Propensity Score Weighting for Causal Subgroup Analysis

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

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

[2]Berry Consultants,Texas, USA

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

[4]Duke Clinical Research Institute, Durham, NC, USA

[5]Department of Statistical Science, Duke University, Durham, NC, USA

ABSTRACT

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 remain underdeveloped, particularly in observational studies. In this article, we develop a suite of analytical methods and visualization tools for causal subgroup analysis. 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 design a new diagnostic graph—the Connect-S plot—for visualizing the subgroup covariate balance. Finally, we propose to use the overlap weighting method to achieve exact balance within subgroups. We further propose a method that combines overlap weighting and LASSO, to balance the bias-variance tradeoff in subgroup analysis. 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.

KEY WORDS: Subgroup analysis, Effect modification, Interaction, Covariate balance, Causal inference, Propensity score, Balancing weights, Overlap weights

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 (Kent and Hayward, 2007; Kent et al., 2010), and in recent years the CER literature has increasingly shifted attention to heterogeneous treatment effects (HTE) (Hill, 2011; Imai and Ratkovic, 2013; Schnell et al., 2016; Wager and Athey, 2018; Lee et al., 2018). 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) (Chipman et al., 2010; Hill, 2011), Causal Forest (Wager and Athey, 2018), and Causal boosting (Powers et al., 2018). 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 *pre-specified*—usually defined using pre-treatment covariates—subgroups of patients. There is an extensive literature on SGA methods in randomized controlled trials (Assmann et al., 2000; Pocock et al., 2002; Wang et al., 2007; Varadhan and Wang, 2014; Alosh et al., 2017). However, causal inference methods for SGA with observational data remain underdeveloped (Radice et al., 2012; Dong et al., 2020; Ben-Michael et al., 2020).

In the context of ATE, covariate balance has been shown to be crucial to unbiased estimation of causal effects(Imai and Ratkovic, 2014; Zubizarreta, 2015). Propensity score methods(Rosenbaum and Rubin, 1983) are the most popular method for achieving covariate balance, but have seldom been discussed in SGA(Radice et al., 2012; Dong et al., 2020). Compared to 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 robustness to mis-specification can be checked through balance diagnostics(Rubin, 2008). In this paper we focus on the propensity score weighting approach(Robins and Rotnitzky, 1995; Robins et al., 2000; Hirano and Imbens, 2001; Hirano et al., 2003; Li et al., 2018; Zhao, 2019). Dong et al. (2020) 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 pre-specified 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 (JAMA)* 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 (Green and Stuart, 2014; Wang et al., 2018). 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 pre-specifying necessary interactions.

For example, generalized boosted models (GBM) have been advocated as a flexible, data-adaptive method(McCaffrey et al., 2004), and random forest was superior to many other tree-based methods for propensity score estimation in extensive simulation studies(Lee et al., 2010). BART have been used to estimate the propensity score model and outperformed GBM on some metrics of balance (Hill et al., 2011). However, it is unclear whether these methods achieve adequate balance and precision in causal SGA. Moreover, when important subgroups are pre-specified, 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). We show that balance of covariates within a subgroup bounds the bias of the estimator of S-WATE (Section 3). Second, we design a new diagnostic graph, which we call the Connect-S plot, for visualizing the subgroup covariate balance (Section 4). Finally, we propose a method that combines LASSO (Tibshirani, 1996) and overlap weighting (Li et al., 2018, 2019; Thomas et al., 2020a), and balances the bias-variance tradeoff in causal SGA (Section 5). Specifically, we treat the pre-specified 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 overlap weighting with a logistic regression to achieve good covariate balance *both* overall and within subgroups, thus reducing bias and variance in causal SGA. We conduct extensive simulation studies to compare the proposed method with several alternative methods (Section 6), and illustrate its application in a motivating example (Section 7)

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 (Stewart et al., 2018). Our goal is to evaluate alternative management options for uterine fibroids for relief of symptoms and quality of life. Subgroup analysis was a primary aim to determine whether certain types of patient subgroups should receive myomectomy versus hysterectomy procedures. Investigators pre-specified 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 subgroup analysis

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 $Z = 0$. 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 pre-treatment covariates, $\boldsymbol{X}_i = (X_{i1}, ..., X_{iP})^T$.

We denote the subgroups of interest by indicator variables $\boldsymbol{S_i} = (S_{i1}, ..., S_{iR})^T$, where $S_{ir} = 1$ if the i^{th} unit is a member of the $r^{th}(r = 1, ..., R)$ subgroup and 0 otherwise (e.g. 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 subgroup analysis where the groups compared are defined as $S_{ir} = 0$ and $S_{ir} = 1$ for each r, while averaging over the levels of $\{S_{i1}, ..., S_{iR}\} \setminus \{S_{ir}\}$. 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 sub-grouping variables $\boldsymbol{S}_i$ may not all be a function of $\boldsymbol{X}_i$. Further, 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 sub-grouping variables $\boldsymbol{S}_i$ are explicitly denoted.

2.2 The estimand: Subgroup weighted average treatment effect

Traditional causal inference methods focus on the average treatment effect (ATE), $\mathbb{E}_f[Y(1) - Y(0)]$, where the expectation is over the sampled population with probability density $f(\boldsymbol{x}, \boldsymbol{s})$ for the covariates and subgroups. Corresponding subgroup analysis 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(Crump et al., 2009; Li et al., 2018; Tao and Fu, 2019; Zhao, 2019; Thomas et al., 2020b). We extend the weighted average treatment effect to the context of subgroup analysis.

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* (Li and Li, 2019), which re-weights 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 subgroup weighted average treatment effect (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 pre-specifying 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 subgroup analysis.

