Statistical Science
2017, Vol. 32, No. 3, 432–454
DOI: 10.1214/17-STS612

Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas

Michael J. Lopez and Roee Gutman

Abstract. The propensity score is a common tool for estimating the causal effect of a binary treatment in observational data. In this setting, matching, subclassification, imputation or inverse probability weighting on the propensity score can reduce the initial covariate bias between the treatment and control groups. With more than two treatment options, however, estimation of causal effects requires additional assumptions and techniques, the implementations of which have varied across disciplines. This paper reviews current methods, and it identifies and contrasts the treatment effects that each one estimates. Additionally, we propose possible matching techniques for use with multiple, nominal categorical treatments, and use simulations to show how such algorithms can yield improved covariate similarity between those in the matched sets, relative the pre-matched cohort. To sum, this manuscript provides a synopsis of how to notate and use causal methods for categorical treatments.

Key words and phrases: Causal inference, propensity score, multiple treatments, matching, observational data.

1. INTRODUCTION

The primary goal of many scientific applications is to identify the causal effect of exposure $T \in \{t_1, \ldots, t_Z\}$ on outcome Y. Randomized experiments are the gold standard for estimating a causal relationship, however, they are sometimes infeasible due to logistical, ethical or financial considerations. Further, randomized experiments may not be as generalizable as observational studies due to the restricted population used in the experiments.

When assignment to treatment is not randomized, those that receive one level of the treatment may differ from those that receive another with respect to covariates, $\boldsymbol{X}$, that may also influence the outcome. For example, in a study estimating the causal effects of neighborhood choice on employment, persons who live in deprived neighborhoods differ from those who live in privileged ones on a variety of characteristics, such as socioeconomic status and education levels (Hedman and Van Ham, 2012). As such, it may be difficult to distinguish between neighborhood effects and the differences between subjects which existed before they chose their neighborhoods. In such settings, establishing causes and effects requires more sophisticated statistical tools and additional assumptions.

Methods such as matching (Dehejia and Wahba, 2002), weighting (Robins, Hernan and Brumback, 2000), subclassification (Rosenbaum and Rubin, 1984), and imputations (Gutman and Rubin, 2015) have been proposed to adjust for the differences in $\boldsymbol{X}$ across the exposure groups. These approaches attempt to obtain covariate balance across treatment groups, where balance refers to equality in the distributions of $\boldsymbol{X}$. By ensuring that the distribution of units receiving different treatments are similar on average, these methods attempt to reproduce a randomized trial, thus reducing the effects of treatment assignment bias on causal estimates.

When $\boldsymbol{X}$ is a scalar, it is relatively straight-forward to perform matching (Rubin, 1976). However, it is more

Michael J. Lopez is Assistant Professor of Statistics, Department of Mathematics, Skidmore College, 815 N. Broadway, Saratoga Springs, New York 12866, USA (e-mail: mlopez1@skidmore.edu). Roee Gutman is Assistant Professor of Biostatistics, Department of Biostatistics, Brown University, 121 S. Main Street, Providence, Rhode Island 02912, USA (e-mail: roee_gutman@brown.edu).

complex to match, subclassify or weight when $\boldsymbol{X}$ is composed of many covariates. With a binary treatment, matching, subclassification, weighting and imputation using the propensity score have been proposed for estimating causal effects from observational studies with binary treatment (Rosenbaum and Rubin, 1983, Stuart, 2010, Gutman and Rubin, 2015). Propensity score is defined as the probability of receiving the treatment conditional on a set of observed covariates. It has been shown in theory (Rubin and Thomas, 1996) and practice (D'Agostino, 1998, Caliendo and Kopeinig, 2008) that under certain assumptions, matching on propensity scores results in unbiased unit-level estimates of the treatment's causal effect (Rosenbaum and Rubin, 1983).

Generalizations and applications of propensity score methods for multiple treatments, however, remain scattered in the literature, in large part because the advanced techniques are unfamiliar and inaccessible. Our first goal is to provide a unifying terminology that will enable researchers to coalesce and compare existing methods. Our second goal is to describe current methods for estimating causal effects with multiple treatments, with a specific focus on approaches for nominal categorical exposures (e.g., a comparison of painkillers Motrin, Advil and Tylenol). We contrast these methods' assumptions and define the causal effects they each attempt to estimate. In doing so, potential pitfalls in the commonly used practice of applying binary propensity score tools to multiple treatments are identified.

Third, we explain the elevated importance of defining a common support region when studying multiple treatments, where differences in the implementation of certain approaches can vary the causal estimands as well as change the study population to which inference is generalizable. Our final goal is to provide a technique for generating matched sets when there are more than two treatments that addresses some of the pitfalls of the current methods, as well as to compare the performance of the new and previously proposed algorithms in balancing covariates' distributions using extensive simulation analysis.

The remainder of Section 1 introduces the notation and identifies existing causal methods for multiple treatments. Section 2 proposes a new algorithm for matching with multiple treatments. Section 3 uses simulations to contrast the new and previously proposed approaches for generating well-matched subgroups of subjects. Section 4 discusses and concludes with a set of practical recommendations.

1.1 Notation for Binary Treatment

Our notation is based on the potential outcomes framework, originally proposed by Neyman for randomized based inference, and extended by Rubin to observational studies and Bayesian analysis, also known as the Rubin Causal Model (RCM) (Splawa-Neyman, Dabrowska and Speed, 1990 [1923], Rubin, 1975, Holland, 1986). Let Y_i, $\boldsymbol{X}_i$, and T_i be the observed outcome, set of covariates and binary treatment assignment, respectively, for each subject $i = 1, \ldots, N$, with $N \leq \mathcal{N}$, where $\mathcal{N}$ is the population size which is possibly infinite. With $T_i \in \mathcal{T}$, let $\mathcal{T}$ be the treatment space. For a binary treatment, $\mathcal{T} = \{t_1, t_2\}$, and let n_{t_1} and n_{t_2} be the number of subjects receiving treatments t_1 and t_2, respectively.

The RCM relies on the Stable Unit Treatment Value Assumption (*SUTVA*) to define the potential outcomes $Y_i(t_1)$ and $Y_i(t_2)$, which would have been observed had unit i simultaneously received t_1 and t_2, respectively (Rubin, 1980). SUTVA specifies no interference between subjects and no hidden treatment versions, entailing that the set of potential outcomes for each subject does not vary with the treatment assignment of others. Because each individual receives only one treatment at a specific point in time, only $Y_i(t_1)$ or $Y_i(t_2)$ is observed for each subject, which is known as the fundamental problem of causal inference (Holland, 1986).

Two commonly used estimands for describing superpopulation effects are the population average treatment effect, $PATE_{t_1,t_2}$, and the population average treatment effect among those receiving t_1, $PATT_{t_1,t_2}$:

$$PATE_{t_1,t_2} = E\big[Y_i(t_1) - Y_i(t_2)\big], \tag{1}$$

$$PATT_{t_1,t_2} = E\big[Y_i(t_1) - Y_i(t_2) | T_i = t_1\big]. \tag{2}$$

Letting $I(T_i = t_1)$ be the indicator function for an individual receiving treatment t_1, $PATE_{t_1,t_2}$ and $PATT_{t_1,t_2}$ are generally approximated by the sample average treatment effects:

$$SATE_{t_1,t_2} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(t_1) - Y_i(t_2)\big), \tag{3}$$

$$SATT_{t_1,t_2} = \frac{1}{n_{t_1}}\sum_{i=1}^{N}\big(Y_i(t_1) - Y_i(t_2)\big) \times I(T_i = t_1). \tag{4}$$

Because only one of the potential outcomes is observed for every unit, an important piece of information to estimate (3) and (4) is the assignment mechanism, $P(\boldsymbol{T}|\boldsymbol{Y}(t_1), \boldsymbol{Y}(t_2), \boldsymbol{X})$, where $\boldsymbol{T} = \{T_i\}$, $\boldsymbol{Y}(t_1) =$

$\{Y_i(t_1)\}$, $\boldsymbol{Y}(t_2) = \{Y_i(t_2)\}$ and $\boldsymbol{X} = \{\boldsymbol{X}_i\}$ (Imbens and Rubin, 2015). Three commonly made restrictions of the assignment mechanism are individualistic, probabilistic and unconfoundedness (Imbens and Rubin, 2015). In the super population, a random sample of N units automatically results in an individualistic assignment mechanism. A super-population probabilistic assignment mechanism entails that

$$0 < f_{\boldsymbol{T}|\boldsymbol{Y}(0),\boldsymbol{Y}(1),\boldsymbol{X}}\big(t_1|Y_i(0), Y_i(1), \boldsymbol{X}_i, \phi\big) < 1$$

for each possible $\boldsymbol{X}_i$, $Y_i(0)$ and $Y_i(1)$, where ϕ is a vector of parameters controlling this distribution.

Finally, a super-population assignment mechanism is unconfounded if

$$f_{\boldsymbol{T}|\boldsymbol{Y}(0),\boldsymbol{Y}(1),\boldsymbol{X}}(t|y_0, y_1, \boldsymbol{x}, \phi) = f_{\boldsymbol{T}|\boldsymbol{X}}(t|\boldsymbol{x}, \phi)$$
$$\forall y_o, y_1, \boldsymbol{x}, \phi \text{ and } t \in \{0, 1\}.$$

For notational convenience, we will drop ϕ throughout.

Under an individualistic assignment mechanism, the combination of a probabilistic and unconfounded treatment assignment has been referred to both as strong unconfoundedness and strong ignorability (Stuart, 2010). The class of assignment mechanisms that are individualistic, probabilistic, and unconfounded, but whose control does not lie in the hands of an investigator, are referred to as regular assignment mechanisms, and are most commonly identified with observational data. Weaker versions of unconfoundedness are sufficient for some estimation techniques and estimands (Imbens, 2000), and are discussed in Section 1.5.4.

Let $e_{t_1,t_2}(\boldsymbol{X}) = P(T = t_1|\boldsymbol{X})$ be the propensity score (PS), and let $\hat{e}_{t_1,t_2}(\boldsymbol{X})$ be the estimated PS, traditionally calculated using logistic or probit regression. If treatment assignment is regular, then it is possible to estimate unbiased unit-level causal effects between those at different treatment assignments with equal PSs (Rosenbaum and Rubin, 1983). Propensity scores are often used for either matching, inverse probability weighting or subclassification to estimate (3) and (4).

1.1.1 *Description of estimands*. It is useful to describe how estimands are affected by the distribution of $\boldsymbol{X}$ in treatment groups t_1 and t_2. Figure 1 shows different sets of overlap in the covariates' distributions between those receiving t_1 and t_2. Each circle in Figure 1 represents a hypothetical distribution of $\boldsymbol{X}$ among those exposed to each treatment, allowing for an infinitesimally small number of units outside of it. For example, each circle could represent the 99th percentiles of a two-dimensional multivariate normal distribution. In Figure 1, shaded regions correspond to the distribution of covariates in a population of interest, S_{t_1}. When $0 < P(T = t_2|X = x^*) < 1\ \forall x^* \in S_{t_1}$, $PATT_{t_1,t_2}$ reflects the ATT of those treated on t_1 (Figure 1, Scenario a).

In Scenario b of Figure 1, $PATT_{t_1,t_2}$ also intends to reflect the ATT of those receiving t_1. However, there exists an $x^* \in S_{t_1}$ such that $P(T = t_2|X = x^*) \approx 0$. Thus, the assignment mechanism is not regular and it is impossible to approximate $PATT_{t_1,t_2}$ without making unassailable assumptions due to individuals with covariates lying outside the intersection of the two treatment groups.

One advice to handle this issue is to use a common support region, where those with either $\boldsymbol{X}$ or $\hat{e}_{(t_1,t_2)}(\boldsymbol{X})$ beyond the range of $\boldsymbol{X}$ or $\hat{e}_{(t_1,t_2)}(\boldsymbol{X})$ of those receiving the other treatment are excluded from the analysis phase (Dehejia and Wahba, 1998, Crump et al., 2009). A different advice to reduce differences between matched subjects is by using a caliper matching procedure, and dropping units without eligible matches with similar $\hat{e}_{t_1,t_2}(\boldsymbol{X})$ in the other group (Caliendo and Kopeinig, 2008, Stuart, 2010). With either of these advices, the treatment effect only generalizes to those receiving t_1 who were eligible to be treated with treatment t_2 (i.e., the intersection of the treatment groups in Figure 1, Scenario c). Let E_{1i} be an indicator for subject i having a propensity score within the common support of $\hat{e}_{(t_1,t_2)}(\boldsymbol{X})$. Defensible estimands of interest

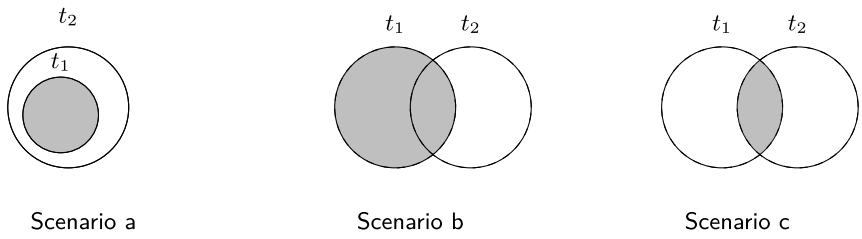

FIG. 1. *Three scenarios of covariate overlap for binary treatment: shaded areas represent subjects included in a matched analysis.*

are now

$$PATE_{E_1(t_1,t_2)} = E\big[Y_i(t_1) - Y_i(t_2)|E_{1i} = 1\big], \tag{5}$$

$$PATT_{E_1(t_1,t_2)} = E\big[Y_i(t_1) - Y_i(t_2)|E_{1i} = 1, T_i = t_1\big]. \tag{6}$$

Although estimands (5) and (6) share the same common support, they may differ if the covariates' distributions of the treated in E_{1i} differ from that of the control.

