Received: 12 October 2020 | Revised: 20 March 2021 | Accepted: 25 April 2021

DOI: 10.1002/sim.9029

RESEARCH ARTICLE

Statistics in Medicine WILEY

Propensity score weighting for causal subgroup analysis

Siyun Yang[1] | Elizabeth Lorenzi[2] | Georgia Papadogeorgou[3] | Daniel M. Wojdyla[4] | Fan Li[5] | Laine E. Thomas [1,4]

[1] Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina

[2] Berry Consultants, Austin, Texas

[3] Department of Statistics, University of Florida, Gainesville, Florida

[4] Duke Clinical Research Institute, Duke University School of Medicine, Durham, North Carolina

[5] Department of Statistical Science, Duke University, Durham, North Carolina

Correspondence

Laine Thomas, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.
Email: laine.thomas@duke.edu

Funding information

Patient-Centered Outcomes Research Institute, Grant/Award Number: ME-2018C2-13289

A common goal in comparative effectiveness research is to estimate treatment effects on prespecified subpopulations of patients. Though widely used in medical research, causal inference methods for such subgroup analysis (SGA) remain underdeveloped, particularly in observational studies. In this article, we develop a suite of analytical methods and visualization tools for causal SGA. First, we introduce the estimand of subgroup weighted average treatment effect and provide the corresponding propensity score weighting estimator. We show that balancing covariates within a subgroup bounds the bias of the estimator of subgroup causal effects. Second, we propose to use the overlap weighting (OW) method to achieve exact balance within subgroups. We further propose a method that combines OW and LASSO, to balance the bias-variance trade-off in SGA. Finally, we design a new diagnostic graph—the Connect-S plot—for visualizing the subgroup covariate balance. Extensive simulation studies are presented to compare the proposed method with several existing methods. We apply the proposed methods to the patient-centered results for uterine fibroids (COMPARE-UF) registry data to evaluate alternative management options for uterine fibroids for relief of symptoms and quality of life.

KEYWORDS

balancing weights, causal inference, covariate balance, effect modification, interaction, overlap weights, propensity score, subgroup analysis

1 | INTRODUCTION

Comparative effectiveness research (CER) aims to estimate the causal effect of a treatment(s) in comparison to alternatives, unconfounded by differences between characteristics of subjects. CER has traditionally focused on the average treatment effect (ATE) for the overall population. However, different subpopulations of patients may respond to the same treatment differently,[1,2] and in recent years the CER literature has increasingly shifted attention to heterogeneous treatment effects (HTE).[3-7] In particular, recent research employs machine learning methods to directly model the outcome function and consequently identify the subpopulations with significant HTEs *post analysis*. Popular examples include the Bayesian additive regression trees (BART),[3,8] Causal Forest,[6] and Causal boosting.[9] In this article, we focus on a different type of HTE analysis, widely used in medical research: the causal *subgroup analysis* (SGA) which estimates treatment effects in *prespecified*—usually defined using pretreatment covariates—subgroups of patients. There is an extensive literature on SGA methods in randomized controlled trials.[10-14] However, causal inference methods for SGA with observational data remain underdeveloped.[15-17]

In the context of ATE, covariate balance has been shown to be crucial to unbiased estimation of causal effects.[18,19] Propensity score methods[20] are the most popular method for achieving covariate balance, but have seldom been discussed in SGA.[15,16] Compared with the aforementioned machine learning methods that directly model the outcomes, propensity score methods are design-based in the sense that they avoid modeling the outcome, and the quality of the analysis can be checked through balance diagnostics.[21] In this article, we focus on the propensity score weighting approach.[22-27] Dong et al[16] shows that the true propensity score balances the covariates in expectation between treatment groups in both the overall population and any subgroup defined by covariates. However, the propensity scores are usually unknown in observational studies and must be first estimated from the study sample, leading to estimated propensity scores that rarely coincide with their true values. Moreover, good balance in the overall sample does not automatically translate in good subgroup balance. In fact, our own experience suggests that severe covariate imbalance in subgroups is common in real applications, which may consequently lead to bias in estimating the subgroup causal effects. Despite routinely reporting effects in prespecified subgroups, medical studies rarely check subgroup balance, partially due to the lack of visualization tools. Indeed, we conducted a literature review of all propensity-score-based comparative effectiveness analyses published in the *Journal of American Medical Association* between January 1, 2017 and August 1, 2018. Of 16 relevant publications, half reported SGA (2-22 subgroups per paper) but *none* reported any metrics of balance within subgroups.

The limited literature on propensity score methods in SGA suggests that the propensity score model should be iteratively updated to include covariate-subgroup interactions until subgroup balance is achieved.[28,29] But this procedure has not been implemented in practice, perhaps because it is cumbersome to manually check interactions. More importantly, it may amplify the classic bias-variance tradeoff: increasing complexity of the propensity score model may help to reduce bias but is also expected to increase variance. Therefore, an effective approach would automatically achieve covariate balance in subgroups while preserving precision. Machine learning methods offer a potential solution for estimating the propensity scores without prespecifying necessary interactions. For example, generalized boosted models (GBM) have been advocated as a flexible, data-adaptive method,[30] and random forest was superior to many other tree-based methods for propensity score estimation in extensive simulation studies.[31] BART have been used to estimate the propensity score model and outperformed GBM on some metrics of balance.[32] However, it is unclear whether these methods achieve adequate balance and precision in causal SGA. Moreover, when important subgroups are prespecified, a more effective approach would incorporate prior knowledge about the subgroups.

In this article, we develop a suite of analytical and visualization tools for causal SGA. First, we introduce the estimand of subgroup weighted average treatment effect (S-WATE) and provide the corresponding propensity score weighting estimator (Section 2). Second, we propose a method that combines LASSO[33,34] and overlap weighting (OW),[26,35,36] and balances the bias-variance tradeoff in causal SGA (Section 2.4). Specifically, we treat the prespecified subgroups as candidates for interactions with standard covariates in a logistic propensity score model and use LASSO to select important interactions. We then capitalize on the exact balance property of OW with a logistic regression to achieve covariate balance *both* overall and within subgroups, thus reducing bias and variance in causal SGA. Then, we show analytically that balancing covariates within a subgroup bounds the bias in estimating subgroup causal effects (Section 3). Finally, we device a new diagnostic graph, which we call the Connect-S plot, for visualizing the subgroup covariate balance (Section 4). We conduct extensive simulation studies to compare the proposed method with several alternative methods (Section 5), and illustrate its application in a motivating example (Section 6).

Our methodology is motivated from an observational comparative effectiveness study based on the comparing options for management: patient-centered results for uterine fibroids (COMPARE-UF) registry.[37] Our goal is to evaluate alternative management options for uterine fibroids for relief of symptoms and quality of life. SGA was a primary aim to determine whether certain types of patient subgroups should receive myomectomy vs hysterectomy procedures. Investigators prespecified 35 subgroups of interest based on categories of 16 variables including race, age, and baseline symptom severity. In addition, 20 covariates were considered as potential confounders, including certain demographics, disease history, quality of life and symptoms. The total sample size is 1430, with 567 patients in the myomectomy group and 863 patients in the hysterectomy group. There are in total 700 subgroup-confounder combinations, which pose great challenges to check and ensure balance for causal analyses.

2 | ESTIMANDS AND ESTIMATION IN CAUSAL SGA

2.1 | Notation

Consider a sample of N individuals, where N_1 units belong to the treatment group, denoted by $Z = 1$, and N_0 to the control group, denoted by $Z = 0$. We maintain the stable unit treatment value assumption (SUTVA),[38] which includes two subassumptions: there is (i) no different versions of the treatment (also known as consistency[39]), and (ii) no interference between units. Under SUTVA, each unit i has two potential outcomes $Y_i(1)$ and $Y_i(0)$ corresponding to the two possible treatment levels, of which only the one corresponding to the actual treatment assigned is observed, $Y_i = Z_iY_i(1) + (1 - Z_i)Y_i(0)$. We also observe a vector of P pretreatment covariates, $\boldsymbol{X}_i = (X_{i1}, \dots, X_{iP})^T$.

We denote the subgroups of interest by indicator variables $\boldsymbol{S_i} = (S_{i1}, \dots, S_{iR})^T$, where $S_{ir} = 1$ if the i^{th} unit is a member of the $r^{\text{th}}(r = 1, 2, \dots, R)$ subgroup and 0 otherwise (eg, Black race, male gender, and younger age). Usually, $S_{ir} = f_r(\boldsymbol{X}_i)$ for some function f_r that defines categories based on $\boldsymbol{X}_i$. The R groups are not required to be mutually exclusive, and a unit i can belong to multiple subgroups. In fact, we are particularly interested in one-at-a-time SGA where the groups compared are defined as $S_{ir} = 0$ and $S_{ir} = 1$ for each r, while averaging over the levels of $\{S_{i1}, \dots, S_{iR}\} \setminus \{S_{ir}\}$. Nonetheless, to simplify notation in Section 2.2, we assume mutually exclusive subgroups so that $\sum_{r=1}^{R} S_{ir} = 1$ hereafter.

The propensity score is $e(\boldsymbol{X}_i, \boldsymbol{S}_i) = \Pr(Z_i = 1|\boldsymbol{X}_i, \boldsymbol{S}_i)$. When the components of $\boldsymbol{S}_i$ are functions of $\boldsymbol{X}_i$, the dependence of the propensity score on the subgroup indicators could be dropped. However, the subgrouping variables $\boldsymbol{S}_i$ may not all be a function of $\boldsymbol{X}_i$. Furthermore, subgroups are most often defined based on physicians' and patients' prior knowledge with respect to which covariates are important for selecting treatment or with respect to the outcome. For this reason the true propensity score may be subgroup-specific in that relationships between $\boldsymbol{X}_i$ and Z_i depend on $\boldsymbol{S}_i$. For this reason, both the typical covariates $\boldsymbol{X}_i$ and the subgrouping variables $\boldsymbol{S}_i$ are explicitly denoted.