To identify the S-WATE from observational data, we make two standard assumptions(Rosenbaum and Rubin, 1983): (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 non-zero probability of being assigned to each treatment condition. Then, we can estimate the S-WATE in subgroup r, $\tau_{r,h}$, using the Hájek estimator

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

where the weights w are the balancing weights corresponding to the specific tilting function $h(\boldsymbol{x}, \boldsymbol{s})$ (equivalently the target population $g(\boldsymbol{x}, \boldsymbol{s})$)(Li et al., 2018):

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

The most widely used balancing weights are the inverse probability weights (IPW)(Robins et al., 2000), $(w_1 = 1/e(\boldsymbol{x}, \boldsymbol{s}), w_0 = 1/(1 - e(\boldsymbol{x}, \boldsymbol{s}))$, 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 subgroup average treatment effect

(S-ATE). Another balancing weight, which will play a key role in this paper (in Section 5), are the overlap weights (OW), ($w_1 = 1 - e(\boldsymbol{x}, \boldsymbol{s}), w_0 = e(\boldsymbol{x}, \boldsymbol{s})$), corresponding to $h(\boldsymbol{x}, \boldsymbol{s}) = e(\boldsymbol{x}, \boldsymbol{s})(1 - e(\boldsymbol{x}, \boldsymbol{s}))$(Li et al., 2018). The target population of OW is the population with the most overlap in covariates between the treatment and control groups, and the subgroup-specific estimand is the subgroup average treatment effect of the overlap population (S-ATO). These 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 the Web Appendix 1.1 that $\widehat{\tau}_{r,h}$ is consistent for $\tau_{r,h}$.

As we shall show in the next section, covariate balance in the subgroups is crucial for unbiased estimation of the S-WATE. In practice, the 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 $\widehat{w}_i$ based on the estimated propensity score $\widehat{e}(\boldsymbol{X}_i, \boldsymbol{S}_i)$. While balancing the true propensity score would balance the covariates in all covariate-defined subgroups in expectation, the estimated weights $\widehat{w}_i$ based on an estimated propensity score often fail to achieve covariate balance, particularly within subgroups(Dong et al., 2020). Therefore, it may be beneficial to choose weights that guarantee balance. In Section 5 we adapt the overlap weights for this purpose. Before we dive into the question of *how to balance*, we first need address *what to balance*. Specifically, it is necessary to first consider what functions of covariates (e.g. moments) should be balanced in estimating subgroup ATEs and how this differs from estimation of the overall ATE. We address this question in the next Section.

3 Bounding Bias for Subgroup Causal Effects

When focusing on additive models, Zubizarreta (Zubizarreta, 2015) showed that the weighting estimator for the population mean is unbiased when the covariate means are balanced. We extend this work to subgroup analysis by showing that balance of covariates within a subgroup leads to minimal bias of the estimator $\widehat{\tau}_{r,h}$. In Proposition (1), we show this result when the treatment effect is homogeneous within a subgroup ($\tau_{r,h} = \tau_r$), and in Proposition (2) we extend it to allow for within-subgroup effect heterogeneity. In both cases, treatment effects are allowed to vary between subgroup levels.

Proposition 1 *Suppose that the outcome surface satisfies an additive model, e.g.* $Y_i(z) = \sum_{r=1}^{R} \beta_r S_{ir} + \sum_{r=1}^{R} \sum_{p=1}^{P} \beta_{rp} S_{ir} X_{ip} + \sum_{r=1}^{R} \tau_r S_{ir} z + \epsilon_i(z)$, *with* $\mathbb{E}[\epsilon_i(z)|\boldsymbol{X}_i, \boldsymbol{S}_i] = 0$. *For any weight* w_i *that is normalized within subgroups (i.e.* $\sum_{i=1}^{N} Z_i S_{ir} w_i = \sum_{i=1}^{N}(1 - Z_i) S_{ir} w_i = 1$*), if mean balance holds in the* r^{th} *subgroup, expressed as*

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

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

Therefore, any weight for which $\delta \approx 0$ will eliminate bias for SGA when the outcome satisfies an additive model. Proposition (1) illustrates that mean balance in the overall sample, $\left| \sum_{i=1}^{N} Z_i w_i X_{ip} - \sum_{i=1}^{N} (1 - Z_i) w_i X_{ip} \right| < \delta$, 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 $\widehat{\tau}_{r,h}$.

Proposition 2 *Suppose the additive model is relaxed to allow treatment effect heterogeneity by covariates* $\boldsymbol{X}_i$ *within subgroups:* $Y_i(z) = \sum_{r=1}^{R} \beta_r S_{ir} + \sum_{r=1}^{R} \sum_{p=1}^{P} \beta_{rp} S_{ir} X_{ip} + \sum_{r=1}^{R} \tau_r S_{ir} z + \sum_{p=1}^{P} \gamma_{rp} S_{ir} X_{ip} z + \epsilon_i(z)$, *with* $E[\epsilon_i(z)|\boldsymbol{X}_i, \boldsymbol{S}_i] = 0$. *If Condition (4) holds and additionally,*

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

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

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

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 heterogeneous treatment effects this second condition is needed. In addition, this reveals a potential risk of using weights that balance covariates without defining a tilting function and target estimand (S-WATE) (Imai and Ratkovic, 2014; Zubizarreta, 2015; Li et al., 2018; Zhao, 2019). 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 non-linear 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 pre-specify 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 non-linear 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) (Austin and Stuart, 2015), 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, weighted standard deviation. That is

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

where $s_{r,p}$ is the weighted, 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 ASMD$_{r,p}$ should be less than 0.10 or 0.20 (i.e. an acceptable δ is ¡0.10 to 0.20) (Austin and Stuart, 2015).

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 variance inflation due to model complexity. Li et al. (2018) suggested to use the following statistic akin to the "design effect" approximation of Kish (1965) in survey literature to approximate the *variance inflation* (VI):

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

where N_z is the sample size of treatment group z. For the unadjusted estimator, $w_{iz} = 1$ for all units. It is straightforward to define the subgroup-specific version of the variance inflation statistic.

The Connect-S plot for S subgroups resembles the rectangular grid of a Connect4 game: each row represents a subgroup variable, (e.g. 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 (e.g. 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 variance inflation.