1.2 Notation for Multiple Treatments

The choice of estimands grows with increasing treatment options. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_Z\}$ be the treatment support for Z total treatments, with $\boldsymbol{\mathcal{Y}}_i = \{Y_i(t_1), Y_i(t_2), \ldots, Y_i(t_Z)\}$ the set of potential outcomes for subject i.

To define potential outcomes and estimate treatment effects with multiple treatments, our assumptions are expanded as follows. First, the SUTVA expands across a subject's vector of potential outcomes. Second, a regular treatment assignment mechanism requires that individualistic, probabilistic and unconfoundedness hold for multiple exposures. Like in the binary case, a random sample of N units from an infinite super-population results in an individualistic assignment mechanism. Assignment mechanisms are super-population probabilistic if

$$0 < f_{\boldsymbol{T}|\boldsymbol{Y}(t_1),\ldots,\boldsymbol{Y}(t_Z),\boldsymbol{X}}\big(t|Y_i(t_1), \ldots, Y_i(t_Z), \boldsymbol{X}_i, \phi\big) < 1$$
$$\forall t \in \{t_1, \ldots, t_Z\},$$

for each possible $\boldsymbol{X}_i, Y_i(t_1), \ldots, Y_i(t_Z)$. With multiple treatments, a super-population unconfounded assignment mechanism requires that

$$f_{\boldsymbol{T}|\boldsymbol{Y}(t_1),\ldots,\boldsymbol{Y}(t_Z),\boldsymbol{X}}(t|y_{t_1}, \ldots, y_{t_Z}, \boldsymbol{x}, \phi) = f_{\boldsymbol{T}|\boldsymbol{X}}(t|\boldsymbol{x}, \phi)$$
$$\forall y_{t_1}, \ldots, y_{t_Z}, \boldsymbol{x}, \phi \text{ and } t \in \{t_1, \ldots, t_Z\}.$$

We first present a broad definition of the possible contrasts that may be of interest with multiple treatments. Define w_1 and w_2 as two subgroups of treatments such that $w_1, w_2 \subseteq \mathcal{T}$ and $w_1 \cap w_2 = \varnothing$. Next, let $|w_1|$ and $|w_2|$ be the cardinality of w_1 and w_2, respectively. Possible estimands of interest are $PATE_{w_1,w_2}$ and $PATT_{w_1|w_1,w_2}$, where

$$PATE_{w_1,w_2} = E\left[\frac{\sum_{t\in w_1} Y_i(t)}{|w_1|} - \frac{\sum_{t\in w_2} Y_i(t)}{|w_2|}\right], \tag{7}$$

$$PATT_{w_1|w_1,w_2} = E\left[\frac{\sum_{t\in w_1} Y_i(t)}{|w_1|} - \frac{\sum_{t\in w_2} Y_i(t)}{|w_2|}\Big|T_i \in w_1\right]. \tag{8}$$

In (7) and (8), the expectation is over all units, $i = 1, \ldots, \mathcal{N}$, and the summation is over the potential outcomes of a specific unit.

An example of when (7) and (8) are scientifically meaningful is in a setting with two conventional and three atypical antipsychotic drugs, where physicians first choose drug type (conventional or atypical) before choosing an exact prescription (Tchernis, Horvitz-Lennon and Normand, 2005). In this case, an investigator could be interested in the general treatment effect between conventional treatments, $w_1 = \{t_1, t_2\}$, and atypical ones, $w_2 = \{t_3, t_4, t_5\}$, and an estimand of interest could be $PATE_{w_1,w_2} = E[\frac{Y_i(t_1)+Y_i(t_2)}{2} - \frac{Y_i(t_3)+Y_i(t_4)+Y_i(t_5)}{3}]$.

The most traditional estimands with multiple treatments contrast all treatments using simultaneous pairwise comparisons, where w_1 and w_2 are each composed of one treatment. Using equation (7), there are $\binom{Z}{2}$ possible *PATE*'s of interest. It is important to note that pairwise *PATE*'s are transitive. Formally, for $w_1 = \{t_1\}$, $w_2 = \{t_2\}$, and $w_3 = \{t_3\}$, $PATE_{w_1,w_3} - PATE_{w_1,w_2} = PATE_{w_2,w_3}$.

For reference group $w_1 = \{t_1\}$, researchers are commonly interested in $Z - 1$ pairwise *PATT*'s, one for each of the treatments which the reference group did not receive (McCaffrey et al., 2013). In order to compare among the $Z - 1$ treatments, the *PATT*'s should also be transitive, such that $PATT_{w_1|w_1,w_3} - PATT_{w_1|w_1,w_2} = PATT_{w_1|w_2,w_3}$. This property generally does not extend when conditioning on a population eligible for different treatment groups. For example, unless the super populations of those receiving treatments w_1 and w_2 are identical, $PATT_{w_1|w_1,w_2} - PATT_{w_2|w_2,w_3}$ is generally not equal to $PATT_{w_1|w_1,w_3}$.

For the remainder of the manuscript, we assume that pairwise contrasts between treatments are the estimands of interest, so that $|w_1| = |w_2| = \cdots = |w_z| = 1$.

1.3 The Generalized Propensity Score

The generalized propensity score (GPS), $r(t, \boldsymbol{X}) = \Pr(T = t|\boldsymbol{X} = \boldsymbol{x})$, extends the PS from a binary treatment setting to the multiple treatment setting (Imbens, 2000, Imai and van Dyk, 2004).

With a binary treatment, knowing $e_{t_1,t_2}(\boldsymbol{X})$ is equivalent to knowing $1 - e_{t_1,t_2}(\boldsymbol{X})$. Thus, two individuals with the same PS are also identical with respect to their probability of receiving t_2. Conditioning with multiple treatments, however, often must be done on a vector of GPSs, defined as $\boldsymbol{R}(\boldsymbol{X}) = (r(t_1, \boldsymbol{X}), \ldots, r(t_Z, \boldsymbol{X}))$, or a function of $\boldsymbol{R}(\boldsymbol{X})$ (Imai and van Dyk, 2004).

Two individuals with the same $r(t, \boldsymbol{X})$ for treatment t may have differing $\boldsymbol{R}(\boldsymbol{X})$'s. For example, for $\mathcal{T} = \{t_1, t_2, t_3\}$, let $\boldsymbol{R}(\boldsymbol{X}_i)$, $\boldsymbol{R}(\boldsymbol{X}_j)$, and $\boldsymbol{R}(\boldsymbol{X}_k)$ be the GPS vectors for subjects i, j, and k, respectively, where $T_i = t_1$, $T_j = t_2$ and $T_k = t_3$, with

$$\boldsymbol{R}(\boldsymbol{X}_i) = (0.30, 0.60, 0.10),$$
$$\boldsymbol{R}(\boldsymbol{X}_j) = (0.30, 0.35, 0.35),$$
$$\boldsymbol{R}(\boldsymbol{X}_k) = (0.30, 0.10, 0.60).$$

Even though $r(t_1, \boldsymbol{X}_i) = r(t_1, \boldsymbol{X}_j) = r(t_1, \boldsymbol{X}_k) = 0.30$, because $r(t_2, \boldsymbol{X}_i) \neq r(t_2, \boldsymbol{X}_j) \neq r(t_2, \boldsymbol{X}_k)$ and $r(t_3, \boldsymbol{X}_i) \neq r(t_3, \boldsymbol{X}_j) \neq r(t_3, \boldsymbol{X}_k)$, differences in outcomes between these subjects would generally not provide unbiased causal effect estimates (Imbens, 2000). In part due to this limitation, Imbens (2000) called individual matching less 'well-suited' to multiple treatment settings. Only under the scenario of $\boldsymbol{R}(\boldsymbol{X}_i) = \boldsymbol{R}(\boldsymbol{X}_j) = \boldsymbol{R}(\boldsymbol{X}_k)$ would contrasts in the outcomes of subjects i, j, and k provide unbiased unit-level estimates of the causal effects between all three treatments (Imbens, 2000, Imai and van Dyk, 2004).

For nominal treatment, the multinomial logistic and the multinomial probit models have been proposed to estimate $\boldsymbol{R}(\boldsymbol{X})$, and for ordinal treatment, the proportional odds model has been suggested (Imbens, 2000, Imai and van Dyk, 2004). Alternatively, researchers have also used models designed for binary outcomes to estimate $\boldsymbol{R}(\boldsymbol{X})$, including logistic and probit regression models on different subsets of subjects receiving each pair of treatments. Although a multinomial model is more intuitive, in practice, Lechner (2002) identified correlation coefficients of roughly 0.99 when comparing the conditional treatment assignment probabilities from a set of binary probit models to those from a multinomial probit. As another option, McCaffrey et al. (2013) used generalized boosted models to independently estimate $P(I_i(t)|\boldsymbol{X})$, where $I_i(t) = \{1 \text{ if } T_i = t, 0 \text{ otherwise}\}$. The probabilities estimated using generalized boosted models may not add up to unity. To address this issue, McCaffrey et al. (2013) proposed an additional procedure that selects one treatment as a holdout and estimates $P(I_i(t)|\boldsymbol{X})$ using the estimated odds ratios of the probability of being assigned to each treatment versus the probability of being assigned to the holdout treatment. The choice of the holdout treatment may result in different estimated probabilities, and because it relies on binary estimation of subsamples of the population, it may not be able to adjust for the entire $\boldsymbol{R}(\boldsymbol{X})$. In our review below, we specify the model that is used to estimate the treatment assignment probabilities suggested by each method.

1.4 Ordinal Treatments

With ordinal treatments, such as scales (e.g., never–sometimes–always) or doses (e.g., low–medium–high), it is sometimes possible to condition on a scalar balancing score in place of conditioning on a vector. This can be done by estimating the assignment mechanism as a function of $\boldsymbol{X}$ using the proportional odds model (McCullagh, 1980), such that

$$\log\left(\frac{P(T_i < t)}{P(T_i \geq t)}\right) = \theta_t - \boldsymbol{\beta}^T \boldsymbol{X}_i, \quad t = 1, \ldots, Z-1. \tag{9}$$

Letting $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)^T$, Joffe and Rosenbaum (1999) and Imai and van Dyk (2004) showed that after using this model for the assignment mechanism, differences in outcomes between units with different exposure levels but equal $\boldsymbol{\beta}^T \boldsymbol{X}$ scores can provide unbiased unit-level estimates of causal effects at that $\boldsymbol{\beta}^T \boldsymbol{X}$.

The balancing property of $\boldsymbol{\beta}^T \boldsymbol{X}$ can be used to match or subclassify subjects receiving different levels of an ordinal exposure. Lu et al. (2001) used nonbipartite matching to form matched sets based on a function of $\boldsymbol{\beta}^T \boldsymbol{X}$ and the relative distance between exposure levels. While this method does not specify an exact causal estimand, it is used for testing the hypothesis of whether or not a dose-response relationship exists between T and Y (see Armstrong, Jagolinzer and Larcker, 2010, Frank, Akresh and Lu, 2010, Snodgrass et al., 2011 to name a few).

Imai and van Dyk (2004), Zanutto, Lu and Hornik (2005), Yanovitzky, Zanutto and Hornik (2005) and Lopez and Gutman (2014) used equation (9) to estimate treatment assignment by subclassifying subjects with similar $\boldsymbol{\beta}^T \boldsymbol{X}$ values. After subclassification on $\boldsymbol{\beta}^T \boldsymbol{X}$, the distribution of $\boldsymbol{X}$ across treatments is roughly equivalent for units in the same subclass. Unbiased causal effects can be estimated within each subclass, and aggregated across subclasses using a weighted average to estimate either *PATE*'s or *PATT*'s (Zanutto, Lu and Hornik, 2005). Lopez and Gutman (2014) found that combining regression adjustment with subclassification yielded more precise estimates.

A different strategy for estimating the causal effects of ordinal exposures is to dichotomize the treatment using a pre-specified cutoff and binary propensity score methods (Chertow, Normand and McNeil, 2004, Davidson et al., 2006, Schneeweiss et al., 2007). This procedure may result in a loss of information, as all subjects on one side of the cutoff are treated as having

the same exposure level, and could violate the component of SUTVA which requires no hidden treatment. Royston, Altman and Sauerbrei (2006) identified a loss of power, residual confounding of the treatment assignment mechanism, and possible bias in estimates as the results of dichotomization. Moreover, dichotomization makes identification of an optimal exposure level impossible. Thus, matching or subclassifications methods which maintain all exposure levels while balancing on $\boldsymbol{\beta}^T \boldsymbol{X}$ are preferred for causal inference with ordinal exposures (Imai and van Dyk, 2004). Inverse probability weighting can also be used to estimate causal effects from ordinal treatments (Imbens, 2000).

1.5 Nominal Treatments

Nominal treatments do not follow a specific order. Thus, it is harder to identify a 'sensible' function that reduces $\boldsymbol{R}(\boldsymbol{X})$ to a scalar. Several methods have been proposed to estimate causal effects with multiple treatments from observational data. We provide an overview of these methods and explicate on their assumptions and estimands.

1.5.1 *Series of binomial comparisons*. Lechner (2001, 2002) estimated *PATT*'s between multiple treatments using a series of binary comparisons (*SBC*). *SBC* implements binary propensity score methods within each of the $\binom{Z}{2}$ pairwise population subsets. For example, a treatment effect comparing t_1 to t_2 uses only subjects receiving either t_1 or t_2, ignoring subjects that received t_3. Lechner advocates matching on either $\hat{e}_{(t_1,t_2)}(\boldsymbol{X})$, estimated using logistic or probit regression, or $\hat{r}(t_1, \boldsymbol{X})/(\hat{r}(t_1, \boldsymbol{X}) + \hat{r}(t_2, \boldsymbol{X}))$, where $\hat{r}(t_1, \boldsymbol{X})$ and $\hat{r}(t_2, \boldsymbol{X})$ are estimated using a multinomial regression model.