2.2 | The estimand: Subgroup weighted average treatment effect

Traditional causal inference methods focus on the ATE, $\mathbb{E}_f[Y(1) - Y(0)]$, where the expectation is over the population with probability density $f(\boldsymbol{x}, \boldsymbol{s})$ for the covariates and subgroups. Corresponding SGA would evaluate the subgroup average treatment effect (S-ATE), $\tau_r = \mathbb{E}_f[Y(1) - Y(0)|S_r = 1]$. Recently there has been increasing focus on weighted average treatment effects which represent average causal effects over a different, potentially more clinically relevant populations.[26,27,40-42] We extend the weighted average treatment effect to the context of SGA.

Let $g(\boldsymbol{x}, \boldsymbol{s})$ denote the covariate/subgroup density of the clinically relevant target population. The ratio $h(\boldsymbol{x}, \boldsymbol{s}) = g(\boldsymbol{x}, \boldsymbol{s})/f(\boldsymbol{x}, \boldsymbol{s})$ is called a *tilting function*,[43] which reweights the distribution of the observed sample to represent the target population. Denote the conditional expectation of the potential outcome in subgroup r with treatment z by $\mu_{rz}(\boldsymbol{x}) = \mathbb{E}_f\{Y(z)|\boldsymbol{X} = \boldsymbol{x}, S_r = 1\}$ for $z = 0, 1$. Then, we can represent the S-WATE over the target population by:

$$\tau_{r,h} = \mathbb{E}_g[Y(1) - Y(0)|S_r = 1] = \frac{\mathbb{E}\{h(\boldsymbol{X}, \boldsymbol{S})(\mu_{r1}(\boldsymbol{X}) - \mu_{r0}(\boldsymbol{X}))|S_r = 1\}}{\mathbb{E}\{h(\boldsymbol{X}, \boldsymbol{S})|S_r = 1\}}. \tag{1}$$

In practice, we specify the target population by prespecifying the tilting function $h(\boldsymbol{x}, \boldsymbol{s})$. Different choices of the function h lead to different estimands of interest. For example, for $h(\boldsymbol{x}, \boldsymbol{s}) = 1$ the S-WATE collapses to the S-ATE: $\tau_{r,h} \equiv \tau_r$. Another special case arises under homogeneity when $\mu_{r1}(\boldsymbol{x}) - \mu_{r0}(\boldsymbol{x})$ is constant for all $\boldsymbol{x}$ and $\tau_{r,h} \equiv \tau_r$ for all h. Several common tilting functions will be discussed subsequently within the context of SGA.

To identify the S-WATE from observational data, we make two standard assumptions:[20] (i) *Unconfoundedness*: $Z \perp\!\!\!\perp \{Y(1), Y(0)\} | \{\boldsymbol{X}, \boldsymbol{S}\}$, which implies that the treatment assignment is randomized given the observed covariates, and (ii) *Overlap (or positivity)*: $0 < e(\boldsymbol{X}_i, \boldsymbol{S}_i) < 1$, which requires that each unit has a nonzero probability of being assigned to either treatment condition. Then, we can estimate the S-WATE in subgroup r, $\tau_{r,h}$, using the Hájek estimator

$$\hat{\tau}_{r,h} = \frac{\sum_{i=1}^{N} Z_i S_{ir} w_{i1} Y_i}{\sum_{i=1}^{N} Z_i S_{ir} w_{i1}} - \frac{\sum_{i=1}^{N} (1 - Z_i) S_{ir} w_{i0} Y_i}{\sum_{i=1}^{N} (1 - Z_i) S_{ir} w_{i0}}, \tag{2}$$

where the weights w are the balancing weights corresponding to the specific tilting function $h(\boldsymbol{x}, \boldsymbol{s})$ (equivalently the target population $g(\boldsymbol{x}, \boldsymbol{s})$):[26]

$$\begin{cases} w_{i1} = \frac{h(\boldsymbol{X}_i, \boldsymbol{S}_i)}{e(\boldsymbol{X}_i, \boldsymbol{S}_i)} & \text{for } Z_i = 1, \\ w_{i0} = \frac{h(\boldsymbol{X}_i, \boldsymbol{S}_i)}{1 - e(\boldsymbol{X}_i, \boldsymbol{S}_i)} & \text{for } Z_i = 0. \end{cases} \tag{3}$$

The most widely used balancing weights are the inverse probability weights (IPW),[23] ($w_1 = 1/e(\boldsymbol{x}, \boldsymbol{s}), w_0 = 1/(1 - e(\boldsymbol{x}, \boldsymbol{s}))$, corresponding to $h(\boldsymbol{x}, \boldsymbol{s}) = 1$. The target population of IPW is the combination of treated and control patients that are represented by the study sample, and the subgroup-specific estimand is the S-ATE. The balancing weights which will play a key role in this article (Sections 2.3 and 2.4) are the overlap weights (OW), ($w_1 = 1 - e(\boldsymbol{x}, \boldsymbol{s}), w_0 = e(\boldsymbol{x}, \boldsymbol{s})$), corresponding to $h(\boldsymbol{x}, \boldsymbol{s}) = e(\boldsymbol{x}, \boldsymbol{s})(1 - e(\boldsymbol{x}, \boldsymbol{s}))$.[26] Balancing weights are defined on the entire sample and are applicable to subgroups where the value of $\boldsymbol{S}_i$ is fixed and defines the subgroup of interest. We show in Web Appendix 1.1 that $\hat{\tau}_{r,h}$ is consistent for $\tau_{r,h}$.

In practice, the true propensity score, $e(\boldsymbol{X}_i, \boldsymbol{S}_i)$, is usually not known and is estimated from the data. Then, the weights w_i in (2) are replaced with $\hat{w}_i$ based on the estimated propensity score $\hat{e}(\boldsymbol{X}_i, \boldsymbol{S}_i)$. While balancing the true propensity score would balance the covariates in all covariate-defined subgroups in expectation, the estimated weights $\hat{w}_i$ based on an estimated propensity score often fail to achieve covariate balance, particularly within subgroups.[16] As we show in Section 3, covariate balance in the subgroups is crucial for unbiased estimation of the S-WATE. Therefore, it is beneficial to choose weights that guarantee balance.

2.3 | Exact subgroup balance via OW

We propose to use OW to achieve exact balance on the subgroup-specific covariate means. As noted above, the overlap weight of each unit is the probability of being assigned to the opposite group: $w_1 = 1 - e(\boldsymbol{x}, \boldsymbol{s})$ and $w_0 = e(\boldsymbol{x}, \boldsymbol{s})$, arising from tilting function $h(\boldsymbol{x}, \boldsymbol{s}) = e(\boldsymbol{x}, \boldsymbol{s})(1 - e(\boldsymbol{x}, \boldsymbol{s}))$. This tilting function is maximized for individuals with propensity scores close to 0.5, that is, those who are equally likely to be treated or not, and minimized for individuals with propensity scores close to 0 or 1, that is, those who are nearly always treated or never treated. Consequently, the target population of OW emphasizes covariate profiles with the most overlap between treatment groups and the subgroup-specific estimand is the subgroup average treatment effect of the overlap population (S-ATO). Though statistically defined, this represents a target population of intrinsic substantive interest.[26,35,36] Specifically, the overlap population mimics the characteristics of a pragmatic randomized trial that is highly inclusive, excluding no study participants from the available sample, but emphasizing the comparison of patients at clinical equipoise. The resulting target population can be empirically described through a weighted baseline characteristics table. When the S-ATO is clinically relevant, its corresponding weighting estimator has attractive properties regarding balance and variance.

First, OW have a unique finite-sample property of exact balance. Specifically, outside the context of SGA, Li et al[26] show that when the propensity score is estimated by a logistic regression, OW lead to exact balance on the weighted covariate means. We extend this property to subgroups as follows.

Proposition 1. *If the postulated propensity score model is logistic regression with subgroup-covariate interactions, that is,* $\hat{e}(\boldsymbol{X}_i, \boldsymbol{S}_i) = \text{logit}^{-1}(\hat{\alpha}_0 + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}_{\boldsymbol{x}} + \boldsymbol{S}_i^T \hat{\boldsymbol{\alpha}}_{\boldsymbol{s}} + (\boldsymbol{X}_i \cdot \boldsymbol{S}_i)^T \hat{\boldsymbol{\alpha}}_{\boldsymbol{xs}})$, *where* $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}_{\boldsymbol{x}}^T, \hat{\boldsymbol{\alpha}}_{\boldsymbol{s}}^T, \hat{\boldsymbol{\alpha}}_{\boldsymbol{xs}}^T)^T$ *is the maximum likelihood (ML) estimator and* $(\boldsymbol{X}_i \cdot \boldsymbol{S}_i)$ *denotes all pairwise interactions between* $\boldsymbol{X}_i$ *and* $\boldsymbol{S}_i$*, then the OW lead to exact mean balance in the subgroups and overall:*

$$\sum_{i=1}^{N} Z_i S_{ir} X_{ip} \hat{w}_{i1} - \sum_{i=1}^{N} (1 - Z_i) S_{ir} X_{ip} \hat{w}_{i0} = 0, \quad \text{for all } r = 1, 2, \dots, R, \text{ and } p = 1, 2, \dots, P.$$

Again the weights need to be normalized such that $\sum_i^N Z_i S_{ir} \hat{w}_{i1} = \sum_i^N (1 - Z_i) S_{ir} \hat{w}_{i0} = 1$ *(Web Appendix 1.5).*

Proposition 1 implies that when a logistic model for propensity scores is augmented to include $(\boldsymbol{X}_i \cdot \boldsymbol{S}_i)$ and paired with OW, exact balance is achieved *both overall and within subgroups*. In addition, the approach can be motivated by focusing on correct specification of the propensity score model in the scientific context. When subgroups are defined *a priori* it is usually based on clinical knowledge of which patient characteristics are most likely to alter the treatment effect. Thus, treatment decisions in the observational data may already be different in these subgroups, corresponding to covariate-subgroup interactions in the true propensity score model. This motivates the inclusion of prespecified subgroups as candidates for interactions with standard covariates in the propensity score model. However, as the propensity score model approaches saturation, the estimated propensity scores will converge to 0 and 1, thus causing variance inflation (VI) in the treatment effect estimates.