Panel (a) of Figure 1 presents the Connect-S plot for COMPARE-UF after adjustment by IPW where the propensity score for myomectomy versus 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 0.25. 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 generalized boosted models and results were similar for random forest and BART methods (Web Appendix 2.3)

5 Combining overlap weighting with LASSO for causal subgroup analysis

To achieve the balance constraints in Section 3 and maintain precision, we propose a method that combines overlap weighting and LASSO for variable selection (Tibshirani, 1996) in the propensity score model. Overlap weighting was defined in Section 2.2, where $h(\boldsymbol{x}, \boldsymbol{s}) = e(\boldsymbol{x}, \boldsymbol{s})(1 - e(\boldsymbol{x}, \boldsymbol{s}))$ in equation (3) and $\tau_{r,h}$ is the subgroup average treatment effect in the overlap population (S-ATO)(Li et al., 2018). The overlap population arises because $h(\boldsymbol{x}, \boldsymbol{s})$ approaches 0 for individuals who are nearly always treated ($e(\boldsymbol{x}, \boldsymbol{s}) = 1$), or never treated ($e(\boldsymbol{x}, \boldsymbol{s}) = 0$) and is maximized for those who are equally likely to be treated or not ($e(\boldsymbol{x}, \boldsymbol{s}) = 0.5$) given their covariates. Thus, the tilting function emphasizes covariate profiles that most overlap between treatment groups. The overlap target 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. When the S-ATO is clinically relevant, its corresponding weighting estimator has attractive properties regarding variance and balance as described below.

First, OWs are naturally bounded between 0 and 1, thus can avoid the issues of extreme weights and large variability that can occur when $h(\boldsymbol{x}, \boldsymbol{s}) = 1$. 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 $\widehat{\tau}_{r,h}$ over all possible h under homoscedasticity (Web Appendix 1.1). This property is particularly attractive for subgroup analysis to mitigate the potential variance inflation that arises from a more complex propensity score model that includes subgroup-covariate interactions, and where subgroup-specific sample sizes are small.

Second, OWs have a desirable small-sample property of exact balance. Specifically, outside the context of SGA, Li et al. (2018) show that when the propensity score is estimated by a logistic regression, overlap weighting leads to exact balance on the weighted covariate means. We extend this property to subgroups as follows.

Proposition 3 *If the postulated propensity score model is logistic regression with subgroup-covariate interactions, i.e.* $\hat{e}(\boldsymbol{X}_i, \boldsymbol{S}_i) = logit^{-1}(\hat{\alpha}_0 + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}_{\boldsymbol{x}} + \boldsymbol{S}_i^T \hat{\boldsymbol{\alpha}}_{\boldsymbol{s}} + (\boldsymbol{X}_i \cdot \boldsymbol{S}_i)^T \hat{\boldsymbol{\alpha}}_{\boldsymbol{xs}})$*, 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 OWs lead to exact mean balance in the subgroups and overall:*

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

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

Proposition 3 implies that when logistic regression is augmented to include $(\boldsymbol{X}_i \cdot \boldsymbol{S}_i)$ and paired with OW, exact balance is achieved overall and within subgroups, i.e. $\delta = 0$ in Proposition 1. Additionally, 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 pre-specified 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 in the treatment effect estimates. This problem is mitigated by the OW. Nonetheless, when the number of covariates and/or subgroups is large, variable selection in the propensity score model is necessary.

We propose starting with a propensity model that has all pairwise interactions between covariates and prespecified subgroup candidates, and then use LASSO to select important interactions. This approach helps achieve covariate balance in the subgroups and mitigate the over-fitting issue in propensity score model. Note, in causal settings regularization inadvertently biases treatment effect estimates by over-shrinking regression coefficients (Hahn et al., 2018). Hence, we adopt the Post-LASSO approach(Belloni and Chernozhukov, 2013; James et al., 2013): we refit the logistic regression with LASSO selected covariate-subgroup pairs to maintain the overlap weights' exact balance property.

6 Simulations

We compare the proposed method (referred to as Post-LASSO OW hereafter) with a number of popular machine learning propensity score methods via simulations under different levels of confounding, sparsity and heterogeneity in causal SGA.

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

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 (i.e. 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 (e.g. 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 the Web Appendix 2.1 we show that these specifications lead to treated and control units with various overlapping true propensity score distributions.

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

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

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 maximum likelihood (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(Tibshirani, 1996) 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 ten-fold cross validation is used for hyperparameter tuning;(Friedman et al., 2010) (4) Post-LASSO: Logistic regression model fitted via ML with the variables selected from the preceding LASSO(Belloni et al., 2013); (5) RF-Main: Random Forest (RF) (Breiman, 2001; Wager and Athey, 2018) with the design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i)$, implemented by R package *ranger* with default hyperparameters and 1000 trees(Wright and Ziegler, 2015); (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 RFs performance depends little on the provided design matrix. For simplicity, we omit results on RF-All; (7) GBM: Generalized boosted model (GBM) (Bühlmann and Yu, 2003; McCaffrey et al., 2004) 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(Ridgeway et al., 2017); (8) BART: Bayesian additive regression trees (Chipman et al., 2010) with the design matrix $(\boldsymbol{X}_i, \boldsymbol{S}_i)$, using the R function *pbart* in package *BART* with default hyperparameters(McCulloch et al., 2019).

Each of the preceding propensity score models is paired with (a) inverse probability of treatment weighting (IPW) and (b) overlap weighting (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 data sets 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 data sets, 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.

6.2 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, Figure 2, and Figure 3, respectively.

Balance. From Web Figure 2, we can see that 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 true confounders in all subgroups. This is expected given OW's exact balance property for any included covariate-subgroup interactions (proposition 3). 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 (i.e. $S_1 = 0$ 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 Post-LASSO OW 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 RFs model. Web Figure 3-4 provides more details of subgroup bias across a range of settings. Specifically, we find that the Logistic-Main IPW is much more sensitive to the simulation parameter specification compared to the Post-LASSO OW. For example, it leads to substantial bias in estimating S-ATE under scenarios with more confounders and stronger confounding effects (i.e. 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. In 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 5-6 demonstrate the RMSE of Post-LASSO OW 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 (Post-LASSO OW) leads to low bias and high efficiency in estimating subgroup causal effects, suggesting LASSO successfully selects the important subgroup-covariate interactions across simulation scenarios. In 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.

