | | | |
| $q$ = 0 | 96.1 | 97.3 | 97.0 | 95.7 |
| $q$ = 0.01 | 89.9 | 74.9 | 63.5 | 57.0 |
| $q$ = 0.05 | 67.6 | 22.6 | 13.1 | 10.6 |

Abbreviation: IPW, inverse probability weighting.

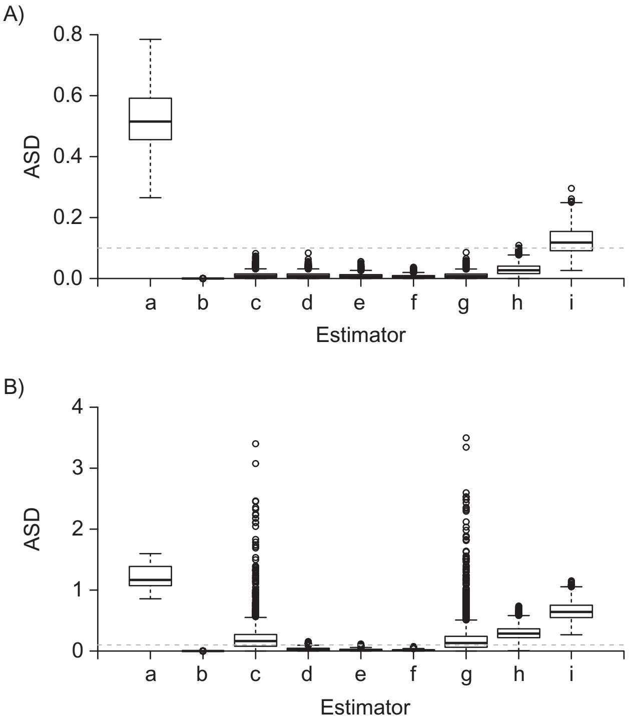

**Figure 2.** Box plots of the absolute standardized difference (ASD) for all covariates in the target population when treatment prevalence equals 0.4 and sample size equals 2,000. A) ASD when γ = 1; B) ASD when γ = 4. The horizontal dashed line indicates adequate balance (ASD = 0.1). Estimators: a, crude; b, overlap weighting; c, inverse probability weighting without trimming; d, symmetric trimming with a threshold of 0.05; e, symmetric trimming with a threshold of 0.1; f, symmetric trimming with a threshold of 0.15; g, asymmetric trimming with a threshold of 0; h, asymmetric trimming with a threshold of 0.01; i, asymmetric trimming with a threshold of 0.05.

this challenging scenario, the confidence interval for OW still outperforms that for symmetric trimming in terms of coverage. To examine how much we could improve the asymptotic results with a smaller sample size, we also use the bootstrap to estimate the variance of OW and obtain the confidence interval. With a small sample size of 500, both our asymptotic confidence interval and the bootstrap confidence interval show nominal coverage under moderate treatment prevalence (Web Table 4). Further, under strong PS tails (γ = 4) and a low treatment prevalence, the bootstrap confidence interval only mildly improves the coverage of the asymptotic confidence interval for OW, suggesting the already satisfactory performance of our proposed variance estimator under strong PS tails.

## DISCUSSION

In this study, we explored the performance of weighting methods used to adjust for confounding in observational treatment comparisons, where the distribution of propensity scores reflects increasing separation between treatment arms and extreme tails. Across this spectrum, standard IPW (without trimming) exhibits increasing bias, large variance, and poor coverage. Symmetric trimming substantially reduces these problems, whereas asymmetric trimming can exaggerate them because of posttrimming nonoverlap. OW produces an unbiased treatment effect estimate, with lower variance and good coverage.

The properties of OW have been demonstrated theoretically, and our findings are not surprising (11). However, it is interesting to note how quickly the problems with asymmetric trimming arise. Bias can occur without extreme tails, but it gets worse as the PS distribution separates towards the tails. Symmetric trimming is relatively safe but sometimes half as efficient as OW. The difference in efficiency is greater when the PS distribution is less extreme.

