

Propensity score weighting with multilevel data

Fan Li ^{1*}, Alan M. Zaslavsky ², Mary Beth Landrum ²

1. Department of Statistical Science, Duke University

Durham, NC 27708, USA

**fli@stat.duke.edu*

2. Department of Health Care Policy, Harvard Medical School

Boston, MA 02115, USA

December 20, 2011

ABSTRACT. Propensity score methods are being increasingly used as a less parametric alternative to traditional regression to balance observed differences across groups in medical and health policy research. Data collected in these disciplines are often multilevel in analytically relevant ways. The propensity score, however, has been developed and used primarily with unstructured data. We present and compare several propensity-score-weighted estimators in the context of hierarchically structured data, including marginal, cluster-weighted and doubly-robust estimators. Using both analytical derivations and Monte Carlo simulations, we illustrate bias arising when the usual assumptions of propensity score analysis do not hold for multilevel data. We show that exploiting the multilevel structure, either parametrically or nonparametrically, in at least one stage of the propensity score analysis can greatly reduce these biases. These methods are applied to a study of racial disparities in breast cancer screening among beneficiaries in Medicare health plans.

KEY WORDS: causal inference, confounding, multilevel, propensity score, racial disparity, weighting.

1 Introduction

Population-based observational studies often are the best methodology for studies of access to, patterns of, and outcomes from medical care. Observational data are increasingly being used for causal inferences, such as comparative effectiveness studies where the goal is to estimate the causal effect of receiving alternative treatment on patient outcomes. However, the propensity score methodology is not limited to causal analysis. For example, a common objective in descriptive studies is to conduct a controlled and unconfounded comparison, without a necessarily causal or randomization interpretation, e.g. comparing outcomes in two different years or different races. Here the propensity score is a tool to achieve balance in distributions of covariates in different groups.

Whether the purpose of the study is description or causal, comparisons between groups can be biased, however, when the groups are unbalanced with respect to confounders. Standard analytic methods adjust for observed differences between groups by stratifying or matching patients on a few observed covariates or by regression adjustments. But if groups differ greatly in observed characteristics, estimates of differences between groups from regression models rely on model extrapolations that can be sensitive to model misspecification [1]. Propensity score methods [2, 3] have been proposed as a less parametric alternative to regression adjustment and are being increasingly used in health policy studies [4, 5, and references therein]. These methods weight, stratify or match subjects according to their propensity for group membership (i.e. to receive treatment or to be in a minority racial group) to balance the distributions of observed characteristics across groups.

Propensity score methods were developed and have been applied in settings with unstructured data. However, data collected in medical care and health policy studies are typically clustered in ways that may be relevant to the analysis, for example by geographical area, treatment center (hospital or physician), or in the example we consider in this paper, health plan.

The unknown mechanism that assigns subjects to clusters may be associated with measured subject characteristics that we are interested in (e.g., race, age, clinical characteristics), measured subject characteristics that are not of intrinsic interest and are believed to be unrelated to outcomes except through their effects on assignment to clusters (e.g., location), and unmeasured subject characteristics (e.g., unmeasured severity of disease, aggressiveness in seeking treatment).

Such clustering or multilevel structure raises several issues. First, if clusters are randomly sampled, standard error calculations that ignore clustering will be inaccurate. A more interesting set of issues arises because measured and unmeasured factors may create cluster-level variation in treatment quality and/or patient outcomes. Multilevel regression models have been developed to give a more comprehensive description than non-hierarchical models provide for such data [6]. Despite the increasing popularity of propensity score analyses and the vast literature regarding regional and provider variation in medical care and health policy research [7, 8], the implications of such data structures for propensity score analyses have not been intensively studied, except for a few unpublished manuscripts and conference papers [9, 10]. An exception is Arpino and Mealli (2011) [11], who addressed this issue explicitly through extensive Monte Carlo simulations that illustrated the benefit of respecting the clustered structure in propensity score matching to protect against bias due to unmeasured cluster-level confounders. Similarly, in some settings one individual's outcome may depend on the treatment assignment of other individuals within the cluster (as in studies of behavioral outcomes or infectious disease where units in the same cluster may influence each other [12]), violating the commonly invoked *stable unit treatment value assumption* (SUTVA) [13].

We focus on propensity score weighting strategies that are widely used in medical care, health policy research and economics [14, 15, 16, 17] and argue that ignoring the multilevel structure can lead to biased estimates. To simplify, we focus on two-level structures. We

investigate the performance of different modeling and weighting strategies in the presence of model misspecification due to cluster-level confounders, both measured and unmeasured, and violations of the SUTVA.

In Section 2, we introduce a study of racial disparities in receipt of breast cancer screening and motivate the use of propensity score methods to account for multilevel structure. In Section 3, we present propensity-score-weighting analogues to some standard regression models for clustered data, including marginal, cluster-weighted and doubly-robust estimators. Section 4 analytically illustrates the bias caused by ignoring clustering in a simple scenario without observed covariates. Section 5 presents a comprehensive Monte Carlo simulation study to examine the performance of the estimators under model misspecification due to observed and unobserved cluster-level covariates. We then apply the methods to the racial disparities study in Section 6. Section 7 concludes with a general discussion.

2 Motivating application

Our motivating application is based on the HEDIS[®] measures of care provided by Medicare health plans. Each of these measures is an estimate of the rate at which a guideline-recommended clinical service is provided to the appropriate population. We obtained individual-level data from the Centers for Medicare and Medicaid Services (CMS) on breast cancer screening of women in these plans [18]. We focused on the difference between whites and blacks, and subjects of other races, for whom racial identification is unreliable in this dataset, were excluded. We also restricted the analysis to plans with at least 25 eligible white enrollees and 25 eligible blacks, leaving 64 plans. To avoid domination of the results by a few very large plans, we drew a random subsample of size 3000 from each of the three large plans with more than 3000 eligible subjects, leaving a total sample size of 56,480.

