Biometrics 72, 1055–1065
December 2016
DOI: 10.1111/biom.12505

Propensity Score Matching and Subclassification in Observational Studies with Multi-Level Treatments

Shu Yang,[1,*] Guido W. Imbens,[2] Zhanglin Cui,[3] Douglas E. Faries,[3] and Zbigniew Kadziola[4]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
[2]Graduate School of Business, Stanford University and NBER, Stanford, California 94305, U.S.A.
[3]Real World Analytics, Eli Lilly and Company, Indianapolis, Indiana 46285, U.S.A.
[4]Real World Analytics, Eli Lilly and Company, Vienna, Austria
email: syang24@ncsu.edu

Summary. In this article, we develop new methods for estimating average treatment effects in observational studies, in settings with more than two treatment levels, assuming unconfoundedness given pretreatment variables. We emphasize propensity score subclassification and matching methods which have been among the most popular methods in the binary treatment literature. Whereas the literature has suggested that these particular propensity-based methods do not naturally extend to the multi-level treatment case, we show, using the concept of weak unconfoundedness and the notion of the generalized propensity score, that adjusting for a scalar function of the pretreatment variables removes all biases associated with observed pretreatment variables. We apply the proposed methods to an analysis of the effect of treatments for fibromyalgia. We also carry out a simulation study to assess the finite sample performance of the methods relative to previously proposed methods.

Key words: Generalized propensity score; Matching; Multi-level treatments; Potential outcomes; Subclassification; Unconfoundedness.

1. Introduction

There is an extensive theoretical and empirical literature on estimating average causal effects of binary treatments in observational studies based on the assumption of unconfoundedness or ignorable treatment assignment. Under this assumption differences in outcomes for units with different treatment levels, but the same values for pretreatment variables, can be interpreted as estimates of causal effects. Much of the literature builds on the seminal article by Rosenbaum and Rubin (1983)(RR83 from here on) which clarified the central role of the propensity score (the conditional probability of receiving the treatment given the pretreatment variables or covariates) in analyses of causal effects in such settings, and which proposed a number of widely used estimators. See Imbens and Rubin (2015) for a textbook treatment.

Although important for empirical practice, much less theoretical work has been done on the setting with more than two treatment levels (exceptions include Imbens, 2000; Robins, Hernan, and Brumback, 2000; Lechner, 2001; Foster, 2003; Hirano and Imbens, 2004; Imai and Van Dyk, 2004; Cole and Frangakis, 2009; Cadarette et al., 2010; Cattaneo, 2010; McCaffrey et al., 2013; Rassen et al., 2013). Because in settings with multi-level treatments, there is no scalar function of the covariates that has all the properties that RR83 presents for the propensity score in the binary treatment case, it has been claimed that there is no natural analog to matching and subclassification on the propensity score (Imbens, 2000; Lechner, 2001; Rassen et al., 2013).

In the main contribution of the current article we show that, contrary to these claims, the essence of the results in RR83 generalizes to the setting with multi-level treatments. In particular, we develop methods for matching and subclassification on scalar functions of the covariates that are valid irrespective of the number of distinct levels for the treatment. The key insight is that we do not construct sets of units with the balancing property that within these sets the treatment level is independent of the covariates. Doing so would require adjusting for the vector of propensity scores with length equal to the number of treatment levels minus one. Instead, we focus on estimating the average of the potential outcomes separately for each treatment level, which requires adjusting only for the probability of receiving that particular level of the treatment. This insight allows us to extend some of the most widely used methods for the binary treatment case to the multi-level treatment case without giving up the dimension reducing property of the propensity score. We provide some simulation evidence that demonstrates the relevance of concerns with the previously proposed estimators and the promise of the new methods.

2. Set Up

Following Rubin (1974) and RR83 we use the potential outcome set up, generalized to the case with more than two, unordered, levels for the treatment in Imbens (2000), Lechner (2001), Imai and Van Dyk (2004), and Cattaneo (2010). The treatment is denoted by $W_i \in \mathbb{W} = \{1, \ldots, T\}$. In the standard binary treatment case $T = 2$, the two treatments are often labeled treatment and control. For each unit i there are T potential outcomes, one for each treatment level, denoted by $Y_i(w)$, for $w \in \mathbb{W}$. Implicitly in this notation is the assumption that

there is no interference between units and no versions of each treatment level (the stable-unit-treatment-value assumption, or sutva, Rubin, 1978). The observed outcome for unit i is the potential outcome corresponding to the treatment received:

$$Y_i^{\text{obs}} = Y_i(W_i).$$

We also observe a vector-valued covariate or pretreatment variable, denoted by X_i. These pretreatment variables are known a priori not to be affected by the treatment, typically measured prior to the determination of the treatment level, and so there are no multiple versions of these covariates corresponding to the different levels of the treatment. These pretreatment variables may include fixed attributes of the units, or measurements prior to the treatment assignment that predict the outcome, for example, prior health status. Although we do not stress this in the notation, there is implicitly a temporal aspect to the study with three stages: first, the pretreatment variables are measured, or at least they assume their values prior to the assignment of the treatment; second, the treatment is assigned or selected; third, the outcome assumes its value and is measured.

We assume the sequence $(W_1, X_1, Y_1(1), \ldots, Y_1(T)), \ldots, (W_N, X_N, Y_N(1), \ldots, Y_N(T))$ with the potential outcomes is i.i.d., so that the sequence of realized values $(W_1, X_1, Y_1^{\text{obs}}), \ldots, (W_N, X_N, Y_N^{\text{obs}})$ is also i.i.d.

Following the literature (e.g., RR83), we focus on average treatment effects as the causal estimands. This is less restrictive than it may appear at first because we can first take transformations of the outcomes and pretreatment variables. For the comparison between treatment levels w and w', the average effect is

$$\tau(w, w') = \mathbb{E}[Y_i(w') - Y_i(w)]. \tag{1}$$

The expectation is taken with respect to the same population (called the target population, Frolich, 2004a) for different treatment-level pairs (w, w'). Some researchers, when analyzing data with multi-level treatments, have used conventional methods for comparing two treatment levels at a time. Often such analyses use only information on units exposed to one of those two treatment levels, which would lead to estimates of $\mathbb{E}[Y_i(w') - Y_i(w) \mid W_i \in \{w, w'\}]$. If the subpopulation of units with treatment levels w or w' is different in terms of potential outcome distributions, these estimands are generally different from the $\tau(w, w')$ defined in (1), because the latter do not condition on $W_i \in \{w, w'\}$. As a result, such binary-comparison analyses make it difficult to compare $\mathbb{E}[Y_i(w') - Y_i(w) \mid W_i \in \{w, w'\}]$ and $\mathbb{E}[Y_i(w'') - Y_i(w) \mid W_i \in \{w, w''\}]$ because they refer to different populations.