7 Application to COMPARE-UF

We now apply the proposed method to our motivating study of myomectomy versus hysterectomy in the 35 pre-specified subgroups of COMPARE-UF. In panel (c) Figure 1 the balance based on ASMD is substantially improved by OW with Post-LASSO though still not perfect. To save space in the comparison of methods we only show 6 subgroups. Additional results for all subgroups were similar and are available in the Web Appendix 2.3. The only subgroup for which good balance was not achieved is age less than 35, though it was improved compared to other methods. The challenge in balancing this subgroup is not surprising given the limited sample size and extreme imbalances that were initially present. We would recommend that comparative statements about this subgroup should be made very cautiously.

Figure 4 displays estimated treatment effects for the primary quality of life endpoint, UFS-QOL score one year after the procedures. The proposed method, Post-LASSO OW is compared to the standard Logistic-Main IPW. In some subgroups, including many of those not shown, the results of Post-LASSO OW confirm those of Logistic-Main IPW. However, some potentially important signals arise. Post-LASSO OW 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 Post-LASSO OW. A similar pattern was observed for the subgroups based on uterine volume. Post-LASSO OW 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 Post-LASSO OW, 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. In contrast, the distributions in Figure 6 (resulting from Post-LASSO OW) are mostly overlapping for hysterectomy versus 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, Post-LASSO OW 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 2.3.

8 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 subgroup analysis with observational data. In this paper 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 subgroup analysis. Our method combines Post-LASSO for selecting the propensity score model and overlap weighting for achieving exact balance within each subgroup. We conducted extensive simulations to examine the operating characteristics of the proposed method. We found that pairing Post-LASSO with overlap weighting performed superior to several other commonly used methods in terms of balance, precision and stability. Our method automatically provides weights that can be used across complimentary analyses of population ATE and subgroup-specific effects. It is particularly relevant to clinically meaningful subgroups which physicians have the subject matter knowledge. The coupling of prior information, 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 pre-specified subgroups to build the propensity score model may reduce bias in a range of propensity-score-based HTE methods.

We emphasized SGA with pre-specified subgroups in observational studies, while alternative methods and settings for HTE are rapidly developing. For example, Luedtke and van der Laan (2017) 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 (VanderWeele et al., 2019). 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 (Shortreed and Ertefaie, 2017). 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 comparative effectiveness research.

The Web appendix and R code with implementation details used in this paper are provide at: `https://github.com/siyunyang/OW_SGA`.

Acknowledgments

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.

References

Alosh, M., Huque, M. F., Bretz, F., and D'Agostino, R. B. (2017), "Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials," *Statistics in Medicine*, 36, 1334–1360.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000), "Subgroup analysis and other (mis)uses of baseline data in clinical trials," *Lancet*, 355, 1064–1069.

Austin, P. C. and Stuart, E. A. (2015), "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies," *Statistics in medicine*, 34, 3661–3679.

Belloni, A., Chernozhukov, V., et al. (2013), "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A. and Chernozhukov, V. J. B. (2013), “Least squares after model selection in high-dimensional sparse models,” 19, 521–547.

Ben-Michael, E., Feller, A., and Rothstein, J. (2020), “Varying impacts of letters of recommendation on college admissions: Approximate balancing weights for subgroup effects in observational studies,” *arxiv*, 2008.04394.

Breiman, L. (2001), “Random forests,” *Machine learning*, 45, 5–32.

Bühlmann, P. and Yu, B. (2003), “Boosting with the L2 loss: regression and classification,” *Journal of the American Statistical Association*, 98, 324–339.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), “BART: Bayesian additive regression trees,” *The Annals of Applied Statistics*, 4, 266–298.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), “Dealing with limited overlap in estimation of average treatment effects,” *Biometrika*, 96, 187–199.

Dong, J., Zhang, J. L., Zeng, S., and Li, F. (2020), “Subgroup Balancing Propensity Score,” *Statistical Methods in Medical Research*, 29, 659–676.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of Statistical Software*, 33, 1–22, URL http://www.jstatsoft.org/v33/i01/.

Green, K. M. and Stuart, E. A. (2014), “Examining Moderation Analyses in Propensity Score Methods: Application to Depression and Substance Use,” *Journal of Consulting and Clinical Psychology*, 82, 773–783, URL <GotoISI>://WOS:000342502100004.

Hahn, P. R., Carvalho, C. M., Puelz, D., He, J., et al. (2018), “Regularization and confounding in linear regression for treatment effect estimation,” *Bayesian Analysis*, 13, 163–182.

Hill, J., Weiss, C., and Zhai, F. (2011), “Challenges with propensity score strategies in a high-dimensional setting and a potential alternative,” *Multivariate Behavioral Research*, 46, 477–513.

Hill, J. L. (2011), “Bayesian nonparametric modeling for causal inference,” *Journal of Computational and Graphical Statistics*, 20, 217–240.

Hirano, K. and Imbens, G. W. (2001), “Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization,” *Health Services and Outcomes Research Methodology*, 2, 259–278.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161–1189.

Imai, K. and Ratkovic, M. (2013), “Estimating treatment effect heterogeneity in randomized program evaluation,” *The Annals of Applied Statistics*, 7, 443–470.

— (2014), “Covariate balancing propensity score,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 243–263.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An introduction to statistical learning*, volume 112, Springer.

Kent, D. M. and Hayward, R. A. (2007), “Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification,” *Journal of American Medical Association*, 298, 1209–1212.

Kent, D. M., Rothwell, P. M., Ioannidis, J. P., Altman, D. G., and Hayward, R. A. (2010), “Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal,” *Trials*, 11, 85.

Kish, L. (1965), *Survey sampling*, number 04; HN29, K5.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010), "Improving propensity score weighting using machine learning," *Statistics in medicine*, 29, 337–346.

Lee, K., Small, D. S., Hsu, J. Y., Silber, J. H., and Rosenbaum, P. R. (2018), "Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 535–546.

Li, F. and Li, F. (2019), "Propensity score weighting for causal inference with multiple treatments," *The Annals of Applied Statistics*, 13, 2389–2415.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018), "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113, 390–400, URL `<GotoISI>://WOS:000438960500039`.

Li, F., Thomas, L. E., and Li, F. (2019), "Addressing extreme propensity scores via the overlap weights," *American journal of epidemiology*, 188, 250–257.

