Propensity score analysis with hierarchical data

Fan Li, Alan M. Zaslavsky, Mary Beth Landrum
Department of Health Care Policy, Harvard Medical School
180 Longwood Avenue, Boston, MA 02115

October 29, 2007

Abstract

Propensity score (Rosenbaum and Rubin, 1983) methods are being increasingly used as a less parametric alternative to traditional regression methods in medical care and health policy research. Data collected in these disciplines are often clustered or hierarchically structured, in the sense that subjects are grouped together in one or more ways that may be relevant to the analysis. However, propensity score was developed and has been applied in settings with unstructured data. In this report, we present and compare several propensity-score-weighted estimators of treatment effect in the context of hierarchically structured data. For the simplest case without covariates, we show the "double-robustness" of those weighted estimators, that is, when both of the true underlying treatment assignment mechanism and the outcome generating mechanism are hierarchically structured, the estimator is consistent as long as the hierarchical structure is taken into account in at least one of the two steps in the propensity score procedure. This result holds for any balancing weight. We obtain the exact form of bias when clustering is ignored in both steps. We apply those methods to study racial disparity in the service of breast cancer screening among elders who participate Medicare health plans.

KEY WORDS: double robustness, health policy research, hierarchical data, propensity score, racial disparity, weighting.

1. Introduction

Population-based observational studies often are the best methodology for obtaining generalizable results on access to, patterns of, and outcomes from medical care when large-scale controlled experiments are infeasible. Comparisons between groups can be biased, however, when the groups are unbalanced with respect to measured and unmeasured confounders. Standard analytic methods adjust for observed differences between treatment groups by stratifying or matching patients on a few observed covariates or with regression analysis in the case of many observed confounders. But if treatment groups differ greatly in observed characteristics, estimates of treatment effects from regression models rely on model extrapolations and the resulting conclusions can be very sensitive to model mis-specification (Rubin, 1979). Propensity score methods (Rosenbaum and Rubin, 1983, 1984) have been proposed as a less parametric alternative to regression adjustment and are being increasingly used in health policy studies (Connors et al., 1996; D'Agostino, 1998, and references therein). This approach, which involves comparing subjects weighted (or stratified, matched) according to their propensity to receive treatment (i.e., propensity score), attempts to balance subjects in treatment groups in terms of observed characteristics as would occur in a randomized experiment. Propensity score methods permit control of all observed confounding factors that might influence both choice of treatment and outcome using a single composite measure, without requiring specification of the relationships between the control variables and outcome.

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 or hierarchically structured, in the sense that subjects are grouped together in one or more ways that may be relevant to the analysis. For example, subjects maybe grouped by geographical area, treatment center (e.g., hospital or physicians), or in the example we consider in this paper, health plan. Generally, subjects are assigned to clusters by an unknown mechanism that 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).

When subjects are hierarchically structured, a number of issues appear that are not present with an unstructured collection of subjects. First of all, standard error calculations that ignore the hierarchical structure will be inaccurate, leading to incorrect inferences. A more interesting set of issues arises because there may be both measured and unmeasured factors at the cluster level that create variation among clusters in quality of treatment and hence in outcomes. Hierarchical regression models have been developed to give a more comprehensive description than non-hierarchical models provide for such data (e.g., Gatsonis et al., 1993). Despite the increasing popularity of propensity score analyses and the vast literature regarding regional and provider variation in medical care and health policy research (e.g., Nattinger et al., 1992; Farrow et al., 1996), however, to our knowledge, the implications of such data structures for propensity score analyses have been rarely studied. Huang et al. (2005) applied propensity score methods to clustered health service data. But their goal was to rank the performance of multiple health service providers (clusters) in-

stead of to estimate an overall treatment effect from data with clustered structure, which is the goal of this paper. Specifically, we will present several propensity score models analogues to many of the commonly used regression models for clustered data in Section 2; investigate the behavior of those estimators, especially the bias when clustering information is ignored in the analysis in Section 3; and apply the methods to study racial disparities in the service of breast cancer screening among elders in Section 4. Summaries and remarks will be provided in Section 5. Our discussion concerns the case where a binary treatment is assigned at individual level. Also, to illustrate the major point and yet without loss of generality, we focus on data with two-level hierarchical structure.

2. Estimators

The class of estimands considered in this paper is generally referred as a "treatment effect",

$$\Delta = E_x[E(Y|X, Z = 1)] - E_x[E(Y|X, Z = 0)], \quad (1)$$

i.e., the average difference in outcome between two treatment groups that have same distribution of covariates.

The propensity score e is defined as the conditional probability of being assigned to a particular treatment z given measured covariates x: $e(x) = P(z = 1|x)$. In most observational studies, the propensity score is not known and thus needed to be estimated. Therefore propensity score analysis usually involves two steps. The first step is to estimate the propensity score, typically by a logistic regression. The second step is to estimate the treatment effect by incorporating (e.g., by weighting or matching) the estimated propensity score. Hierarchical structure leads to a range of different choices of modeling in both steps. In this section, we will introduce several most widely used models.