In this article, we mainly focus on the case where the different treatment levels are qualitatively distinct. In that case, the interest is typically in average effects of the form $\tau(w, w')$. In other cases, however, the treatment levels may measure the quantity of a dose. In such cases, the researcher may be interested in weighted combinations of average effects. For example, one might be interested in $\sum_{w=1}^{T-1} \lambda_w \tau(w, w+1)$, with the weights adding up to one, which would correspond to a weighted average of unit increases in the dose. One advantage of such estimands is that their variance may be lower than that for particular $\tau(w, w')$. In this case there may also be particular interest in $\tau(1, T)$, the effect of the maximum dose. Our results also apply to all such estimands.

3. Weak and Strong Unconfoundedness and the Generalized Propensity Score

Our focus is on observational studies where assignment to treatment is not completely random. Instead, following a large strand of the observational studies literature, we assume that assignment to treatment is unconfounded so that, within subpopulations that are homogenous in observed pretreatment variables, assignment to treatment is as good as random. This is strictly weaker than complete randomization by allowing for general associations between the treatment level and the pretreatment variables.

3.1. *The Generalized Propensity Score*

In this section, we discuss the generalization of the notion of the propensity score, introduced in the causality literature by RR83 for the binary treatment case, to our setting with multi-level treatments. In the binary treatment case, RR83 defines the propensity score as the conditional probability of receiving the active treatment rather than the control treatment, $p(x) = pr(W_i = 1 \mid X_i = x)$. Here, we generalize that to the multi-level treatment case, following Imbens (2000):

Definition 1. *(Generalized Propensity Score) The generalized propensity score is the conditional probability of receiving each treatment level:*

$$p(w \mid x) = pr(W_i = w \mid X_i = x).$$

3.2. *Overlap*

Before formally discussing unconfoundedness assumptions, let us assume that there is overlap in the covariate distributions:

Assumption 1. *(Overlap) For all values of x, the probability of receiving any level of the treatment is positive:*

$$p(w \mid x) > 0 \quad \textit{for all } w, x.$$

Without this assumption, there will be values of x for which we cannot estimate the average effect of some treatments relative to others without relying on extrapolation. In Section 6, we discuss methods for constructing a subsample with better overlap for cases where this assumption is (close to) violated.

3.3. *Strong Unconfoundedness*

We start by generalizing the conventional RR83 version of the unconfoundedness assumption to the case with multi-level treatments. We refer to this as strong unconfoundedness to distinguish it from the weaker condition of weak unconfoundedness.

Definition 2. *(Strong Unconfoundedness) The assignment mechanism is strongly unconfounded if*

$$W_i \perp\!\!\!\perp \Big(Y_i(1), \ldots, Y_i(T)\Big) \mid X_i.$$

Here, we use the $\perp\!\!\!\perp$ notation introduced by Dawid (1979) to denote (conditional) independence.

The assumption of strong unconfoundedness has no testable implications. In a particular application the assumption is a substantive one, and often a controversial one. Often it can be made more plausible by collecting detailed information at baseline on characteristics of the units that are related to treatment and outcome. As a result, the dimension of X_i may be high.

One implication of strong unconfoundedness is the following extension of the propensity score result in RR83 to the multi-level treatment case:

LEMMA 1. *(RR83) Suppose the assignment mechanism is strongly unconfounded. Then,*

$$W_i \perp\!\!\!\perp \Big(Y_i(1), \ldots, Y_i(T)\Big) \Big| \Big(p(1 \mid X_i), \ldots, p(T-1 \mid X_i)\Big) .$$

Because $\sum_{w=1}^{T} p(w|x) = 1$, it follows that $p(T|x)$ is a linear combination of $p(1|x), \ldots, p(T-1|x)$, and so we do not need to include $p(T|x)$ in the conditioning set. If there are two levels of the treatment, the result in the lemma reduces to the result in RR83. As pointed out in Imbens (2000) and Rassen et al. (2013), the dimension reduction property of the lemma depends on the number of distinct levels for the treatment, and therefore the result is less useful in settings with a substantial number of treatment levels. The problem is that without additional assumptions, there is in general no scalar function $b(x)$ of the covariates such that $W_i \perp\!\!\!\perp (Y_i(1), \ldots, Y_i(T)) \mid b(X_i)$, suggesting that the advantages of the propensity score approach do not carry over to the multi-level treatment case. Joffe and Rosenbaum (1999); Lu et al. (2001); Imai and Van Dyk (2004); Zanutto, Lu, and Hornik (2005) discuss additional assumptions under which functions $b(\cdot)$ exist with this property and whose dimension is lower than $T-1$. In particular, Lu et al. (2001) assume that a scalar balancing function $b(\cdot)$ exists and propose a matching estimator based on $b(\cdot)$, and Zanutto, Lu, and Hornik (2005) propose a subclassification estimator under this assumption. Nevertheless, in general such functional form assumptions may be controversial.

3.4. *Weak Unconfoundedness*

We improve the dimension reduction property of the generalized propensity score by weakening the requirement of strong unconfoundedness condition to weak unconfoundedness. Define the T indicator variables $D_i(w) \in \{0, 1\}$:

$$D_i(w) = \begin{cases} 1 & \text{if } W_i = w, \\ 0 & \text{otherwise.} \end{cases}$$

In terms of these indicator variables, strong unconfoundedness is equivalent to

$$\Big(D_i(1), \ldots, D_i(T-1)\Big) \perp\!\!\!\perp \Big(Y_i(1), \ldots, Y_i(T)\Big) \mid X_i.$$

Now, we can formulate the weak unconfoundedness notion, introduced in Imbens (2000).

DEFINITION 3. *(Weak unconfoundedness) The assignment mechanism is weakly unconfounded if for all $w \in \mathbb{W}$,*

$$D_i(w) \perp\!\!\!\perp Y_i(w) \mid X_i.$$

Although formally it is obviously weaker, we do not wish to argue that weak unconfoundedness is substantively weaker than strong unconfoundedness. In fact neither have testable implications, and there appear to be no interesting estimands that are identified under the stronger assumption but not under the weaker assumption. Rather, the two key insights, and the motivation for distinguishing between the two unconfoundedness assumptions, are, one, that, as shown in Lemma 2 below, weak unconfoundedness is preserved if we condition on a scalar function of the pretreatment variables, whereas preserving strong unconfoundedness requires conditioning on a set of $T-1$ functions of the pretreatment variables, as shown in Lemma 1, and two, that weak unconfoundedness is sufficient for identifying average causal effects, as formalized in Lemma 3 below.