Luedtke, A. R. and van der Laan, M. J. (2017), "Evaluating the impact of treating the optimal subgroup," *Statistical methods in medical research*, 26, 1630–1640.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity score estimation with boosted regression for evaluating causal effects in observational studies," *Psychological methods*, 9, 403.

McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer, C., and Pratola, M. (2019), *BART: Bayesian Additive Regression Trees*, URL `https://CRAN.R-project.org/package=BART`. R package version 2.7.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002), "Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems," *Statistics in Medicine*, 21, 2917–2930.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018), "Some methods for heterogeneous treatment effect estimation in high dimensions," *Statistics in medicine*, 37, 1767–1787.

Radice, R., Ramsahai, R., Grieve, R., Kreif, N., Sadique, Z., and Sekhon, J. S. (2012), "Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach," *The international journal of biostatistics*, 8.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., and Burgette, L. (2017), *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*, URL `https://CRAN.R-project.org/package=twang`. R package version 1.5.

Robins, J. and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rubin, D. B. (2008), "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2, 808–840.

Schnell, P. M., Tang, Q., Offen, W. W., and Carlin, B. P. (2016), "A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects," *Biometrics*, 72, 1026–1036.

Shortreed, S. M. and Ertefaie, A. (2017), "Outcome-adaptive lasso: Variable selection for causal inference," *Biometrics*, 73, 1111–1122.

Stewart, E. A., Lytle, B. L., Thomas, L., Wegienka, G. R., Jacoby, V., Diamond, M. P., Nicholson, W. K., Anchan, R. M., Venable, S., Wallace, K., et al. (2018), "The Comparing Options for Management: PAtient-centered REsults for Uterine Fibroids (COMPARE-UF) registry: rationale and design," *American journal of obstetrics and gynecology*, 219, 95–e1.

Tao, Y. and Fu, H. (2019), "Doubly robust estimation of the weighted average treatment effect for a target population," *Statistics in medicine*, 38, 315–325.

Thomas, L. E., Li, F., and Pencina, M. J. (2020a), "Overlap weighting: A propensity score method that mimics attributes of a randomized clinical trial," *Journal of the American Medical Association*, 323, 2417–2418.

— (2020b), "Using propensity score methods to create target populations in observational clinical research," *Journal of the American Medical Association*, 323, 466–467.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

VanderWeele, T. J., Luedtke, A. R., van der Laan, M. J., and Kessler, R. C. (2019), "Selecting optimal subgroups for treatment using many covariates," *Epidemiology*, 30, 334–341.

Varadhan, R. and Wang, S.-J. (2014), "Standardization for subgroup analysis in randomized controlled trials," *Journal of biopharmaceutical statistics*, 24, 154–167.

Wager, S. and Athey, S. (2018), “Estimation and inference of heterogeneous treatment effects using random forests,” *Journal of the American Statistical Association*, 113, 1228–1242.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007), “Statistics in medicine - Reporting of subgroup analyses in clinical trials,” *New England Journal of Medicine*, 357, 2189.

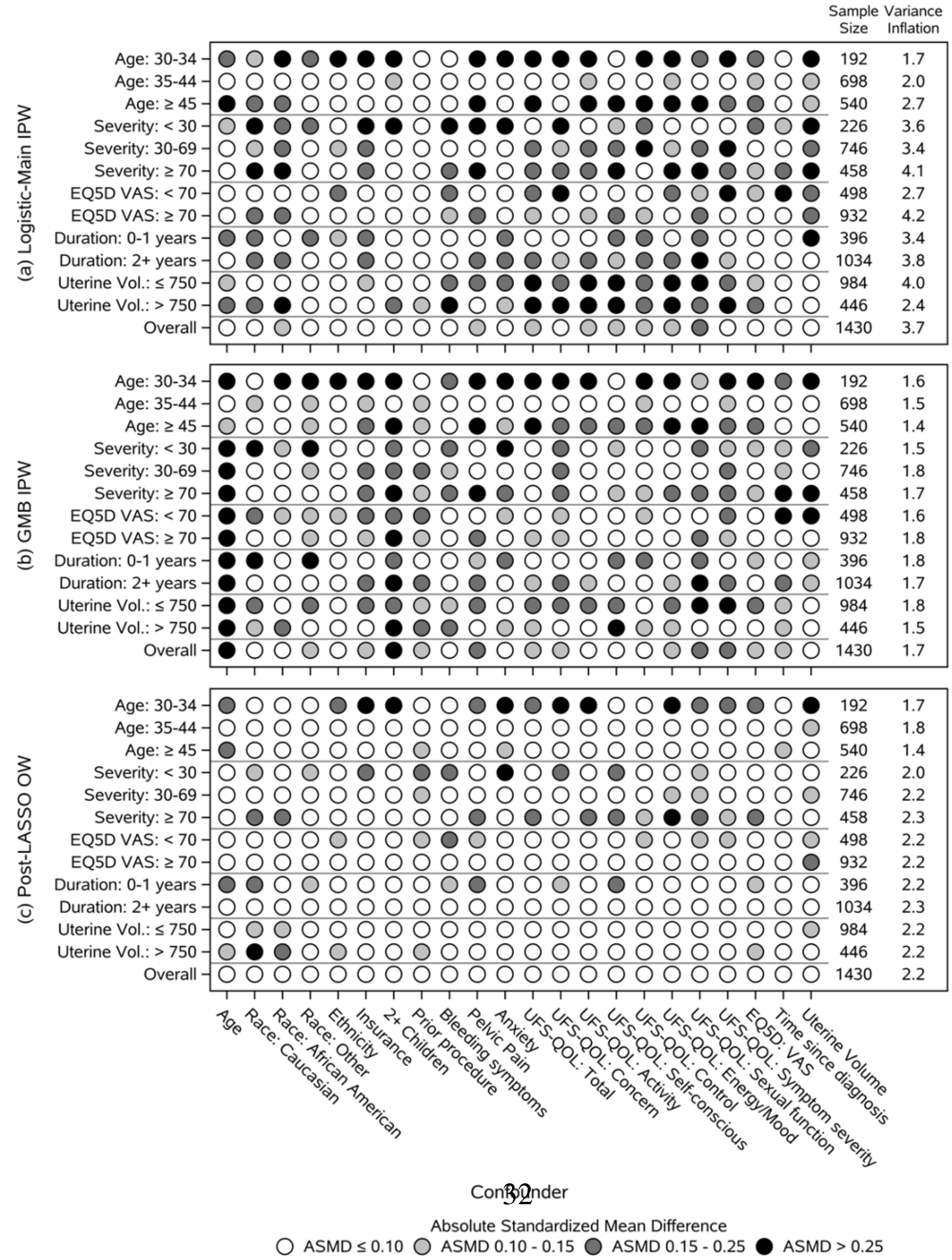
Wang, S. V., Jin, Y., Fireman, B., Gruber, S., He, M., Wyss, R., Shin, H., Ma, Y., Keeton, S., Karami, S., Major, J. M., Schneeweiss, S., and Gagne, J. J. (2018), “Relative Performance of Propensity Score Matching Strategies for Subgroup Analyses,” *American Journal of Epidemiology*, 187, 1799–1807, URL `http://dx.doi.org/10.1093/aje/kwy049https://watermark.silverchair.com/kwy049.pdf`.