Before going into more details, here we make a note regarding the targeted estimand "treatment effect" defined above, which is slightly different from those "causal treatment effect" defined using the conventional potential outcomes framework. The propensity score originated from and has been widely used in causal inference, but its use is certainly not restricted to studying causal effects. For instance, in many health policy studies, the major interest is to compare the difference in the average of a feature (e.g., access to care) between two groups (e.g., races, social economical status), rather than to make a causal statement. Moreover, the "treatment" is often a non-manipulable variable, e.g., race or gender, which does not gives a well-defined casual effect in the sense of Rubin (1978) (more discussion in Section 4). Nevertheless, propensity score is still a valid and powerful tool to balance the covariates distribution between groups for studies with non-causal purposes. Therefore, we avoid the subtle issue of causality throughout the paper and note the results obtained here are applicable for studies with more general (non-causal) purposes. For ease of description, we still refer to our estimands discussed as "treatment effects" even though they are not necessarily causal.

Henceforth, let m denote the total number of clusters; n_h the number of subjects in cluster h; y_{hk} the outcome for subject k in cluster h (e.g., a clinical diagnosis); x_{hk} the corresponding covariates (typically vector-valued, e.g., age, stage of detection, comorbidity scores, etc.); v_h the cluster-level covariates (e.g., teaching status or measures of technical capacity of a hospital); z_{hk} the treatment assignment for the subject, $z_{hk} \in \{0, 1\}$; and e_{hk} the propensity score.

2.1 Step 1. Estimating the propensity score

To estimate the propensity score, several logistic regression models are available with various treatment of the hierarchical structure.

2.1.1 Marginal model

As the name suggests, marginal regression models ignore clustering information. A typical marginal propensity score model would be

$$\log\left(\frac{e_{hk}}{1 - e_{hk}}\right) = \beta^e x_{hk} + \kappa^e v_h, \quad (2)$$

where $e_{hk} = P(z_{hk} = 1 \mid x_{hk}, v_h)$. This model in fact assumes the treatment assignment mechanism is the same across all clusters. In other words, it assumes that two subjects are exchangeable in terms of treatment propensity if they have the same vector of covariates, whether or not they come from the same cluster.

This propensity score model can be thought of as a nonparametric alternative to a regression-based adjustment for individual and cluster covariates. The analogous marginal regression model would be,

$$y_{hk} = \gamma z_{hk} + \beta^y x_{hk} + \kappa^y v_h + \epsilon_{hk}, \quad (3)$$

where $\epsilon_{hk} \sim N(0, \delta^2_\epsilon)$, and γ is the treatment effect. As model (2), estimates derived from this regression model rely on the assumption that the outcome generating mechanism is the same across all clusters.

Models (2) and (3) have a manifest similarity of form. A deeper connection is that the sufficient statistics to estimate the treatment effect that are balanced under propensity score estimator are the same that must be balanced under model (3).

2.1.2 Pooled within-cluster model

A pooled within-cluster model for propensity score conditions on both the covariates and the cluster indicators,

$$\log(\frac{e_{hk}}{1 - e_{hk}}) = \delta^e_h + \beta^e x_{hk}, \quad (4)$$

where δ^e_h is a cluster-level main effect, $\delta^e_h \sim N(0, \infty)$, and $e_{hk} = P(z_{hk} = 1 \mid x_{hk}, h)$. This model implies the treatment assignment mechanism differs among clusters, and the difference is controlled by a cluster-level main effect δ^e_h. Model (4) involves a more general assumption (weaker) on the treatment assignment mechanism than the marginal model (2), because the cluster-level covariate v_h is a function of the cluster indicator h.

In the above model, if we assume the cluster-specific main effects δ_h^e follow a distribution, $\delta_h^e \sim N(0, \sigma_\delta^2)$, then we have a new propensity score model with random effects,

$$\log(\frac{e_{hk}}{1-e_{hk}}) = \delta_h^e + \beta^e x_{hk} + \kappa^e v_h.$$

More generally, β^e can be allowed to vary across clusters and follow a distribution. In practice, results from the above random effects model are usually similar to those from the pooled within-cluster model when the number of clusters is big.

A corresponding pooled within-cluster outcome model adjusting for cluster-level main effects and covariates is of the form:

$$y_{hk} = \gamma z_{hk} + \delta_h^y + \beta^y x_{hk} + \epsilon_{hk}, \quad (5)$$

where δ_h^y is a cluster-level main effect, $\delta_h^y \sim N(0, \infty)$. Under this model, all information is obtained by comparisons within clusters, since the δ_h^y term absorbs all between-cluster information.