VI with increasing PS model complexity is partially mitigated by OW. OWs are naturally bounded between 0 and 1, thus can avoid the issues of extreme weights and large variability that can occur when $h(\boldsymbol{x}, \boldsymbol{s}) = 1$.[26] In fact, the overlap tilting function $h(\boldsymbol{x}, \boldsymbol{s}) = e(\boldsymbol{x}, \boldsymbol{s})(1 - e(\boldsymbol{x}, \boldsymbol{s}))$ gives the smallest large-sample variance of the weighted estimator $\hat{\tau}_{r,h}$ over all possible h under homoscedasticity (Web Appendix 1.1). For SGA, the optimal efficiency helps to mitigate the potential VI arising from a more complex propensity score model. Nonetheless, when the number of covariates and/or subgroups is large, variable selection in the propensity score model is necessary. Therefore, we propose a new method for causal SGA to accommodate considerations on both variable selection and covariate balance.

2.4 | Combining OW with Post-LASSO for causal SGA: The OW-pLASSO algorithm

We propose the *OW-pLASSO* algorithm for causal SGA, which combines two main components. The first component uses the Post-LASSO approach[34,44] to select covariate-subgroup interactions and estimate the propensity scores. In causal settings regularization inadvertently biases treatment effect estimates by overshrinking regression coefficients.[45] Hence, we adopt the Post-LASSO approach instead of the original LASSO.[33] The second component uses OW to achieve covariate balance in the subgroups.

The *OW-pLASSO* algorithm consists of the following steps:

S1. Fit a logistic propensity score model with all prespecified covariates and subgroup variables along with pairwise covariate-subgroup interactions, that is, design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{X}_i \cdot \boldsymbol{S}_i)$, *and perform LASSO to select covariate-subgroup interactions (without penalizing the main effects in the model).*

S2. Estimate the propensity scores by refitting the logistic regression with all main effects and selected covariate-subgroup interactions from S1.

S3. Calculate the OW based on the propensity scores estimated from S2, and check subgroup balance using the Connect-S plot (Section 4) before and after weighting.

S4. Estimate the causal effects for all prespecified subgroups using Estimator (2) *with the OW from S3.*

From extensive simulation studies (Section 5), we find the *OW-pLASSO* algorithm outperforms combinations of IPW and other popular machine learning models for propensity scores in estimating the S-WATE estimands. One of the key reasons of OW-pLASSO's advantage is that it achieves within-subgroup exact mean balance, which is crucial for bias reduction, as we show analytically in the following section.

To estimate the variance of the (overall and subgroup) treatment effects, we suggest two methods: (i) the robust sandwich estimator, as recently described for IPW;[46] this approach is known to be slightly conservative as it does not take into account the uncertainty in estimating the propensity scores, but has been shown to work well in practice. (ii) Bootstrapping: estimate propensity scores in the original sample using Post-LASSO and treat the estimated propensity scores as fixed when estimating the causal effects in each bootstrap sample. Note that in the bootstrap method, we caution against the practice of refitting the propensity score using Post-LASSO in each bootstrap sample, because the bootstrap is *in*consistent for LASSO estimators.[47] Since uncertainty for LASSO estimators is hard to quantify, none of the variance estimation approaches we consider aims to incorporate the variability of propensity score estimates. However, ignoring the uncertainty of the propensity score is justifiable in causal inference studies[25] as the propensity score is often viewed as part of the "design" phase of a study.[48] Our simulation studies in Web Appendix 2.3 validate these variance estimation methods coupled with the proposed OW-pLASSO algorithm.

3 | BOUNDING BIAS FOR SUBGROUP CAUSAL EFFECTS

When focusing on additive models, Zubizarreta[19] showed that the weighting estimator for the population mean is unbiased when the covariate means are balanced. We extend this work to SGA by showing that balance of covariates within a subgroup leads to minimal bias of the estimator $\hat{\tau}_{r,h}$. In Proposition 2, we show this result when the treatment effect is homogeneous within a subgroup ($\tau_{r,h} = \tau_r$), and in Proposition 3 we extend it to allow for within-subgroup effect heterogeneity. In both cases, treatment effects are allowed to vary between subgroup levels.

Proposition 2. *Suppose that the outcome surface satisfies an additive model, for example,* $Y_i(z) = \sum_{r=1}^{R} \beta_r S_{ir} + \sum_{r=1}^{R} \sum_{p=1}^{P} \beta_{rp} S_{ir} X_{ip} + \sum_{r=1}^{R} \tau_r S_{ir} z + \epsilon_i(z)$, *with* $\mathbb{E}[\epsilon_i(z) | \boldsymbol{X}_i, \boldsymbol{S}_i] = 0$. *For any weight* w_i *that is normalized within subgroups (ie,*

$\sum_{i=1}^{N} Z_i S_{ir} w_{i1} = \sum_{i=1}^{N} (1 - Z_i) S_{ir} w_{i0} = 1$), if mean balance holds in the r$^{\text{th}}$ *subgroup, expressed as*

$$\left| \sum_{i=1}^{N} Z_i S_{ir} w_{i1} X_{ip} - \sum_{i=1}^{N} (1 - Z_i) S_{ir} w_{i0} X_{ip} \right| < \delta, \quad \text{for all} \;\; p = 1, 2, \ldots , P, \tag{4}$$

then the bias for the r$^{\text{th}}$ *subgroup is bounded, $|E[\hat{\tau}_{r,h} - \tau_r]| < \delta \sum_{p=1}^{P} |\beta_{rp}|$ (Web Appendix 1.2).*

Therefore, any weighting scheme for which $\delta \approx 0$ will eliminate bias for SGA when the outcome satisfies an additive model. Proposition 2 illustrates that mean balance in the overall sample, $\left|\sum_{i=1}^{N} Z_i w_{i1} X_{ip} - \sum_{i=1}^{N} (1 - Z_i) w_{i0} X_{ip}\right| < \delta$, is *not* sufficient, and balance is required *within the subgroup*. Even in the special case where the true response surface is additive in the covariates and the treatment effect is constant ($\beta_{rp} = \beta_p$, and $\tau_r = \tau$), the subgroup-specific Condition (4) is still necessary to ensure minimal bias of $\hat{\tau}_{r,h}$.

Proposition 3. *Suppose the additive model is relaxed to allow treatment effect heterogeneity by covariates $\boldsymbol{X}_i$ within subgroups: $Y_i(z) = \sum_{r=1}^{R} \beta_r S_{ir} + \sum_{r=1}^{R} \sum_{p=1}^{P} \beta_{rp} S_{ir} X_{ip} + \sum_{r=1}^{R} \tau_r S_{ir} z + \sum_{p=1}^{P} \gamma_{rp} S_{ir} X_{ip} z + \varepsilon_i(z)$, with $E[\varepsilon_i(z)|\boldsymbol{X}_i, \boldsymbol{S}_i] = 0$. If Condition (4) holds and additionally,*

$$\left| \sum_{i=1}^{N} Z_i S_{ir} w_i X_{ip} - \frac{\sum_{i=1}^{N} h(\boldsymbol{X}_i, \boldsymbol{S}_i) S_{ir} X_{ip}}{\sum_{i=1}^{N} h(\boldsymbol{X}_i, \boldsymbol{S}_i) S_{ir}} \right| < \delta_2, \quad \text{for all} \;\; p = 1, 2, \ldots , P, \tag{5}$$

then the bias for the r$^{\text{th}}$ *subgroup is bounded, $|E[\hat{\tau}_{r,h} - \tau_{r,h}]| < \delta \sum_{p=1}^{P} |\beta_{rp}| + \delta_2 \sum_{p=1}^{P} |\gamma_{rp}|$ (Web Appendix 1.3).*

Condition (5) requires the weighted sample covariate mean of treated patients within the subgroup to be close to the subgroup target population covariate mean. This condition can be verified when h is a predefined function, but not when $h(\boldsymbol{X}_i, \boldsymbol{S}_i)$ depends on an unknown propensity score $e(\boldsymbol{X}_i, \boldsymbol{S}_i)$. However, this term is expected to be small unless the model for the propensity score is severely misspecified. In Web Appendix 1.4, we show that an alternative, verifiable condition: $\left|\sum_{i=1}^{N} Z_i S_{ir} w_i X_{ip} - \frac{\sum_{i=1}^{N} \hat{h}(\boldsymbol{X}_i, \boldsymbol{S}_i) S_{ir} X_{ip}}{\sum_{i=1}^{N} \hat{h}(\boldsymbol{X}_i, \boldsymbol{S}_i) S_{ir}}\right| < \delta_2$, is sufficient if we are willing to estimate a slightly different estimand, namely, the subgroup-sample weighted average treatment effect, $\tau_{r,\hat{h}} = \frac{\sum_i \hat{h}(\boldsymbol{X}_i, \boldsymbol{S}_i)[\mu_{r1}(\boldsymbol{X}_i, \boldsymbol{S}_i) - \mu_{r0}(\boldsymbol{X}_i, \boldsymbol{S}_i)] S_{ir}}{\sum_i \hat{h}(\boldsymbol{X}_i, \boldsymbol{S}_i) S_{ir}}$. Therefore, verifiable mean balance conditions are sufficient for $\hat{\tau}_{r,h}$ to have a causal interpretation, but the propensity score model must be approximately correct in order for the weighted population to correspond to the target population and estimate $\tau_{r,h}$. Similarly to Condition (4), Condition (5) can be checked by the Connect-S plot (Section 4).