Figure 2 (Scenario d) depicts the unique common support regions for $Z = 3$ when using SBC, where treatment effects reflect different subsets of the population. Let $\boldsymbol{e}_{(t_1,t_2)}(\boldsymbol{X}, \boldsymbol{T} = \boldsymbol{t}_1)$ and $\boldsymbol{e}_{(t_1,t_2)}(\boldsymbol{X}, \boldsymbol{T} = \boldsymbol{t}_2)$ be the vector of all binary propensity scores among subjects receiving t_1 and t_2, respectively. We define $E_{2i}(t_1, t_2)$ as the indicator for subject i having a binary propensity score for treatments t_1 and t_2 within the common support:

$$E_{2i}(t_1, t_2) = \begin{cases} 1 & \text{if } e_{(t_1,t_2)}(\boldsymbol{X}_i) \in \boldsymbol{e}_{(t_1,t_2)}(\boldsymbol{X}, \boldsymbol{T} = \boldsymbol{t}_1) \cap \\ & \quad \boldsymbol{e}_{(t_1,t_2)}(\boldsymbol{X}, \boldsymbol{T} = \boldsymbol{t}_2), \\ 0 & \text{otherwise.} \end{cases}$$

SBC estimates the causal effect of treatment t_1 versus treatment t_2, among those on t_1, as

$$\begin{aligned} &PATT_{E_2(t_1|t_1,t_2)} \\ &\quad = E\big[Y_i(t_1) - Y_i(t_2) | T_i = t_1, E_{2i}(t_1, t_2) = 1\big]. \end{aligned} \tag{10}$$

Each pairwise treatment effect from *SBC* generalizes only to subjects eligible for that specific pair of treatments, as opposed to those eligible for all treatments. Such pairwise treatment effects are not transitive, and cannot generally inform which treatment is optimal when applied to the entire population. For example, $PATT_{E_2(t_1|t_1,t_2)}$ and $PATT_{E_2(t_1|t_1,t_3)}$ may generalize to separate subsets of units who received t_1 [i.e., the super population where $E_{2i}(t_1|t_1, t_2) = 1$ could differ from the super population where $E_{2i}(t_1|t_1, t_3) = 1$].

Despite this major limitation, versions of *SBC* have been applied in economics, politics and public health (Bryson, Dorsett and Purdon, 2002, Dorsett, 2006, Levin and Alvarez, 2009, Drichoutis, Lazaridis and Nayga Jr, 2005, Kosteas, 2010).

1.5.2 *Common referent matching*. With three treatments, Rassen et al. (2011) proposed common referent matching (*CRM*) to create sets with one individual from each treatment type. For $\mathcal{T} = \{t_1, t_2, t_3\}$, the treatment t_1 such that $n_{t_1} = \min\{n_{t_1}, n_{t_2}, n_{t_3}\}$, is used as the reference group.

CRM is composed of 3 steps. (1) Among those receiving each pair of treatments, $\{t_1, t_2\}$ or $\{t_1, t_3\}$, logistic or probit regression is used to estimate $e_{t_1,t_2}(\boldsymbol{X})$ and $e_{t_1,t_3}(\boldsymbol{X})$, respectively; (2) Using $1:1$ matching, pairs of units receiving t_1 or t_2 are matched using $\hat{e}_{t_1,t_2}(\boldsymbol{X})$ and pairs of units receiving t_1 or t_3 are matched using $\hat{e}_{t_1,t_3}(\boldsymbol{X})$; (3) These two cohorts are used to construct $1:1:1$ matched triplets using the patients receiving t_1 who were matched to both a unit receiving t_2 and a unit receiving t_3, along with their associated matches. Matched pairs from treatments t_1 and t_3 are discarded if the unit receiving t_1 was not matched with a unit on treatment t_2, and pairs of units receiving t_1 and t_2 are discarded when there is no match for the reference unit to a unit receiving t_3.

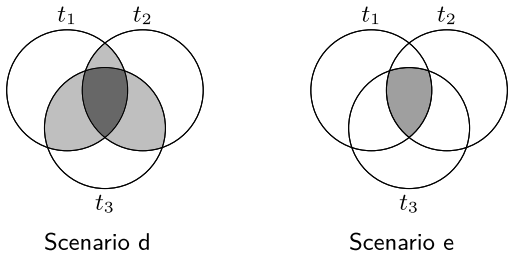

FIG. 2. *Two scenarios of eligible subjects with three treatments: shaded areas represent subjects included in a matched analysis.*

Let E_{3i} be the indicator for having two pairwise binary PSs within their respective common supports, such that

$$E_{3i} = \begin{cases} 1 & \text{if } E_{2i}(t_1, t_2) = 1 \text{ and } E_{2i}(t_1, t_3) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

CRM attempts to estimate the following treatment effects:

$$PATT_{E_3(t_1|t_1,t_2)} = E\big[Y_i(t_1) - Y_i(t_2)|T_i = t_1, E_{3i} = 1\big],$$

$$PATT_{E_3(t_1|t_1,t_3)} = E\big[Y_i(t_1) - Y_i(t_3)|T_i = t_1, E_{3i} = 1\big],$$

$$PATT_{E_3(t_1|t_2,t_3)} = E\big[Y_i(t_2) - Y_i(t_3)|T_i = t_1, E_{3i} = 1\big],$$

$PATT_{E_3(t_1|t_2,t_3)}$ is the average difference in the potential outcomes of receiving treatments t_2 and t_3 among the population of subjects who received t_1.

Rassen et al. (2011) relied on common sampling variance estimates produced by the SAS statistical software (SAS Institute Inc., 2003) to make inference. These estimates may underestimate the sampling variance, because they ignore the variability induced by the matching procedure. The next section will explain the possible issues that arise from *CRM* and similar procedures.

1.5.3 *Interlude*: *Binary PS applications to multiple treatments*. The following hypothetical example with $Z = 3$ illustrates issues with the implementation of binary PS tools, as in *SBC* and *CRM*, when there are multiple treatments.

Let $X_i = \binom{x_{1i}}{x_{2i}}$ be a vector of covariates for subject i, and we will assume that $X_i|T_i = t \sim N(\mu_t, \mathbb{1})$, where μ_t is a 2×1 mean vector and $\mathbb{1}$ is the 2×2 identity matrix. Further, we let $\mu_1 = (0, 0)$, $\mu_2 = (0, a)$, and $\mu_3 = (a, 0)$.

An arbitrary linear combination of X can be expressed as the sum of components along the standardized linear discriminant, $\mathcal{Z}$, and orthogonal to it, $\mathcal{W}$ (Rubin and Thomas, 1992a). Matching on the true or estimated propensity score does not introduce any bias in $\mathcal{W}$ when $X_i|T_i$ follows a multivariate normal distribution. In addition, after matching, $\mathcal{W}$ will have the same expected second moment (Rubin and Thomas, 1992b). Specifically, when matching treatment 1 to treatment 2 with $a = 2$, $Z_{12} = \binom{0}{2}' X_1/\sqrt{2} = \sqrt{2}X_2$ and $\mathcal{W}_{12} = X_1$. After matching, Rubin and Thomas (1992b) showed that

$$E\big(\overline{\mathcal{Z}}_{12}^{m_2}\big) = 2 - \Omega(N_{t_2}, n_{t_2}) \cong 2 - 2\pi \log\left(\frac{N_{t_2}}{n_{t_2}}\right),$$

$$E\big(\overline{\mathcal{Z}}_{12}^{m_1}\big) = 0 + \Omega(N_{t_1}, n_{t_1}) \cong 2 + 2\pi \log\left(\frac{N_{t_1}}{n_{t_1}}\right),$$

where $\overline{\mathcal{Z}}_{12}^{m_1}$ and $\overline{\mathcal{Z}}_{12}^{m_2}$ are the averages of the standardized linear discriminate in the matched treatments 1 and 2, respectively, $\Omega(N_t, n_t)$ is the average expectation of the n largest of the N randomly sampled standard normal variables, and its approximation was depicted in Rubin (1976).

In our example with $a = 2$, $\mu_m = E(\overline{\mathcal{Z}}_{12}^{m_1}) = E(\overline{\mathcal{Z}}_{12}^{m_2})$ when $\frac{N_{t_2}}{n_{t_2}}$ and $\frac{N_{t_1}}{n_{t_1}}$ are bigger than 3. Similar results can be derived when matching treatments 1 and 3 with $Z_{13} = \sqrt{2}X_1$ and $\mathcal{W}_{13} = X_2$.

Matching units that received either treatment 1 or 2 separate from units that received either treatment 1 or 3 generates two subpopulations, one with mean $\binom{0}{\mu_m}$ and another with mean $\binom{\mu_m}{0}$. Note that $\overline{\mathcal{W}}_{12}^{m_1}$ and $\overline{\mathcal{W}}_{12}^{m_2}$ are independent and have similar means (Rubin and Thomas, 1992b). Similarly, $\overline{\mathcal{W}}_{13}^{m_1}$ and $\overline{\mathcal{W}}_{13}^{m_3}$ are independent and have similar means. Lastly, $\overline{\mathcal{W}}_{12}^{m_1}$ is independent from $\overline{\mathcal{W}}_{13}^{m_1}$. When using *CRM*, the units that are kept as matches that received treatment 1 will have the high values of X_1 and X_2. However, because of the independence, group 2 will still have $\overline{\mathcal{W}}_{12}^{m_2}$ that has a mean close to zero and group 3 will still have $\overline{\mathcal{W}}_{13}^{m_3}$ that has a mean close to zero. Thus, in certain settings *CRM* may perform worse than without matching.

This analysis can be observed in a simple simulation where, letting $a = 2$, $n_{t_1} = 400$, and $n_{t_2} = n_{t_3} = 800$, we calculate the sample means among those matched after using a binary matching algorithm [with caliper $0.25 \times \text{SD}(e_{t_1,t_2}(\boldsymbol{X}))$]. Table 1 shows the median covariate values among those receiving each treatment, using only the subjects that remain after matching.

Among the matched set, those receiving t_1 are similar to those receiving t_2 on X_1 but not X_2, and similar to those receiving t_3 on X_2 but not X_1.

Figure 3 depicts one iteration. The ellipses represent 95% quantiles of the bivariate distribution of X_1

TABLE 1
Median covariate values among those matched using a binary algorithm with $Z = 3$

T	X_1	X_2
t_1	0.71 (0.56, 0.82)	0.72 (0.57, 0.83)
t_2	0.70 (0.56, 0.84)	0.01 (−0.16, 0.20)
t_3	0.01 (−0.19, 0.18)	0.72 (0.58, 0.85)

2.5th, 97.5th percentiles shown in parenthesis.

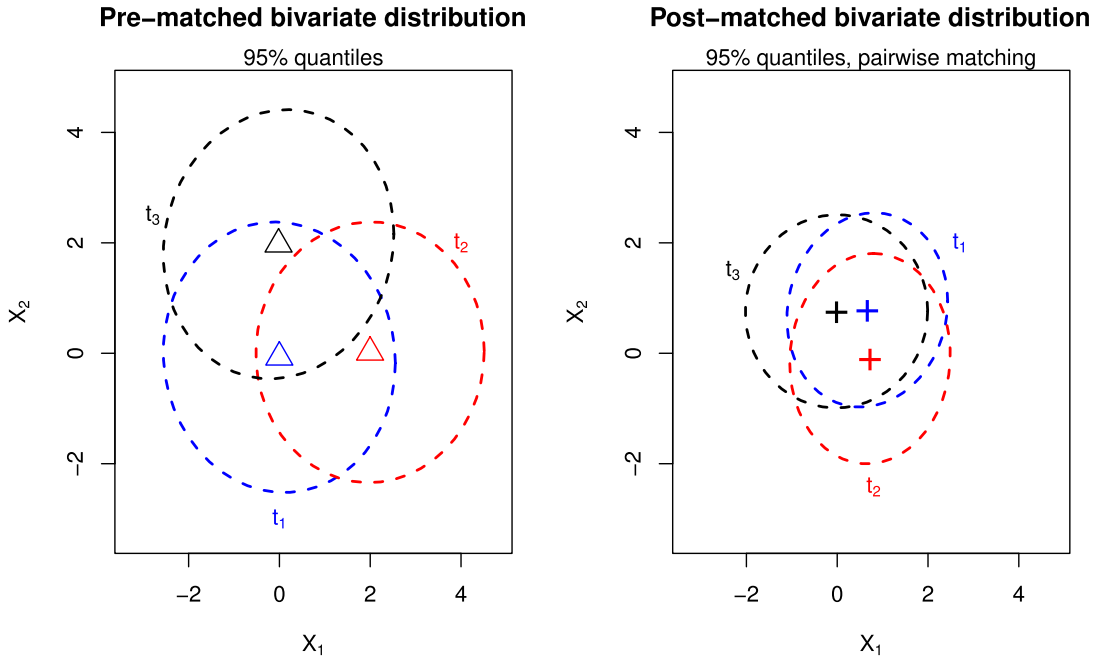

FIG. 3. *95% quantiles of bivariate X_1 and X_2 distribution among subjects matched for $Z = 3$, for pre-matched (left) and post-matched (right) cohorts. Means of the pre- and post-matched covariates' distributions depicted by symbols.*

and X_2, with one ellipse for subjects receiving each treatment both before and after matching. The triangles represent the pre-matched sample mean among those receiving each treatment, while the '+' signs are the mean covariate values among those matched. While matching reduced the covariates' bias relative to the pre-matched sample, the covariate spaces of those receiving each treatment remain unique in the post-matched cohort, and there is limited overlap between subjects receiving t_2 and t_3.