2.1.3 Surrogate indicator model

When there are a large number of clusters with large sample size, the computational task of fitting the pooled within-cluster model can get demanding for standard software. Alternatively, define $d_h = \sum_{k\in h} \frac{z_{hk}}{n_h}$, the cluster-specific proportion of being treated, we can consider the following propensity score model

$$\log(\frac{e_{hk}}{1-e_{hk}}) = \lambda \log(\frac{d_h}{1-d_h}) + \beta^e x_{hk} + \kappa^e v_h. \quad (6)$$

In the simplest situation where there is no covariates, $e_{hk} = d_h$ for any h, k. Therefore, comparing models (4) and (6), the logit of d_h maybe expected to be a reasonable surrogate for the cluster indicator in the pooled within-cluster model with the coefficient λ being around 1. The inference is same as in the marginal model with an additional covariate logit(d_h). Usually the coefficients of the cluster-level covariates κ^e are very small since most of their effects have been absorbed by λ. The surrogate indicator model reduces the m parameters (δ_h's) in the pooled within-cluster model to a single parameter λ, thus greatly reducing the computation required for model fitting. However, this reduction is based on the assumption that logit of the empirical cluster-specific proportion of being treated, logit(d_h), is linearly correlated with logit of the true propensity score. When the underlying truth is far from this assumption, the surrogate indicator model could perform poorly.

The goodness of fit of these models can be checked by conventional diagnostic procedures (e.g., Rosenbaum and Rubin, 1984). For example, one can check both the overall and within-cluster balance of the distribution of covariates weighted by the estimated propensity score in different groups.

2.2 Step 2. Estimating the treatment effect

Common approaches estimate treatment effects using propensity score involve weighting, matching and stratification. We will focus on weighting in this report.

2.2.1 Marginal estimator

Similar to the marginal model in step 1, the marginal estimator ignores clustering. A specific nonparametric estimator is the difference of the weighted overall means of the outcome of two treatment groups,

$$\hat{\Delta}_{.,marg} = \frac{\sum_{h,k}^{z_{hk}=1} w_{hk} y_{hk}}{\sum_{h,k}^{z_{hk}=1} w_{hk}} - \frac{\sum_{h,k}^{z_{hk}=0} w_{hk} y_{hk}}{\sum_{h,k}^{z_{hk}=0} w_{hk}}, \quad (7)$$

where the weight w_{hk} is a function of the estimated propensity score. The choice of weight will be discussed in Section 2.3.

Assume y_{hk} is homoscedastic and $var(y_{hk}) = \sigma^2$, then the large sample variance of the marginal estimator is,

$$\begin{aligned} s^2_{.,marg} &= var(\hat{\Delta}_{.,marg}) \\ &= \frac{\sigma^2 \sum_{h,k}^{z_{hk}=1} w_{hk}^2}{(\sum_{h,k}^{z_{hk}=1} w_{hk})^2} + \frac{\sigma^2 \sum_{h,k}^{z_{hk}=0} w_{hk}^2}{(\sum_{h,k}^{z_{hk}=0} w_{hk})^2}. \end{aligned} \quad (8)$$

In practice σ^2 can be estimated from the sample variance of y_{hk}.

2.2.2 Clustered estimator

A second estimator is to first obtain the cluster-specific weighted difference and then calculate the weighted average of these differences based on the sum of weights in each cluster. That is, for cluster h,

$$\hat{\Delta}_h = \frac{\sum_{k\in h}^{z_{hk}=1} w_{hk} y_{hk}}{\sum_{k\in h}^{z_{hk}=1} w_{hk}} - \frac{\sum_{k\in h}^{z_{hk}=0} w_{hk} y_{hk}}{\sum_{k\in h}^{z_{hk}=0} w_{hk}}.$$

The variance of the cluster-specific estimator $\hat{\Delta}_h$ under the independent homoscedastic assumption of y_{hk} within cluster h is

$$\begin{aligned} s_h^2 &= var(\hat{\Delta}_h) \\ &= \frac{\sigma_h^2 \sum_{k\in h}^{z_{hk}=1} w_{hk}^2}{(\sum_{k\in h}^{z_{hk}=1} w_{hk})^2} + \frac{\sigma_h^2 \sum_{k\in h}^{z_{hk}=0} w_{hk}^2}{(\sum_{k\in h}^{z_{hk}=0} w_{hk})^2}. \end{aligned}$$

Similarly, σ_h^2 can be estimated from its empirical counterpart within each cluster.

Let w_h be a function of the weights in cluster h, e.g., the sum of weights $w_h = \sum_{k\in h} w_{hk}$, or the precision of the estimator $\hat{\Delta}_h$, $w_h = s_h^{-2}$. The overall clustered estimator is then an average of the $\hat{\Delta}_h$'s weighted by w_h,

$$\hat{\Delta}_{.,clu} = \frac{\sum_h w_h \hat{\Delta}_h}{\sum_h w_h}. \quad (9)$$

And the overall variance is

$$s^2_{.,clu} = var(\hat{\Delta}_{.,clu}) = \frac{\sum_h (\sum_{k\in h} w_{hk})^2 s_h^2}{(\sum_{h,k} w_{hk})^2}. \quad (10)$$

Standard errors of estimators $s^2_{.,marg}$ and $s^2_{.,clu}$ also be obtained from resampling methods such as the bootstrap.