Wright, M. N. and Ziegler, A. (2015), “Ranger: a fast implementation of random forests for high dimensional data in C++ and R,” *arXiv:1508.04409*.

Zhao, Q. (2019), “Covariate balancing propensity score by tailored loss functions,” *The Annals of Statistics*, 47, 965–993.

Zubizarreta, J. R. (2015), “Stable weights that balance covariates for estimation with incomplete outcome data,” *Journal of the American Statistical Association*, 110, 910–922.

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 generalized boosted models with IPW; (c) Post-LASSO OW, propensity score estimated by post-LASSO with OW. Select subgroups are displayed in rows and all confounders are displayed in columns.



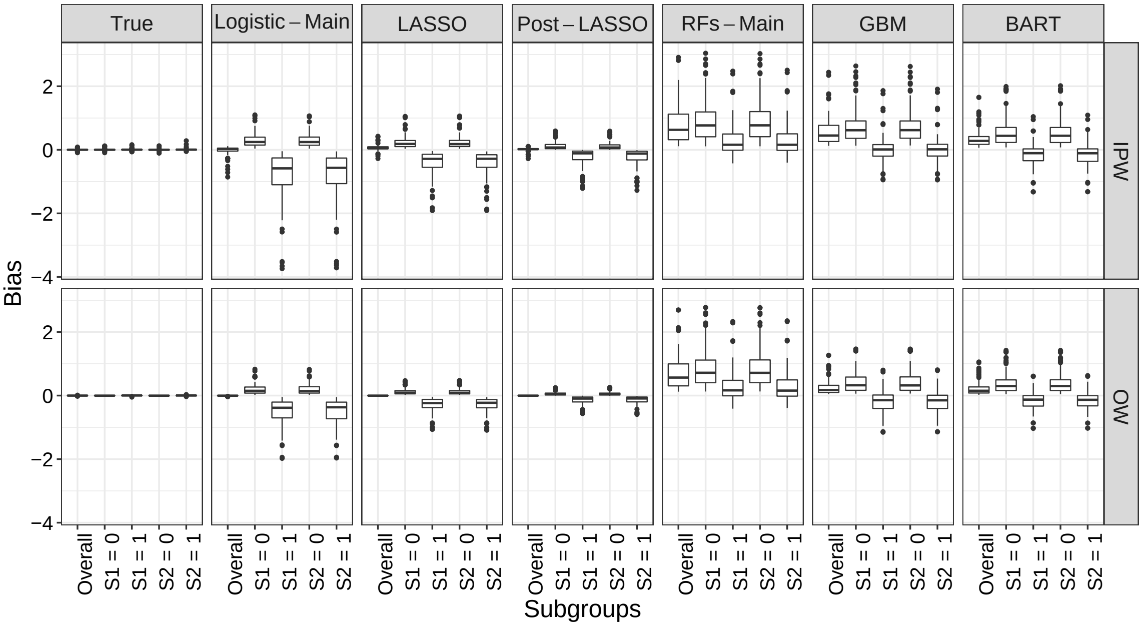


Figure 2: Bias in estimating the overall WATE and the four subgroup S-WATE across different postulated propensity models and weighting schemes. Each dot represents one of the 72 simulation scenarios.