1.5.4 *Inverse probability weighting for multiple treatments*. One common approach for estimating causal effects with multiple treatments uses the inverse probability of treatment assignment as weights (Imbens, 2000, Feng et al., 2012, McCaffrey et al., 2013). When estimating the *PATE* and *PATT* with IPW, a relaxed version of the assumption of a regular treatment assignment can be adopted. IPW requires only that $\forall t \in \mathcal{T}$, $P(I_i(t) = 1|Y_i(t), X_i) = P(I_i(t) = 1|X_i)$ to estimate $PATE_{t_1,t_2}$ and $PATT_{t_1,t_2}$. This condition is referred to as weak unconfoundedness instead of strong unconfoundedness (Imbens, 2000). Imbens (2000) acknowledges that the contrast between weak unconfoundedness and strong unconfoundedness is 'not very different.'

Feng et al. (2012) implemented *IPW* to estimate *PATE*'s between each pair of treatments, such that to contrast $t_1, t_2 \in \mathcal{T}$,

$$\widehat{PATE}_{t_1,t_2} = \widehat{E[Y_i(t_1)]} - \widehat{E[Y_i(t_2)]} \quad \text{where}$$

$$\widehat{E[Y_i(t_1)]} = \left(\sum_{i=1}^{N} \frac{I(T_i = t_1)Y_i}{r(t_1, \boldsymbol{X}_i)}\right) \times \left(\sum_{i=1}^{N} \frac{I(T_i = t_1)}{r(t_1, \boldsymbol{X}_i)}\right)^{-1} \quad \text{and} \tag{11}$$

$$\widehat{E[Y_i(t_2)]} = \left(\sum_{i=1}^{N} \frac{I(T_i = t_2)Y_i}{r(t_2, \boldsymbol{X}_i)}\right) \times \left(\sum_{i=1}^{N} \frac{I(T_i = t_2)}{r(t_2, \boldsymbol{X}_i)}\right)^{-1}.$$

When using *IPW*, extreme weights that are close to 0 can yield erratic causal estimates with large sample variances (Little, 1988, Kang and Schafer, 2007, Stuart and Rubin, 2008), an issue which is increasingly likely as Z increases, where treatment assignment probabilities for some treatments may become quite small. For example, in an analysis of rare treatment decisions with $Z = 7$, Kilpatrick et al. (2013) found weights greater than 10^4 and resulting confidence intervals that were sensitive to model specification. A possible solution to the unstable estimates that has been applied in the binary treatment setting is to trim subjects with extreme

weights (Lee, Lessler and Stuart, 2011). Kilpatrick et al. (2013) observed increased precision with weight removal, relative to the inclusion of all subjects; however, dropping extreme weights also yielded increased bias. This observation reveals a subtle point that is not always recognized. As shown in Section 1.3, in contrast to binary propensity scores, the comparison of units with similar $r(t, \boldsymbol{X})$ and different $\boldsymbol{R}(\boldsymbol{X})$ that receive different treatments has no causal interpretation (Imbens, 2000). Instead, only a comparison of the $r(t, \boldsymbol{X})$ weighted averages has such interpretation. As a result, trimming units with $r(t, \boldsymbol{X})$ that are close to 0 or 1 may actually drop units with different covariates' distributions, which could ultimately increase the bias.

For binary treatment, other approaches have been suggested to limit the effects of large weights. These include a doubly robust approach (Tan, 2010), a covariate balancing propensity score (Imai and Ratkovic, 2014), and generalized boosted models (McCaffrey, Ridgeway and Morral, 2004, McCaffrey et al., 2013), with the latter two methodologies also extending to a multiple treatments framework. To provide confidence intervals for (11), Feng et al. (2012) use the 2.5 and 97.5 quantiles from a nonparametric bootstrap algorithm (Efron and Tibshirani, 1994) to obtain a 95% confidence interval, while McCaffrey et al. (2013) approximate the standard errors by using robust (or so-called 'sandwich') procedure. However, McCaffrey et al. (2013) acknowledge that there is currently no theory that guarantees that these will result in proper confidence intervals when using generalized boosted models, and this is an area for further statistical research.

1.5.5 *Matching for multiple treatments*. Recently, attempts have been made to group several subjects together who have similar $\boldsymbol{R}(\boldsymbol{X})$, including at least one subject receiving each treatment. With $Z = 3$, Rassen et al. (2013) proposed 'within-trio' matching (*WithinTrio*) to form triplets of subjects. *WithinTrio* uses the KD-tree algorithm (Moore, 1991) to optimize triplet similarities based on units' GPSs for treatments t_1 and treatments t_2, by using a distance function between all possible pairs of triplets (Hott, Brunelle and Myers, 2012). Using simulations, Rassen et al. (2013) found that triplets produced using *WithinTrio* generally yielded lower standardized covariate bias when compared to *CRM* and *SBC*.

One limitation of *WithinTrio* is that it uses only t_1 as the reference treatment, where $n_{t_1} = \min\{n_{t_1}, n_{t_2}, n_{t_3}\}$, and so *PATT*'s generalizable to those receiving treatment t_2 or t_3 cannot yet be estimated. Because all subjects receiving t_1 are matched, there is also the potential to form dissimilar triplets, if, for example, all close matches to a subject who received t_1 are already taken as matches by other subjects. At this stage in its development, *WithinTrio* has focused on $Z = 3$ treatment types. An additional limitation is that there is no known procedure for sampling variance estimates, and application of the bootstrap method may be computationally intensive.

Tu, Jiao and Koh (2012) examined a clustering algorithm to bin units into subclasses based on their $\widehat{\boldsymbol{R}(\boldsymbol{X})}$'s using simulations. The authors showed that *K-means* clustering (*KMC*, Johnson et al., 1992) on the logit transformation of the GPS vector, $\text{logit}(\widehat{\boldsymbol{R}(\boldsymbol{X})}) = (\log(\widehat{r(t_1, \boldsymbol{X})}/(1 - \widehat{r(t_1, \boldsymbol{X})})), \ldots, \log(\widehat{r(t_Z, \boldsymbol{X})}/(1 - \widehat{r(t_Z, \boldsymbol{X})}))$, generally provided the highest within subclass covariate similarity between those receiving different treatments. Although the authors do not provide guidelines regarding which units should be included in generating the clusters (e.g., a common support), if all subjects were subclassified, causal effects could be estimated within each subclass and then aggregated across subclasses using a weighted average to estimate either *PATE*s or *PATT*s. One possible issue with clustering on $\boldsymbol{R}(\boldsymbol{X})$ is that some subclasses may not include units from all treatment groups, which will require extrapolation to that subclass. We know of no implementations of *KMC* to estimate causal effects for a nominal exposure with real data. Moreover, there is no known procedure for estimating the sampling variance, and randomization based sampling variance estimates may be too small (Gutman and Rubin, 2015).

2. MATCHING ON A VECTOR OF GENERALIZED PROPENSITY SCORES

In observational studies that intend to compare multiple treatments, matching algorithms attempt to eliminate extraneous variation due to observed covariates. In other words, matching attempts to replicate a multiarm randomized trial where the covariates' distributions of units in each arm are similar. When the number of covariates is significantly larger than the number of treatments, matching on the GPS can reduce the complexity of the algorithms in comparison to matching on the complete set of covariates.

As was shown in Section 1.5.3, relying on standard matching tools for two treatments may result in treatment groups with different distributions of covariates, because matching on a single treatment assignment probability does not ensure similarity across the GPS vector. Additionally, approaches like *SBC* and *CRM*

generalize to specific pairwise subsets of the population, which may be insufficient for clinicians and policy makers, who are generally looking to compare three or more active treatments at once (Rassen et al., 2011, Hott, Brunelle and Myers, 2012). Meanwhile, current approaches designed to match for multiple treatments tend to be either inaccessible or limited in scope.

To address these limitations, we propose a new algorithm, called vector matching (*VM*), which can match subjects with similar $\boldsymbol{R}(\boldsymbol{X})$ using available software. *VM* is designed to generalize to subjects 'eligible' for all treatments simultaneously, which is representative of the multi-arm clinical trial that we are hoping to replicate. We begin by describing the treatment effect that we estimate using *VM*.

Estimands and common support. We expand the work of Dehejia and Wahba (1998) to identify a common support for multiple treatments as follows. Estimate $\boldsymbol{R}(\boldsymbol{X})$ using, for example, a multinomial regression model. For each treatment $t \in \mathcal{T}$, let

$$r(t,\boldsymbol{X})^{(\text{low})} = \max\big(\min\big(\boldsymbol{r}(t,\boldsymbol{X}|\boldsymbol{T}=\boldsymbol{t}_1)\big),\ldots,\min\big(\boldsymbol{r}(t,\boldsymbol{X}|\boldsymbol{T}=\boldsymbol{t}_Z)\big)\big), \tag{12}$$

$$r(t,\boldsymbol{X})^{(\text{high})} = \min\big(\max\big(\boldsymbol{r}(t,\boldsymbol{X}|\boldsymbol{T}=\boldsymbol{t}_1)\big),\ldots,\max\big(\boldsymbol{r}(t,\boldsymbol{X}|\boldsymbol{T}=\boldsymbol{t}_Z)\big)\big), \tag{13}$$

where $\boldsymbol{r}(\boldsymbol{t},\boldsymbol{X}|\boldsymbol{T}=\boldsymbol{\ell})$ is the treatment assignment probability for t among those who received treatment ℓ. This is a rectangular common support region that may drop some units that could be included in the analysis. A more complex common support region based on multidimensional ellipsoids or convex hull regions provide areas for further research.

Subjects with $r(t,\boldsymbol{X}) \notin (r(t,\boldsymbol{X})^{(\text{low})}, (r(t,\boldsymbol{X})^{(\text{high})})$ $\forall t \in \mathcal{T}$ may have $\boldsymbol{X}$ values that are not observed for some treatment groups, and should be discarded. After using this exclusion criterion, it is recommended to re-fit the GPS model, to ensure that estimated GPSs are not disproportionately impacted by those dropped (adapted from the binary treatment scenario in Imbens and Rubin, 2015). Re-fitting is generally done once; unless the minimum and maximum estimated GPSs are identical among each group receiving each treatment, there will always be subjects outside the boundaries in a continuously re-fit model.

Let E_{4i} be the indicator for all treatment eligibility, where

$$E_{4i} = \begin{cases} 1 & \text{if } r(t,\boldsymbol{X}_i) \in \big(r(t,\boldsymbol{X})^{(\text{low})}, r(t,\boldsymbol{X})^{(\text{high})}\big) \\ & \quad \forall t \in \mathcal{T}, \\ 0 & \text{otherwise.} \end{cases}$$

The shaded region in Figure 2, Scenario e, depicts the subset of those eligible for all three treatments.

Using t_1 as a reference treatment, *PATT*'s among subjects eligible for all treatments are defined as follows:

$$\begin{aligned} &PATT_{E_4(t_1|t_1,t_2)} \\ &\quad = E\big[Y_i(t_1) - Y_i(t_2)|T_i = t_1, E_{4i} = 1\big], \\ &PATT_{E_4(t_1|t_1,t_3)} \\ &\quad = E\big[Y_i(t_1) - Y_i(t_3)|T_i = t_1, E_{4i} = 1\big], \\ &\cdots = \cdots \end{aligned} \tag{14}$$

$$PATT_{E_4(t_1|t_1,t_Z)} = E\big[Y_i(t_1) - Y_i(t_Z)|T_i = t_1, E_{4i} = 1\big]. \tag{15}$$

There are two benefits to our definition of eligibility. First, all estimands in (14) are transitive; $PATT_{E_4(t_1|t_1,t_2)}$ and $PATT_{E_4(t_1|t_1,t_3)}$, for example, could be contrasted to compare t_2 and t_3 in the population of subjects who received t_1. Second, because all subjects included have $r(t,\boldsymbol{X})^{(\text{low})} < r(t,\boldsymbol{X}) < r(t,\boldsymbol{X})^{(\text{high})}$ $\forall t$, extrapolation to subjects that did not received a specific treatment is reduced.

2.1 Vector Matching

As described in Section 1.3, when comparing multiple treatments, the GPS is a vector composed of $Z-1$ independent components; ultimately, our goal is similarity across this vector. One possible matching algorithm for $\boldsymbol{R}(\boldsymbol{X})$ begins by creating K1 intervals based on $r(t_1,\boldsymbol{X})$ so that there is at least one unit from each treatment group in each interval. The algorithm continues by subclassifying units into K2 intervals within each of the K1 intervals with similar $r(t_2,\boldsymbol{X})$ such that each new interval includes at least one unit from each treatment group. This proceeds until all of the components of $\boldsymbol{R}(\boldsymbol{X})$ have been subclassified. Such an algorithm may be influenced by the order that the components of $\boldsymbol{R}(\boldsymbol{X})$ are subclassified. Some orderings of the components may lead to declaring a large set of units as unmatchable and may result in estimates that have limited use in practice.

To handle these difficulties, vector matching consists of two steps that can be implemented using common software. First, place subjects into clusters using *KMC* such that subjects within each cluster are roughly similar on one or more GPS components and there is at least one subject from each treatment in each cluster. Second, match pairs of subjects together only if they appear in the same subclass.