2.2.3 Doubly-robust estimators

The weighted mean can be regarded as a weighted regression without covariates. Therefore in step 2, we can replace the nonparametric weighted mean (7) or (9) by a parametric regression (e.g., model (3) or (5)) weighted by the estimated propensity score. And the coefficient of the treatment assignment γ is the targeted estimand of treatment effect. This is essentially the class of doubly-robust estimators proposed by Scharfstein et al. (1999). Doubly-robust estimators allow flexible model choices in both steps, which can be very beneficial in applications. These estimators are coined "doubly-robust" in the sense that they are proven to be consistent if one but not necessarily both of the step 1 and 2 models are correctly specified under the Horvitz-Thompson weight (see below). Detailed discussion of this property with hierarchical data is presented in the next section.

2.3 Choice of weights

We now consider the choice of weights. We call the class of weights which balances the distribution of covariates between treatment groups *balancing weights*. The most widely used balancing weight is the Horvitz-Thompson (inverse probability) weight

$$w_{hk} = \begin{cases} \frac{1}{e_{hk}}, \text{for } z_{hk} = 1 \\ \frac{1}{1-e_{hk}}, \text{for } z_{hk} = 0. \end{cases}$$

The H-T weight is a balancing weight because $E\left[\frac{XZ}{e(X)}\right] = E\left[\frac{X(1-Z)}{1-e(X)}\right]$. The H-T estimator compares the expected outcome of the subjects placed in $z = 0$ versus that of the subjects placed in $z = 1$, averaging over the distribution of covariates in the combined population. That is,

$$E\left[\frac{YZ}{e(X)} - \frac{Y(1-Z)}{1-e(X)}\right] = E[(Y|Z=1) - (Y|Z=0)].$$

In fact, the doubly-robust estimators in Scharfstein et al. (1999) are restricted to using the H-T weight because of this clear causal interpretation. However, the H-T estimator has been well known to have excessively large variance when there are subjects with extremely small propensity score. Nevertheless, the same idea is readily extended to any balancing weight, although alternative weights might define different estimands. For example, we can consider the population-overlap weight,

$$w_{hk} = \begin{cases} 1 - e_{hk}, \text{for } z_{hk} = 1 \\ e_{hk}, \text{for } z_{hk} = 0. \end{cases}$$

where each subject is weighted by the probability of being assigned to the other treatment group. It is also a balancing weight because $E[XZ\{1-e(X)\}] = E[X(1-Z)e(X)]$. In theory, the population-overlap weight gives the smallest variance under a homoscedastic model for Y given X. But it defines a different estimand than the the Horvitz-Thompson weight. Specifically, we call this the population-overlap weight because it results in an average treatment effect that is averaged over the distribution of covariates in the population where the two treatment groups overlap

$$\begin{aligned} & E[YZ\{1-e(X)\} - Y(1-Z)e(X)] \\ = \; & E[\{(Y|Z=1) - (Y|Z=0)\}e(X)\{1-e(X)\}]. \end{aligned}$$

This population-overlap estimator can be calculated with acceptable variance when the H-T estimator cannot be practically estimated, because $e(x)$ can approach 0 or 1 for some part of x space such that $\frac{1}{e(x)}$ or $\frac{1}{1-e(x)}$ would become extremely large. In effect the H-T estimator attempts to estimate a treatment effect for types of cases which are essentially unrepresented in one or the other group, while the population-overlap weighting focuses on the types of cases with a more balanced distribution of "treatment". In addition to its statistical advantage, the latter analysis may be more scientifically relevant since it focuses attention on comparison of outcomes among the kinds of cases which both "treatments" are currently observed, for example those in clinical equipoise between treatments.

3. Bias of Estimators

In this section, we investigate the bias of each of the estimators proposed in the previous section. We first look at the simplest case with two level-hierarchical structure and no covariates.

Let $n_{h1}(n_{h0})$ denote the number of subjects with $z = 1(z = 0)$ in cluster h; and $n_{+1} = \sum_h n_{h1}, n_{+0} = \sum_h n_{h0}, n_{++} = n_{+1} + n_{+0}$.

Assume the *outcome generating mechanism* for a continuous outcome follows a random effects model with cluster-level random intercepts and random treatment effects,

$$y_{hk} = \delta_h + \gamma_h z_{hk} + \alpha d_h + \epsilon_{hk}, \tag{11}$$

where $\delta_h \sim N(0, \sigma_\delta^2), \epsilon_{hk} \sim N(0, \sigma_\epsilon^2)$, α is the effect of the cluster-specific proportion of being treated d_h on the outcome, and the true treatment effect is γ_h with $\gamma_h \sim N(\gamma_0, \sigma_\gamma^2)$.