It is instructive to consider the special case were $h(\boldsymbol{X}_i, \boldsymbol{S}_i) = 1$ and the target population is the sampled population. In this case, h is known and Condition (5) can be empirically verified. However, it will not necessarily be satisfied for weights based on an estimated propensity score. To the best of our knowledge, Condition (4) is typically checked but Condition (5) is not. Under HTEs this second condition is needed. This reveals a potential risk of using weights that balance covariates without defining a tilting function and target estimand (S-WATE).[18,19,26,27] The implicit estimand is the S-ATE with $h(\boldsymbol{X}_i, \boldsymbol{S}_i) = 1$. While these methods are designed to satisfy Condition (4), Condition (5) does not play a role in the construction of the weights and may be violated.

The assumption of linearity in the covariates can be relaxed and the nonlinear case is addressed in Web Appendix 1.6 (Proposition 4). We find that mean balance remains an important condition for unbiasedness, but various higher order moments are potentially important, depending on the true model. Whether it would be practically feasible to prespecify and interpret the corresponding, higher order balance checks, particularly in finite samples, requires future investigation. We do not undertake that here, but instead focus on correct estimation of the propensity score model, coupled with mean balance which is sufficient in linear models (above) and necessary in nonlinear models.

4 | VISUALIZING SUBGROUP BALANCE: THE CONNECT-S PLOT

In practice, it is often difficult to assess whether existing propensity score methods achieve the balance conditions defined in Section (3). For example, in the motivating application of COMPARE-UF, there are 700 combinations of subgroups and covariates for which to check Condition (4). In this section, we introduce a new graph for visualizing subgroup balance—the Connect-S plot. We first introduce two important metrics that will be presented in the plot.

The first statistic is the *absolute standardized mean difference* (ASMD),[49] which is widely used for measuring covariate balance. The ASMD is the difference in weighted means, defined in Condition (4), further scaled by the pooled, unweighted standard deviation. That is

$$\text{ASMD}_{r,p} = \frac{\sum_{i=1}^{N} Z_i S_{ir} w_{i1} X_{ip} - \sum_{i=1}^{N} (1 - Z_i) S_{ir} w_{i0} X_{ip}}{s_{r,p}} \tag{6}$$

where $s_{r,p}$ is the unweighted, pooled standard deviation for the r^{th} subgroup and the p^{th} covariate (See Web Appendix 1.5 for details). Scaling by $s_{r,p}$ facilitates a practical interpretation of the weighted mean difference, relative to the standard deviation of the variable X_p. Various rules of thumb suggest that the $\text{ASMD}_{r,p}$ should be less than 0.10 or 0.20 (ie, an acceptable δ is <0.10 to 0.20).[49]

The second metric concerns variance. In the context of SGA, the propensity score model is typically complex, including many interaction terms. Therefore, a particularly important consideration in propensity score weighting is the VI due to model complexity. Li et al[26] suggested to use the following statistic akin to the "design effect" approximation of Kish[50] in survey literature to approximate the *VI*:

$$\text{VI} = (1/N_1 + 1/N_0)^{-1} \sum_{z=0,1} \frac{\sum_{i=1}^{N_z} w_{iz}^2}{\left(\sum_{i=1}^{N_z} w_{iz}\right)^2}, \tag{7}$$

where N_z is the sample size of treatment group z. For the unadjusted estimator, $w_{iz} = 1$ for all units and $\text{VI} = 1$. Increasing values of VI imply increasingly worse efficiency for alternative weighting algorithms. It is straightforward to define the subgroup-specific version of the VI statistic.

The Connect-S plot for S subgroups resembles the rectangular grid of a Connect4 game: each row represents a subgroup variable (eg, a race group), and the name and subgroup sample size is displayed at the beginning and the end of each row, respectively; each column represents a confounder that we want to balance (eg, age). Therefore, each dot corresponds to a specific subgroup S and confounder X, and the shade of the dot is coded based on the ASMD of confounder X in subgroup S, with darker color meaning more severe imbalance. The end of each row also presents subgroup-specific approximate VI.

Panel (a) of Figure 1 presents the Connect-S plot for COMPARE-UF after adjustment by IPW where the propensity score for myomectomy vs hysterectomy is estimated by main effects logistic regression. The bottom row of this panel shows that this method does a good job of balancing the confounders, overall. However, it does a poor job of achieving balance within subgroups. For example, subgroups based on age, symptom severity, EQ5D quality of life score, and uterine volume have many ASMDs greater than 0.10 and often greater than 0.20. These are not generally acceptable and motivate alternative methodology. A potential solution would be to use a more flexible model for the propensity score that does not assume main effects. Panel (b) of Figure 1 shows that balance in COMPARE-UF is not improved by estimating the propensity score with GBM and results were similar for random forest and BART methods (Web Appendix 2.4).

5 | SIMULATIONS

We compare the proposed OW-pLASSO method with a number of popular machine learning propensity score methods via simulations under different levels of confounding, sparsity, and heterogeneity in causal SGA.

5.1 | Simulation design

Data generating process. In alignment with the COMPARE-UF study we generate $N = 3000$ patients, with $P \in \{18, 48\}$ independent covariates $\boldsymbol{X}_i$, half of which drawn from a standard normal distribution $N(0, 1)$, and the other half from Bernoulli(0.3). Two subgroup variables $\boldsymbol{S}_i = (S_{i1}, S_{i2})$ are independently drawn from Bernoulli(0.25). The treatment indicator Z_i is generated from Bernoulli($e(\boldsymbol{X}_i, \boldsymbol{S}_i)$), with the *true propensity score model*:

$$\text{logit}(e(\boldsymbol{X}_i, \boldsymbol{S}_i)) = \alpha_r + \boldsymbol{S}_i^T \boldsymbol{\alpha}_s + \boldsymbol{X}_i^T \boldsymbol{\alpha}_x + (\boldsymbol{X}_i \cdot \boldsymbol{S}_i)^T \boldsymbol{\alpha}_{\boldsymbol{xs}}, \tag{8}$$

with coefficients $\boldsymbol{\alpha} = (\alpha_r, \boldsymbol{\alpha}_{\boldsymbol{s}}^T, \boldsymbol{\alpha}_{\boldsymbol{x}}^T, \boldsymbol{\alpha}_{\boldsymbol{xs}}^T)^T$.

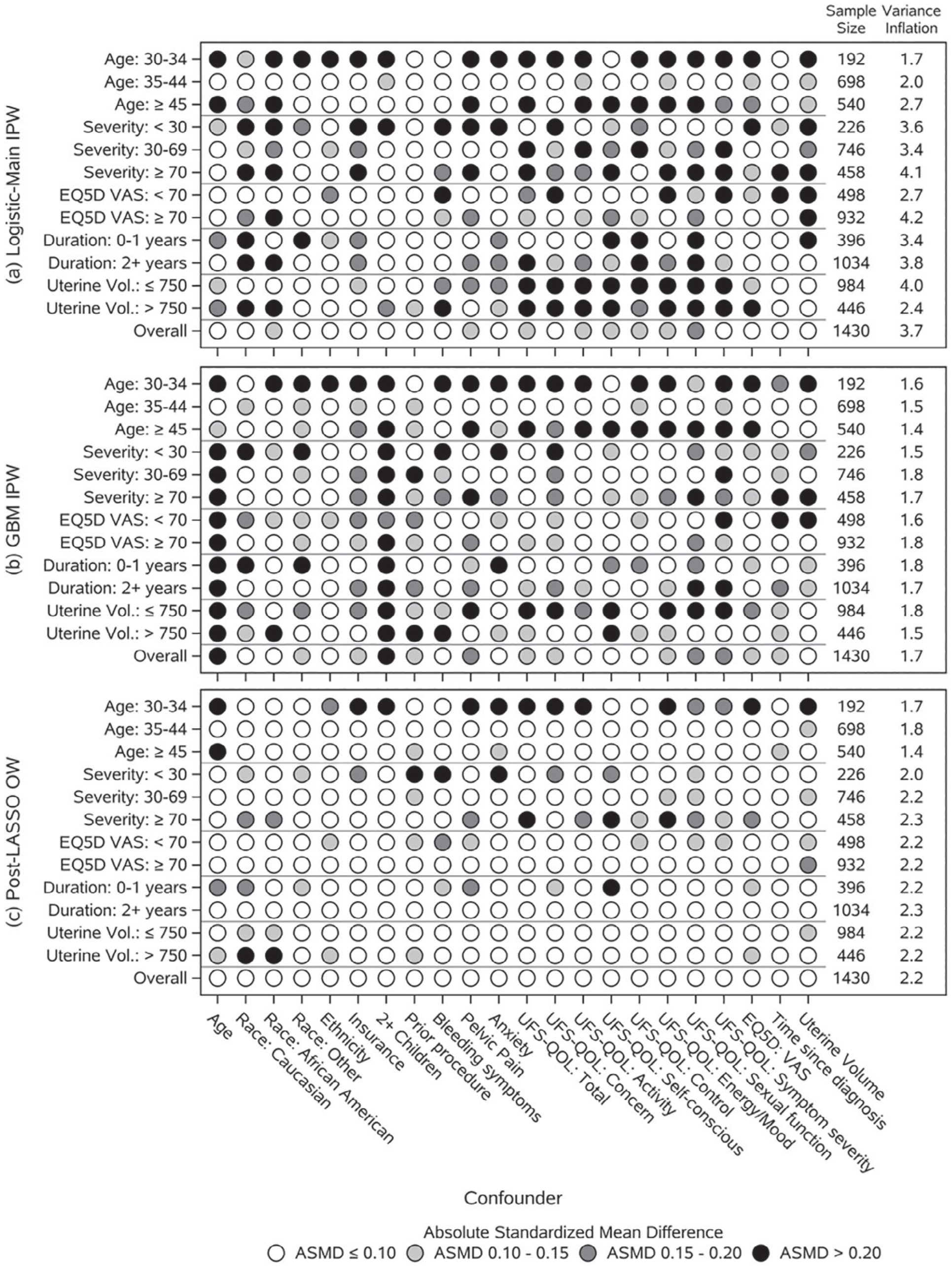


FIGURE 1 The Connect-S plot of the subgroup ASMD and approximate variance inflation in COMPARE-UF after applying balancing weights for adjustment by a) Logistic-Main IPW, propensity score estimated by main effects logistic regression with IPW; b) GBM IPW, propensity score estimated by GBM with IPW; c) OW-pLASSO, propensity score estimated by Post-LASSO with OW. Select subgroups are displayed in rows and all confounders are displayed in columns