LEMMA 2. *(Weak Unconfoundedness) Suppose the assignment mechanism is weakly unconfounded. Then for all $w \in \mathbb{W}$,*

$$D_i(w) \perp\!\!\!\perp Y_i(w) \mid p(w|X_i).$$

LEMMA 3. *(Average Causal Effects under Weak Unconfoundedness) Suppose the assignment mechanism is weakly unconfounded. Then,*

$$\mathbb{E}[Y_i(w') - Y_i(w)] = \mathbb{E}\Big[\mathbb{E}[Y_i^{\text{obs}} \mid W_i = w', p(w' \mid X_i)]\Big] - \mathbb{E}\Big[\mathbb{E}[Y_i^{\text{obs}} \mid W_i = w, p(w \mid X_i)]\Big].$$

Lemma 3 is the key result. For its interpretation, it is useful to compare it to the standard result under strong unconfoundedness. Under the strong unconfoundedness assumption, we create subpopulations where we can simultaneously compare units with all different levels of the treatment, leading to

$$\begin{aligned} &\mathbb{E}[Y_i(w') - Y_i(w)] \\ &= \mathbb{E}\Big[\mathbb{E}[Y_i^{\text{obs}} \mid W_i = w', p(1 \mid X_i), \ldots, p(T-1|X_i)]\Big] \\ &\quad - \mathbb{E}\Big[\mathbb{E}[Y_i^{\text{obs}} \mid W_i = w, p(1 \mid X_i), \ldots, p(T-1 \mid X_i)]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}[Y_i(w') - Y_i(w) \mid p(1 \mid X_i), \ldots, p(T-1|X_i)]\Big]. \end{aligned}$$

To allow for comparisons of all treatments, these subpopulations were defined by common values for the full set of $T-1$ propensity scores $(p(1 \mid X_i), \ldots, p(T-1 \mid X_i))$. Under

weak unconfoundedness we do not, and in fact cannot, construct such subpopulations. However, in order to estimate the average effect $E[Y_i(w') - Y_i(w)]$, it is not necessary to do so. Instead, we construct, for each treatment level w separately, subpopulations where we can estimate the average value of the potential outcomes, but only for that single treatment level. For treatment level w, these subpopulations are defined by the value of a single score, $p(w|X_i)$, leading to the equality

$$\mathbb{E}[Y_i(w)] = \mathbb{E}\Big[\mathbb{E}[Y_i^{\text{obs}} \mid W_i = w,\, p(w \mid X_i)]\Big].$$

That difference in focus allows us to reduce the dimension of the conditioning variable to a scalar, irrespective of the number of treatment levels.

4. Matching

In this section, we discuss matching methods. First, we discuss conventional matching on the full set of pretreatment variables. This is not a new method, but it will be useful to contrast with the proposed methods. Then, we discuss how the generalized propensity score can be used to develop a new matching estimator that matches only on a scalar function of the pretreatment variables.

4.1. *Matching*

Frölich (2004b) demonstrates covariate matching in multi-level treatments. Here, we focus on nearest neighbor matching. Other modifications include multiple nearest neighbors matching, kernel matching, and so forth. Reviews of general matching methods can be found in Hirano and Imbens (2004); Huber, Lechner, and Wunsch (2013); Imbens and Rubin (2015). Define the covariate matching function $m_{\text{cov}} : \mathbb{W} \times \mathbb{X} \mapsto \{1, \ldots, N\}$ as the index for the unit with treatment level w that is closest to x in terms of covariates (ignoring ties):

$$m_{\text{cov}}(w, x) = \arg\min_{j:W_j=w} ||X_j - x||.$$

Here, we use $||\cdot||$ to denote a generic metric. In practice, one would typically use the Mahalabonis metric, where $||x - x'|| = \{(x - x')^T V^{-1}(x - x')\}^{1/2}$, with $V = \sum_i (X_i - \overline{X})(X_i - \overline{X})^T/N$, and $\overline{X} = \sum_i X_i/N$. Note that the set of indices we search over includes all units, including unit i itself, so that for all i, $m_{\text{cov}}(W_i, X_i) = i$. Given the covariate matching function $m_{\text{cov}}(w, x)$, the potential outcomes for unit i are imputed as

$$\hat{Y}_i(w) = Y^{\text{obs}}_{m_{\text{cov}}(w,X_i)},$$

for $w = 1, \ldots, T$. Now, we estimate $\tau(w, w')$ as

$$\hat{\tau}_{\text{cov}}(w, w') \;= N^{-1} \sum_{i=1}^{N} \left(Y^{\text{obs}}_{m_{\text{cov}}(w',X_i)} - Y^{\text{obs}}_{m_{\text{cov}}(w,X_i)}\right). \qquad (2)$$

Note that to estimate $\tau(w, w')$, we impute potential outcomes $Y_i(w)$ and $Y_i(w')$ even for units who did not receive either treatment level w or w'. This ensures comparability of average treatment effects for different pairs of treatments.

4.2. *Matching on the Generalized Propensity Score*

Just as in the binary treatment setting, matching on all covariates is not an attractive procedure in the multi-level treatment setting if the number of covariates is substantial (*e.g.*, Imai and Van Dyk, 2004; Abadie and Imbens, 2006; Imbens and Rubin, 2015). In the binary treatment case, RR83 proposed matching on the propensity score to reduce the dimensionality of the matching problem. If $p(1|x)$ is the Rosenbaum–Rubin propensity score, the matching function for the binary treatment case would be

$$m^{\text{binary}}_{\text{ps}}(w, p) = \arg\min_{j:W_j=w} ||p(1|X_j) - p||. \qquad (3)$$

One could generalize that to the multi-level treatment case by matching on the full set of scores, leading to

$$m^{\text{multilvl}}_{\text{gps}}(w, p_1, \ldots, p_{T-1}) = \arg\min_{j:W_j=w} \left\| \begin{pmatrix} p(1|X_j) - p_1 \\ \vdots \\ p(T-1|X_j) - p_{T-1} \end{pmatrix} \right\|. \qquad (4)$$

Here, we generalize this to the case with multi-level treatments in a way that allows for a scalar matching variable. In this case, matching is conceptually quite different from matching on covariates. We separate the estimation of $\tau(w, w') = \mathbb{E}[Y_i(w')] - \mathbb{E}[Y_i(w)]$ into the two terms. First, we focus on the problem of estimating $\mathbb{E}[Y_i(w)]$. Define the generalized propensity score matching function as

$$m_{\text{gps}}(w, p) = \arg\min_{j:W_j=w} ||p(w|X_j) - p||. \qquad (5)$$

Here, the treatment level w enters into the matching function not only by limiting the set of potential matches to the set of units with $W_j = w$, but also in the function of the covariates that is being matched, $p(w|X_j)$. In covariate and conventional propensity score matching, the treatment level only affects the set of potential matches.