In a simple comparison, 39.3% of eligible black women did not undergo breast cancer screening compared to 33.5% of white women. Suppose, however, that we are interested in comparing these rates for black and white women with similar distributions of as many covariates as possible. The unadjusted difference in receipt of recommended services ignores observed differences in individual (for example, age and eligibility for Medicaid) and cluster (geographic region, tax status, provider practice model) characteristics between black and white women. Standard propensity score analyses would account for these observed differences, but there may also be unobserved differences among plans related to quality. When there are such unmeasured confounders, the propensity score model is likely to be misspecified, leading to inaccurate estimates of the propensity score and differences across groups. For example, analyses that ignore variation in minority enrollment across the plans would attribute these plan effects to race-based difference in treatment. Misspecification can also arise from assuming an incorrect functional form. Sensitivity to misspecification of the propensity score model for unclustered data was examined in [19]. Clustering widens the range of model choices in each step of a propensity-score analysis, as described in the next section.

The goal of the following analyses is to assess the difference in the proportion undergoing breast cancer screening between whites and blacks, controlled for individual and plan level effects. Race is not a “treatment” in the conventional sense of causal inference, because it is not manipulable [20]. Hence our goal is not to establish a causal relationship between race and health service utilization, but simply to estimate the difference in rates under balanced distributions of covariates between the two races. Causal inference is the goal in many other applications, such as comparative effectiveness studies that seek to estimate causal effect of receiving alternative treatments on patient outcomes. The propensity score is a powerful tool to balance the covariate distribution between groups for studies with either causal or non-causal purposes, and our propensity score methods are applicable in both settings.

3 Propensity score weighted estimators for multilevel data

3.1 Basics

Consider a sample or population of n units, from cluster h ($h = 1, \dots, H$), each including n_h units indexed by $k = 1, \dots, n_h$, and $n = \sum_h n_h$. Each unit belongs to one of two groups for which covariate-balanced comparisons are of interest, possibly defined by a treatment; in either case we will use the terms “treatment” and “control” to refer to the groups. Let $Z_{hk} = z_{hk}$ be the binary variable indicating whether subject k in cluster h is assigned to the treatment ($z_{hk} = 1$) or the control ($z_{hk} = 0$), and let $\mathbf{z} = (z_{11}, \dots, z_{h,n_h}, \dots, z_{H,n_H})'$. Also let \mathbf{X}_{hk} be a vector of unit-level covariates, \mathbf{V}_h be a vector of cluster-level covariates, and $\mathbf{U}_{hk} = (\mathbf{X}_{hk}, \mathbf{V}_h)$.

Under the potential outcome framework for causal inference [21], each unit has a potential outcome $Y_{hk}(\mathbf{z})$ under each possible vector of treatment assignment indicators \mathbf{z} . A typical assumption of causal analysis is the Stable Unit Treatment Value Assumption (SUTVA), which states that the outcomes for each unit are unaffected by the treatment assignments of other units (whether within or across clusters). Formally, for any k and h , and any \mathbf{z} and \mathbf{z}' with $z_{hk} = z'_{hk}$, $Y_{hk}(\mathbf{z}) = Y_{hk}(\mathbf{z}')$. Hence, the potential outcomes can be written as a functions of z_{hk} rather than the entire vector \mathbf{z} , and there are only two potential outcomes for each unit [13].

To estimate a causal effect from observed data, the key identifying assumption is *strong ignorability of treatment assignment* [2], consisting of two sub-assumptions. First, *unconfoundedness* represents randomization of treatment assignment within cells defined by the values of observed covariates, $\Pr(Y(z)|\mathbf{U}) = \Pr(Y|\mathbf{U}, Z = z)$, for $z = 0, 1$. Therefore, the effects can be identified based on the observed distributions. Second, the *overlap* assumption requires that the study population be restricted to values of covariates for which there can be both control and treated units; otherwise the data cannot support an inference about comparisons of outcomes under the two treatments.

Under these assumptions, a causal treatment effect can be defined by comparison of the potential outcomes under treatment versus control on *a common set of units*, which may be summarized by the “Average Treatment Effect” (ATE),

$$\pi = E\{Y(1) - Y(0)\}, \quad (1)$$

that is, the population average difference in the potential outcomes under treatment and control. Alternative estimands, such as the “Average Treatment Effect on the Treated” (ATT) $E(Y(1) - Y(0)|T = 1)$ may also be of interest, as in [11]. What is common across these analyses is that means of $Y(1)$ and $Y(0)$ are compared under the *same* hypothesized distribution of the covariates. Descriptive comparisons without a causal interpretation but controlled for covariates similarly involve construction of populations with the same balance property.

The propensity score is defined as $e(\mathbf{U}) = P(Z = 1 | \mathbf{U})$, the conditional probability of being in (treatment or descriptive) group $Z = 1$ given covariates \mathbf{U} . The utility of the propensity score resides in the fact that under strong ignorability [see, e.g. 16],

$$E\left\{\frac{ZY}{e(\mathbf{U})} - \frac{(1-Z)Y}{1-e(\mathbf{U})}\right\} = E\{Y(1) - Y(0)\}. \quad (2)$$

In essence, this estimator weights both groups to a common distribution of covariates, namely that of the combined population. Therefore, the ATE (or controlled descriptive comparison) can be estimated by comparing weighted averages of the observed outcomes using the inverse-probability (Horvitz-Thompson or HT) weights $w_{hk} = 1/e(\mathbf{U}_{hk})$ for units with $Z_{hk} = 1$ and $w_{hk} = 1/(1 - e(\mathbf{U}_{hk}))$ for units with $Z_{hk} = 0$. It can readily be verified that this weighting balances (in expectation) the weighted distribution of covariates in the two groups. The validity of this method depends, however, on the correctness of the specification of the propensity score. In the following subsections we consider modifications of this approach appropriate to clustered data.

3.2 Models for the propensity score

We consider three alternative propensity score models for clustered data, corresponding to different assumptions about the assignment mechanism. A *marginal model* uses cluster membership only as a link to *observed* cluster-level covariates. The propensity score thus is a function of the observed covariates, $e_{hk} = e(\mathbf{U}_{hk})$, as in the logistic model

$$\text{logit}(e_{hk}) = \delta_0 + \mathbf{U}_{hk}\boldsymbol{\alpha}. \quad (3)$$

Such a model yields a valid balancing score under the assumption that all cluster-level covariates that are associated with group (treatment) assignment are observed.