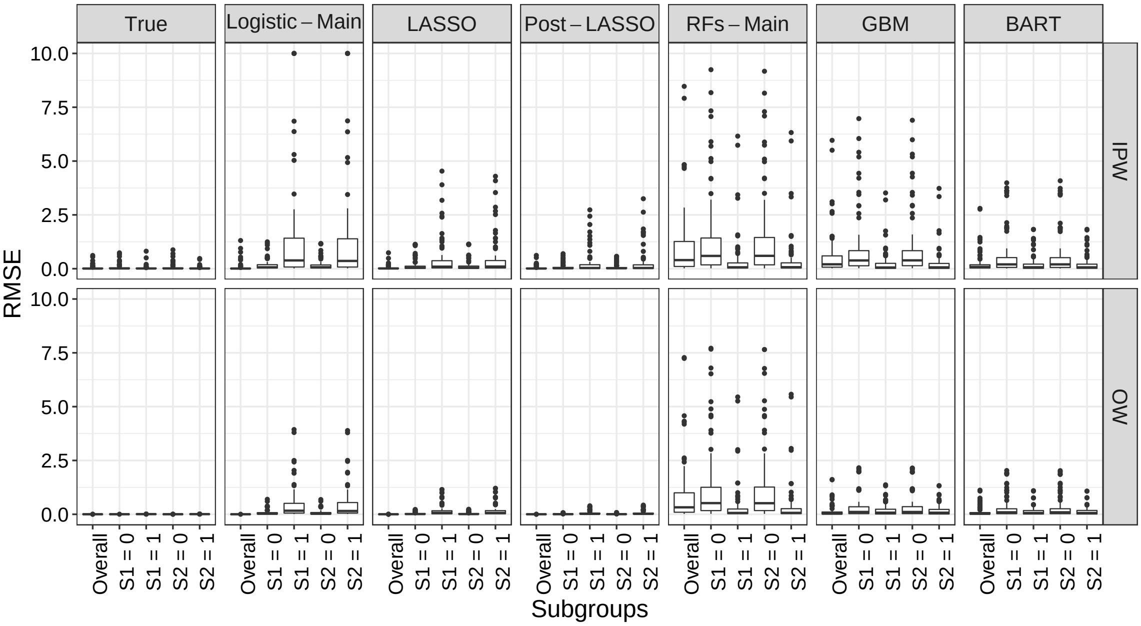
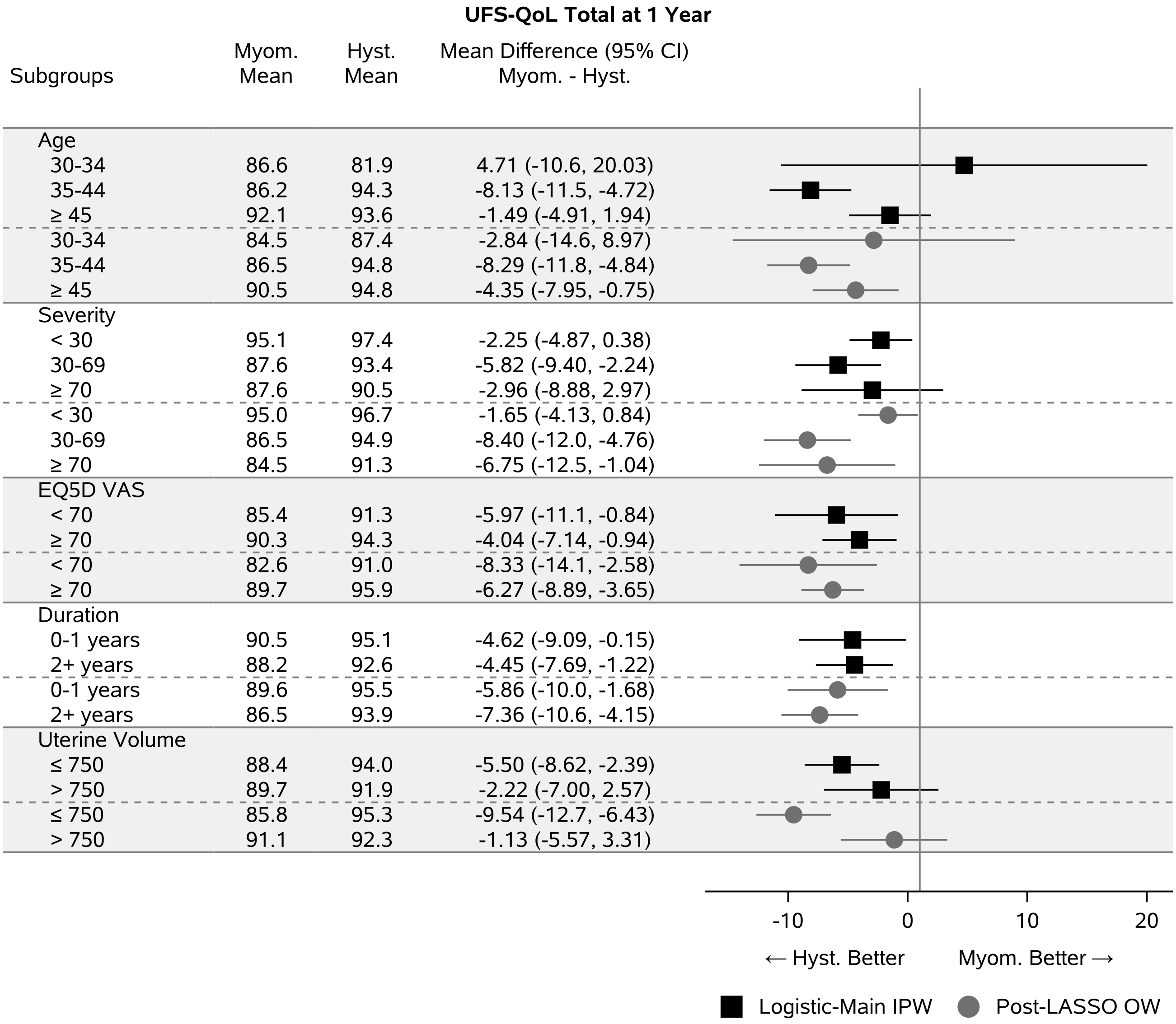


Figure 3: RMSE in estimating the overall WATE and the four subgroup S-WATE across different propensity models and weighting schemes. Values greater than 10 are truncated at 10. Each dot represents one of the 72 simulation scenarios.



UFS-QoL Total at 1 Year

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.