We first look at the situation where clustering information is ignored in both steps. For the marginal model in step 1, it is easy to show that the estimated propensity score is the same for each subject $\hat{e}_{hk} = \frac{n_{+1}}{n_{++}}$. Consequently, the marginal estimator is

$$\begin{aligned} & \hat{\Delta}_{marg,marg} \\ = \; & \frac{\sum_{h,k}^{z_{hk}=1} y_{hk}}{n_{+1}} - \frac{\sum_{h,k}^{z_{hk}=0} y_{hk}}{n_{+0}} \\ = \; & \sum_h \frac{n_{h1}}{n_{+1}}\gamma_h + \sum_h (\frac{n_{h1}}{n_{+1}} - \frac{n_{h0}}{n_{+0}})\delta_h \\ & + (\frac{\sum_{h,k}^{z_{hk}=1} \epsilon_{hk}}{n_{+1}} - \frac{\sum_{h,k}^{z_{hk}=0} \epsilon_{hk}}{n_{+0}}) \\ & + \alpha \frac{\frac{n_{++}}{n_{+1}n_{+0}} - \sum_h n_h d_h (1-d_h)}{\frac{n_{++}}{n_{+1}n_{+0}}} \end{aligned}$$

Assume the common regularity conditions $\sum_h^\infty \frac{n_{h1}^2}{n_{+1}^2} < \infty$ and $\sum_h^\infty \frac{n_{h0}^2}{n_{+0}^2} < \infty$ hold, then by the weak law of large num-

bers for the weighted sum of independent and identically-distributed random variables (e.g., Chow and Lai, 1973), $\sum_h \frac{n_{h1}}{n_{+1}}\gamma_h$ converges to γ_0 as the number of clusters goes to infinity, and $\sum_h(\frac{n_{h1}}{n_{+1}} - \frac{n_{h0}}{n_{+0}})\delta_h$ goes to 0, so does the third term in the above formula. In the fourth term, $\frac{n_{++}}{n_{+1}n_{+0}}$ is in fact the variance of the total number of treated subjects, $var(n_{+1})$, if all clusters are exchangeable, i.e., if all subjects regardless of the clusters follow the same treatment assignment mechanism, $z \sim Bernoulli(\frac{n_{+1}}{n_{++}})$. Furthermore, $\sum_h n_h d_h(1-d_h)$ is the sum of the variance of the number of treated subjects within each cluster, $\sum_h var(n_{h1})$, if each cluster separately follows a treatment assignment mechanism, $z_{k\in h} \sim Bernoulli(\frac{n_{h1}}{n_h})$. Therefore, bias of the marginal estimator with propensity score estimated from the marginal model is

$$Bias(\hat{\Delta}_{marg,marg}) = \alpha\left[\frac{var(n_{+1}) - \sum_h var(n_{h1})}{var(n_{+1})}\right]. \tag{12}$$

The size of the bias is controlled by two factors: (1) the ratio of the variance of the total number of treated subjects under a homogeneous versus a cluster-heterogeneous treatment assignment mechanism; and (2) the effect that the cluster-specific proportion of being treated d_h has in the response, i.e., $|\alpha|$. This is intuitive because the first factor measures the variation in the treatment assignment mechanism among clusters and the second measures the variation in the outcome generating mechanism, both of which are ignored in the analysis with marginal models in both steps. When either but not necessarily both of the two mechanisms is homogenous across clusters, the marginal estimator, $\hat{\Delta}_{marg,marg}$, is also consistent. However, in reality, it is most likely that both of the mechanisms are heterogenous among clusters.

We now look at the opposite situation where clustering information is taken into account in both steps. For the pooled within-cluster model in step 1, it is easy to show that the estimated propensity score is $\hat{e}_{hk} = \frac{n_{h1}}{n_h}$. Then the clustered weighted estimator is

$$\begin{aligned}
&\hat{\Delta}_{pool,clu} \\
=\ & \frac{\sum_h(\sum_{k\in h}^{z_{hk}=1}\frac{y_{hk}}{n_{h1}})}{m} - \frac{\sum_h(\sum_{k\in h}^{z_{hk}=0}\frac{y_{hk}}{n_{h0}})}{m} \\
=\ & \frac{\sum_h \gamma_h}{m} + \frac{\sum_h(\sum_{k\in h}^{z_{hk}=1}\frac{\epsilon_{hk}}{n_{h1}})}{m} - \frac{\sum_h(\sum_{k\in h}^{z_{hk}=0}\frac{\epsilon_{hk}}{n_{h0}})}{m} \\
\overset{n_h,m\to\infty}{\to}\ & \gamma_0
\end{aligned} \tag{13}$$

which is asymptotically unbiased. The result is free of the form of weight. Simple calculation shows that the clustered weighted estimator combining the marginal model in step 1, $\hat{\Delta}_{marg,clu}$, is of exactly the same form as that in (13) and thus also unbiased. Furthermore, the marginal estimator with propensity score estimated from the pooled within-cluster model, $\hat{\Delta}_{pool,clu}$, follows the same form as in (13), but only under H-T weight and a balanced design (i.e., each cluster has same number of subjects). Under H-T weight but an unbalanced design, the estimator is also consistent (assume $\sum_h^\infty \frac{n_h^2}{n_{++}^2} < \infty$) as

$$\hat{\Delta}_{pool,marg} = \frac{\sum_h n_h\gamma_h}{n_{++}} \overset{n_h,m\to\infty}{\to} \gamma_0.$$