A *fixed effects model* is augmented with a cluster-level main effect δ_h , as in the following logistic model:

$$\text{logit}(e_{hk}) = \delta_h + \mathbf{X}_{hk}\boldsymbol{\alpha}, \quad (4)$$

The δ_h term absorbs the effects of both observed and unobserved cluster-level covariates \mathbf{V}_h , protecting against misspecification due to cluster-level confounders. With maximum likelihood estimation, the observed and predicted numbers of treated cases *within each cluster* will agree, guaranteeing balance of the HT estimator on both observed and unobserved cluster level covariates. The fixed effects model estimates a balancing score without requiring knowledge of \mathbf{V}_h , but might lead to larger variance than the propensity score (the coarsest balancing score) estimated under a correct model with fully observed \mathbf{V}_h .

When there are many small clusters, the fixed effects model can make propensity score estimates unstable due to the large number of free parameters and the possibility of separation (representation of only one group) in some clusters. In the latter case, strong ignorability would require exclusion of those clusters from the inferential population. An alternative is to assume a *random effects model*, augmenting (4) with a prior distribution $\delta_h \sim N(\delta_0, \sigma_\delta^2)$ on the cluster-

specific main effects

$$\text{logit}(e_{hk}) = \delta_h + \mathbf{U}_{hk}\boldsymbol{\alpha}. \quad (5)$$

More generally, the random effects may include random coefficients of some individual-level covariates. The distributional assumption on δ_h in the random effects model greatly reduces the number of parameters compared to the fixed effects model. To estimate propensity scores from (5), one can plug in a point estimate, such as the posterior mode or the posterior mean of the inverse-probability weight. However, the random effects model does not guarantee balance within each cluster, due to the shrinkage of random effects toward zero, and therefore is somewhat reliant on inclusion of important cluster-level covariates \mathbf{V}_h as regressors. Thus it represents a compromise between the marginal and fixed effects models, with results converging to those from a corresponding fixed effects model as the sample size per cluster increases. (Differences among these models are examined by simulations in Section 5.)

The goodness of fit of these models can be checked by conventional diagnostic procedures [e.g. 3]. For example, one can check both the overall and within-cluster balance of the distribution of covariates weighted by the inverse estimated propensity score in the two groups.

3.3 Estimators for the average treatment effect

In general there are two types of propensity-score-weighted estimators for the ATE, applying the weights to either the observed outcomes (*nonparametric* estimators), as in (2), or the fitted outcomes from a parametric model (*parametric* estimators). The clustered data structure offers possibilities for several variations on these inverse-probability-weighted estimators, which we consider here.

A nonparametric *marginal estimator* is the difference of the weighted overall means of the

outcome between the treatment and control groups, ignoring clustering,

$$\hat{\pi}^{\text{ma}} = \sum_{Z_{hk}=1} \frac{Y_{hk}w_{hk}}{w_1} - \sum_{Z_{hk}=0} \frac{Y_{hk}w_{hk}}{w_0}, \quad (6)$$

where w_{hk} is the inverse-probability weight of subject k in cluster h based on the estimated propensity score (e.g., from one of the three models in Section 3.2) with $w_{hk} = 1/\hat{e}_{hk}$ for units with $Z_{hk} = 1$ and $w_{hk} = 1/(1 - \hat{e}_{hk})$ for units with $Z_{hk} = 0$, and $w_z = \sum_{h,k:Z_{hk}=z} w_{hk}$ for $z = 0, 1$.

A nonparametric *clustered estimator* first estimates the ATE within each cluster:

$$\hat{\pi}_h = \frac{\sum_{k \in h}^{z_{hk}=1} Y_{hk}w_{hk}}{w_{h1}} - \frac{\sum_{k \in h}^{z_{hk}=0} Y_{hk}w_{hk}}{w_{h0}},$$

where $w_{hz} = \sum_{k \in h}^{z_{hk}=z} w_{hk}$ for $z = 0, 1$, and then takes their mean weighted by cluster sample sizes:

$$\hat{\pi}^{\text{cl}} = \sum_h \hat{\pi}_h n_h / n. \quad (7)$$

The numerator and denominator of one of the terms of (7) will be zero for clusters where all units are assigned to the same group, violating the overlap assumption. Therefore, to implement the clustered estimator (7) with propensity scores estimated from a fixed effects model, one need to excludes clusters with $n_{h0} = 0$ or $n_{h1} = 0$.

An attractive parametric weighted estimator is the *doubly-robust (DR) estimator* proposed by Robins and colleagues [e.g. 14, 15, 22]:

$$\hat{\pi}^{\text{DR}} = n^{-1} \sum_{h,k} \left[\frac{Z_{hk}Y_{hk}}{\hat{e}_{hk}} - \frac{(Z_{hk} - \hat{e}_{hk})\hat{Y}_{hk}(1)}{\hat{e}_{hk}} \right] - n^{-1} \sum_{h,k} \left[\frac{(1 - Z_{hk})Y_{hk}}{1 - \hat{e}_{hk}} + \frac{(Z_{hk} - \hat{e}_{hk})\hat{Y}_{hk}(0)}{1 - \hat{e}_{hk}} \right], \quad (8)$$

where $\hat{Y}(z)$ is the fitted potential outcome (we discuss potential outcome models in Section 3.4). The name ‘‘doubly-robust’’ refers to the large sample property that $\hat{\pi}^{\text{DR}}$ is a consistent estimator of the ATE π if either the propensity score model or the potential outcome model is correctly specified, but not necessary both. In large samples, if e is modeled correctly, $\hat{\pi}^{\text{DR}}$

has smaller variance than the nonparametric inverse weighted estimators, while if the potential outcome model is correctly specified, $\hat{\pi}^{DR}$ may have larger variance than the direct regression estimator, but it provides protection when the model is misspecified.