Given the generalized propensity score matching function, we impute $Y_i(w)$ as

$$\hat{Y}_i(w) = Y^{\text{obs}}_{m_{\text{gps}}(w,p(w|X_i))}.$$

The average effect is estimated as

$$\hat{\tau}_{\text{gps}}(w, w') = N^{-1} \sum_{i=1}^{N} \left(Y^{\text{obs}}_{m_{\text{gps}}(w',p(w'|X_i))} - Y^{\text{obs}}_{m_{\text{gps}}(w,p(w|X_i))}\right). \qquad (6)$$

Note that the difference $Y^{\text{obs}}_{m_{\text{gps}}(w',p(w'|X_i))} - Y^{\text{obs}}_{m_{\text{gps}}(w,p(w|X_i))}$ in (6) is *not* generally an estimate of an average causal effect, whereas in (2) the difference $Y^{\text{obs}}_{m_{\text{cov}}(w',X_i)} - Y^{\text{obs}}_{m_{\text{cov}}(w,X_i)}$ *is* an estimate of the average causal effect $\mathbb{E}[Y_i(w') - Y_i(w) \mid X_i]$. In the binary treatment case, this distinction between covariate and propensity score matching does not matter: in

that case, there are only two values for w, $w = 1, 2$, so that $p(1 \mid x) = 1 - p(2 \mid x)$, and therefore matching on $p(1 \mid x)$ is the same as matching on both $p(1 \mid x)$ and $p(2 \mid x)$, and the same as matching on $p(2 \mid x)$.

In Web Appendix, we provide mathematical details for inference and show that under mild regularity conditions, the matching estimator based on the generalized propensity score or the estimated generalized propensity score is asymptotically normally distributed.

5. Subclassification on the Generalized Propensity Score

In the binary treatment literature, a common alternative to matching is subclassification or stratification on the propensity score, originally proposed by RR83. To put our proposed methods for the multi-level treatment case in perspective, let us briefly summarize their approach for the binary treatment case in our current notation to show why it does not directly extend to the multivalued treatment case. Divide the sample into a number of subclasses by the value of the propensity score $p(1|x)$. Based on Cochran (1968) who shows that this removes much of the bias, researchers often use five subclasses. To be specific, let $q_j^{p(1|x)}$ be the jth quintile of the empirical distribution of $p(1|X_i)$, for $j = 1, \ldots, 4$, and define $q_0^{p(1|x)} = 0$ and $q_5^{p(1|x)} = 1$. Then, we construct the five subclasses, based on the propensity score being between $q_{j-1}^{p(1|x)}$ and $q_j^{p(1|x)}$. For subclass j, one can estimate the average causal effect of treatment 1 versus treatment 2 as

$$\tau_j(1,2) = \frac{1}{N_{j2}} \sum_{i: q_{j-1}^{p(1|x)} < p(1|X_i) \le q_j^{p(1|x)}, W_i = 2} Y_i^{\text{obs}} - \frac{1}{N_{j1}} \sum_{i: q_{j-1}^{p(1|x)} < p(1|X_i) \le q_j^{p(1|x)}, W_i = 1} Y_i^{\text{obs}},$$

where N_{jw} is the number of units in subclass j with treatment level w. The overall average treatment effect is then estimated by averaging over the subclasses:

$$\hat{\tau}(1,2) = \sum_{j=1}^{5} \frac{N_{j1} + N_{j2}}{N} \cdot \hat{\tau}_j(1,2).$$

Because all $N_{j1} + N_{j2}$ are close to equal, at most differ by 1 (assuming there are no ties), this is essentially a simple arithmetic mean of the J estimates $\hat{\tau}_j(1,2)$.

Now, consider the multi-level treatment case. We are interested in $\tau(w, w')$ for some pair of treatment levels w and w'. Again, and this is a cornerstone of our approach, we write this as a difference of two expectations, $\tau(w, w') = \mathbb{E}[Y_i(w')] - \mathbb{E}[Y_i(w)]$ and separately estimate the two terms $\mathbb{E}[Y_i(w')]$ and $\mathbb{E}[Y_i(w)]$. To estimate the second term, $\mathbb{E}[Y_i(w)]$ we construct subclasses or strata based on $p(w|x)$. Let $q_j^{p(w|x)}$ be the quintiles of $p(w \mid X_i)$ in the sample. Then, the average value of $Y_i(w)$ in subclass j is estimated as

$$\hat{\mu}_{jw} = \frac{1}{N_{jw}} \sum_{i: q_{j-1}^{p(w|x)} < p(w|X_i) \le q_j^{p(w|x)}, W_i = w} Y_i^{\text{obs}},$$

where N_{jw} is the number of units with $q_{j-1}^{p(w|x)} < p(w \mid X_i) \le q_j^{p(w|x)}$ and $W_i = w$. The overall average of $Y_i(w)$ is then estimated as

$$\hat{\mathbb{E}}[Y_i(w)] = \sum_{j=1}^{5} \frac{N_w}{N} \cdot \hat{\mu}_{jw}.$$

The key is that, in contrast to what is done in the binary treatment case, we do not construct subclasses defined by similar values for the $T - 1$ propensity scores such that we can estimate causal effects within the subclasses. Instead, we construct subclasses defined by similar values for a single propensity score at a particular treatment level so that we can estimate the average potential outcome for that treatment level within the subclasses, and we do so separately for each treatment level, with different subclasses for each treatment level. In the binary treatment case, this amounts to the same thing because the two propensity scores $p(1|x)$ and $p(2|x)$ are linearly related, but in the multi-level case these two approaches are different.

6. Assessing and Ensuring Overlap

6.1. *Assessing Balance*

We focus on assessing balance in the covariate distributions in terms of the propensity score as well as directly in terms of the covariates, following the discussion in Imbens and Rubin (2015) for the binary case. For each treatment level w, we calculate the average values for each component of the covariate vectors and their corresponding sample variances:

$$\overline{X}_w = N_w^{-1} \sum_{i: W_i = w} X_i, \quad \text{and} \quad S_{X,w}^2 = (N_w - 1)^{-1} \sum_{i: W_i = w} (X_i - \overline{X}_w)^2.$$

Define also for each treatment level, the average value of the covariates for units with a treatment level different from w and the average variance:

$$\overline{X}_{\overline{w}} = (N - N_w)^{-1} \sum_{i: W_i \neq w} X_i, \quad \text{and} \quad S_{X|W}^2 = T^{-1} \sum_{w=1}^{T} S_{X,w}^2,$$

respectively. The first approach to assessing the covariate balance is to inspect the normalized differences for each covariate and each treatment level:

$$nd_w^{\text{COV}} = (\overline{X}_w - \overline{X}_{\overline{w}}) / S_{X|W}. \tag{7}$$

We can also assess balance by looking at the generalized propensity score. For each treatment level, we can calculate the normalized difference for the generalized propensity score

for that treatment level:

$$nd_w^{\text{GPS}} = \left(\overline{p(w \mid X)}_w - \overline{p(w \mid X)}_{\overline{w}}\right) / S_{p(w|X)|W}, \qquad (8)$$

where $\overline{p(w \mid X)}_w = N_w^{-1} \sum_{i:W_i=w} p(w \mid X_i)$ and $\overline{p(w \mid X)}_{\overline{w}} = (N - N_w)^{-1} \sum_{i:W_i \neq w} p(w \mid X_i)$. Finally, one may wish to plot a histogram of $p(w \mid X_i)$ for the N_w units with $W_i = w$ and a histogram of $p(w \mid X_i)$ for the $N - N_w$ units with $W_i \neq w$ in the same figure.