However, the same estimator under the population-overlap weight is

$$\begin{aligned}
&\hat{\Delta}_{pool,marg} \\
=\ & \frac{\sum_h n_{h0}(\sum_{k\in h}^{z_{hk}=1}\frac{y_{hk}}{n_h}) - \sum_h n_{h1}(\sum_{k\in h}^{z_{hk}=0}\frac{y_{hk}}{n_h})}{\sum_h \frac{n_{h1}n_{h0}}{n_h}} \\
=\ & \frac{\sum_h \frac{n_{h1}n_{h0}}{n_h}(\gamma_h + \sum_{k\in h}^{z_{hk}=1}\frac{\epsilon_{hk}}{n_{h1}} - \sum_{k\in h}^{z_{hk}=0}\frac{\epsilon_{hk}}{n_{h0}})}{\sum_h \frac{n_{h1}n_{h0}}{n_h}} \\
\overset{n_h,m\to\infty}{\to}\ & \gamma_0.
\end{aligned}$$

Even though this estimator is also asymptotically unbiased, its small sample behavior can be quite different from that of the estimator under H-T weight.

Under the homoscedasticity assumption of outcome, the three H-T estimators $\hat{\Delta}_{pool,marg}$, $\hat{\Delta}_{marg,clu}$, and $\hat{\Delta}_{pool,clu}$ that take into account clustering in at least one step have the same variance,

$$s^2 = \sum_h \frac{\sigma_\epsilon^2 n_h^2}{n_{++}^2}(\frac{1}{n_{h1}} + \frac{1}{n_{h0}}).$$

Similarly as the discussion on bias, this result is generally not applicable for other type of weights. Specifically, the variance of $\hat{\Delta}_{pool,marg}$ is usually larger than that of $\hat{\Delta}_{marg,clu}$ and $\hat{\Delta}_{pool,clu}$.

When there are no covariates, the surrogate indicator model gives the estimated propensity score as the pooled within-cluster model. Thus the results obtained above regarding the pooled within-cluster model automatically hold for the surrogate indicator model. But this is not the case for the general situation with covariates.

The proofs are analogous for data with a higher order of hierarchical levels. For the simplest case without covariates, above we have shown the "double-robustness" of those propensity score estimators, that is, when both of the true underlying treatment assignment mechanism and outcome generating mechanism are hierarchically structured, the estimator using a balancing weight is consistent as long as the hierarchical structure is taken into account in at least one of the two steps in the propensity score procedure. This can be viewed as both a special case and an extension of the "double-robustness" property of the estimator in Scharfstein et al. (1999). The extension lies in that our conclusion is instead free of the form of weight.

In the more general cases with covariates, usually there is no closed-form solution to the logistic models for estimating the propensity score. Consequently, there is no closed-form of the bias of those estimators as above. Nevertheless, this situation can be explored either by large-scale simulations, or by adopting a probit (instead of logistic) link for estimating the propensity score. Intuitively, the "double-robustness" property still holds. But the bias of a marginal estimator $\hat{\Delta}_{marg,marg}$ is

expected to also be affected by the size of the true treatment effect γ (negative correlated) and the ratio of between-cluster and within-cluster variance $g = \frac{\sigma_\delta^2}{\sigma_\epsilon^2}$ (positively correlated), in addition to α and $\frac{var(n_{+1}) - \sum_h var(n_{h1})}{var(n_{+1})}$ in (12). A comprehensive discussion is beyond the scope this report and is subject to further research.

4. Application

We now apply the above methods to study racial disparity in health services. Disparity refers to racial differences in care attributed to operations of health care system. Our application concerns the HEDIS® measures of health care provided in Medicare health plans. Each of these measures is an estimate of the rate at which a guideline-recommended clinical service is provided within the appropriate population. We obtained individual-level data from the Centers for Medicare and Medicaid Services (CMS) on breast cancer screening of women in Medicare managed care health plans (Schneider et al., 2002). Our main interest is the disparity between whites and blacks, so we exclude subjects of other races for whom racial identification is unreliable in this dataset. We focus on plans with at least 25 whites and 25 blacks, leaving 64 plans with a total sample size of 75012. For practical reasons, we drew a random subsample of size 3000 from each of the three large plans with more than 3000 subjects, leaving a total sample size of 56480.