Below, we explicate and summarize the procedure for a reference treatment $t \in \mathcal{T} = \{t_1, \ldots, t_Z\}$:

1. Estimate $\boldsymbol{R}(\boldsymbol{X}_i), i = 1, \ldots, N$ using, for example, a multinomial logistic model.
2. Drop units outside the common support (e.g., those with $E_{4i} = 0$), and re-fit the model once.
3. $\forall t' \neq t$

   (a) Classify all units using *KMC* on the logit transform of $\widehat{\boldsymbol{R}}_{t,t'}(\boldsymbol{X})$, where $\widehat{\boldsymbol{R}}_{t,t'}(\boldsymbol{X}) = (\widehat{r(\ell, \boldsymbol{X})}$ $\forall \ell \neq t, t')$. This forms K strata of subjects, with similar $Z - 2$ GPS scores [not including $\widehat{r(t, \boldsymbol{X})}$ or $\widehat{r(t', \boldsymbol{X})}$] in each $k \in K$.

   - *Example*: with $Z = 5$, $\mathcal{T} = \{t_1, \ldots, t_5\}$, reference treatment t_1 and letting $t' = t_2$, *VM* would use *KMC* on $\text{logit}(\widehat{r(t_3, \boldsymbol{X}_i)}, \widehat{r(t_4, \boldsymbol{X}_i)}, \widehat{r(t_5, \boldsymbol{X}_i)})$

   (b) Within each strata $k \in K$, use $1:1$ matching to match those receiving t to those receiving t' on $\text{logit}(\widehat{r(t, \boldsymbol{X}_i)})$. Matching is performed with replacement using a caliper of $\varepsilon \times \text{SD}(\text{logit}(\widehat{r(t, \boldsymbol{X}_i)}))$, where $\varepsilon = 0.25$.

   - *Example*: this matches subjects receiving t_1 to those receiving t_2 within each of the strata produced by *KMC*

4. Subjects receiving t who were matched to subjects receiving all treatments $\ell \neq t$, along with their matches receiving the other treatments, compose the final matched cohort.

Up to $n_{t_1, E_4=1}$ sets can be generated using vector matching, where $n_{t_1, E_4=1}$ is the number of subjects receiving t_1 with $E_{4i} = 1$.

For $Z = 3$, vector matching reduces to:

1. Those receiving t_1 are matched to those receiving t_2 using $\text{logit}(\widehat{r(t_1, \boldsymbol{X}_i)})$ within *K-means* strata of $\text{logit}(\widehat{r(t_3, \boldsymbol{X}_i)})$
2. Those receiving t_1 are matched to those receiving t_3 using $\text{logit}(\widehat{r(t_1, \boldsymbol{X}_i)})$ within *K-means* strata of $\text{logit}(\widehat{r(t_2, \boldsymbol{X}_i)})$
3. Extract the subjects receiving t_1 who were matched to both subjects receiving t_2 and t_3, as well as their matches.

After the completion of VM, we are left with many sets that include a unit from the reference treatment and matched units from each of the other $Z - 1$ treatments. By matching within a subclass, we have ensured that matched units are close on one component of the GPS and roughly similar on the other components. As a result, *VM* improves the balance in covariates' distributions between those receiving different treatments relative to matching on a single element of the GPS. *VM* is relatively efficient computationally, and is not as affected by the ordering of the GPS elements.

We implemented *VM* by matching on $\text{logit}(\widehat{r(t, \boldsymbol{X})})$ as well as $\widehat{r(t, \boldsymbol{X})}$ within strata estimated using KMC. The logit transformation produced smaller biases, which parallels findings observed with binary treatment (Rosenbaum and Rubin, 1985). Additionally, while the recommendation for binary treatment uses $\varepsilon = 0.25$ (Austin, 2011), we examined $\varepsilon \in \{0.25, 0.50, 1.0\}$. Based on the simulation design that is described in Section 3, *VM* performed best in terms of bias and percent of matched eligible subjects with $\varepsilon = 0.25$ (data not shown). The in strata matching procedure is implemented using the *Matching* (Sekhon, 2011) package in *R* statistical software (R Core Team, 2014).

Figure 4 shows the 95% quantiles of the bivariate X_1 and X_2 distribution after implementing vector matching on the same iteration as the one shown in Figure 3 (Section 1.5.3). Whereas binary procedures were insufficient for identifying similar matched sets, the circles are near perfect overlaps after using vector matching.

2.2 Post-Matching Analysis

Although our focus is on the design phase of matching for multiple treatments, it is important to consider how matched sets could be used to make inferences. Point estimates for (14)–(15) using *VM* matches can be obtained by contrasting those matched using a weighted average, with weights proportional to ψ_i, where ψ_i is the number of times subject i is part of a matched set. Let n_{trip} be the number of matched sets. Point estimates for (14)–(15) can be obtained using (16)–(17), where

$$SATT_{E_4(t_1|t_1,t_2)} = \frac{\sum_{i \in E_4} Y_i I(T_i = t_1)\psi_i - Y_i I(T_i = t_2)\psi_i}{n_{\text{trip}}}, \tag{16}$$

$$SATT_{E_4(t_1|t_1,t_3)} = \frac{\sum_{i \in E_4} Y_i I(T_i = t_1)\psi_i - Y_i I(T_i = t_3)\psi_i}{n_{\text{trip}}},$$

$$\cdots = \cdots$$

$$SATT_{E_4(t_1|t_1,t_Z)} = \frac{\sum_{i \in E_4} Y_i I(T_i = t_1)\psi_i - Y_i I(T_i = t_Z)\psi_i}{n_{\text{trip}}}. \tag{17}$$

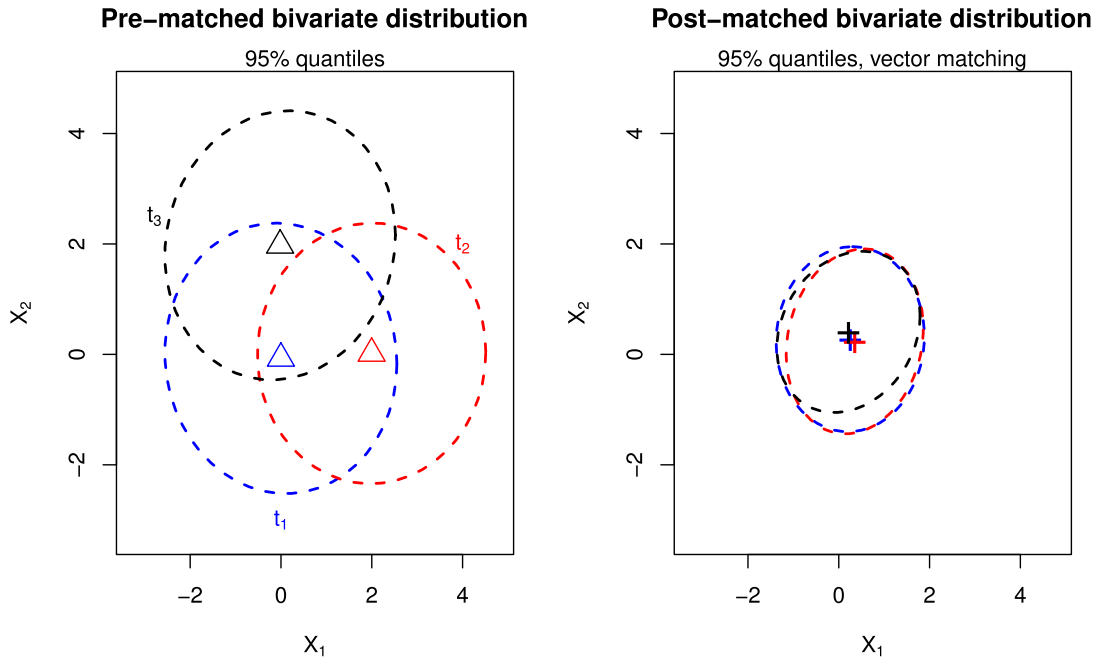

FIG. 4. *95% quantiles of bivariate X_1 and X_2 distribution among subjects matched for $Z = 3$, for pre-matched (left) and post-matched (right) cohorts. Means of the pre- and post-matched distributions depicted by symbols.*

As highlighted earlier, an advantage of these estimands is that they generalize to subjects 'eligible' for all treatments.

Like other approaches that match with multiple treatments, estimating the standard error of these point estimates is still an open research question. One approach for estimating the sampling variances of (16)–(17) is to use functions of the sample variances of $\boldsymbol{Y}$ within those matched at each treatment group. Hill and Reiter (2006) provide weighted variance formulas where the variance in each treatment group is weighted to account for multiplicities in the matched units. Abadie and Imbens (2006) derived a different weighted consistent estimator for the sampling variance of the *PATE* and the *PATT* for a binary treatment. Their estimator matches units with similar covariates' values within each treatment group to estimate the variability of the unit level effects. In general, weighted variance estimators may overestimate the true sampling variance, because they do not account for the correlation between subjects that are matched to one another. Deriving closed form solutions for multiple treatments is an area for further research.

Bootstrapping was proposed as a possible technique to estimate the standard errors of matching estimators of the *PATE* and *PATT* in a binary treatment setting. For matching without replacement, Austin and Small (2014) identified that a bootstrap algorithm that sampled the matched pairs resulted in estimates of the standard error that were close to the empirical standard deviation of the sampling distribution of the estimated treatment effect. For matching with replacement, Hill and Reiter (2006) proposed a more complex form of the bootstrap algorithm. In the complex bootstrap algorithm, bootstrap samples from the original sample are drawn, and within each bootstrap sample, a separate propensity score model is fit and unique sets of matches are identified. In a simulation analysis, the complex bootstrap method was shown to be statistically valid without having extremely large average interval lengths. A similar strategy could be employed with multiple treatments by using *VM* within each iteration of the bootstrap. However, we caution against use of a similar procedure, because in the binary treatment setting, the bootstrap procedure can either overestimate or underestimate the asymptotic variance given that there can be a high degree of consistency in subjects that are matched to one another after using with-replacement matching (Abadie and Imbens, 2008).

A different computationally intensive strategy is to use randomization-based approaches, in which the distributions of treatment effects under the null are formed using different permutations of treatment assignments.

Rosenbaum (2002) described such a permutation approach in the context of matching with a binary treatment and nonoverlapping sets of matches. Hill and Reiter (2006) implemented a similar strategy when matching with replacement by using the Hodges–Lehmann aligned rank test. In simulation analysis, they showed that this test outperformed both the bootstrap and the weighted variance estimators for with replacement matching. Extending this approach for multiple treatments is possible; each matched set obtained by *VM* would be permuted independently, with the observed test statistic compared to the randomization distribution obtained by these permutations. For multiple treatments with matched cohorts, the Friedman test statistic (Sprent and Smeeton, 2007) or the Quade test statistic (Quade, 1979) can be used as alternatives to the Hodges–Lehman aligned rank test statistic.

3. SIMULATIONS

We examine the performance of the methods described in Section 1.5 and the newly proposed method in reducing the bias on observed $\boldsymbol{X}$ using simulations. SBC is not included in the analysis because it cannot be used to contrast three or more treatments simultaneously. Additionally, we assume no natural ordering to the treatment, and thus methods designed for ordinal treatments (Section 1.4) are excluded.

3.1 Evaluating Balance of Matched Sets by Simulation

In order to provide advice to investigators and following Rubin (2001), we generated simulation configurations that are either known or can be estimated from the data. A P-dimensional $\boldsymbol{X}$ was generated for $N = n_{t_1} + n_{t_2} + n_{t_3}$ subjects receiving one of three treatments, $\mathcal{T} \in \{t_1, t_2, t_3\}$, with n_{t_1}, $n_{t_2} = \gamma n_{t_1}$ and $n_{t_3} = \gamma^2 n_{t_1}$ the sample size of subjects receiving treatments t_1, t_2 and t_3. For a similar set of simulations using $Z = 5$, see the Appendix. The values of $\boldsymbol{X}$ were generated from multivariate symmetric distributions such that

$$\begin{aligned} &T_i = t_1, \quad i = 1, \ldots, n_{t_1}, \\ &T_i = t_2, \quad i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1}, \\ &T_i = t_3, \quad i = n_{t_1} + \gamma n_{t_1} + 1, \ldots, \\ &\qquad\qquad n_{t_1} + \gamma n_{t_1} + \gamma^2 n_{t_1}, \end{aligned} \tag{18}$$

$$\boldsymbol{X}_i | \{T_i = t_1\} \sim f(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad i = 1, \ldots, n_{t_1}, \tag{19}$$

$$\boldsymbol{X}_i | \{T_i = t_2\} \sim f(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1}, \tag{20}$$

$$\boldsymbol{X}_i | \{T_i = t_3\} \sim f(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3), \quad i = n_{t_1} + \gamma n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1} + \gamma^2 n_{t_1}, \tag{21}$$

$$\begin{aligned} \mu_1 &= \big((b,0,0), \ldots, (b,0,0)\big)^T, \\ \mu_2 &= \big((0,b,0), \ldots, (0,b,0)\big)^T, \quad \text{and} \\ \mu_3 &= \big((0,0,b), \ldots, (0,0,b)\big)^T, \end{aligned} \tag{22}$$

$$\Sigma_1 = \begin{pmatrix} 1 & \tau & \ldots & \tau \\ \tau & 1 & \ldots & \tau \\ \cdot & \cdot & \ldots & \cdot \\ \tau & \tau & \ldots & 1 \end{pmatrix},$$

$$\Sigma_2 = \begin{pmatrix} \sigma_2 & \tau & \ldots & \tau \\ \tau & \sigma_2 & \ldots & \tau \\ \cdot & \cdot & \ldots & \cdot \\ \tau & \tau & \ldots & \sigma_2 \end{pmatrix}, \quad \text{and} \tag{23}$$

$$\Sigma_3 = \begin{pmatrix} \sigma_3 & \tau & \ldots & \tau \\ \tau & \sigma_3 & \ldots & \tau \\ \cdot & \cdot & \ldots & \cdot \\ \tau & \tau & \ldots & \sigma_3 \end{pmatrix}.$$

The following design implicitly assumes a regular assignment mechanism (Imbens and Rubin, 2015) that depends on eight factors (Table 2). The distance between treated groups, b, is defined in terms of standardized bias B, where

$$B = \frac{b}{\sqrt{\frac{1+\sigma_2^2+\sigma_3^2}{3}}} \tag{24}$$

in order to evaluate the reduction in initial bias somewhat independently of the variance ratios σ_2^2 and σ_3^2.