3.4 Models for potential outcomes

We now consider several models for potential outcomes. An additive *marginal (potential) outcome model* has the form

$$Y_{hk}(z) = \eta_0 + z\gamma + \mathbf{U}_{hk}\boldsymbol{\beta} + \epsilon_{hk}, \quad (9)$$

where $\epsilon_{hk} \sim N(0, \delta_\epsilon^2)$, and γ is the *constant* treatment effect. Analogous to the marginal propensity model (3), the marginal outcome model assumes that the cluster effect on the potential outcomes is only through the covariates. The deeper connection is that the sufficient statistics that are balanced under the marginal propensity score estimator are the same that must be balanced to eliminate confounding differences under model (9), namely the treatment and control group means of U .

A *fixed-effects outcome model* adjusts for cluster-level main effects and covariates:

$$Y_{hk}(z) = \eta_h + z\gamma + \mathbf{X}_{hk}\boldsymbol{\beta} + \epsilon_{hk}, \quad (10)$$

where η_h is the cluster-specific main effect. Under this model, all information is obtained by comparisons within clusters, since the η_h term absorbs all between-cluster information. The corresponding *random effects outcome model* is:

$$Y_{hk}(z) = \eta_h + z\gamma + \mathbf{U}_{hk}\boldsymbol{\beta} + \epsilon_{hk}, \quad (11)$$

with $\eta_h \sim N(0, \sigma_\eta^2)$. A natural extension is a random additive treatment effect, replacing γ by cluster-specific treatment effect γ_h with $(\eta_h, \gamma_h)' \sim N(0, \Sigma_{\eta\gamma})$.

Interactions between covariates and treatment can be added to models (9), (10) and (11) to allow nonadditive relationships. Analogous generalized linear models (GLM) or generalized linear mixed models (GLMM) can be used for binary or ordinal outcomes.

4 Large-sample properties with unobserved cluster-level confounding

In this section, we investigate the large sample bias of these estimators under a generating model representing intra-cluster influence that causes violations of unconfoundness and SUTVA, in a simple setting with no observed covariates. Let n_{hz} denote the number of subjects with $Z = z$ in cluster h ; and $n_z = \sum_h n_{hz}$, $n_h = n_{h1} + n_{h0}$, $n = n_1 + n_0$. We assume a Bernoulli treatment assignment mechanism with varying rates by cluster:

$$Z_{hk} \sim \text{Bernoulli}(p_h). \quad (12)$$

We assume a continuous outcome model with cluster-specific random intercepts η_h and constant treatment effect π ,

$$Y_{hk}(z) = \eta_h + z\pi + \tau d_h + \epsilon_{hk}, \quad (13)$$

where $\eta_h \sim N(\eta_0, \sigma_\eta^2)$, $\epsilon_{hk} \sim N(0, \sigma_\epsilon^2)$, and τ is the coefficient of the cluster-specific proportion treated $d_h = n_{h1}/n_h$. SUTVA does not hold and τ measures the magnitude of the influence (contamination) within cluster. This model could also result from common unmeasured cluster traits that affect both treatment assignment and outcome. Nonetheless, it is easy to show the average treatment effect is a well-defined additive constant π .

Under the marginal propensity score model, $\hat{e}_{hk} = n_1/n$ is the same for every subject. Let $\hat{\pi}_{\text{ma}}^{\text{ma}}$ denote the marginal estimator (6), where the subscript ma indicates using the marginal model for the propensity score and superscript ma indicates using the marginal estimator for

ATE. Then

$$\hat{\pi}_{\text{ma}}^{\text{ma}} = \pi + \sum_h \eta_h \left(\frac{n_{h1}}{n_1} - \frac{n_{h0}}{n_0} \right) + \left(\sum_{h,k}^{z_{hk}=1} \frac{\epsilon_{hk}}{n_1} - \sum_{h,k}^{z_{hk}=0} \frac{\epsilon_{hk}}{n_0} \right) + \tau \frac{\mathcal{V}}{\mathcal{V}_0}, \quad (14)$$

where $\mathcal{V}/\mathcal{V}_0 = (\sum n_h (d_h^2 - (n_1/n)^2)) / (n_1 n_0 / n^2)$, that is, the weighted sample variance of the $\{d_h\}$ divided by its maximum possible value. The maximum is attainable if each cluster is assigned to either all treatment or all control, corresponding to a cluster randomized design where the cluster-level treatment effect is $\pi + \tau$. The second and third terms of (14) have expectation zero under the model and approach zero as the number of clusters approaches infinity under mild regularity conditions on n_{h1} and n_{h2} . Therefore, under the generating model (12) and (13), the large sample bias of the marginal estimator with propensity score estimated from the marginal model is $\tau \mathcal{V}/\mathcal{V}_0$. Intuitively, τ measures the variation in the outcome generating mechanism between clusters and $\mathcal{V}/\mathcal{V}_0$ measures the variation in the treatment assignment mechanism between clusters, both of which are ignored in $\hat{\pi}_{\text{ma}}^{\text{ma}}$.

We next consider estimation when clustering information is taken into account in both steps. Using the fixed effects model (4), the estimated propensity score is $\hat{e}_{hk} = n_{h1}/n_h$. Then the clustered estimator (7) is $\hat{\pi}_{\text{fe}}^{\text{cl}}$, where the subscript fe refers to the fixed effects model for the propensity score, and superscript cl refers to using the clustered estimator for ATE, given by

$$\hat{\pi}_{\text{fe}}^{\text{cl}} = \pi + \frac{\sum_h (\sum_{k \in h}^{z_{hk}=1} \frac{\epsilon_{hk}}{n_{h1}})}{H} - \frac{\sum_h (\sum_{k \in h}^{z_{hk}=0} \frac{\epsilon_{hk}}{n_{h0}})}{H}, \quad (15)$$

which converges to π as H and n_h increase. Simple calculation shows that the clustered weighted estimator combining the marginal propensity score model with the clustered estimator, $\hat{\pi}_{\text{ma}}^{\text{cl}}$, is equivalent to that in (15) and thus also consistent. Furthermore, the marginal estimator with propensity score estimated from the fixed effects model, $\hat{\pi}_{\text{fe}}^{\text{ma}}$, is equivalent to (15) only under a balanced design (clusters of equal size). Under an unbalanced design, the estimator remains consistent under the standard regularity condition of $\sum_{h=1}^H n_h^2/n^2$ being bounded as H goes to infinity, but its small-sample behavior can be quite different.