All the covariates considered in the analysis are binary. The individual-level covariates x_{hk} include two indicators of age category (70-80,>80) with reference group being 60-70; eligibility for Medicaid (1 yes); neighborhood status indicator (1 poor). The plan-level covariates v_h include nine geographical code indicators; non/for-profit status (1 for-profit); and the practice model of providers (1 staff-group model; 0 network-independent practice model). The outcome y is a binary variable equal to 1 if the enrollee underwent breast cancer screening and equal to 0 otherwise, and the "treatment" z here is race (1 black, 0 white). We want to estimate the difference in the proportion of undergoing breast cancer screening between whites and blacks. As mentioned before, race is not a valid "treatment" in conventional sense in causal inference, because it is not manipulable (Holland, 1986). However, in this particular application, our goal is not to study the causal pathway between race and health service utilization, but simply to estimate the magnitude of disparity under balanced distributions of covariates between the two races. Hence, the propensity score in this application is merely an analytical tool to achieve this goal, and it should not be taken as having the explicit meaning of the probability of being black.

We first estimate the propensity score using the three models introduced in Section 2.1 with all the above covariates included. Details of the fitted models are omitted here since the focus is the fitted values (estimated propensity score). All models suggest that living in poor neighborhood, being eligible for Medicaid and enrollment in for-profit insurance plan are significantly associated with being black race. Figures 1 and 2 show histograms of the estimated propensity score for whites and blacks. Different models clearly give quite different estimates of propensity score in this data, where the marginal model departs mostly from the other two models. The variance of the estimated propensity score of blacks is much bigger than that of whites, regardless of the model. We checked the weighted distributions of covariates. Each model leads to good balance of the overall weighted covariates distributions between groups. However, the marginal model in general does poorly in balancing covariates between races within each cluster, while the surrogate indicator model does better, and the pooled within-cluster model does the best. This suggests that there is important between-cluster variation.

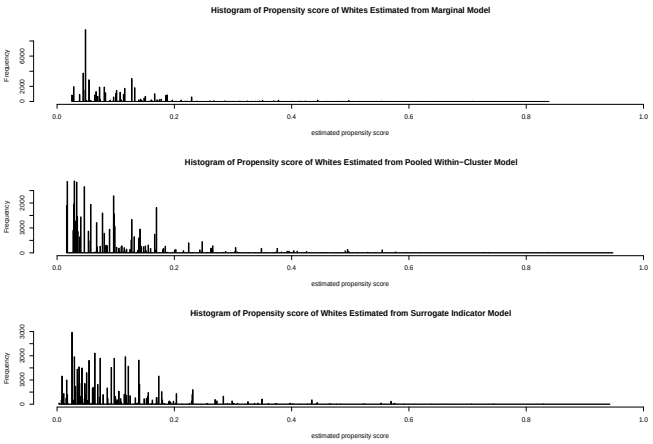

Figure 1: Histogram of propensity score estimated from different models for whites.

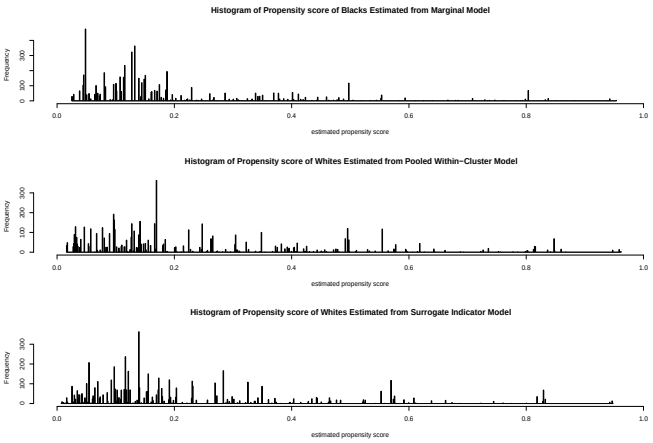

Figure 2: Histogram of propensity score estimated from different models for blacks.

Using the estimated propensity score, we estimate racial disparity in breast cancer screening among the elder women participating Medicare health plans by the estimators proposed in Section 2.2. Although the outcome is binary in this case, the probabilities of outcome are in a range where the linear probability model is an acceptable fit. Hence, for the doubly-robust estimators, we adopt the combinations of the three propensity score models (2), (4) and (6) in step 1 and

the two outcome models (3) and (5) in step 2. Table 1 shows the estimates using the H-T weight. Each row represents one step 1 model, and each column represents one type of step 2 model/estimator. Analogous results using the population-overlap weight are given in Table 2.

	weighted		doubly-robust	
	marginal	clustered	marginal	pooled within
marginal	-0.050	-0.020	-0.042	-0.021
	(0.008)	(0.008)	(0.004)	(0.004)
pooled	-0.024	-0.021	-0.018	-0.022
within	(0.009)	(0.008)	(0.004)	(0.004)
surrogate	-0.017	-0.015	-0.012	-0.015
indicator	(0.009)	(0.008)	(0.004)	(0.004)

Table 1: Difference in the proportion of getting breast cancer screening between blacks and whites using Horvitz-Thompson weight

All models show the proportion of receiving breast cancer screening is significantly lower among blacks than among whites with similar characteristics. The estimates are similar except for the analyses that ignore clustering in both steps, which overestimate the treatment effect. This pattern matches the double-robustness property. Results from the surrogate indicator model in step 1 are slightly different from the others, suggesting the cluster-specific proportion of being treated d_h is correlated with certain covariates. The doubly-robust estimates have smaller standard errors because the extra variation is explained by covariates in step 2. Not surprisingly, the estimates using H-T weight have much larger variances than those using the population-overlap weight. We also 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.