Due to the small number of eligible subjects remaining when $P = 6$ and $n_{t_1} = 500$, these simulations are discarded, leaving 1080 simulation configurations. For each simulation condition, 200 data sets are generated, and on each data set, *VM* (using $K = 5$ strata), *CRM*, *IPW* and *KMC* are used to identify matched, weighted or subclassified sets. For *CRM*, we used $\varepsilon = 0.25$ (Austin, 2011).

TABLE 2
Simulation factors

Factor	Levels of factor
n_{t_1}	$\{500, 2000\}$
$\gamma = \frac{n_{t_2}}{n_{t_1}} = \frac{n_{t_3}}{n_{t_2}}$	$\{1, 2\}$
f	$\{t_7, Normal\}$
b	$B = \frac{b}{\sqrt{\frac{1+\sigma_2^2+\sigma_3^2}{3}}}$ takes levels $\{0, 0.25, 0.50, 0.75, 1.00\}$
τ	$\{0, 0.25\}$
σ_2^2	$\{0.5, 1, 2\}$
σ_3^2	$\{0.5, 1, 2\}$
P	$\{3, 6\}$

3.2 Simulation Metrics

While several metrics have been proposed for evaluating the success of matching with binary treatments (see, e.g., Austin, Grootendorst and Anderson, 2007, Austin, 2009), assessments for multiple treatments are not as well formalized (Stuart, 2010).

For *VM* or *CRM*, let n_{trip} be the number of triplets formed, and let ψ_i be the number of times subject i is part of a triplet. The weighted mean of covariate p, $p = 1, \ldots, P$, at treatment t, is defined as $\bar{X}_{pt}$, such that

$$\bar{X}_{pt} = \frac{\sum_{i=1}^{N} X_{pi} I_i(t) \psi_i}{n_{\text{trip}}}. \tag{25}$$

For *IPW*, $\psi_i = \frac{1}{r(t, \boldsymbol{X}_i)}$ is each subject's weight, where $r(t, \boldsymbol{X})$ is estimated using multinomial logistic regression, and n_{trip} is simply the number of matched subjects receiving each treatment t. With *KMC*, $\bar{X}_{pt}$'s are calculated within each subclass, and weighted across subclasses, with weights proportional to the number of subjects in each subclass.

For a binary treatment, Rubin and Thomas (1996) and Rubin (2001) suggest that the standardized bias between $\boldsymbol{X}_p$ in the treatment (t_1) and control groups (t_2), SB_{p12}, should be less than 0.25 to make defensible causal statements, where

$$SB_{p12} = \frac{\bar{X}_{p1} - \bar{X}_{p2}}{\delta_{p1}} \tag{26}$$

and δ_{p1} is the standard deviation of $\boldsymbol{X}_p$ in t_1.

In our simulations, we calculated three such biases for each covariate p for each pair of treatments, SB_{p12}, SB_{p13} and SB_{p23}. As in Hade (2012), we extract the maximum absolute standardized pairwise bias at each covariate, $Max2SB_p$, such that

$$Max2SB_p = \max\bigl(|SB_{p12}|, |SB_{p13}|, |SB_{p23}|\bigr). \tag{27}$$

For all of the matching algorithms and at each p, δ_{p1}, the standard deviation of $\boldsymbol{X}_p$ in the full sample among those receiving reference t_1, is used for standardization, to ensure that observed differences in the similarity of those matched are easily contrasted (as in Stuart and Rubin, 2008.

With three treatment pairs, $Max2SB_p$ reflects the largest discrepancy in estimated covariate means between any two treatment groups for a specific covariate. Using a similar metric to assess covariate balance, McCaffrey et al. (2013) advocated using a standardized bias cutoff of 0.20 for multiple treatments. We also examined average absolute standardized biases, $\frac{|SB_{p12}|+|SB_{p13}|+|SB_{p23}|}{3}$, finding similar results to those with $Max2SB_p$.

In addition to bias, for *VM* and *CRM* we also estimated the fraction of units from the entire population who received t_1 and were eligible to receive the other two treatments which were included in the final matched set, *%Matched*. This metric provides a sense of the similarity between those matched and the population that we are interested in generalizing to. Simulations with *%Matched* ≈ 1 and relatively low $Max2SB_p$ $\forall p$ are optimal in the sense that almost all subjects who received t_1 are matched with subjects receiving t_2 and t_3 and the distributions of their covariates are similar. *%Matched* is not relevant for *IPW*, because weights are estimated for all subjects that meet the eligibility criteria.

At each simulation configuration and for each matching algorithm, $Max2SB_p$ $\forall p$ and *%Matched* are obtained, and averaged across 200 replications. For simplicity, we summarize $Max2SB_p$ $\forall p$ by averaging over p, where $\overline{Max2SB} = \sum_{p=1,\ldots,P} Max2SB_p / P$.

3.3 Determinants of Matching Performance

Figure 5 shows boxplots of $\overline{Max2SB}$ and *%Matched* across each of the simulation factors. $\overline{Max2SB}$ was calculated for *VM*, *CRM*, *IPW*, *KMC* and in the pre-matched cohort of eligible subjects. Each point in each of the boxplots represents the bias at one factors' configuration. In Figure 5, $\overline{Max2SB}$ exceeds a cutoff of 0.20 in 57% of combinations when using *KMC*, compared to 25% when using *IPW*, 19% when using *CRM* and to 4% when using *VM*. There are 16 simulation configurations for which *IPW* yields a $\overline{Max2SB}$ greater than 1.5. In general, *KMC* has done the worst, with $\overline{Max2SB}$ in more than 75% of configurations lying

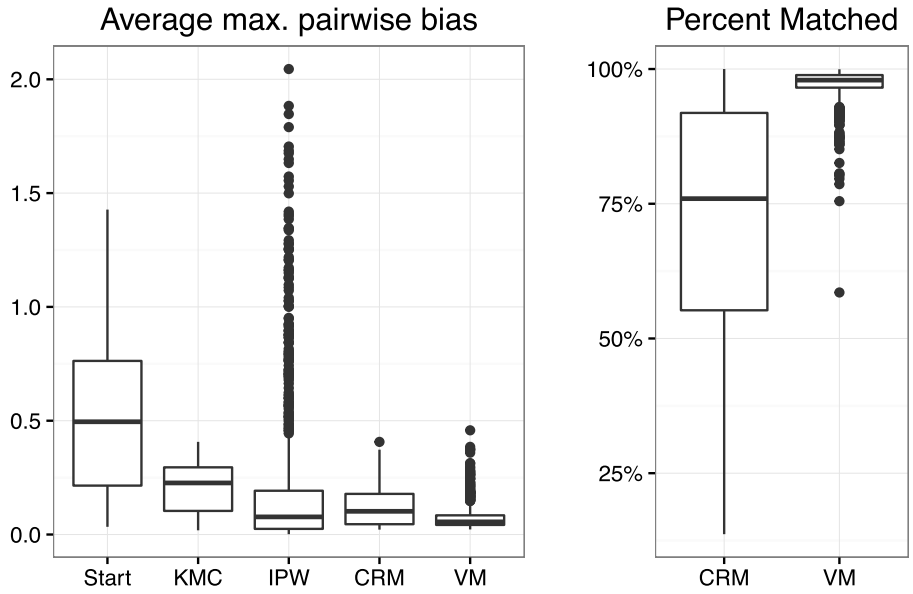

FIG. 5. *$\overline{Max2SB}$ for pre-matched cohort and by matching algorithms (left), and %Matched for VM and CRM.*

above the median $\overline{Max2SB}$ for each of the other algorithms. This corresponds to results for a binary treatment assignment which suggest that subclassification alone may not sufficiently to reduce bias in the covariates' distributions (Gutman and Rubin, 2013), as well as the problem that some clusters may not include units from all treatment groups. Given its poor performance, *KMC* is not shown in the tables below.

VM matched at least 85% of eligible reference subjects in a matched triplet in 99% of the configurations, while only 37% of the configurations for *CRM* reached the 85% cutoff. *VM* matched at least 95% of the eligible reference group subjects on more than 85% of the configurations.

To identify factors with the largest influence on the performance of using *VM*, *CRM* and *IPW*, we rank them by their MSE for both $\overline{Max2SB}$ as well as *%Matched* (as in Rubin, 1979, Cangul et al., 2009). Because *%Matched* was highly skewed, we used the Box–Cox power transformation (Sakia, 1992) to make this metric approximately normally distributed.

Initial covariate bias B drives the highest proportion of variation in $\overline{Max2SB}$, accounting for roughly 85%, 70% and 45% of the variability for *CRM*, *VM* and *IPW*, respectively (Table 3). Compared to *VM* and *CRM*, *IPW* biases' are substantially driven by the distribution type (f) and the variance terms σ_2 and σ_3. While γ, the rate of those receiving t_2 and t_3 relative to the number of subjects receiving t_1, is not an important factor for *IPW*, it is the second and third most important factors of *CRM* and *VM*, respectively. This is also noticed with matching methods for binary treatment (Rubin, 1973). B also drives nearly 100% of the variability in *%Matched* for *VM* and *CRM* (not shown). The second most influential factor for the ANOVA of *%Matched* using those matched via *CRM* is γ; for a binary matching approach, the increased number of available matches on t_2 and t_3 increases the likelihood that a subject receiving t_1 is matched.

Having identified the principal determinants of bias and matching size, we average over the other factors in order to further detail the effects of the principal ones. Tables 4 to 7 show $\overline{Max2SB}$ based on different biases (B), distributions of $\boldsymbol{X}$ (f), number of parameters (P),

TABLE 3
ANOVA for VM, CRM and IPW $\overline{Max2SB}$: most influential factors

VM		***IPW***		***CRM***	
Variable	**MSE**	**Variable**	**MSE**	**Variable**	**MSE**
B	16,046	B	154,397	B	41,354
p	3776	f	115,220	γ	4089
γ	1259	$B \times f$	42,628	σ_t	2271
σ_t	934	σ_s	22,903	σ_s	469
σ_s	791	p	14,746	$B : \sigma_t$	243
τ	692	n_{t_1}	11,707	$B : \gamma$	233
$B \times p$	692	σ_t	10,137	$B \times \sigma_s$	147
$B \times \sigma_t$	223	$B \times \sigma_s$	5280	p	64
$p \times \sigma_s$	220	$B \times p$	3838	$B \times n_{t_1}$	48
$p \times \sigma_t$	216	$B \times n_{t_1}$	3776	f	28

TABLE 4
$\overline{Max2SB}$, *small/equal sample sizes*: $n_{t_1} = 500$, $n_{t_2} = 500$, $n_{t_3} = 500$

	$P = 3$					
	$f =$ Normal			$f = t_7$		
B	*VM*	*CRM*	*IPW*	*VM*	*CRM*	*IPW*
0.00	0.06	0.06	0.01	0.05	0.06	0.01
0.25	0.06	0.08	0.03	0.06	0.08	0.04
0.50	0.07	0.13	0.07	0.07	0.13	0.11
0.75	0.10	0.20	0.11	0.09	0.19	0.30
1.00	0.15	0.25	0.17	0.14	0.26	0.58

TABLE 5
$\overline{Max2SB}$, *small/unequal sample sizes*: $n_{t_1} = 500$, $n_{t_2} = 1000$, $n_{t_3} = 2000$

	$P = 3$					
	$f =$ Normal			$f = t_7$		
B	*VM*	*CRM*	*IPW*	*VM*	*CRM*	*IPW*
0.00	0.04	0.05	0.01	0.05	0.05	0.01
0.25	0.05	0.05	0.04	0.05	0.05	0.04
0.50	0.05	0.08	0.08	0.06	0.07	0.11
0.75	0.07	0.13	0.12	0.07	0.13	0.29
1.00	0.10	0.20	0.17	0.10	0.19	0.60

number of subjects receiving t_1 (n_{t_1}), and the ratio of units receiving t_2 to those receiving t_1 (γ).

In settings with low bias and normally distributed covariates, all three matching approaches appear to properly balance covariates. The average $\overline{Max2SB}$ using *IPW* is less than 0.05 across each simulation configuration with $B = 0$ and $f =$ Normal. As B increases, $\overline{Max2SB}$ for *CRM* rises faster than for *VM*. *IPW* bias also rises with higher B, but in most settings with normally distributed covariates, *IPW* yields $\overline{Max2SB}$ less than 0.25, but higher than *VM*.

VM and *CRM* produce better matched groups than *IPW* with heavy tailed covariates. When the covariates are distributed as multivariate t_7, the maximum pairwise bias's using *IPW* vary substantially (e.g., Table 7). While γ is not a major determinant of $\overline{Max2SB}$ for *IPW*, *VM* and *CRM* perform better in settings with $\gamma = 2$ (Tables 5 and 7).

Table 8 shows the *%Matched* for different values of n_{t_1} and γ, averaging over P, f, τ, σ_2 and σ_3. For low bias and with a larger number of controls ($\gamma = 2$), *CRM* generally matches as many triplets as *VM*. With increasing B, however, the fraction of eligible units that were matched is much smaller for *CRM*. With $\gamma = 1$, $B = 1$ and $n_{t_1} = 1000$, for example, *CRM* matches only 36% of eligible subjects on average, compared to 93% of the subjects using *VM*.

To account for the smaller number of subjects matched using *VM*, which is a possible unfair advantage for *VM*, we also measured bias in the covariates' distributions for *IPW* using only the subjects that were utilized by *VM*. In more than 98% of configurations, the biases observed were larger than those using *IPW* with all units.