In summary, in this simple case violations of standard propensity score assumptions associated with unobserved cluster effects or violation of SUTVA, ignoring the clustered structure in both propensity score and potential outcome modeling stages induce bias in estimating ATE, while exploiting the structure in at least one of the models gives consistent estimates.

5 Simulation studies

We now examine a propensity score weighting analysis with observed covariates when there is an unmeasured cluster-level confounder. Generally there is no closed-form expression for the propensity score weighted estimators for clustered data with covariates, so we instead conduct a simulation study.

5.1 Simulation design

The simulation design is similar to but more general than that of [11]. We assume both treatment assignment and outcome generating mechanisms follow two-level random effects models with two individual-level covariates X_1, X_2 and a cluster level covariate V . Specifically, the treatment assignment follows

$$\text{logit Pr}(Z_{hk} = 1 | \mathbf{U}_{hk}) = \delta_h + \mathbf{U}_{hk} \boldsymbol{\alpha}, \quad (16)$$

where $\mathbf{U}_{hk} = (X_{1,hk}, X_{2,hk}, V_h)$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_3)'$ are the coefficients for the fixed effects, and $\delta_h \sim \text{N}(0, 1)$ are the cluster-specific random intercepts. The potential outcomes are generated from a random effects model as (10), with an extra interaction term between treatment and V :

$$Y_{hk}(z) = \eta_h + \mathbf{U}_{hk} \boldsymbol{\beta} + z(V_h \kappa + \gamma_h) + \epsilon_{hk}, \quad \epsilon_{hk} \sim \text{N}(0, \sigma_y^2), \quad (17)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ are the coefficients of the covariates, κ is the coefficient of the interaction, and $\eta_h \sim \text{N}(0, 1), \gamma_h \sim \text{N}(0, 1)$ are the cluster-specific random intercepts and random

slopes of treatment, respectively.

The individual-level continuous covariate X_1 is simulated from $N(1, 1)$ and binary X_2 from $\text{Bernoulli}(0.4)$. The cluster-level covariate V is generated in two ways that are plausible in real applications: (i) uncorrelated with X , with $V \sim N(1, 2)$; (ii) correlated with the cluster-specific intercept in the propensity score model, with $V_h = 1 + 2\delta_h + u$, where u is an error term. An extreme case of (ii) is that V is a linear function of the cluster-average propensity score, when $u \approx 0$. Here we let $u \sim N(0, 0.5)$. Case (ii) introduces an interaction between treatment and cluster-level random effects. This is expected to cause differences between the fitting of fixed and random effects models when V is omitted, because generally the former cannot readily accommodate the interaction while the later with random slopes for z can. Another common situation not examined here is that V is correlated with X ; omitting V in the analysis in that case is expected to lead to smaller bias than case (i), as including X as a covariate partially compensates for excluding V .

We compare the three propensity score models (3), (4) and (5) in Section 3.2, all fitted with only the individual-level covariates X_1, X_2 , but omitting the cluster-level covariate V . For comparison, we also estimate the propensity score by the true model (16). With the propensity score estimated from each of these models, we calculate the ATE by the marginal (6), clustered (7) and DR (8) estimators. For the DR estimators, we fit the three potential outcome models (9), (10) (with cluster-specific intercepts) and (11) (with random intercepts and random slopes for z) in Section 3.4, each fitted with only the individual-level covariates X_1, X_2 , but not the cluster-level covariate V . As a benchmark, we also estimate the potential outcomes by the true model (17) that includes V and its interaction with z . In total, we compare four models for propensity score and six ATE estimators (including four DR estimators), giving twenty-four combinations.

We first simulate under a two-level balanced design, where all the clusters have the same

sample sizes with $H = 30, n_h = 400$, which is approximately half of the size of our real application. We then also simulate unbalanced designs, with n_h randomly generated from $\text{Uniform}(200, 800)$.

We fix the parameters for the propensity score model: $\delta_h \sim \text{N}(0, 1)$ and $\alpha = (-1, -0.5, \alpha_3)$, and vary $\alpha_3 = -1$ or 1 to give low (around 0.2) and medium (around 0.5) overall rates of treatment assignment, respectively, across the simulated samples. Similarly, for the potential outcome model, we set the interaction $\kappa = 2$ and $\beta = (1, 0.5, \beta_3)$, while changing β_3 to control the magnitude of the effect of V on the outcome.

Under each simulation scenario, 500 replicates from models (16) and (17) are generated. The random effects models are fitted using the `lmer` command in the `lme4` package in R2.13.2 [23]. For each simulated dataset, we calculate the relative error (RE) (measure of bias) $E(|\hat{\pi} - \pi|/\pi)$ and variance $\text{Var}(\hat{\pi})$ of each of the twenty-four estimators for the ATE. The true ATE π is calculated by $\sum_{h,k} \{Y_{hk}(1) - Y_{hk}(0)\}/n$ from the simulated units - this is in fact a finite sample ATE rather than the large population ATE defined in (1). But the difference between the two is usually very small given the sample size we consider here.

5.2 Results

Table 1 presents the REs of the estimators with V generated from $\text{N}(1, 2)$, $\alpha_3 = 1, \beta_3 = 2$, resulting in an average propensity score of 0.5. Several patterns emerge in this example and in other variations of our simulations. First, as expected, estimators that ignore clustering in both propensity score and outcome models lead to much larger errors than approaches that consider clustering in at least one stage: the biases are over 104%, which are over thirty times larger than the average bias. Second, we observed few differences between the propensity score models that account for clustering; both fixed- and random-effects propensity score models performed as well as the true benchmark models. Third, we did observe variation in performance between

parametric and non-parametric weighted estimators and across choices for the outcome model in parametric estimates (i.e across the columns). The DR estimators with the benchmark outcome model (column 3) and the random effects outcome model (column 6) perform the best; there is no virtual difference in the results between these two, matching our expectation that exploiting cluster structure protects against misspecification due to unobserved V . The DR estimators with the fixed effects outcome model (column 5) appear to give significantly worse results than the above two. However, this is largely due to the simulation design, where the outcomes are generated from an underlying random effects model and the number of clusters is large. If we instead simulate a small number of clusters with a fixed effects outcome model as the benchmark where the cluster-specific coefficients do not follow a distribution, the fixed effects model would be expected to perform better than the random effects model. The nonparametric clustered estimators (column 2) also lead to small REs, with the marginal propensity score model leading to the largest errors. The nonparametric marginal estimators (column 1) and the DR estimators with the misspecified marginal outcome model (column 4) have the largest REs.