We set the coefficients in model (8) as follows: $\alpha_r = -2$, $\boldsymbol{\alpha}_s^T = (1, 1)$. Out of the P coefficients in $\boldsymbol{\alpha}_x$, ψ portion of them have nonzero coefficients (ie, true confounders in our simulation). The coefficients for the continuous and binary confounders take equally distanced values between $(0.25\gamma, 0.5\gamma)$, separately, and the rest are zeros. Last, we set $\boldsymbol{\alpha}_{xs} = -\boldsymbol{\alpha}_x\kappa$. To create a range of realistic scenarios in SGA we vary the three hyperparameters (ψ, γ, κ) in the true propensity score model: (1) $\psi \in \{0.25, 0.75\}$ controls the proportion of covariates $\boldsymbol{X}_i$ that are true confounders; (2) $\gamma \in \{1, 1.25, 1.5\}$ controls the scale of the regression coefficients for $\boldsymbol{X}_i$, and (3) $\kappa \in \{0.25, 0.5, 0.75\}$ scales the regression coefficients for $(\boldsymbol{X}_i \cdot \boldsymbol{S}_i)$. For example, for $P = 18, \gamma = 1, \psi = 0.25$, and $\kappa = 0.5$, the above setting specifies $\boldsymbol{\alpha}_x^T = (0.25, 0.5, \mathbf{0}_7, 0.25, 0.5, \mathbf{0}_7)$, $\boldsymbol{\alpha}_{xs}^T = (-0.125, -0.25, \mathbf{0}_7, -0.125, -0.25, \mathbf{0}_7)$, where $\mathbf{0}_k$ is a k-vector of zeros. The above simulation settings mimic a common SGA situation in clinical studies. Specifically, when $S_1 = 1, S_2 = 1$, the two subgroup variables represent high risk conditions associated with the outcome (eg, risk score) and increase the likelihood of being treated. In the presence of these high risk conditions, other patient characteristics $\boldsymbol{X}_i$ play a lesser role in driving treatment decisions; this is reflected by the fact that magnitude of $\boldsymbol{\alpha}_x$ in the propensity model is smaller than $\boldsymbol{\alpha}_s$. In Web Appendix 2.1, we show that these specifications lead to treated and control units with various amounts of overlap for the true propensity score distributions.

Next, a continuous outcome Y_i (eg, risk score) is generated from a linear regression model:

$$Y_i = \beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}_x + \boldsymbol{S}_i^T\boldsymbol{\beta}_s + \beta_z Z_i + (\boldsymbol{S}_i \cdot Z_i)^T\boldsymbol{\beta}_{sz} + \varepsilon_i, \tag{9}$$

where $(\boldsymbol{S}_i \cdot Z_i)$ is a vector of all possible interactions between subgroup variables and treatment assignment, and ε_i is independently sampled from $N(0, 1)$. We fix the model parameter $\beta_0 = 0$, $\boldsymbol{\beta}_x = \boldsymbol{\alpha}_x$, $\boldsymbol{\beta}_s^T = (0.8, 0.8)$, $\beta_z = -1$, and vary $\boldsymbol{\beta}_{sz}^T = (\beta_{1z}, \beta_{2z})^T \in \{(0, 0), (0.5, 0.5)\}$. When $\boldsymbol{\beta}_{sz}^T = (0, 0)$, the treatment effect is homogeneous, and $\tau_r = \beta_z = -1$ for all subgroups. When $\boldsymbol{\beta}_{sz}^T = (0.5, 0.5)$, the underlying treatment effect is heterogeneous within subgroups and between different subgroup levels. For example, when $P = 18$, $\psi = 0.25$, $\gamma = 1$, $\kappa = 0.75$, the true causal effect $\tau_h = -0.67$ for ATO, and -0.75 for ATE; $\tau_{\{S_1=0,h\}} = \tau_{\{S_2=0,h\}} = -0.83$ for S-ATO, and -0.87 for S-ATE; $\tau_{\{S_1=1,h\}} = \tau_{\{S_2=1,h\}} = -0.35$ for S-ATO, and -0.37 for S-ATE.

Postulated propensity score models. To estimate the propensity scores, we compare Post-LASSO with several popular alternatives in the literature: (1) True model: Logistic regression fitted via ML with the correctly specified propensity score (8), representing the oracle reference; (2) Logistic-Main: logistic regression with only main effects of the predictors $(\boldsymbol{X}_i, \boldsymbol{S}_i)$ fitted via ML, representing the standard practice; (3) LASSO: LASSO[33] with the design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{X}_i \cdot \boldsymbol{S}_i)$, implemented by the R package *glmnet* without penalizing the main effects, and 10-fold cross-validation for hyperparameter tuning;[51] (4) Post-LASSO: Logistic regression model fitted via ML with the covariate-subgroup interactions selected from the preceding LASSO;[34] (5) RF-Main: Random Forest (RF)[6,52] with the design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i)$, implemented by R package *ranger* with default hyperparameters and 1000 trees;[53] (6) RF-All: RF with the augmented design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{X}_i \cdot \boldsymbol{S}_i)$; Among the examined scenarios, we observe no difference between the RF-All and RF-Main PS model, suggesting that RF performance depends little on the provided design matrix. For simplicity, we omit results on RF-All; (7) GBM: GBM[30,54] with the design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i)$, implemented by R package *twang* with 5000 trees, interaction depth equals to 2, and other default hyperparameters;[55] (8) BART: Bayesian additive regression trees[8] with the design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i)$, using the R function *pbart* in package *BART* with default hyperparameters.[56]

Each of the preceding propensity score models is paired with (a) IPW and (b) OW. All the simulation analyses are conducted under R version 3.4.4. In total, there are 72 scenarios examined by the factorial design, with 100 replicate datasets generated per scenario.

Performance metrics. The performance of different approaches is compared overall (averaged over subgroups) and within four subgroups defined by $S_{i1} = 0$, $S_{i1} = 1$, $S_{i2} = 0$, $S_{i2} = 1$. First, we check balance of covariates by the ASMD of each covariate, averaged across the 100 simulated datasets, and calculate the maximum ASMD value across all covariates. Second, we consider the relative bias and root mean squared error (RMSE) to study the precision and stability of various estimators.

5.2 | Simulation results

Covariate balance (AMSD), bias, and RMSE of the various estimators based on different postulated propensity score models and weighting schemes in the simulations are shown in Web Figure 2, Figures 2, and 3, respectively.

Balance. From Web Figure 2, OW estimators achieve better covariate balance than IPW estimators across all propensity score models. The true propensity score model and OW achieves perfect balance for the confounders in all subgroups. This is expected given OW's exact balance property for any included covariate-subgroup interactions (proposition 1). Within the same weighting scheme, the LASSO and Post-LASSO model perform similarly, resulting in smaller ASMDs than the other methods. The Logistic-Main leads to satisfactory balance in the overall sample and the baseline subgroups (ie, $S_1 = 0$ and $S_2 = 0$), but fails to balance the covariates in the $S_1 = 1$ and $S_2 = 1$ subgroups, particularly when paired with IPW. The RF models result in inferior balance performance (measured using ASMDs), occasionally leading to severe subgroup imbalances. BART and GBM perform similarly, which lie between the Logistic-Main and the LASSO models.

Bias. From Figure 2, we can see that OW results in lower bias than IPW, for each propensity score modeling approach, both the overall and the subgroup effects. Between the different propensity score models, the pattern follows closely the degree of covariate imbalance. We find that OW-pLASSO returns the smallest bias within each subgroup and overall. LASSO is slightly inferior to Post-LASSO, likely due to the shrinkage-induced bias. The common practice of using Logistic-Main IPW overestimates treatment effect in the baseline subgroups and greatly underestimates treatment effect in the $S_1 = 1$ and $S_2 = 1$ subgroups. If the same estimated propensity scores are paired with OW, the resulting estimates are much closer to the truth, and the bias for subgroups $S_1 = 1$ and $S_2 = 1$ is reduced to half. BART and GBM perform slightly better than the Logistic-Main and RF model. Web Figure 3 and 4 provide more details of subgroup bias across a range of