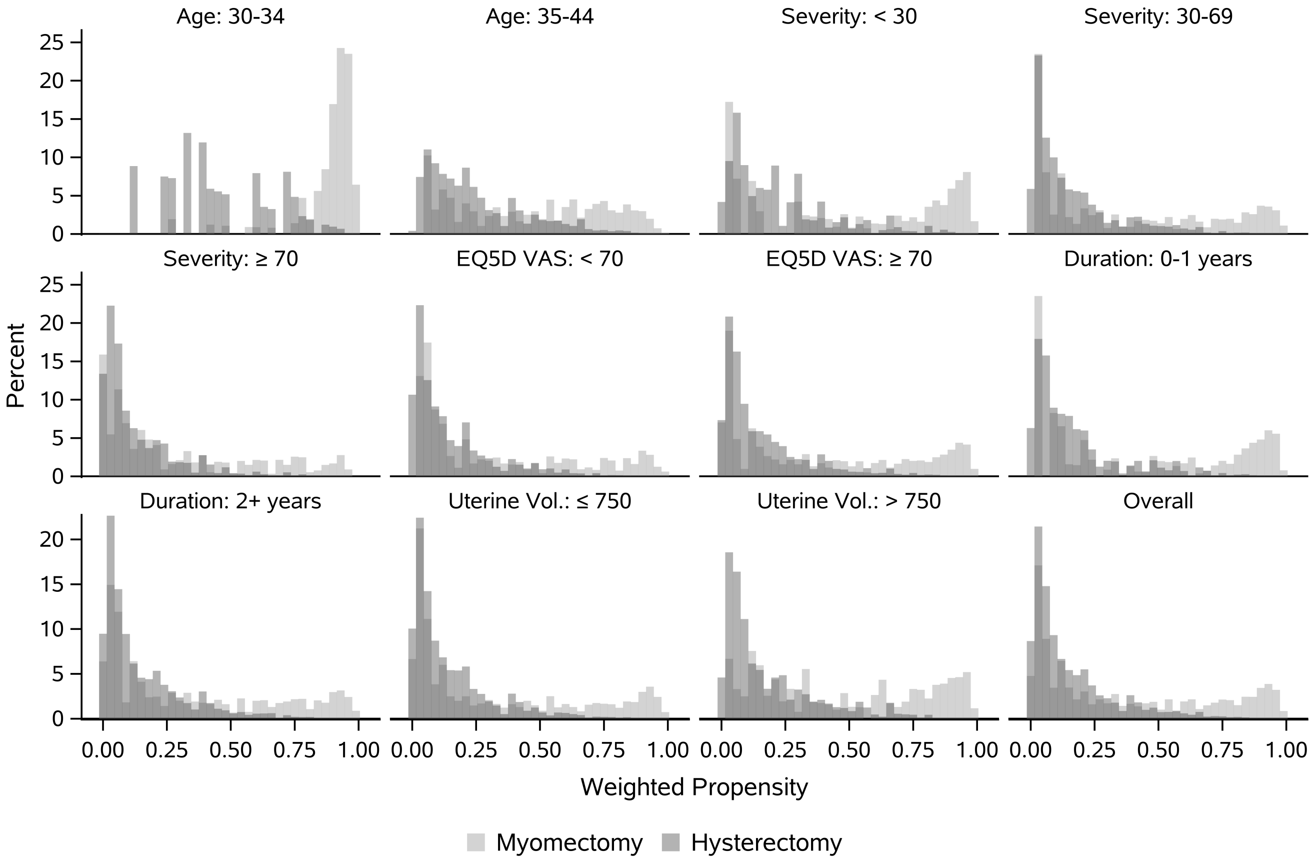


Figure 5: Propensity score distributions by treatment after weighting, by Logistic-Main IPW.

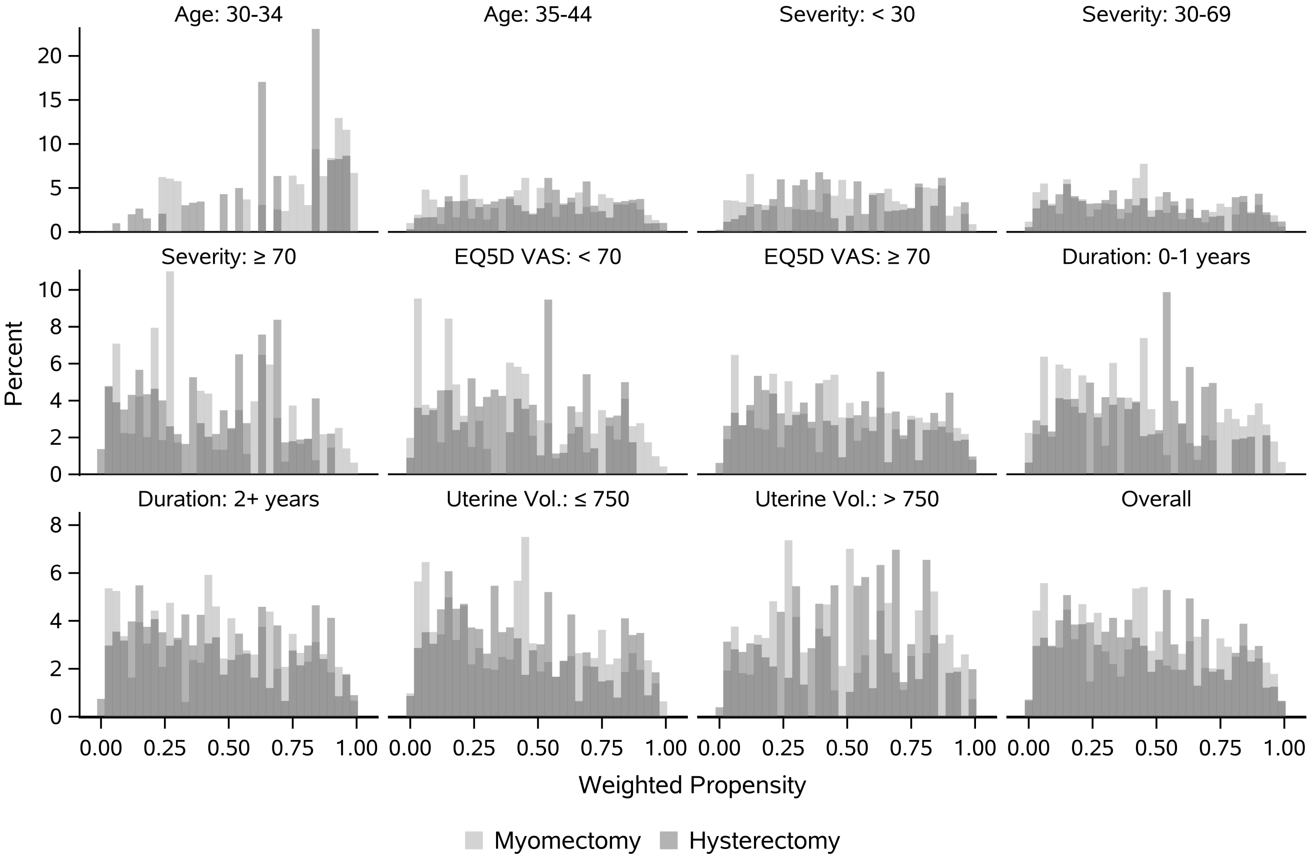


Figure 6: Propensity score distributions by treatment after weighting, by Post-LASSO OW.