6.2. *Improving Overlap*

In many applications, there are regions of the covariate space with low values for the probability of receiving one of the treatments. This is likely in the setting with a binary treatment, but even more likely to be an issue in settings with many treatment levels. Of note, lack of overlap affects the credibility of all methods attempting to estimate all pairwise average causal effects from the common population. In that case, we may wish to modify the estimands to average only over the part of the covariate space with all treatment probabilities away from zero. The question is how to choose the set of covariates with overlap. For the binary treatment case, Crump et al. (2009) proposed a method for improving overlap by trimming the sample. Specifically, they suggest dropping units from the analysis with low and high values of the propensity score. The threshold for dropping units is based on minimizing the variance of the estimated average treatment effect on the trimmed sample. By trimming the sample, this method generally alters the estimand to the so-called feasible estimand, by changing the reference population. Using the feasible estimand is widely recommended in the literature, as long as we are careful to characterize the resulting quantity of interest. Here, we generalize the Crump et al. (2009) approach to the multi-level treatment case. We focus on average treatment effects for subsets of the covariate space. Formally, define the conditional average treatment effect:

$$\tau(w, w'|\mathbb{C}) = \mathbb{E}[Y_i(w') - Y_i(w) \mid X_i \in \mathbb{C}].$$

The semiparametric efficiency bound for $\tau(w, w' \mid \mathbb{C})$ is, building on the work by Hahn (1998) and Hirano, Imbens, and Ridder (2003), under homoscedasticity and constant treatment effects,

$$\mathbb{V}(w, w'|\mathbb{C}) = \frac{\sigma^2}{pr(X_i \in \mathbb{C})} \mathbb{E}\left[\frac{1}{p(w|X_i)} + \frac{1}{p(w'|X_i)} \mid X_i \in \mathbb{C}\right].$$

In the binary case Crump et al. (2009) proposed choosing $\mathbb{C}$ to minimize $\mathbb{V}(w, w' \mid \mathbb{C})$, which leads to dropping units with $p(1 \mid X_i) \leq \alpha$ and units with $p(1 \mid X_i) \geq 1 - \alpha$, with α an estimable function of the marginal distribution of the propensity score.

For the multi-level treatment case, we suggest focusing on the subset of the covariate space $\mathbb{C}$ that minimizes

$$\begin{aligned}\overline{\mathbb{V}}(\mathbb{C}) &= \sum_{w,w'} \mathbb{V}(w, w' \mid \mathbb{C}) \\ &= \frac{2\sigma^2}{pr(X_i \in \mathbb{C})} \mathbb{E}\left[\sum_{w=1}^{T} \frac{1}{p(w \mid X_i)} \mid X_i \in \mathbb{C}\right].\end{aligned}$$

Under homoscedasticity and a constant treatment effect, this will lead to a set $\mathbb{C}$ of the form

$$\mathbb{C} = \left\{ X_i \in \mathbb{X} \,\middle|\, \sum_{w=1}^{T} \frac{1}{p(w \mid X_i)} \leq \lambda \right\},$$

where the threshold λ satisfies

$$\begin{aligned}\lambda \leq \frac{2}{pr\Big(\sum_{w=1}^{T} (p(w \mid X_i))^{-1} \leq \lambda\Big)} \\ \times E\left[\sum_{w=1}^{T} \frac{1}{p(w \mid X_i)} \,\middle|\, \sum_{w=1}^{T} \frac{1}{p(w \mid X_i)} \leq \lambda\right].\end{aligned}$$

To implement the trimming method in practice in the multi-level treatment case, we replace the expectation by an average and then find the largest λ that satisfies the inequality.

7. A Simulation Study

In this section, we assess the performance of the two new estimators in cases of multi-level treatments (matching on the generalized propensity score, GPSM, and subclassification on the generalized propensity score, GPSS) in a Monte Carlo study relative to five previously proposed estimators, first the simple difference in average outcomes (DIF) by treatment status, second pairwise propensity score matching (PPSM) that compares two treatment levels at a time using the binary propensity score matching on the units exposed to one of those two treatment levels, third the estimator based on matching on the set of $T - 1$ propensity score set (PSSM), fourth the estimator based on weighting, and fifth, matching on all covariates (COV). In the binary treatment and missing data case, previous simulations have found that weighting estimators can have high variability, e.g., Kang and Schafer (2007) and Guo and Fraser (2010), especially if the probabilities are close to zero. Frolich (2004a) found that the weighting estimator was inferior to pairwise matching estimators in terms of root mean squared error. This is even more likely to be a concern in settings with multiple treatment levels than in the binary treatment case because, with the probabilities for the T treatment levels adding up to one, with T large some probabilities are likely to be close to zero. Because in the binary treatment case, it has been found that matching on high-dimensional covariates is not practical for commonly found sample sizes (e.g., the theoretical results in Abadie and Imbens, 2006), it is likely that in settings with many treatment levels matching on all scores is not effective either. These results motivate us to compare the seven estimators in settings

Table 1

Simulation results, design I. Estimators: (1) DIF, simple difference in outcomes for units with different treatment levels; (2) PPSM: pairwise comparison using binary propensity score matching; (3) PSSM: matching on the propensity score set; (4) W: weighting estimator; (5) COV: matching on all covariates; (6) GPSM: matching on the generalized propensity score; (7) GPSS: stratification on the generalized propensity score. Variance estimators: (1) bootstrapping variance estimator for DIF, GPSS, and W; (2) Abadie and Imbens (2006) variance estimator for COV matching and PSSM; (3) Abadie and Imbens (2012) variance estimator for PPSM and GPSM.

	Bias			RMSE			Coverage 95% CI		
	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$	$\tau(1,2)$	$\tau(1,3)$	$\tau(2,3)$
DIF	1.34	0.57	−0.77	1.38	0.60	0.83	0.01	0.26	0.41
PPSM	−0.6	−1.1	−0.8	0.77	1.16	0.91	0.80	0.001	0.74
PSSM	0.21	0.19	−0.02	0.27	0.33	0.29	0.90	0.92	0.98
W	0.05	0.02	−0.03	0.55	0.43	0.53	0.91	0.97	0.94
COV	0.29	0.19	−0.11	0.33	0.22	0.20	0.75	0.88	0.99
GPSM	0.14	0.04	−0.10	0.56	0.36	0.61	0.95	0.95	0.95
GPSS	0.31	0.05	−0.27	0.53	0.24	0.54	0.91	0.99	0.94

with a large number of treatment levels, and where some of the treatment levels have low probability for some covariate values. In the simulations, we focus on two designs, one with three treatment levels and one with six treatment levels, and both with six covariates.