A reduced set of simulations using $Z = 5$ showed that both *VM* and *IPW* reduce the initial bias. In some scenarios, *VM* had larger reduction than *IPW*, and in some scenarios the opposite (see the Appendix for additional details).

4. CONCLUSION

Many real world problems involve making a decision among three or more possible interventions. Simultaneous assessment of all of these interventions is

TABLE 6
$\overline{Max2SB}$, *large/equal sample sizes*: $n_{t_1} = 1000$, $n_{t_2} = 1000$, $n_{t_3} = 1000$

	$P = 3$						$P = 6$					
	$f =$ Normal			$f = t_7$			$f =$ Normal			$f = t_7$		
B	*VM*	*CRM*	*IPW*	*VM*	*CRM*	*IPW*	*VM*	*CRM*	*IPW*	*VM*	*CRM*	*IPW*
0.00	0.04	0.03	0.01	0.04	0.04	0.01	0.05	0.04	0.01	0.05	0.04	0.01
0.25	0.04	0.07	0.03	0.04	0.07	0.04	0.05	0.08	0.04	0.05	0.07	0.05
0.50	0.05	0.14	0.07	0.05	0.13	0.11	0.09	0.14	0.08	0.08	0.14	0.21
0.75	0.07	0.20	0.11	0.07	0.20	0.35	0.13	0.19	0.14	0.14	0.19	0.66
1.00	0.10	0.26	0.16	0.10	0.25	0.75	0.23	0.24	0.23	0.24	0.26	1.06

TABLE 7
$\overline{Max2SB}$, *large/equal sample sizes*: $n_{t_1} = 1000$, $n_{t_2} = 2000$, $n_{t_3} = 4000$

	$P = 3$						$P = 6$					
	$f =$ Normal			$f = t_7$			$f =$ Normal			$f = t_7$		
B	***VM***	***CRM***	***IPW***	***VM***	***CRM***	***IPW***	***VM***	***CRM***	***IPW***	***VM***	***CRM***	***IPW***
0.00	0.03	0.03	0.01	0.03	0.03	0.01	0.04	0.04	0.01	0.04	0.04	0.01
0.25	0.03	0.04	0.04	0.03	0.03	0.04	0.04	0.04	0.04	0.05	0.04	0.05
0.50	0.04	0.08	0.08	0.04	0.07	0.12	0.06	0.09	0.09	0.06	0.08	0.20
0.75	0.05	0.14	0.13	0.05	0.12	0.33	0.11	0.16	0.14	0.09	0.15	0.67
1.00	0.08	0.21	0.17	0.07	0.19	0.80	0.17	0.21	0.22	0.16	0.21	1.28

attractive, because it allows for the identification of the best intervention without the need to perform many studies in which each pair of interventions is compared. However, even in a randomized controlled environment, multi-arm trials can be considerably more complex to design, conduct and analyze than two-arm, single-question trials (Vermorken et al., 2005). These complications include sample size requirements, eligibility of all participants for all of the interventions, the comparisons that will be made, as well as the summaries of those comparisons. These problems are exacerbated in nonrandomized settings. While estimating causal effects for binary treatment in randomized and nonrandomized settings has been discussed extensively in the literature, we highlighted how the specification of causal effects for multiple treatments may be complex due to the choice of estimands and the different subsets of the population which investigators are interested in. Different estimands may yield different conclusions with respect to treatment effectiveness, and we advocate that researchers consider carefully the causal effect, or sets of causal effects, of primary interest, as in Dore et al. (2013).

4.1 Discussion of Vector Matching

We demonstrated that matching on a vector can address some of the drawbacks of currently available methods for estimating treatment effects with a nominal treatment assignment. *VM* attempts to replicate a randomized multi-arm trial by generating sets of subjects that are roughly equivalent on measured covariates. Simulations demonstrated that, relative to other available methods, *VM* generally yielded the lowest bias in the covariates' distributions between the different treatment groups, while retaining most of the eligible subjects that received the reference treatment. Under regular assignment mechanism, differences in *VM* matched units' outcomes could be contrasted, providing treatment effects that can be generalized to the population of subjects receiving t_1.

VM is a starting point for algorithms that intend to estimate transitive treatment effects and reduce bias when comparing multiple treatments. It is worth explicating on a few of the algorithm's strengths and weaknesses. *VM* uses with replacement matching because it has been shown to yield lower bias in comparison to matching without replacement with binary treatment

TABLE 8
%Matched: *The percent of eligible subjects receiving* t_1 *who were matched*

	$n_{t_1} = 500$				$n_{t_1} = 1000$			
	$\gamma = 1$		$\gamma = 2$		$\gamma = 1$		$\gamma = 2$	
B	***VM***	***CRM***	***VM***	***CRM***	***VM***	***CRM***	***VM***	***CRM***
0.00	0.99	0.91	0.99	0.99	0.99	0.92	0.99	0.99
0.25	0.97	0.81	0.99	0.97	0.98	0.79	0.99	0.96
0.50	0.95	0.67	0.98	0.87	0.98	0.63	0.99	0.79
0.75	0.94	0.52	0.97	0.72	0.97	0.47	0.98	0.61
1.00	0.91	0.42	0.95	0.55	0.93	0.36	0.96	0.47

(Abadie and Imbens, 2006). Additionally, matching with replacement allows estimation of *PATT*'s which are generalizable to each treatment group, and not just the group with the smallest sample size. One difficulty of matching with replacement is that subjects can be matched multiple times. As a result, although no adjustments are necessary for point estimates, an analysis phase will require adjustments for estimating sampling variances (Abadie and Imbens, 2006).

VM can be used to estimate any causal estimand of interest and is not restricted to differences in averages. While we concentrated on *PATT*s in this manuscript, *VM* can be extended for *PATE*s by forming a matched set for each eligible unit, as opposed to just a set for each unit receiving the reference treatment. If all eligible subjects can be matched to subjects receiving other treatments, contrasts between the matched cohorts would generalize to the population as a whole. As noted in Abadie and Imbens (2006), pair matching for *PATE*s can only be done with replacement, as differences in the sample sizes at each treatments will require some subjects to be matched more often than others. In this respect, *VM* would be preferred to *CRM*, *SBC* and *WithinTrio*, which are limited to only estimating *PATT*s.

While *KMC* is one approach for grouping similar subjects, by restricting the matching to be within the clusters, some possible matches may not be considered by *VM* because they are on the boundaries of the clusters. This could lead to nonoptimal matches, or even to the exclusion of some reference units that will not have a match in the other treatment groups. One plausible extension of *VM* would be to use fuzzy clustering (Bezdek, Ehrlich and Full, 1984), which would allow for units to belong to multiple clusters.

Another downside of *KMC* is the possibility of obtaining clusters where there are no units receiving a certain treatment. However, clustering on $Z - 2$ components of the GPS, as in vector matching, is preferred to clustering on all Z components, as would be done in using *KMC* alone. For large Z, if clustering on $Z - 2$ components yields clusters without at least one unit from each treatment group, one possibility is to re-fit *KMC*, given that *K-means* often returns different partitions.

Finally, *VM* is based on a greedy matching algorithm, which may not be the most optimal procedure to partition the GPS. Among other alternatives to matching on the GPS, coarsened exact matching could be used to pair subjects within each of the *K-means* subclasses (Iacus, King and Porro, 2011). With binary treatment, algorithms like full matching (Rosenbaum, 1991) and mixed integer programming (Zubizarreta, 2012) were proposed to optimally match units such that the difference in the covariates' distributions between the two treatment groups is minimized while retaining most of the units. In contrast to binary treatment matching, optimally matching for multiple treatments, also known as k-dimensional matching, was shown to be a NP-hard problem (Karp, 1972). Further research is required to apply these methods to multiple treatments.

As with other procedures for estimating causal effects with multiple treatments, methods for estimating the sampling variances of estimands when using *VM* are not well established. Variance weighting and resampling are two procedures that have been suggested for estimating the sampling variance of causal estimands with binary treatments, and we proposed that similar procedures could be used with multiple treatments. However, further research is required to identify the operating characteristics of each of these procedures.

One set of strategies that we did not explore is covariate adjustment for the GPS or a function of the GPS using a regression model (Filardo et al., 2007, 2009), Dearing, McCartney and Taylor, 2009, Spreeuwenberg et al., 2010). Such techniques are subject to possible model misspecification and extrapolation problems, as shown in standard regression adjustment for binary treatment (Dehejia and Wahba, 1998, 2002), and simulations have found that these strategies can perform worse than matching, stratification, or weighting with multiple treatments (Hade and Lu, 2014).

4.2 Recommendations

Causal modeling is challenging because it requires estimation of quantities that cannot be measured simultaneously. This problem is exacerbated when comparing multiple treatments, because the proportion of these quantities increases. Methods for multiple treatments continue to evolve, and more work is still needed in several areas, particularly with respect to the estimation of the sampling variance. Below, we provide a list of recommendations for researchers who are looking to estimate causal effects with multiple treatments (see Table 9 for a summary of the acronyms):

1. Comparing multiple treatments in observational studies is similar to comparing multiple interventions in a multi-arm trial. Thus, it is important to ascertain that the data is composed of enough units that are 'eligible' to receive all of the treatments, and units that are not eligible should be removed when attempting to identify the best treatment.

TABLE 9
Summary of acronyms

Acronym	Description
CRM	Common referent matching
GPS	Generalized propensity score
IPW	Inverse probability weighting
KMC	K-means clustering
PATE	Population average treatment effects
PATT	Population average treatment effects among the treated
RCM	Rubin causal model
SATE	Sample average treatment effects
SATT	Sample average treatment effects among the treated
SBC	Series of binary comparisons
SUTVA	Stable unit treatment value assumption
VM	Vector matching

2. Causal estimands of interest and the populations to which these estimands generalize require careful consideration. These decisions become more complex with increased number of treatments.

3. For ordinal treatment assignment such as scales or doses, the linear predictor from a proportional odds model of treatment assignment acts as a scalar balancing score on which to balance the covariates' distributions. Nonbipartite matching (Lu et al., 2001), subclassification (Imai and van Dyk, 2004, Zanutto, Lu and Hornik, 2005), and the combination of subclassification with regression adjustment (Lopez and Gutman, 2014) stand out as approaches for making inferences.

4. For nominal treatment assignment, methods that rely on binary propensity scores that are estimated only on units receiving one of two treatments (such as *SBC* and *CRM*) may result in significant bias in the covariates' distributions between units receiving the different treatments. These may lead to biased and nontransitive estimates and, therefore, should not be applied generally.

5. For nominal or ordinal treatment assignment, a simple implementation of *K-means* clustering (*KMC*) may result in clusters that do not include units receiving all treatments, which results in increased bias. Our simulations show that in comparison to other matching and weighting procedures, it suffers from the smallest bias reduction.

6. For nominal or ordinal treatment assignment, matching on the GPS using vector matching (*VM*) or using inverse probability weighting (*IPW*) are promising approaches.

- *IPW* reduces the bias significantly; however, as our simulations show, it may suffer from extreme weights that yield erratic causal estimates. This problem is exacerbated with increasing number of treatments or covariates that are not normally distributed. Simple trimming of units with GPS components that are close to 0 or 1 may result in increased bias, because units that are similar on a single GPS component may differ on others. Other approaches for estimating the GPS, such as generalized boosted models, may solve this issue. However, more research is needed to derive sampling variance estimates for these procedures and to examine their behavior in a wide range of applications. Lastly, *IPW* estimates are mainly suitable for estimating differences in averages, and are not well suited for comparison of other estimands.
- *VM* uses an in-strata matching algorithm to identify matched sets of subjects in order to estimate treatment effects generalizable to the population of units eligible for each treatment. Across a set of simulation configurations, *VM* tended to yield the largest improvement in balance in the covariates' distributions between units receiving different treatments. Under certain assumptions, this would allow for unbiased comparisons of the effects of multiple interventions. Additional research is needed to identify sampling variance formulas for estimates from the matched cohorts, as well as to explore alternative mechanisms for matching on the GPS. To sum, *VM* is one approach that seems to compare favorably to commonly available methods, but more research is needed to explore other alternatives.