Even though we do not know the underlying truth, the similarity of various estimators suggests our analyses capture the main information regarding disparity in this data. 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, social economical status and health plan characteristics.

	weighted		doubly-robust	
	marginal	clustered	marginal	pooled within
marginal	-0.043	-0.030	-0.043	-0.032
	(0.007)	(0.008)	(0.004)	(0.004)
pooled	-0.030	-0.031	-0.031	-0.031
within	(0.007)	(0.008)	(0.004)	(0.004)
surrogate	-0.035	-0.030	-0.031	-0.030
indicator	(0.007)	(0.008)	(0.004)	(0.004)

Table 2: Difference in the proportion of getting breast cancer screening between blacks and whites using population-overlap weight

5. Summary and Remarks

Since first been proposed twenty-five years ago, 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 rule rather than exception nowadays. However, despite the wide appreciation of propensity score among both statisticians and health policy researchers, there is very limited literature regarding the methodological issues of propensity score methods in the context of hierarchical data, which motivates our exploration in this paper. Specifically, we present three typical models for estimating propensity score and two types of nonparametric weighted (by estimated propensity score) estimators of treatment effect for hierarchically structured data. Furthermore, for the simplest (conceptual) case without covariates, we show the "double-robustness" of those weighted estimators: when both of the true underlying treatment assignment mechanism and outcome generating mechanism are hierarchically structured, the estimator is consistent as long as the hierarchical structure is taken into account in at least one of the two steps in the propensity score procedure. We also quantify the bias of the estimator when clustering is ignored in both steps.

We have focused on the case of treatment being assigned at the individual level in this paper. Treatment assigned at the cluster level (e.g., hospital, health care provider) is also common in medical care and health policy studies, where several new challenging issues can arise. 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, the cluster-level propensity score only balances the cluster-level covariates and the average individual-level coviariates. What are the consequences of the possible imbalance in the overall distributions of individual-level covariates? This also has a strong connection to the ecological inference commonly encountered in political science (e.g., King, 1997) where the estimand has an interpretation as an average effect on individual outcomes. Third, 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 appear to be preferable in this case. But what specific regression model to choose greatly depends on the specific data. Fourth, most interestingly, the foundational stable-unit-treatment-value assumption (SUTVA) – "the observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox 1958, 2.4) often no longer holds under clustered treatment assignment, especially in the studies with, for instance, behavioral outcomes and infectious disease. In that case, correct modeling of the interference

among subjects is crucial for valid analysis. Those 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.

REFERENCES

Connors, A., Speroff, T., Dawson, N., and et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association* **276**, 889-897.

Cox. C.P. (1958). The Analysis of Latin Square Designs with Individual Curvatures in one Direction. *Journal of the Royal Statistical Society. Series B.* **20(1)**, 193-204.

Chow, Y. S. and Lai, T.L. (1973). Limiting behavior of weighted sums of independent random variables. *The Annals of Probability* **1(5)**, 810-824.

D'Agostino, R. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparisons of a treatment to a non-randomized control. *Statistics in Medicine* **17**, 2265-2281.

Farrow, D., Samet, J. and Hunt, W. (1996). Regional variation in survival following the diagnosis of cancer. *Journal of Clinical Epidemiology* **49**, 843-847.

Gatsonis, C., Normand, S., Liu, C., and Morris, C. (1993). Geographic variation of procedure utilization: a hierarchical model approach. *Medical Care* **31**, YS54-YS59.

King, G. (1997). A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton University Press.

Holland, P.W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945-970.

Huang, I.C., Frangakis, C.E., Dominici, F., Diette, G. and Wu, A.W. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research* **40**, 253-278.

Nattinger, A., Gottilieb, M., Veum, J., and et al. (1992). Geographic variation in the use of breast-conserving treatment for breast cancer. *New England Journal of Medicine* **326**, 1102-1127.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70(1)**, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516-524.

Rubin, D.B. (1978) Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* **6**, 34-58.

Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318-324.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94**, 1096-1146.

Schneider E.C., Zaslavsky A.M., Epstein, A.M. (2002). Racial disparities in the quality of care for enrollees in Medicare managed care. *Journal of the American Medical Association* **287(10)**, 1288-1294.