In the first design with three treatment levels, the covariates X_{1i}, X_{2i}, and X_{3i} are multivariate normal with means zero, variances of $(2, 1, 1)$, and covariances of $(1, -1, -0.5)$; $X_{4i} \sim U[-3, 3]$; $X_{5i} \sim \chi^2_1$; and $X_{6i} \sim$ Bernoulli(0.5), with $X_i^T = (1, X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i})$. The three treatment groups are formed using multinomial regression model

$$(D_i(1), D_i(2), D_i(3)) \sim \text{Multinom}(p(1|X_i), p(2|X_i), p(3|X_i)),$$

where $D_i(w)$ is the treatment indicator, i.e., $D_i(w) = 1$, if the unit i belongs to treatment w, and $p(w|X_i) = \exp(X_i^T\beta_w)/\sum_{w'=1}^3 \exp(X_i^T\beta_{w'})$, $\beta_1^T = (0,0,0,0,0,0,0)$, $\beta_2^T = 0.7 \times (0,1,1,1,-1,1,1)$, and $\beta_3^T = 0.4 \times (0,1,1,1,1,1,1)$. The outcome design is $Y_i(w) = X_i^T\gamma_w + \eta_i$ with $\eta_i \sim N(0,1)$, $\gamma_1^T = (-1.5,1,1,1,1,1,1)$, $\gamma_2^T = (-3,2,3,1,2,2,2)$, and $\gamma_3^T = (1.5,3,1,2,\ -1,-1,-1)$. The sample sizes are $N_w = 500$, for $w = 1, 2, 3$.

We compare the seven estimators over 1,000 datasets. The generalized propensity scores are estimated using multinomial logistic regression model with all covariates entering the model linearly. 95% confidence intervals for point estimates were calculated using: (i) 2.5 and 97.5 percentiles from 1000 bootstrap samples for DIF, GPSS, and weighting; (ii) point estimate $\pm 1.96 \times (\text{variance})^{1/2}$ for Abadie and Imbens (2006) variance estimator for COV and PSSM; and for Abadie and Imbens (2012) variance estimator for PPSM and GPSM, which takes into account the uncertainty of the matching procedure and the uncertainty of estimating generalized propensity scores, as in Web Appendix.

Table 1 presents the bias, root mean squared error (RMSE) and coverage of 95% confidence intervals. DIF shows that there are substantial biases associated with the covariates. PPSM compares two treatment levels at a time using the units exposed to one of those two treatment levels, which focuses on different populations of inference each time. This leads to inconsistency for simultaneous comparison of treatment levels. One implication is that $\hat\tau(1,2) + \hat\tau(2,3) + \hat\tau(3,1) \neq 0$. Even with only three treatment levels, and so only two propensity scores to match on, PSSM did not control the bias well. The four remaining procedures, including COV, GPSM, and GPSS, and weighting, do a fairly good job of reducing the bias for all average treatment effects. Among these four, COV has smallest RMSE. For inference, asymptotic 95% confidence intervals provide coverage very close to the nominal coverage for GPSM, which confirms our inference theory in Web Appendix. Asymptotic 95% confidence intervals for GPSS and weighting are also fairly accurate, but COV leads to undercoverage, consistent with the results in Abadie and Imbens (2006) on the bias of matching estimators with multiple covariates.

In the second design with six treatment levels, we consider propensity score design as $p(w \mid X_i) = \exp(X_i^T\beta_w)/\sum_{w'=1}^6 \exp(X_i^T\beta_{w'})$, where $\beta_1^T = (0,0,0,0,0,0,0)$, $\beta_2^T = 0.4 \times (0,1,1,2,1,1,1)$, $\beta_3^T = 0.6 \times (0,1,1,1,1,1,-5)$, $\beta_4^T = 0.8 \times (0,1,1,1,1,1,5)$, $\beta_5^T = 1.0 \times (0,1,1,1,-2,1,1)$, and $\beta_6^T = 1.2 \times (0,1,1,1,-2,-1,1)$. The outcome design is $Y_i(w) = X_i^T\gamma_w + \eta_i$, with $\gamma_1^T = (-1.5,1,1,1,1,1,1)$, $\gamma_2^T = (-3,2,3,1,2,2,2)$, $\gamma_3^T = (3,3,1,2,-1,-1,-4)$, $\gamma_4^T = (2.5,4,1,2,-1,-1,-3)$, $\gamma_5^T = (2,5,1,2,-1,-1,-2)$, and $\gamma_6^T = (1.5,6,1,2,-1,-1,-1)$ with $\eta_i \sim N(0,1)$. The sample sizes are $N_w = 1000$, for all w.

In Figure 1, we present the results for the estimators for the 15 average treatment effects. The simulation setup creates six treatment groups with strong separation in covariate distributions, which makes it fundamentally difficulty removing all biases in estimating 15 treatment effects simultaneously. Overall GPSM outperforms the other methods in terms of bias and coverage rates, with the coverage rate for nominal 95% confidence intervals never going below 0.75. To assess the performance of the weighting estimator, it is useful to look at the weights that underly the estimator. Normalizing the weights so that they average to N_w for each of the treatment levels, the maximum weight for units in each of the treatment levels is 16.9 (treatment level one), 21.2 (treatment level two), 50.0 (treatment level three), 64.9 (treatment level four), 21.2 (treatment level five), and 185.1 (treatment level