In practice, the application of trimming rules may be heterogeneous. Various adaptations of symmetric and asymmetric trimming are employed with little justification. Here we focused on published methods which are most likely to produce favorable results; still, there are uncertainties in how to apply them. For example, Crump et al. (9) recommended refitting the PS model and rederiving the weights after trimming. Stürmer et al. (8) did not suggest this step. Additional simulations (not shown) indicate that results would be worse if we did not refit the PS model. To optimize trimming, we applied this extra step to both methods. This is consistent with the intuition that inverse probability weights provide covariate balance on the sample for which they are derived but not necessarily when a unique subset of patients is excluded. Our results may understate the problems with trimming that could arise if the PS model is not refitted and reapplied after trimming or if these trimming methods are adapted without study. This is a potential advantage to OW, in that only a single step is required without ad hoc variations.

The methods of symmetric and asymmetric trimming are distinct from the concept of excluding regions of nonoverlap in the PS distribution. The latter is commonly applied, regardless of PS method, whereby patients are excluded if their PS is outside the range of PS values seen in the other treatment arm. Stürmer et al. (8) explicitly excluded regions of PS nonoverlap as part of the algorithm. Crump et al. (9) did not address this, perhaps because symmetric trimming would usually exclude this region automatically. When using OW, the ideal approach has been unclear; however, this extra step should not be necessary, because patients with such extreme propensities would typically be down-weighted to nearly 0, having very little influence on the results. Here we applied OW directly, without excluding the region of nonoverlap. We find that the results are unbiased, with low variance even in cases of extreme tails.

Our simulations had a variety of limitations. The scope was narrow, and while observations about OW were supported by theory, trimming rules were ad hoc and performance may vary in cases that we did not consider. Our study was also specific to cross-sectional comparisons, for which the current development of OW applies. We did not consider settings of model misspecification or unmeasured confounding. In addition, we focused on a homogeneous treatment effect, with the advantage that the targets of inference remained comparable across different methods. This allowed for a straightforward assessment of bias. In the presence of heterogeneous individual treatment effects, the different weighing methods target different population average treatment effects. Just as trimming changes the population, so does OW, and this depends on various features of the sample that are not evident a priori. In practice, it is important to describe the target population of any analysis. This can be done by providing a baseline characteristics table of the weighted pseudopopulation. Critical evaluation of the target population for clinical relevance is equally important for IPW, particularly when the sampling scheme does not guarantee representation of an inherently important population. In the latter case, extreme propensity scores near 0 and 1 are common, and OW may target a more clinically relevant population for which treatment decisions remain uncertain. For additional perspectives, see the paper by Mao et al. (17), who provided detailed rationales for considering targets of inference other than the average treatment effect.

This study was designed to evaluate PS weighting methods under the usual assumption of no unmeasured confounding. The asymmetric trimming method of Stürmer et al. (8) was originally proposed to address a specific type of unmeasured confounding—that limited to the tails of the PS distribution. We did not attempt to reevaluate that purpose. Instead, we considered implications of using this approach as a generic trimming method and found that other methods perform better. Stürmer et al. (8) exemplified the generalizable principle that propensity methods will perform better in populations for which no unmeasured confounding is a plausible assumption, not necessarily everyone in the available sample. Careful study design may help support this assumption and the validity of methods considered here.