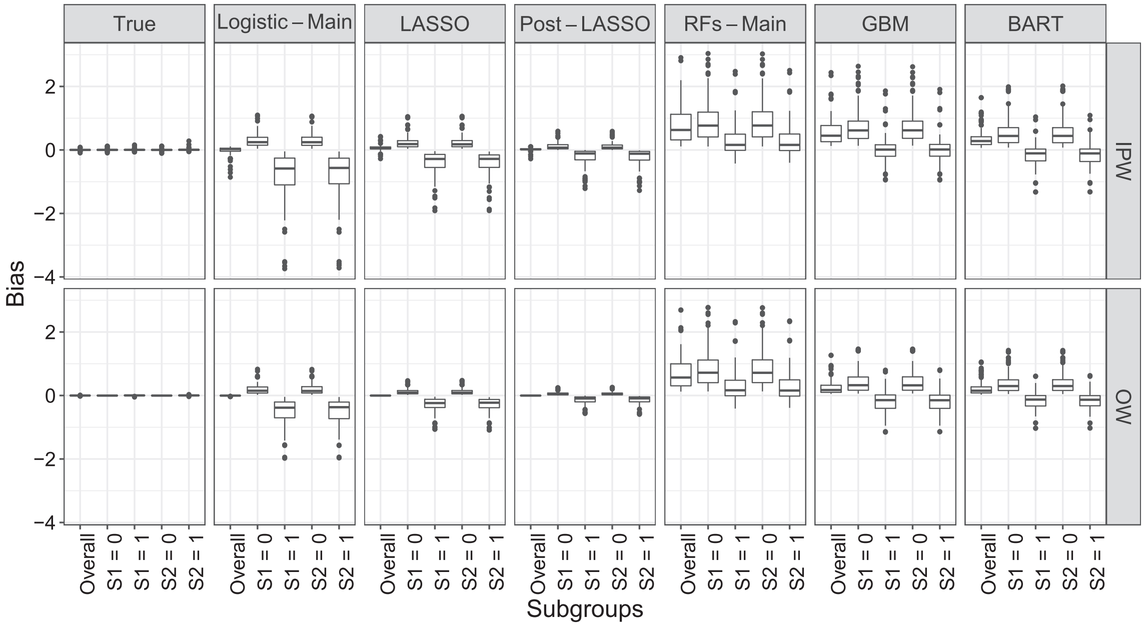


FIGURE 2 Bias in estimating the overall WATE and the four subgroup S-WATE across different postulated propensity models and weighting schemes. Each dot represents one of the 72 simulation scenarios

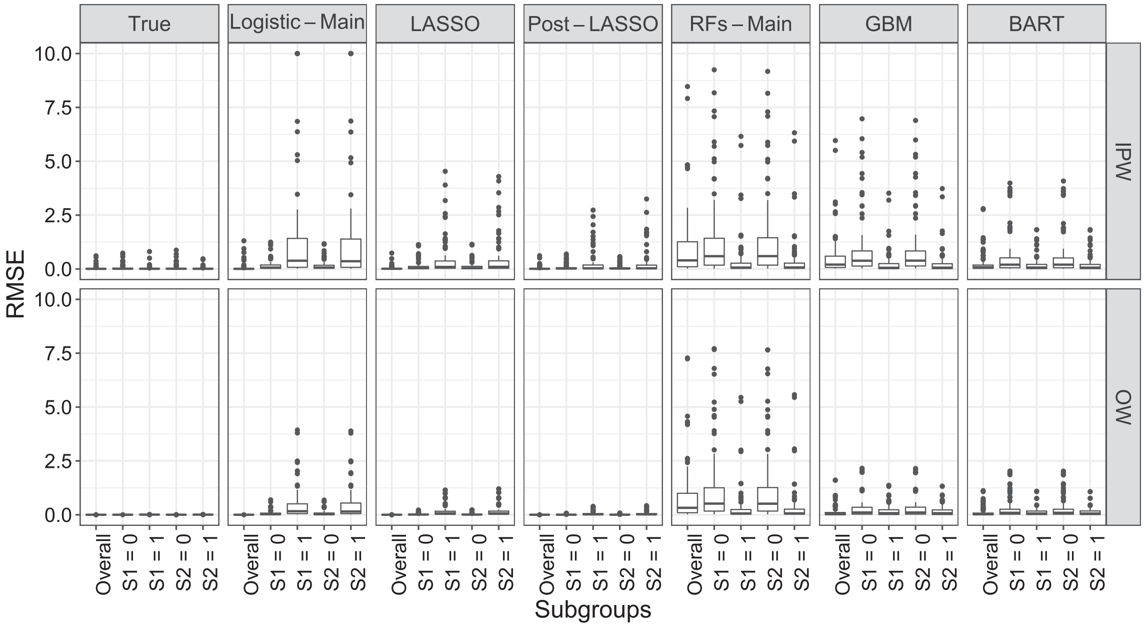


FIGURE 3 Root mean squared error in estimating the overall WATE and the four subgroup S-WATE across different propensity models and weighting schemes. Values greater than 10 are truncated at 10. Each dot represents one of the 72 simulation scenarios

settings. Specifically, we find that the Logistic-Main IPW is much more sensitive to the simulation parameter specification compared with the OW-pLASSO. For example, it leads to substantial bias in estimating S-ATE under scenarios with more confounders and stronger confounding effects (ie, larger P and ψ, larger γ and κ values).

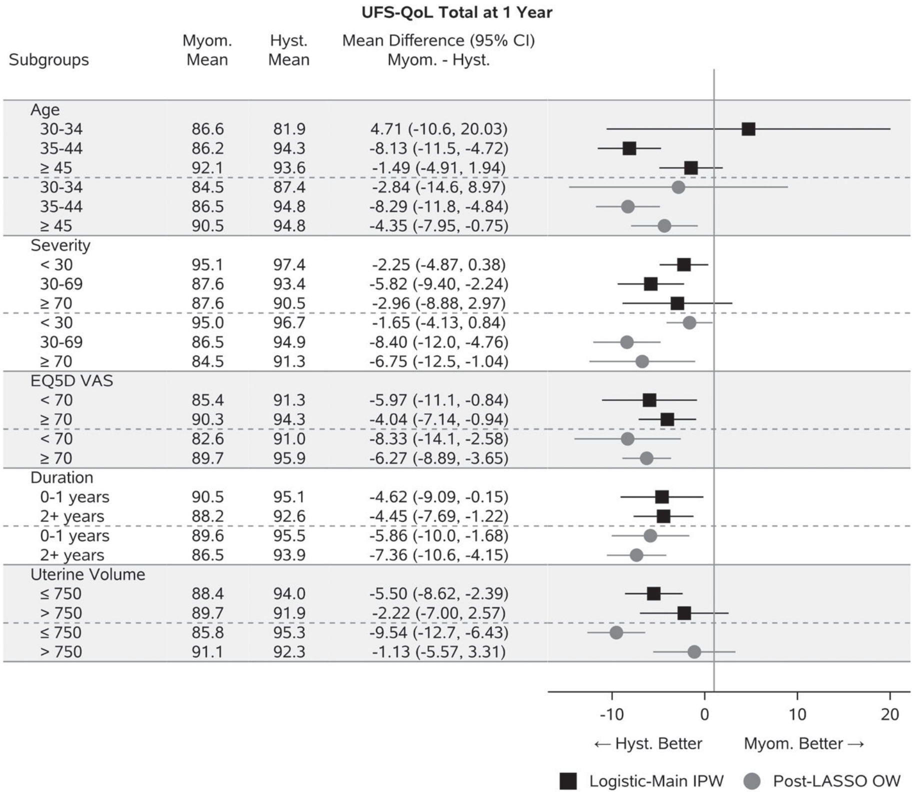
RMSE. From Figure 3 we can see that, with the same propensity score model, the RMSE is generally higher for IPW than for OW. This is expected, due to (i) the improved balance and (ii) the minimum variance property of OW. Neither the Logistic-Main nor the RF models capture the interactions in the true PS model and consequently result in large biases and variances of subgroup effects. This suggests that the RF models under our chosen hyperparameter settings are inadequate in learning the interactions (when given main effects only) or performing variable selection (when given the fully expanded design matrix including subgroup interactions), leading to inaccurate and noisy treatment estimates. By contrast, LASSO coupled with OW provides low bias and high efficiency. Post-LASSO further improves upon LASSO across all the simulation settings we explored. Similarly to the previous observations, magnitude of the RMSE from BART and GBM is between that from the LASSO and Logistic-Main model. Web Figure 5 and 6 demonstrate the RMSE of OW-pLASSO is invariant to regression coefficients, while larger P and ψ, larger γ and κ values greatly increase the RMSE of the IPW main effect model.

To summarize, OW estimators achieve better covariate balance, smaller relative bias and RMSE than IPW estimators across various propensity score models. The proposed method (OW-pLASSO) leads to low bias and high efficiency in estimating subgroup causal effects, suggesting LASSO successfully selects the important subgroup-covariate interactions across simulation scenarios. By contrast, the standard Logistic-Main as well as alternative machine learning models for the propensity scores lead to large bias and RMSE in estimating the subgroup causal effects, particularly under moderate and strong confounding.

6 | APPLICATION TO COMPARE-UF

We now apply the proposed method to our motivating study of myomectomy vs hysterectomy in the 35 prespecified subgroups of COMPARE-UF. In Figure 1C, the balance based on ASMD is substantially improved by OW-pLASSO, though still not perfect. This improvement in balance does not come at the expense of variance. Both overall and within subgroups, the VI metric is lower with OW-pLASSO than the standard main effects logistic regression (panel A). To save space in the comparison of methods we only show six subgroups. Additional results for all subgroups were similar and are available in Web Appendix 2.4. The only subgroup for which good balance was not achieved is age less than 35, though it was improved compared with the other methods. The challenge in balancing this subgroup is not surprising given the limited sample size and extreme imbalances that were initially present. We recommend that comparative statements about this subgroup are made very cautiously.