APPENDIX

We implement *VM* and *IPW* for $Z = 5$, where $\boldsymbol{X}$ is generated for $N = n_{t_1} + n_{t_2} + n_{t_3} + n_{t_4} + n_{t_5}$ subjects receiving one of five treatments, $\mathcal{T} \in \{t_1, t_2, t_3, t_4, t_5\}$, with n_t the sample size of subjects receiving treatment t. Let $\mathbb{1}$ be the 5×5 identify matrix. The values of $\boldsymbol{X}$ were generated from multivariate symmetric distributions such that

$$
\begin{aligned}
&T_i = t_1, \quad i = 1, \ldots, n_{t_1},\\
&T_i = t_2, \quad i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1},\\
&T_i = t_3, \quad i = n_{t_1} + \gamma n_{t_1} + 1, \ldots,\\
&\qquad\qquad n_{t_1} + 2 \times \gamma n_{t_1},\\
&T_i = t_4, \quad i = n_{t_1} + 2 \times \gamma n_{t_1} + 1, \ldots,
\end{aligned}
\tag{28}
$$

$$
\begin{aligned}
&n_{t_1} + 2 \times \gamma n_{t_1} + \gamma^2 \times n_{t_1},\\
T_i = t_5, \quad &i = n_{t_1} + 2 \times \gamma n_{t_1} + \gamma^2 \times n_{t_1} + 1, \ldots,\\
&n_{t_1} + 2 \times \gamma n_{t_1} + 2 \times \gamma^2 \times n_{t_1},
\end{aligned}
$$

$$
\begin{aligned}
&\boldsymbol{X}_i | \{T_i = t_1\} \sim f(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n_{t_1},\\
&\boldsymbol{X}_i | \{T_i = t_2\} \sim f(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}),\\
&\quad i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1},\\
&\boldsymbol{X}_i | \{T_i = t_3\} \sim f(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}),\\
&\quad i = n_{t_1} + \gamma n_{t_1} + 1, \ldots, n_{t_1} + 2 \times \gamma n_{t_1},\\
&\boldsymbol{X}_i | \{T_i = t_4\} \sim f(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}),\\
&\quad i = n_{t_1} + 2 \times \gamma n_{t_1} + 1, \ldots,\\
&\qquad n_{t_1} + 2 \times \gamma n_{t_1} + \gamma^2 \times n_{t_1},\\
&\boldsymbol{X}_i | \{T_i = t_5\} \sim f(\boldsymbol{\mu}_5, \boldsymbol{\Sigma}),\\
&\quad i = n_{t_1} + 2 \times \gamma n_{t_1} + \gamma^2 \times n_{t_1} + 1, \ldots,\\
&\qquad n_{t_1} + 2 \times \gamma n_{t_1} + 2 \times \gamma^2 \times n_{t_1},
\end{aligned} \tag{29}
$$

$$
\begin{aligned}
&\mu_1 = (b, 0, 0, 0, 0)^T, \quad \mu_2 = (0, b, 0, 0, 0)^T,\\
&\mu_3 = (0, 0, b, 0, 0)^T, \quad \mu_4 = (0, 0, 0, b, 0)^T,\\
&\mu_5 = (0, 0, 0, 0, b)^T,
\end{aligned} \tag{30}
$$

$$
\Sigma = \mathbb{1}. \tag{31}
$$

The following design implicitly assumes a regular assignment mechanism that depends on four factors (Table 10).

For each simulation condition, 200 data sets are generated, and on each data set, *VM* (using $K = 5$ strata) and *IPW* are used to identify matched and weighted sets. *CRM* is not considered do to the small number of matches generated.

In all 20 configurations, both *VM* and *IPW* reduced the average $\overline{Max2SB}$ relative to the pre-matched cohort.

TABLE 10
Simulation factors

Factor	Levels of factor
n_{t_1}	{1000}
$\gamma = \frac{n_{t_2}}{n_{t_1}} = \frac{n_{t_3}}{n_{t_1}} = \frac{n_{t_4}}{n_{t_2}} = \frac{n_{t_5}}{n_{t_2}}$	{1, 2}
f	$\{t_7, Normal\}$
b	{0, 0.25, 0.50, 0.75, 1.00}

In cases with large initial bias and with covariates from t distribution, *VM* performed better than *IPW*, but with smaller initial bias, *IPW* performed better. On average, *VM* matched at least 93% of eligible subjects in each configuration. Investigating the performance of different matching methods with five or more treatments is an area of further research.

ACKNOWLEDGMENTS

M. J. Lopez was supported by the National Institute for Health (IMSD Grant #R25GM083270) and the National Institute on Aging (NIA Grant #F31AG046056). R. Gutman was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Award ME-1403-12104. Disclaimer: All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325

ABADIE, A. and IMBENS, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76** 1537–1557.

ARMSTRONG, C. S., JAGOLINZER, A. D. and LARCKER, D. F. (2010). Chief executive officer equity incentives and accounting irregularities. *J. Acc. Res.* **48** 225–271.

AUSTIN, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28** 3083–3107. MR2750408

AUSTIN, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm. Stat.* **10** 150–161.

AUSTIN, P. C., GROOTENDORST, P. and ANDERSON, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Stat. Med.* **26** 734–753. MR2339171

AUSTIN, P. C. and SMALL, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Stat. Med.* **33** 4306–4319.

BEZDEK, J. C., EHRLICH, R. and FULL, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **10** 191–203.

BRYSON, A., DORSETT, R. and PURDON, S. (2002). The use of propensity score matching in the evaluation of active labour market policies.

CALIENDO, M. and KOPEINIG, S. (2008). Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* **22** 31–72.

CANGUL, M. Z., CHRETIEN, Y. R., GUTMAN, R. and RUBIN, D. B. (2009). Testing treatment effects in unconfounded studies under model misspecification: Logistic regression, discretization, and their combination. *Stat. Med.* **28** 2531–2551. MR2750307

CHERTOW, G. M., NORMAND, S. L. T. and MCNEIL, B. J. (2004). "Renalism": Inappropriately low rates of coronary angiography in elderly individuals with renal insufficiency. *J. Am. Soc. Nephrol.* **15** 2462–2468.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144

D'AGOSTINO, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **17** 2265–2281.

DAVIDSON, M. B., HIX, J. K., VIDT, D. G. and BROTMAN, D. J. (2006). Association of impaired diurnal blood pressure variation with a subsequent decline in glomerular filtration rate. *Arch. Intern. Med.* **166** 846–852.

DEARING, E., MCCARTNEY, K. and TAYLOR, B. A. (2009). Does higher quality early child care promote low-income children's math and reading achievement in middle childhood? *Child Dev.* **80** 1329–1349.

DEHEJIA, R. H. and WAHBA, S. (1998). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. Technical report, National Bureau of Economic Research.

DEHEJIA, R. H. and WAHBA, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* **84** 151–161.

DORE, D. D., SWAMINATHAN, S., GUTMAN, R., TRIVEDI, A. N. and MOR, V. (2013). Different analyses estimate different parameters of the effect of erythropoietin stimulating agents on survival in end stage renal disease: A comparison of payment policy analysis, instrumental variables, and multiple imputation of potential outcomes. *J. Clin. Epidemiol.* **66** S42–S50.

DORSETT, R. (2006). The new deal for young people: Effect on the labour market status of young men. *Labour Econ.* **13** 405–422.

DRICHOUTIS, A. C., LAZARIDIS, P. and NAYGA JR., R. M. (2005). Nutrition knowledge and consumer use of nutritional food labels. *Eur. Rev. Agricult. Econ.* **32** 93–118.

EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press, Boca Raton.

FENG, P., ZHOU, X.-H., ZOU, Q.-M., FAN, M.-Y. and LI, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat. Med.* **31** 681–697.

FILARDO, G., HAMILTON, C., HAMMAN, B. and GRAYBURN, P. (2007). Obesity and stroke after cardiac surgery: The impact of grouping body mass index. *Ann. Thorac. Surg.* **84** 720–722.

FILARDO, G., HAMILTON, C., HAMMAN, B., HEBELER JR., R. F. and GRAYBURN, P. A. (2009). Relation of obesity to atrial fibrillation after isolated coronary artery bypass grafting. *Am. J. Cardiol.* **103** 663–666.

FRANK, R., AKRESH, I. R. and LU, B. (2010). Latino immigrants and the US racial order. *Am. Sociol. Rev.* **75** 378–401.

GUTMAN, R. and RUBIN, D. B. (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Stat. Med.* **32** 1795–1814. MR3067363

GUTMAN, R. and RUBIN, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Stat. Med.* **34** 3381–3398. MR3412639

HADE, E. M. (2012). Propensity score adjustment in multiple group observational studies: Comparing matching and alternative methods. Ph.D. thesis, Ohio State University.

HADE, E. M. and LU, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. *Stat. Med.* **33** 74–87. MR3141554

HEDMAN, L. and VAN HAM, M. (2012). *Understanding Neighbourhood Effects*: *Selection Bias and Residential Mobility*. Springer, Berlin.

HILL, J. and REITER, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Stat. Med.* **25** 2230–2256. MR2240098

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HOTT, J. R., BRUNELLE, N. and MYERS, J. A. (2012). KD-tree algorithm for propensity score matching with three or more treatment groups. Division of Pharmacoepidemiology and Pharmacoeconomics, Technical Report Series.

IACUS, S. M., KING, G. and PORRO, G. (2011). Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* mpr013.

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263.

IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics*, *Social*, *and Biomedical Sciences*. Cambridge Univ. Press, Cambridge.

JOFFE, M. M. and ROSENBAUM, P. R. (1999). Invited commentary: Propensity scores. *Am. J. Epidemiol.* **150** 327–333.

JOHNSON, R. A., WICHERN, D. W. et al. (1992). *Applied Multivariate Statistical Analysis* **4**. Prentice Hall, Englewood Cliffs, NJ.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539.

KARP, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations* (*Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y.*, 1972) 85–103. Plenum, New York. MR0378476

KILPATRICK, R. D., GILBERTSON, D., BROOKHART, M. A., POLLEY, E., ROTHMAN, K. J. and BRADBURY, B. D. (2013). Exploring large weight deletion and the ability to balance confounders when using inverse probability of treatment weighting in the presence of rare treatment decisions. *Pharmacoepidemiol. Drug Saf.* **22** 111–121.

KOSTEAS, V. D. (2010). The effect of exercise on earnings: Evidence from the NLSY. *J. Labor Res.* 1–26.

LECHNER, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *Econom. Evaluation Labour Mark. Polic.* 43–58.

LECHNER, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Rev. Econ. Stat.* **84** 205–220.

LEE, B. K., LESSLER, J. and STUART, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE* **6** e18174.

LEVIN, I. and ALVAREZ, R. M. (2009). Measuring the effects of voter confidence on political participation: An application to the 2006 Mexican election. VTP Working Paper 75, Caltech/MIT Voting Technology Project.

LITTLE, R. J. A. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* 287–296.

LOPEZ, M. J. and GUTMAN, R. (2014). Estimating the average treatment effects of nutritional label use using subclassification with regression adjustment. *Stat. Methods Med. Res.* DOI:10.1177/0962280214560046.

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. *J. Amer. Statist. Assoc.* **96** 1245–1253.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.

MCCAFFREY, D. F., GRIFFIN, B. A., ALMIRALL, D., SLAUGHTER, M. E., RAMCHAND, R. and BURGETTE, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **32** 3388–3414. MR3074364

MCCULLAGH, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **42** 109–142. MR0583347

MOORE, A. W. (1991). An introductory tutorial on kd-trees. Extract from PhD thesis. Technical report.

QUADE, D. (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *J. Amer. Statist. Assoc.* **74** 680–683.

R CORE TEAM (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RASSEN, J. A., SOLOMON, D. H., GLYNN, R. J. and SCHNEEWEISS, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: A pragmatic "matrix design." *Pharmacoepidemiol. Drug Saf.* **20** 675–683.

RASSEN, J. A., SHELAT, A. A., FRANKLIN, J. M., GLYNN, R. J., SOLOMON, D. H. 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.

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. R. Stat. Soc., B* **53** 597–610.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

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

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

ROYSTON, P., ALTMAN, D. G. and SAUERBREI, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat. Med.* **25** 127–141. MR2222078

RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29** 159–183.

RUBIN, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the Social Statistics Section of the American Statistical Association* 233–239.

RUBIN, D. B. (1976). Multivariate matching methods that are equal percent bias reducing. II. Maximums on bias reduction for fixed sample sizes. *Biometrics* **32** 121–132. MR0400556

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (1980). Discussion of Basu's paper. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–188.

RUBIN, D. B. and THOMAS, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *Ann. Statist.* **20** 1079–1093.

RUBIN, D. B. and THOMAS, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* **79** 797–809. MR1209479

RUBIN, D. B. and THOMAS, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* **52** 249–264.

SAKIA, R. M. (1992). The Box–Cox transformation technique: A review. *Statistician* **42** 169–178.

SAS INSTITUTE INC. (2003). *SAS/STAT Software*. SAS Institute Inc., Cary, NC.

SCHNEEWEISS, S., SETOGUCHI, S., BROOKHART, A., DORMUTH, C. and WANG, P. S. (2007). Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. *CMAJ, Can. Med. Assoc. J.* **176** 627–632.

SEKHON, J. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *J. Stat. Softw.* **42**, 1–52.

SNODGRASS, G., BLOKLAND, A. A. J., HAVILAND, A., NIEUWBEERTA, P. and NAGIN, D. S. (2011). Does the time cause the crime? An examination of the relationship between time served and reoffending in the Netherlands. *Criminology* **49** 1149–1194.

SPLAWA-NEYMAN, J., DABROWSKA, D. M. and SPEED, T. P. (1990 [1923]). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472.

SPREEUWENBERG, M. D., BARTAK, A., CROON, M. A., HAGENAARS, J. A., BUSSCHBACH, J. J. V., ANDREA, H., TWISK, J. and STIJNEN, T. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: An introduction from a case study in mental health. *Med. Care* **48** 166.

SPRENT, P. and SMEETON, N. C. (2007). *Applied Nonparametric Statistical Methods*. CRC Press, Boca Raton, FL.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

STUART, E. A. and RUBIN, D. B. (2008). Best practices in quasi-experimental designs. *Best Pract. Quant. Methods* 155–176.

TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682. MR2672490

TCHERNIS, R., HORVITZ-LENNON, M. and NORMAND, S. L. T. (2005). On the use of discrete choice models for causal inference. *Stat. Med.* **24** 2197–2212.

TU, C., JIAO, S. and KOH, W. Y. (2012). Comparison of clustering algorithms on generalized propensity score in observational studies: A simulation study. *J. Stat. Comput. Simul.* **83** 2206–2218.

VERMORKEN, J. B., PARMAR, M. K., BRADY, M. F., EISENHAUER, E. A., HOGBERG, T., OZOLS, R. F., ROCHON, J.,RUSTIN, G. J., SAGAE, S., VERHEIJEN, R. H. et al. (2005). Clinical trials in ovarian carcinoma: Study methodology. *Ann. Oncol.* **16** viii20.

YANOVITZKY, I., ZANUTTO, E. and HORNIK, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Eval. Program Plann.* **28** 209–220.

ZANUTTO, E., LU, B. and HORNIK, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *J. Educ. Behav. Stat.* **30** 59–73.

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400