The problem of extreme tails can be addressed in multiple ways that were not considered here. For instance, weight truncation caps the extreme weights at a specified upper threshold and exchanges bias for smaller variance (18). Further, the matching weights also belong to the class of balancing weights and similarly down-weight the tails, although they are not as efficient as OW and do not guarantee exact balance (17, 19). Matching has also been used for this purpose, though prior work suggests that it is less efficient than matching weights (19). More importantly, investigators may reconsider inclusion and exclusion criteria, particularly if a single variable is strongly predictive in the PS model and explains the model separation. In this case, the propensity scores near 0 and 1 may draw attention to a subgroup in which there is no treatment uncertainty, and observational causal inference is not needed. Scientifically justifiable exclusion criteria can reduce the problem. However, this solution is often incomplete when 1) many factors are involved, 2) investigators do not agree on exclusions, 3) expanding the population is a goal, or 4) preserving sample size is a priority. Methods that are robust to extreme propensity scores remain desirable.

The challenge of extreme propensities has been identified as a primary downside of weighting and a reason to switch to alternative PS approaches (5). Our results suggest that the problem is not inherent to weighting but is specific to IPW. Trimming methods initially addressed this problem by excluding people at the tails. OW similarly emphasizes a population at clinical equipoise, where the most uncertainty in treatment decisions remains. When tails in the PS distribution are present, despite good study design and thoughtful inclusion/exclusion criteria, OW removes the arbitrary decisions involved in trimming and improves the characteristics of bias and precision.

## ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics and Bioinformatics, School of Medicine, Duke University, Durham, North Carolina (Fan Li*, Laine E. Thomas); Department of Statistical Science, Trinity College of Arts and Sciences, Duke University, Durham, North Carolina (Fan Li**); and Duke Clinical Research Institute, Durham, North Carolina (Fan Li*, Laine E. Thomas, Fan Li**).

This work was funded in part by grant R03 HS24310 from the Agency for Healthcare Research and Quality and grant SES-1424688 from the National Science Foundation.

The content of this article is solely the view of the authors and does not represent the view of the Agency for Healthcare Research and Quality or the National Science Foundation.

Conflict of interest: none declared.

## REFERENCES

1. 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(28):3661–3679.
2. Bouillon K, Bertrand M, Bader G, et al. Association of hysteroscopic vs laparoscopic sterilization with procedural, gynecological, and medical outcomes. *JAMA*. 2018;319(4):375–387.
3. Brown HK, Ray JG, Wilton AS, et al. Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. *JAMA*. 2017;317(15):1544–1552.
4. Jones PM, Cherry RA, Allen BN, et al. Association between handover of anesthesia care and adverse postoperative outcomes among patients undergoing major surgery. *JAMA*. 2018;319(2):143–153.
5. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
6. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3):e18174.
7. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Methodol*. 2001;2(3-4):259–278.
8. Stürmer T, Rothman KJ, Avorn J, et al. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol*. 2010;172(7):843–854.
9. Crump RK, Hotz VJ, Imbens GW, et al. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187–199.
10. Gagne JJ, Polinski JM, Avorn J, et al. *Standards for Causal Inference Methods in Analyses of Data From Observational and Experimental Studies in Patient-Centered Outcomes Research: Final Technical Report*. Boston, MA: Brigham and Women’s Hospital and Harvard Medical School; 2012. https://www.pcori.org/sites/default/files/Standards-for-Causal-Inference-Methods-in-Analyses-of-Data-from-Observational-and-Experimental-Studies-in-Patient-Centered-Outcomes-Research1.pdf. Accessed April 14, 2018.
11. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2018;113(521):390–400.
12. Hernán MA, Robins JM. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC. In press.
13. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
14. Lip GYH, Skjøth F, Nielsen PB, et al. Effectiveness and safety of standard-dose nonvitamin k antagonist oral anticoagulants and warfarin among patients with atrial fibrillation with a single stroke risk factor: a nationwide cohort study. *JAMA Cardiol*. 2017;2(8):872–881.
15. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29–38.
16. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.
17. Mao H, Li L, Greene T. Propensity score weighting analysis and treatment effect discovery [published online ahead of print June 19, 2018]. *Stat Methods Med Res*. (doi:https://doi.org/10.1177/0962280218781171).
18. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–664.
19. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*. 2013;9(2):215–234.