Figure 4 displays estimated treatment effects for the primary quality of life endpoint, UFS-QOL score 1 year after the procedures with 95% confidence intervals based on the robust sandwich variance estimator.[46] The proposed method,



Subgroups	Myom. Mean	Hyst. Mean	Mean Difference (95% CI) Myom. - Hyst.
Age			
30-34	86.6	81.9	4.71 (-10.6, 20.03)
35-44	86.2	94.3	-8.13 (-11.5, -4.72)
≥ 45	92.1	93.6	-1.49 (-4.91, 1.94)
30-34	84.5	87.4	-2.84 (-14.6, 8.97)
35-44	86.5	94.8	-8.29 (-11.8, -4.84)
≥ 45	90.5	94.8	-4.35 (-7.95, -0.75)
Severity			
< 30	95.1	97.4	-2.25 (-4.87, 0.38)
30-69	87.6	93.4	-5.82 (-9.40, -2.24)
≥ 70	87.6	90.5	-2.96 (-8.88, 2.97)
< 30	95.0	96.7	-1.65 (-4.13, 0.84)
30-69	86.5	94.9	-8.40 (-12.0, -4.76)
≥ 70	84.5	91.3	-6.75 (-12.5, -1.04)
EQ5D VAS			
< 70	85.4	91.3	-5.97 (-11.1, -0.84)
≥ 70	90.3	94.3	-4.04 (-7.14, -0.94)
< 70	82.6	91.0	-8.33 (-14.1, -2.58)
≥ 70	89.7	95.9	-6.27 (-8.89, -3.65)
Duration			
0-1 years	90.5	95.1	-4.62 (-9.09, -0.15)
2+ years	88.2	92.6	-4.45 (-7.69, -1.22)
0-1 years	89.6	95.5	-5.86 (-10.0, -1.68)
2+ years	86.5	93.9	-7.36 (-10.6, -4.15)
Uterine Volume			
≤ 750	88.4	94.0	-5.50 (-8.62, -2.39)
> 750	89.7	91.9	-2.22 (-7.00, 2.57)
≤ 750	85.8	95.3	-9.54 (-12.7, -6.43)
> 750	91.1	92.3	-1.13 (-5.57, 3.31)

FIGURE 4 Estimates and 95% confidence intervals for treatment comparison of Myomectomy to Hysterectomy. Weighted means are reported and then contrasted

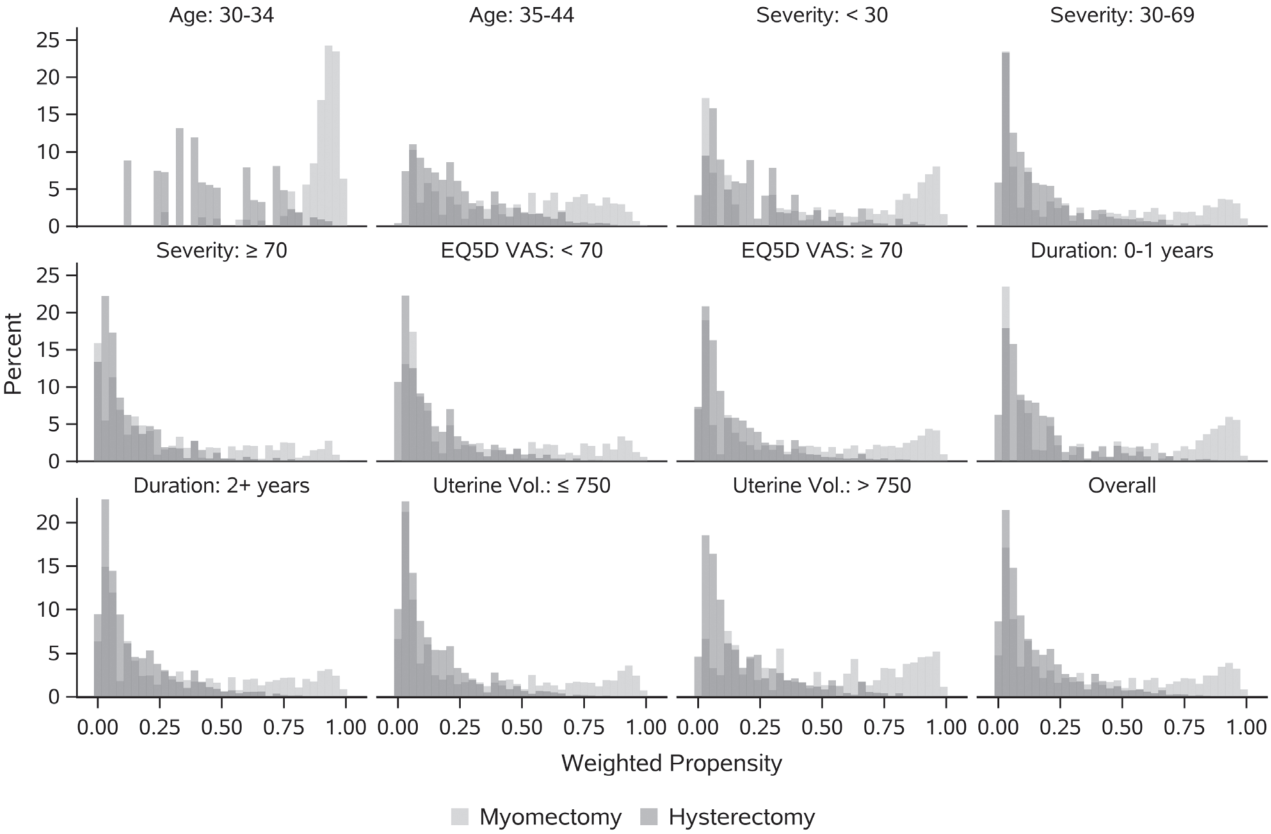


FIGURE 5 Propensity score distributions by treatment after weighting, by Logistic-Main IPW

OW-pLASSO is compared with the standard Logistic-Main IPW. In some subgroups, including many of those not shown, the results of OW-pLASSO confirm those of Logistic-Main IPW. However, some potentially important signals arise. OW-pLASSO reveals different treatment effects in the subgroups defined by baseline symptom severity. Individuals with mild symptom severity (<30) at baseline have similar outcomes with hysterectomy or myomectomy, whereas subgroups with higher initial symptoms (30-69, >70) receive a larger improvement in overall quality of life with hysterectomy. This is expected clinically, as hysterectomy entirely eliminates symptoms whereas symptoms can recur with myomectomy. Those with the greatest initial symptoms would have the most to gain. The results of Logistic-Main IPW did not detect this difference. This is consistent with Figure 1 where covariate imbalances after weighting by Logistic-Main IPW were corrected by OW-pLASSO. A similar pattern was observed for the subgroups based on uterine volume. OW-pLASSO indicated that women with lower uterine volume had significantly larger benefits from hysterectomy. This result is not immediately intuitive, but may be related to the fact that women with lower uterine volume also had higher pain and self-consciousness score at baseline and therefore more to gain from a complete solution. This finding was obscured by Logistic-Main IPW because large imbalances in the baseline covariates favored myomectomy.

The COMPARE-UF data exemplify an additional advantage of OW-pLASSO, in the creation of a clinically relevant target population that emphasizes patients who are reasonably comparable, for all subgroups (S-ATO). To illustrate the shift in target population we display the propensity score distributions by subgroups after weighting. Figure 5 illustrates two features of Logistic-Main IPW: (1) IPW has not made the hysterectomy and myomectomy groups similar; (2) The cohort is dominated by individuals at the extremes, with propensity values near 0 or 1. By contrast, the distributions in Figure 6 (resulting from OW-pLASSO) are mostly overlapping for hysterectomy vs myomectomy and emphasize people with propensity scores away from 0 and 1. While Logistic-Main IPW could be improved by iterative corrections, such as range trimming, or adapting the propensity score model, these steps would be cumbersome in COMPARE-UF to implement manually across 35 subgroups. Instead, OW-pLASSO automatically finds a population at clinical equipoise, for whom comparative data are most essential, across all subgroups. The resulting overlap cohort is displayed through a weighted baseline characteristics table in Web Appendix 2.4.

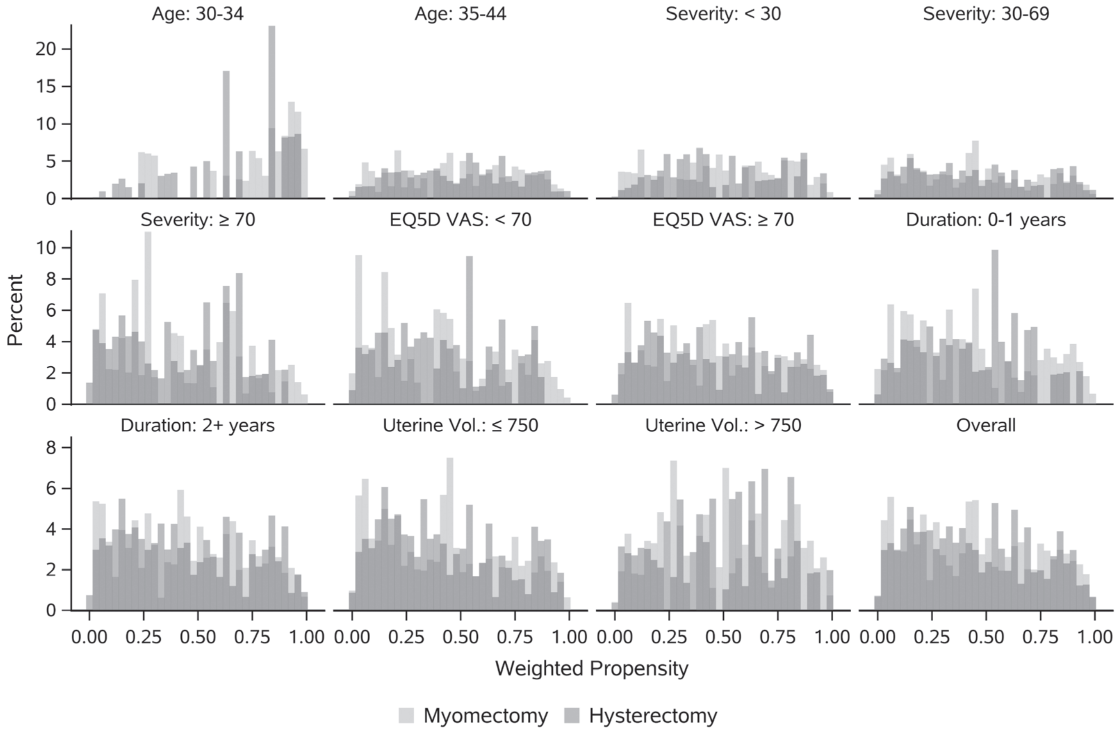


FIGURE 6 Propensity score distributions by treatment after weighting, by OW-pLASSO