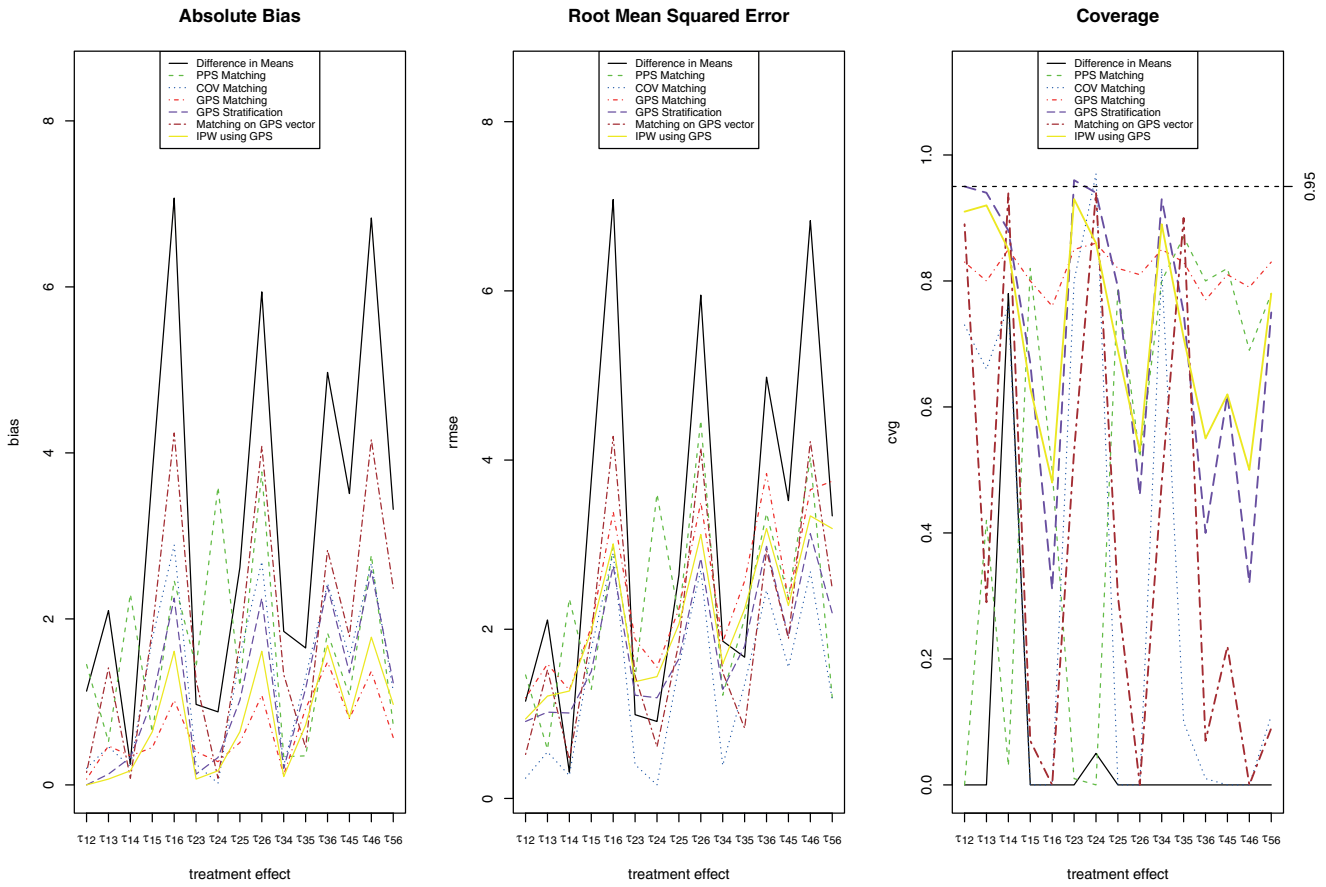

Figure 1. Simulation results, design II.

six). Even in the three treatment level case, these maximum weights are substantial. There the maximum weight for units in each of the treatment levels is 16.5 (treatment level one), 95.8 (treatment level two), and 17.9 (treatment level three).

In an extended simulation (see Web Appendix), we compare the performance of the estimators under the combinations of (w/o) trimming and (correct/incorrect) generalized propensity score model. When the propensity score model is incorrect, the performance for all methods based on the propensity score deteriorates. In particular, the weighting estimator shows huge bias and variance and poor coverage for all parameters. GPSM is inferior to COV in terms of bias and variance; however, it presents better coverage for 10 parameters out of 15 parameters. This suggests that when covariates are high dimensional, the inference for COV is not satisfactory. After trimming, bias and variance are greatly reduced and coverage is improved for all parameters for GPSM, GPSS, and weighting, which suggests that trimming can improve the performance of GPS-based methods.

8. An Application

We re-examined data from the REFLECTIONS (Real-World Examination of Fibromyalgia: Longitudinal Evaluation of Costs and Treatments) study. REFLECTIONS was a 12-month prospective observational study of patients being treated for fibromyalgia at 58 outpatient sites in the United States and Puerto Rico. Patients had to be at least 18 years of age and initiating a new pharmacologic treatment for fibromyalgia. In keeping with the observation nature of the study, inclusion and exclusion criteria were kept to a minimum, no requirements on the nature of the fibromyalgia treatment were made, and physicians decisions regarding the proper treatment and care of patients were made in the course of normal clinical practice. Data from patients were collected at baseline during a standard office visit and at 1, 3, 6, and 12 months post baseline via a computer-assisted telephone interviews. For details on the design of REFLECTIONS see Robinson et al. (2012).

For this example, we focused on the analysis of three fibromyalgia medication cohorts (Peng et al., 2015): patients treated with an opioid (either monotherapy or with other medications), patients treated with tramadol but not an opioid, and patients not treated with tramadol or an opioid (referred to as the Other cohort). The outcome variable utilized here is the total score of Fibromyalgia Impact Questionnaire (FIQ), which is composed of items measuring physical functioning, number of days the patient felt well, number of days the patient felt unable to work due to FM symptoms, and patient ratings of work difficulty, pain intensity, fatigue, morning tiredness, stiffness, anxiety, and depression. The total score ranges from 0 to 80 with lower scores indicating better outcomes, and research suggests a 14% reduction (or 7.6 points in this sample) is clinically relevant (Bennett et al., 2009). The objective is to produce causal inference pairwise comparisons

Table 2

Analysis results: change from Baseline to Endpoint (12 Months) for the FIQ total score. OPI = opioid cohort; TRA = tramadol cohort; OTH = other cohort. TRA–OPI indicates the estimated difference in change from baseline scores for the tramadol cohort minus the same value for the opioid cohort. Thus, negative values indicate greater reduction in symptoms for the first cohort. Analyses are on the trimmed sample. Confidence intervals were calculated using the same methods as for simulated study above.

	Pairwise differences means (95% CI)		
Method	TRA–OPI	OTH–OPI	OTH–TRA
DIF	−1.1 (−3.8, 1.1)	−1.7 (−3.5, 0.3)	−0.6 (−2.6, 2.2)
PPSM	−3.9 (−7.2, −0.6)	−1.4 (−3.4, 0.6)	−1.8 (−4.2, 0.6)
PSSM	−2.5 (−6.6, 1.6)	−1.9 (−4.1, 0.4)	0.7 (−3.1, 4.4)
W	−0.8 (−5.4, 4.7)	−0.3 (−3.8, 5.1)	0.4 (−2.3, 4.0)
COV	−1.6 (−4.8, 1.5)	−1.5 (−3.8, 0.9)	0.2 (−2.5, 2.8)
GPSM	−1.6 (−4.3, 1.1)	−0.9 (−2.8, 1.1)	0.7 (−1.8, 3.2)
GPSS	−1.6 (−5.5, 1.2)	−1.2 (−4.1, 0.9)	0.4 (−2.2, 3.8)

between the cohorts all based on the same population (the population represented by the trimmed sample), in order to test the study hypothesis that there is no difference in FIQ total score among the three cohorts.

The generalized propensity scores (three estimated probabilities for each patient) were computed using a multinomial model with 32 predictors from demographics, baseline clinical characteristics, comorbidities, resource use, prior fibromyalgia treatment, and physician information.

To help address lack of overlap of the populations, the modified Crump (Crump et al., 2009) algorithm of Section 6 was applied. With the REFLECTIONS data, $\lambda = 29.88$, and thus patients were trimmed if their $\sum_{w=1}^{T} 1/p(w|X_i)$ value was greater than 43.52. This resulted in removal of 363 patients (25% of the sample), with 31(9%) from the Opioid cohort (OPI), 17(8%) of the Tramadol cohort (TRA), and 315(34%) from the Other cohort (OTH). Thus, the analysis cohort included 1101 patients (308 OPI, 188 TRA, 605 OTH).