REs from the simulations with low average propensity score (around 0.2 with $\alpha_3 = -1$) display the same pattern in terms of the comparison across the estimators, but with consistently larger values than those in Table 1 (median propensity score). Patterns in variances are also very similar between these two scenarios and thus the details are omitted here.

[Table 1 about here.]

To see how the estimators behave with a larger effect of V on the outcome, we doubled the coefficient β_3 in the outcome model with α_3 set to 1 (Table 2). Comparing with Table 1, only the marginal weighted (column 1) and DR (column 4) estimators with the marginal outcome model appeared to be affected, nearly doubling the biases.

[Table 2 about here.]

Simulation results with V correlated with the cluster-specific random intercept in propensity score (setting (ii) in Section 5.1) are presented in Table 3. Besides the patterns observed from the previous simulation settings, here the random effects outcome model appear to dominate the fixed effects model even more - biases from the former are usually 1/7 of the corresponding ones from the later.

[Table 3 about here.]

We also simulated from various other scenarios, including unbalanced designs, exponentially distributed V , and larger interactions between V and treatment (larger κ). In particular, we simulated a cluster level covariate that is included in the models, but with a misspecified linear instead of the true quadratic relationship with the outcome. Despite all the differences in designs, the patterns described above are consistently observed across the simulations.

Given the arbitrariness of the sample sizes and variances in this simulation design, absolute values of variances are not very informative and thus not shown here. Nevertheless, across the above simulation scenarios, we have consistently observed that most of the estimators have very similar variances, but the ones ignoring cluster structure in the second stage lead to considerably larger variances. Overall, the difference in variances across the estimators is much smaller than in biases.

In summary, ignoring the hierarchical structure in both stages of the propensity score analysis of multilevel data is a bad practice, leading to large bias of the ATE. It appears that exploiting the hierarchical structure in at least one of the two stages, either parametrically or nonparametrically, greatly reduces the estimating errors due to model misspecification. This is in line with the conclusion of [11] in propensity score matching analysis. Moreover, the outcome models have much larger influences on the final estimates than the propensity score models. In practice, when there is a concern over misspecification of the outcome model, a nonparametric clustered

estimator with propensity score estimating from multilevel models might be a safe choice for ATE.

Finally we note two observations regarding the simulations of the DR estimators. First, estimators using a correct outcome models with a misspecified propensity score model are less biased than those using a correct propensity score model with a misspecified outcome model. For example, the DR estimators with the misspecified marginal model (column 4) give even worse results than the nonparametric marginal estimators (column 1), regardless of the propensity score model, suggesting that a DR estimator with a misspecified outcome model can do more “harm” than “protection”. Second, when the outcome model is correct or nearly so, the misspecified marginal propensity score model slightly outperforms all the other models, including the benchmark model. It is known [e.g. 24, 22] that when the outcome model is correct, a DR estimator can have larger variance than a direct regression estimator. With the correct outcome model, a simple propensity score model like the marginal model might produces the least variation in the DR estimates, thus reduce the overall estimation errors relative to those from the more complex models such as the random effects models.

6 Application

We apply our alternative methods to a study of racial differences in care provided in Medicare health plans. The individual-level covariates \mathbf{X}_{hk} considered include two indicators of age category (70-80 years, >80 years with reference group 65-69 years; eligibility for Medicaid (1=yes); neighborhood status indicator (1=poor). The plan-level covariates \mathbf{V}_h include nine geographical region indicators, tax status (1=for-profit), the practice model of providers (1= staff or group model; 0= network-independent practice association model), and affiliation (1= national; 2= Blue Cross/ Blue Shield; 3= independent). The outcome Y is a binary variable

equal to 1 if the enrollee underwent breast cancer screening and 0 otherwise, and the “treatment” z is race (1= black, 0= white). The goal is to assess differences in the proportions undergoing breast cancer screening between whites and blacks. We control for all covariates to illustrate their effects and for consistency with [18], although by some definitions these would not all be controlled for in a policy-oriented disparities calculation [25].

As stated previously, race is not a “treatment” in the conventional sense of causal inference, because it is not manipulable [20]. We define the ATE to be the difference between white and black screening rates among patients with a set of characteristics U (implicitly smoothed by models), averaged over the distribution of U in the combined black and white populations. There may be interest in a variety of estimators that control for different sets of covariates. For example, if U only contains individual level covariates, we interpret the ATE as the difference between groups controlled for differences in these individual characteristics such as age or poverty. Including V in the analysis controls for differences in treatment that result from differences in the types of plans in which the two groups enroll. Similarly, balancing on cluster membership accounts for unobserved differences in the quality of health plans that enroll minorities and provides an estimate of differences in treatments between minorities and whites within individual health plans.

We first estimate the propensity score using the three models introduced in Section 3.2 with all the above covariates included. Details of the fitted models are omitted here since the focus is on the estimated propensity score. All models suggest that living in a poor neighborhood, being eligible for Medicaid and enrollment in a for-profit insurance plan are significantly associated with black race. The distributions of the estimated propensity scores in both blacks and whites from the marginal model are quite different those from the random and fixed effects models, which are similar to each other. A well-estimated propensity score is expected to balance the cluster membership between races in the “pseudo population” obtained by inverse weighting

using that propensity score. Figure 1 shows the histograms of the difference in the weighted numbers of white and black enrollees in each cluster, using propensity scores estimated from different models (and the unweighted difference for comparison). Clearly, the distributions of cluster membership differ significantly between races when unweighted or weighted using the propensity score from the marginal model, but are similar when weighted using the fixed or random effects models. This suggests important between-cluster variation, failing to take in account of which (such as in the marginal model) leads to poor estimates of propensity score in this application.

[Figure 1 about here.]