7 DISCUSSION

As researchers look for real world evidence of comparative effectiveness in increasingly diverse and heterogeneous populations, it is crucial to advance appropriate methods for causal SGA with observational data. In this article, we developed a suite of propensity score weighting methods and visualization tools for such a goal. We showed that it is essential to balance covariates within a subgroup, which bounds the estimation bias of subgroup causal effects. We further proposed a method that aims to balance the bias-variance trade-off in causal SGA. Our method combines Post-LASSO for selecting the propensity score model and OW for achieving exact balance and efficiency within each subgroup. We conducted extensive simulations to examine the operating characteristics of the proposed method. We found that pairing Post-LASSO with OW performed superior to several other commonly used methods in terms of balance, precision and stability. Our method provides one set of weights that can be used for both population average and subgroup-specific treatment effect estimation. The coupling of substantive knowledge about prespecified subgroups, to generate candidate interactions, as well as machine learning for variable selection, may not only improve SGA but also the validity of the propensity score model for population average comparisons. As we move beyond SGA, using the knowledge of prespecified subgroups to build the propensity score model may reduce bias in a range of propensity-score-based HTE methods.

We emphasized SGA with prespecified subgroups in observational studies, while alternative methods and settings for HTE are rapidly developing. For example, Luedtke and van der Laan[57] showed that studying the additive treatment effect in SGA is similar to solving an optimization question when estimating the mean outcome. Recent research further recommends to select optimal subgroups based on the outcome mean difference between the effects and move away from one-covariate-at-a-time type of SGA.[58] Similar to their idea, our method simultaneous uses all important covariates to make decisions.

The proposed methods maintain the causal inference principle of separating study design from analysis of outcomes. These methods allow an analyst to thoroughly investigate the model adequacy and balance without risk of being influenced by observing various treatment effects. Recent developments in causal inference are moving to incorporate information on the outcome in the propensity score estimation.[59] When the candidate list of covariates is large, and investigators are not able to prioritize covariates, using the outcome data may be helpful. Future research could adapt the proposed method to incorporate outcome information.

We also designed a new diagnostic graph—the Connect-S plot—that allows visualizing subgroup balance for a large number of subgroups and covariates simultaneously. We hope the Connect-S plot and the associated programming code would facilitate more routine check of subgroup balance in CER.

The R and SAS code with implementation details used in this article are provide at: https://github.com/siyunyang/OW_SGA.

ACKNOWLEDGEMENTS

This research is supported by the Patient-Centered Outcomes Research Institute (PCORI) contract ME-2018C2-13289. The COMPARE-UF study, which constitutes our motivating example, was supported by the Agency for Healthcare Research and Quality grant RFA-HS-14-006. The contents of this article are solely the responsibility of the authors and do not necessarily represent the view of PCORI nor AHRQ. We appreciate the clinical input and motivating questions from COMPARE-UF PI Evan Myers and COMPARE-UF investigators.

DATA AVAILABILITY STATEMENT

The COMPARE-UF data is not currently available, but will be made publicly available through PCORI in the near future.

ORCID

Siyun Yang https://orcid.org/0000-0003-2895-532X
Georgia Papadogeorgou https://orcid.org/0000-0002-1982-2245
Fan Li https://orcid.org/0000-0002-0390-3673
Laine E. Thomas https://orcid.org/0000-0002-5340-8742

REFERENCES

1. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *J Am Med Assoc*. 2007;298(10):1209-1212.
2. Kent DM, Rothwell PM, Ioannidis JPA, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*. 2010;11(1):85.
3. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20(1):217-240.
4. Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat*. 2013;7(1):443-470.
5. Schnell PM, Tang Q, Offen WW, Carlin BP. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*. 2016;72(4):1026-1036.
6. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113(523):1228-1242.
7. Lee K, Small DS, Hsu JY, Silber JH, Rosenbaum PR. Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *J Royal Stat Soc Ser A (Stat Soc)*. 2018;181(2):535-546.
8. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat*. 2010;4(1):266-298.
9. Powers S, Qian J, Jung K, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med*. 2018;37(11):1767-1787.
10. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069. https://doi.org/10.1016/S0140-6736(00)02039-0.
11. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Stat Med*. 2002;21(19):2917-2930. https://doi.org/10.1002/sim.1296.
12. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine - reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357(21):2189. https://doi.org/10.1056/NEJMsr077003.
13. Varadhan R, Wang S-J. Standardization for subgroup analysis in randomized controlled trials. *J Biopharm Stat*. 2014;24(1):154-167.
14. Alosh M, Huque MF, Bretz F, D'Agostino RB. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Stat Med*. 2017;36(8):1334-1360. https://doi.org/10.1002/sim.7167.
15. Radice R, Ramsahai R, Grieve R, Kreif N, Sadique Z, Sekhon JS. Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *Int J Biostat*. 2012;8(1):25–25.
16. Dong J, Zhang JL, Zeng S, Li F. Subgroup balancing propensity score. *Stat Methods Med Res*. 2020;29(3):659-676. https://doi.org/10.1177/0962280219870836.
17. Ben-Michael E, Feller A, Rothstein J. Varying impacts of letters of recommendation on college admissions: approximate balancing weights for subgroup effects in observational studies; 2020. arXiv preprint arXiv:2008.04394.
18. Imai K, Ratkovic M. Covariate balancing propensity score. *J Royal Stat Soc Ser B (Stat Methodol)*. 2014;76(1):243-263.
19. Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *J Am Stat Assoc*. 2015;110(511):910-922.

20. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1): 41-55.
21. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat*. 2008;2(3):808-840.
22. Robins J, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc*. 1995;90(429):122-129.
23. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.
24. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcome Res Methodol*. 2001;2:259-278.
25. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71(4):1161-1189.
26. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2018;113(521):390-400. https://doi.org/10.1080/01621459.2016.1260466.
27. Zhao Q. Covariate balancing propensity score by tailored loss functions. *Ann Stat*. 2019;47(2):965-993.
28. Green KM, Stuart EA. Examining moderation analyses in propensity score methods: application to depression and substance use. *J Consult Clin Psychol*. 2014;82(5):773-783. https://doi.org/10.1037/a0036515.
29. Wang SV, Jin Y, Fireman B, et al. Relative performance of propensity score matching strategies for subgroup analyses. *Am J Epidemiol*. 2018;187(8):1799-1807. https://doi.org/10.1093/aje/kwy049.
30. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psycholog Methods*. 2004;9(4):403.
31. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346.
32. Hill J, Weiss C, Zhai F. Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivar Behav Res*. 2011;46(3):477-513.
33. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
34. Belloni A, Chernozhukov VJB. Least squares after model selection in high-dimensional sparse models. *Bernoulli*. 2013;19(2):521-547.
35. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol*. 2019;188(1):250-257.
36. Thomas LE, Li F, Pencina MJ. Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. *J Am Med Assoc*. 2020;323(23):2417-2418.
37. Stewart EA, Lytle BL, Thomas L, et al. The comparing options for management: PAtient-centered REsults for uterine fibroids (COMPARE-UF) registry: rationale and design. *Am J Obstet Gynecol*. 2018;219(1):95-e1.
38. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591-593.
39. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Math Modell*. 1986;7(9-12):1393-1512.
40. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199.
41. Tao Y, Fu H. Doubly robust estimation of the weighted average treatment effect for a target population. *Stat Med*. 2019;38(3):315-325.
42. Thomas LE, Li F, Pencina MJ. Using propensity score methods to create target populations in observational clinical research. *J Am Med Assoc*. 2020;323(5):466-467.
43. Li F, Li F. Propensity score weighting for causal inference with multiple treatments. *Ann Appl Stat*. 2019;13(4):2389-2415.
44. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. New York, NY: Springer; 2013.
45. Hahn PR, Carvalho CM, Puelz D, He J. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal*. 2018;13(1):163-182.
46. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Methods Med Res*. 2017;26(4): 1654-1670.
47. Chatterjee A, Lahiri S. Asymptotic properties of the residual bootstrap for lasso estimators. *Proc Am Math Soc*. 2010;138(12):4497-4509.
48. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal*. 2007;15(3):199-236.
49. 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.
50. Kish L. *Survey sampling*. No. 04; HN29, K5; 1965.
51. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1): 1-22.
52. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
53. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R; 2015. arXiv:1508.04409.
54. Bühlmann P, Yu B. Boosting with the L2 loss: regression and classification. *J Am Stat Assoc*. 2003;98(462):324-339.
55. Ridgeway G, McCaffrey D, Morral A, Griffin BA, Burgette L. twang: toolkit for weighting and analysis of nonequivalent groups. R package version 1.5; 2017.
56. McCulloch R, Sparapani R, Gramacy R, Spanbauer C, Pratola M. BART: Bayesian additive regression trees. R package version 2.7; 2019.
57. Luedtke AR, Laan MJ. Evaluating the impact of treating the optimal subgroup. *Stat Methods Med Res*. 2017;26(4):1630-1640.

58. VanderWeele TJ, Luedtke AR, Laan MJ, Kessler RC. Selecting optimal subgroups for treatment using many covariates. *Epidemiology*. 2019;30(3):334-341.
59. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*. 2017;73(4):1111-1122.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Yang S, Lorenzi E, Papadogeorgou G, Wojdyla DM, Li F, Thomas LE. Propensity score weighting for causal subgroup analysis. *Statistics in Medicine*. 2021;40:4294–4309. https://doi.org/10.1002/sim.9029