The trimming primarily removed patients in OTH who had high propensities for being in OTH (and low propensities for either OPI or TRA) and were under-represented in OPT and TRA. Using the trimmed sample, generalized propensity score matching was implemented following the steps in Section 4 to produce counterfactual outcomes (imputed FIQ total scores) for each patient and cohort. The quality of the matches appeared acceptable, with the mean difference in propensity scores for the matched patients ranging from 0.0012 to 0.0014 across the cohorts and the largest matched pair with a difference of 0.035.

The unadjusted mean changes(sd) from baseline to endpoint (12 months post baseline) for the FIQ total score in the trimmed cohort were −2.4(12.3) for OPI, −3.7(14.0) for TRA, and −4.2(13.4) for OTH, indicating small numerical improvement in pain symptoms. Table 2 summarizes the comparative analysis of the FIQ total score improvement among the three cohorts on the trimmed sample using generalized propensity score matching (GPSM) and stratification (GPSS). As a comparison, results using no bias adjustment (Difference in Means), propensity score set matching (PSSM), weighting (W), and covariate matching (COV) are included. Without bias adjustment, no cohort differences reach the level of statistical significance, though OTH demonstrated marginally greater reductions than OPI ($p = 0.058$). Similarly, none of the adjusted differences led to any statistical significant findings, indicating similar health condition improvements at 12 months over baseline among the three cohorts. Note that by using the same population across all comparisons, pairwise differences across the cohorts using the generalized propensity scoring methods (GPSM, GPSS) are consistent, whereas using PPSM this is not true (the PPSM estimates suggest that changing the treatment from OPI to TRA leads to an average effect of −3.9, changing the treatment effect from TRA to OTH leads to an average effect of −1.8, and changing the treatment from OPI to OTH leads to an average effect of −1.4, which cannot all be true at the same time). This illustrates the impact of differing populations can have when using PPSM.

9. Conclusion

In this article, we develop new methods for estimating causal treatment effects using observational data in settings with multiple (more than two) treatment levels. Existing methods require additional assumptions assuming the existence of a scalar balance score, so as to facilitate matching and subclassification. We show that, contrary to claims in the literature, matching and subclassification methods using the propensity score generalize naturally to the multi-level treatment case. We focus on matching and subclassification on the generalized propensity score using the notion of weak unconfoundedness, and show that adjusting for a scalar function of the covariates can remove all biases associated with differences in observed covariates.

As with other propensity-based analyses, this approach depends on correct specification of propensity score modeling, and does not resolve the potential for bias due to unmeasured confounding. An initial simulation study and example demonstrated the potential benefits of the proposed approach at reducing bias and providing causal inference comparisons for multiple cohorts on a common population.

10. Supplementary Materials

Supplementary Web Appendices, referenced in Sections 4 and 7, are available with this article at the *Biometrics* website on Wiley Online Library. R code is available at https://github.com/shuyang1987/multilevelMatching.

Acknowledgements

This article and the preparation of this article were funded in full by Eli Lilly and Company, Indianapolis, IN, USA. Editing support was provided by Casie Polanco, inVentivHealth Clinical, Indianapolis, IN, USA.

Declaration of personal interests: Cui, Faries, and Kadziola are all employees and minor stockholders of Eli Lilly and Company.

References

Abadie, A. and Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–267.

Abadie, A. and Imbens, G. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association* **107**, 833–843.

Bennett, R. M., Bushmakin, A. G., Cappelleri, J. C., Zlateva, G., and Sadosky, A. B. (2009). Minimal clinically important difference in the fibromyalgia impact questionnaire. *The Journal of Rheumatology* **36.6**, 1304–1311.

Cadarette, S., Gagne, J., Solomon, D., Katz, J., and Sturmer, T. (2010). Confounder summary scores when comparing the effects of multiple drug exposures. *Pharmacoepidemiology and Drug Safety* **19**, 2–9.

Cattaneo, M. (2010). Efficient semiparametric estimation of multilevel treatment effects under ignorability. *Journal of Econometrics* **155**, 138–154.

Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**, 295–314.

Cole, S. and Frangakis, C. (2009). The consistency assumption in causal inference: A definition or an assumption?. *Epidemiology* **20**, 3–5.

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

Dawid, A. P. (1979). Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society, Series B* **41**, 249–252.

Frölich, M. (2004a). Finite-sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics* **86**, 77–90.

Frölich, M. (2004b). Programme evaluation with multiple treatments. *Journal of Economic Surveys* **18.2**, 181–224.

Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care* **41.10**, 1183–1192.

Guo, S. and Fraser, M. (2010). *Propensity Score Analysis*, Thousand Oaks, California: Sage.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Hirano, K. and Imbens, G. (2004). The propensity score with continuous treatments. *Applied Bayesian Modelling and Causal Inference from Missing Data Perspectives*, Gelman and Meng (eds), New York: Wiley.

Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **74**, 1161–1189.

Huber, M., Lechner, M., and Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics* **175.1**, 1–21.

Imai, K. and Van Dyk, D. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854–866.

Imai, K. and Ratkovic, M. (2014). Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **76**, 243–246.

Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–710.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* **86**, 4–29.

Imbens, G. and Rubin, D. (2015). *An Introduction to Causal Inference in the Statistical, Biomedical and Social Sciences*. Cambridge: Cambridge University Press.

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology* **150.4**, 327–333.

Kang, J. and Schafer, J. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluations of Active Labor Market Policies in Europe*, M. Lechner and F. Pfeiereds (eds), 43–58. Heidelberg: Physica.

Lee, B. K., Lessker, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* **29.3**, 337–346.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* **96.456**, 1245–1253.

Mccaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgetteb, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* **32**, 3388–3414.

Morgan, S. and Winship, C. (2007). *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Peng, X., Robinson, R. L., Mease, P., Kroenke, K., Williams, D. A., Chen, Y., et al. (2015). Long-term evaluation of opioid treatment in fibromyalgia. *The Clinical Journal of Pain* **31.1**, 7–13.

Rassen, J., Shelat, A., Franklin, J., Glynn, R., and Schneeweiss, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24**, 401–409.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560.

Robinson, R. L., Kroenke, K., Mease, P., Williams, D. A., Chen, Y., D'Souza, D., Wohlreich, M., et al. (2012). Burden of illness and treatment patterns for patients with fibromyalgia. *Pain Medicine* **13**, 1366–1376.

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

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized study. *Journal of Educational Psychology* **65**, 688–701.

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

Setoguchi, S., Schneeweiss, S., Brookhart, A., Glynn, R., and Cook, F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety* **17.6**, 546.

Zanutto, E., Lu, B., and Hirnik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* **30.1**, 59–73.

Received August 2015. *Revised December* 2015.
Accepted January 2016.