Using the estimated propensity score, we estimate racial disparity in breast cancer screening among the elder women participating Medicare health plans by the estimators in Section 3.3. Since the outcome is binary, for the DR estimators, we use the generalized linear models (GLM) corresponding to the three outcome models in Section 3.4 in combination with each of the three propensity score models. Table 4 displays the point estimates with bootstrap standard errors.

[Table 4 about here.]

All estimators show the rate of receipt breast cancer screening is significantly lower among blacks than among whites with similar characteristics. Accounting for differences in individual and plan level covariates, but not plan membership, we estimate the rate of screening for breast cancer is 5 percentage points lower in blacks compared to whites. That is, among the elders who participate in Medicare health plans, blacks on average have a significantly lower chance to receive breast cancer screening than whites, after adjusting for age, geographical region, some socioeconomic status variables and health plan characteristics. Accounting for plan membership in either stage of the analysis decreases this difference by approximately 50%,

suggesting that approximately half of lower rates of breast cancer screening among black in this population is a result of minorities enrolling in plans with low screening rates and half results from lower probability of black women undergoing screening within each plan.

As in the simulation studies we observed few differences across estimators that account for clustering in at least one stage of the analysis. The DR estimates have smaller standard errors because the extra variation is explained by covariates in step 2. Similar to the simulations, we notice that the estimates incorporating clustering in step 2 have less variation than those doing so in step 1. This observation suggests, in application, modeling the hierarchical structure for the outcome generating mechanism leads to more stable estimates, even though in theory correct model specification in both steps are equivalent in terms of their effect on consistency. A possible explanation is the impact of misspecifying propensity score is attenuated through weighting because the ultimate estimand is a function of the outcome, rather than of the propensity score.

7 Summary and Remarks

Since first proposed in 1983, propensity score methods have gained increasing popularity in observational studies in multiple disciplines. One example is health care policy research, where data with hierarchical structure are the norm rather than the exception. However, despite the wide appreciation of propensity score methods among both statisticians and health policy researchers, only a limited literature deals with the methodological issues of propensity score methods in the context of multilevel data. We compared three models for estimating the propensity score and three types of propensity-score-weighted estimators for the ATE for multilevel data. Consequences of the violation to the key assumptions of SUTVA and unconfoundedness in propensity score analysis of multilevel data were explored using both analytical

deviations and Monte Carlo simulations. We found that exploiting the hierarchical structure, either parametrically or nonparametrically, in at least one stage of the propensity score analysis can greatly reduce the errors of estimating ATE in presence of violations to those assumptions.

We focused here on treatments assigned at the individual level. Treatment assigned at the cluster level (e.g., hospital, health care provider) are also common in medical care and health policy studies, engendering new challenges. First, the number of clusters is often relatively small despite a large total sample size. This could lead to poorly estimated propensity scores with excessively large standard errors. Second, unobserved cluster-level covariates cannot be controlled for even with fixed effects models, unlike the cases discussed in the paper. This has a strong connection to the ecological inference commonly encountered in political science [e.g. 26] where the estimand has an interpretation as an average effect on individual outcomes. Treatments assigned at the cluster level has also been discussed in [27, 28], among others.

All the nonparametric weighted estimators discussed in this paper do not make use of the individual-level covariates, which often contain crucial information. The doubly-robust estimators with flexible regression model choice in the second step might be preferable in this case, with specification depending on the particulars of the data.

Violations to both SUTVA and unconfoundedness often coexist. We have shown that accounting for clustering in propensity score analyses can account for violations that occur at the cluster level. However, there may be additional violations within clusters (for example, high volume surgeons may have improved outcomes that would not be accounted for in an analysis that balances on treating hospital but not surgeon). Correct modeling of the interference among subjects is crucial for valid analysis, so that model selection and checking is important. In addition, in some situations there may be direct interest in spill-over effects, such as in volume-outcome studies. [29] describe the use of a two-stage propensity score model to estimate the propensity for elementary schools to retain low performing kindergarten students and then the

propensity for students nested within schools to be retained. These authors relax the SUTVA and allow the effectiveness of retention to vary according to how many peers within a child's school were also retained.

These issues are among a range of open questions remained to be explored on this topic. Further systematic research efforts are desired to shed insight to the methodological issues and to provide guidelines for practical applications.

8 Acknowledgements

This work was funded by National Cancer Institute (NCI) grant to the Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium (U01 CA093344), and by grant 138255 from the Academy of Finland.

References

- [1] Rubin D. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 1979; **74**:318–324.
- [2] Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Journal of the Royal Statistical Society: Series B* 1983; **70**(1):41–55.
- [3] Rosenbaum P, Rubin D. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
- [4] Connors A, Speroff T, Dawson N, Thomas C, Harrell Jr F, Wagner D, Desbiens N, Goldman L, Wu A, Califf R, *et al.*. The effectiveness of right heart catheterization in the initial

care of critically ill patients. *Journal of the American Medical Association* 1996; **276**:889–897.

- [5] D’Agostino R. Tutorial in biostatistics: propensity score methods for bias reduction in the comparisons of a treatment to a non-randomized control. *Statistics in Medicine* 1998; **17**:2265–2281.
- [6] Gatsonis C, Normand S, Liu C, Morris C. Geographic variation of procedure utilization: a hierarchical model approach. *Medical Care* 1993; **31**:YS54–YS59.
- [7] Nattinger A, Gottlieb M, Veum J, Yahnke D, Goodwin J. Geographic variation in the use of breast-conserving treatment for breast cancer. *New England Journal of Medicine* 1992; **326**:1102–1127.
- [8] Farrow D, Samet J, Hunt W. Regional variation in survival following the diagnosis of cancer. *Journal of Clinical Epidemiology* 1996; **49**:843–847.
- [9] Lingle J. Evaluating the performance of propensity scores to address selection bias in a multilevel context: A monte carlo simulation study and application using a national dataset. *Educational Policy Studies Dissertations, Paper 56*. 2009.
- [10] Su YS, Cortina J. What do we gain? combining propensity score methods and multilevel modeling. *APSA 2009 Toronto Meeting Paper*, 2009.
- [11] Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis* 2011; **55**:1770–1780.
- [12] Hudgens M, Halloran M. Towards causal inference with interference. *Journal of the American Statistical Association* 2008; **103**:832–842.

- [13] Rubin D. Comment on ‘Randomization analysis of experimental data: The Fisher randomization test’ by D. Basu. *Journal of the American Statistical Association* 1980; **75**:591–593.
- [14] Robins J, Rotnitzky A, Zhao L. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**(429):106–121.
- [15] Robins J, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**(429):122–129.
- [16] Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology* 2001; **2**:259–278.
- [17] Hirano K, Imbens G, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 2003; **71**:1161–1189.
- [18] Schneider E, Zaslavsky A, Epstein A. Racial disparities in the quality of care for enrollees in medicare managed care. *Journal of the American Medical Association* 2002; **287**(10):1288–1294.
- [19] Zhao Z. Sensitivity of propensity score methods to the specifications. *Economics Letters* 2008; **98**(3):309–319.
- [20] Rubin D. Which ifs have causal answers: comment on ‘Statistics and causal inference’ by P.W. Holland. *Journal of the American Statistical Association* 1986; **81**:961–962.
- [21] Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**(1):688–701.

- [22] Bang H, Robins J. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**:962–972.
- [23] Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using s4 classes 2011. URL <http://cran.r-project.org>, R Package Version 0.999375-42.
- [24] Lunceford J, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
- [25] McGuire T, Alegria M, Cook B, Wells K, Zaslavsky A. Implementing the Institute of Medicine definition of disparities: An application to mental health care. *Health Services Research* 2006; **41**:1979–2005.
- [26] King G. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press: Princeton, New Jersey, 1997.
- [27] Oakes J. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Social Science and Medicine* 2004; **58**:1929–1952.
- [28] VanderWeele T. Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine* 2008; **27**:1934–1943.
- [29] Hong G, Raudenbush S. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association* 2006; **101**:901–910.

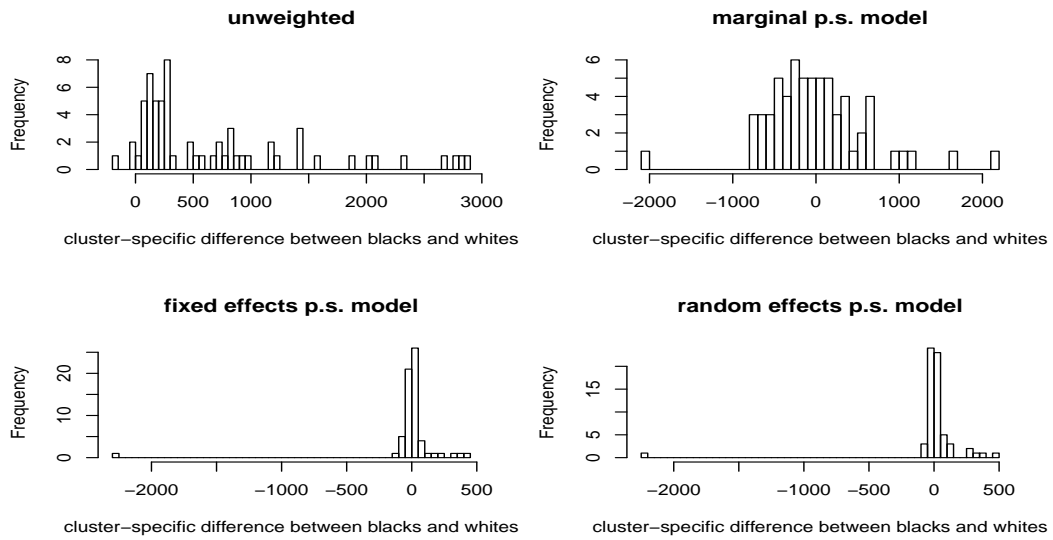


Figure 1: Histogram of cluster-specific differences in the weighted numbers of white and black enrollees using propensity scores estimated from different models. Closer to zero values indicate better balance in cluster membership between races.

	weighted		bench	doubly-robust		
	marginal	clustered		marginal	fixed	random
bench	3.4	1.3	0.9	4.8	2.3	0.9
marginal	104.6	8.5	0.6	104.7	7.0	0.6
fixed	3.5	1.2	0.9	5.0	2.4	0.9
random	4.0	1.3	0.9	5.4	2.5	0.6

Table 1: Relative error (RE) in percentage of different estimators, with $V \sim N(1, 2)$, $\alpha_3 = 1$ and $\beta_3 = 2$. Different rows correspond to different models to estimate propensity score, different columns correspond to different outcome models.

	weighted		bench	doubly-robust		
	marginal	clustered		marginal	fixed	random
bench	7.1	1.5	0.8	10.6	2.6	0.8
marginal	185.4	9.1	0.6	185.6	7.2	0.6
fixed	6.9	1.5	0.9	10.3	2.8	0.8
random	8.1	1.5	0.8	11.5	2.5	0.8

Table 2: Relative error (RE) in percentage with larger magnitude of V on the outcome: $V \sim \text{N}(1, 2)$, $\alpha_3 = 1$ and $\beta_3 = 4$.

	weighted		bench	doubly-robust		
	marginal	clustered		marginal	fixed	random
bench	24.4	4.0	1.7	34.2	13.5	1.6
marginal	229.9	15.5	1.4	229.9	10.9	1.3
fixed	19.0	3.9	1.7	26.3	10.4	1.6
random	24.4	4.0	1.6	33.4	9.8	1.5

Table 3: Relative error (RE) in percentage with $V = 1 + 2\delta_h + N(0, 0.5)$.

	weighted		doubly-robust		
	marginal	clustered	marginal	fixed eff	random eff
marginal	-4.96 (.79)	-1.73 (.83)	-4.43 (.85)	-2.15 (.41)	-1.65 (.43)
fixed eff	-2.49 (.92)	-1.78 (.81)	-1.93 (.82)	-2.21 (.42)	-1.96 (.41)
random eff	-2.56 (.91)	-1.78 (.82)	-2.00 (.44)	-2.22 (.39)	-1.95 (.39)

Table 4: Adjusted difference in percentage with standard error (in parenthesis) in the proportion of getting breast cancer screening between blacks and whites. Different rows correspond to different propensity score models and different columns correspond to different outcome models.