Estimation of causal effects with multiple treatments: a review and new ideas

Michael J. Lopez*, and Roe Gutman

Skidmore College (e-mail: mlopez1@skidmore.edu)
Brown University (e-mail: roee.gutman@brown.edu)

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. Further, we propose a matching technique for use with multiple, nominal categorical treatments, and use simulations to show how this algorithm yields improved covariate similarity between those in the matched sets compared to both the pre-matched cohort and to other current matching strategies. 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, \( X \), 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), subclassification (Rosenbaum and Rubin, 1984), weighting (Robins, Hernan and Brumback, 2000), and imputations (Gutman and Rubin, 2015) have been proposed to adjust for the differences in $X$ across the exposure groups. These approaches attempt to obtain covariate balance across treatment groups, where balance refers to equality in the distributions of $X$. By ensuring that the distribution of units receiving different treatments are similar on average, these methods reduce the effects of treatment selection bias on causal estimates.

When $X$ is a scalar, it is relatively straight-forward to perform matching (Rubin, 1976). However, it is more complex to match, subclassify, or weight when $X$ is composed of many covariates. With 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 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 pain-killers 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.

Further, 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 third goal is to provide a technique for generating matched sets when there are more than two treatments, and to compare the performance of the new and previously proposed algorithms in balancing covariate 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.
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 \( Y_i, X_i, \) and \( T_i \) be the observed outcome, set of covariates, and binary treatment assignment, respectively, for each subject \( i = 1, \ldots, N, \) with \( N \leq N, \) where \( N \) is the population size which is possibly infinite. With \( T_i \in T, \) let \( T \) be the treatment space. For a binary treatment, \( T = \{t_1, t_2\}, \) 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 is independent of 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 super-population effects are the population average treatment effect, \( PATE_{t_1,t_2}, \) and the population average treatment effect among those receiving \( t_1, \) \( PATT_{t_1,t_2}. \)

\[
\begin{align*}
PATE_{t_1,t_2} &= E[Y_i(t_1) - Y_i(t_2)] \\
PATT_{t_1,t_2} &= E[Y_i(t_1) - Y_i(t_2)|T_i = t_1]
\end{align*}
\]

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.

\[
\begin{align*}
SATE_{t_1,t_2} &= \frac{1}{N} \sum_{i=1}^{N} (Y_i(t_1) - Y_i(t_2)) \\
SATT_{t_1,t_2} &= \frac{1}{n_{t_1}} \sum_{i=1}^{N} (Y_i(t_1) - Y_i(t_2)) \times I(T_i = t_1)
\end{align*}
\]

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, \( P(T|Y(t_1), Y(t_2), X) \) (Imbens and Rubin, 2015). Three commonly made restrictions of the assignment mechanism are individualistic, probabilistic, and unconfounded (Imbens and Rubin, 2015). Let \( q \) be a function, \( q : (X_i, Y_i(t_1), Y_i(t_2)) \to [0, 1]. \) An assignment mechanism is individualistic for binary treatment if

\[
P(T|X, Y(t_1), Y(t_2)) = c \prod_{i=1}^{N} q(X_i, Y_i(t_1), Y_i(t_2))^{T_i} (1 - q(X_i, Y_i(t_1), Y_i(t_2)))^{1-T_i}
\]

for \( (T, X, Y(t_1), Y(t_2)) \in \mathcal{A}, \) for some set \( \mathcal{A}, \) and zero elsewhere, and where \( c \) is a normalizing constant that ensures the probabilities sum to 1.
A probabilistic assignment mechanism entails

\[ 0 < \sum_{T_i = t_1} P(T_i | X, Y(t_1), X(t_2)) < 1 \]

\[ \forall i \in N \text{ and each possible } X, Y(t_1), X(t_2) \]. Finally, a treatment assignment is unconfounded if \( P(T | Y(t_1), Y(t_2), X) = P(T | X) \). Under individualistic assignment, the combination of 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 is most commonly identified with observational data. Weaker versions of unconfoundedness are sufficient for some estimation techniques (Imbens, 2000), and are discussed in Section 1.5.5.

Let \( e_{t_1,t_2}(X) = P(T = t_1 | X) \) be the propensity score (PS), and let \( \hat{e}_{t_1,t_2}(X) \) 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 PS’s (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 \( X \) in the treatment groups \( t_1 \) and \( t_2 \). Figure 1 shows different sets of covariates’ distributions overlap between those receiving \( t_1 \) and \( t_2 \). Each circle in Figure 1 represents a hypothetical distribution of \( X \) among those on 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 covariates distribution of 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 $X$ or $\hat{e}_{t_1,t_2}(X)$ beyond the range of $X$ or $\hat{e}_{t_1,t_2}(X)$ 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 $\hat{e}_{t_1,t_2}(X)$ 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 $E_{i1}$ be an indicator for subject $i$ having a propensity score within the common support of $\hat{e}_{t_1,t_2}(X)$. Defensible estimands of interest are now

$$PATE_{E_{i1},t_1,t_2} = E[Y_i(t_1) - Y_i(t_2)|E_{i1} = 1]$$  
(5)
$$PATT_{E_{i1},t_1,t_2} = E[Y_i(t_1) - Y_i(t_2)|E_{i1} = 1, T_i = t_1].$$  
(6)

### 1.2 Notation for multiple treatments

The choice of estimands grows with increasing treatment options. Let $T = \{t_1, t_2, ..., t_Z\}$ be the treatment support for $Z$ total treatments, with $Y_i = \{Y_i(t_1), Y_i(t_2), ..., Y_i(t_Z)\}$ 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. Let $q_t, t \in \{t_1, t_2, ..., t_Z\}$, be a set of functions $q_t : (X_i, Y_i(t_1), Y_i(t_2)) \to [0, 1]$, such that $\sum_{t \in \{t_1, t_2, ..., t_Z\}} q_t = 1$. An assignment mechanism is individualistic in the multi-treatment setting if

$$P(T|X, Y(t_1), \ldots, Y(t_Z)) = c \prod_{i=1}^{N} \prod_{t \in \{t_1, t_2, ..., t_Z\}} q_t(X_i, Y_i(t_1), \ldots, Y_i(t_Z))^{T_i = t}$$

for $(T, X, Y(t_1), \ldots, Y(t_Z)) \in A$, for some set $A$, and zero elsewhere, and where $c$ is a normalizing constant that ensures the probabilities sum to 1.

If $\forall i \in N, t \in \{t_1, \ldots, t_Z\}$ and each possible $X, Y(t_1), \ldots, Y(t_Z)$, assignment mechanisms with multiple treatments are probabilistic if $0 < \sum_{T_i=t} P(T|X, Y(t_1), \ldots, Y(t_Z)) < 1$. With multiple treatments, an unconfounded assignment mechanism requires that $P(T|Y(t_1), \ldots, Y(t_1), X) = P(T|X)$. See Imbens (2000) and Imai and Van Dyk (2004) for expanded discussions of these assumptions.

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 $w_1, w_2 \subseteq T$ and $w_1 \cap w_2 = \emptyset$. 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

$$PATE_{w_1,w_2} = E \left[ \frac{\sum_{t \in w_1} Y_i(t)}{|w_1|} - \frac{\sum_{t \in w_2} Y_i(t)}{|w_2|} \right],$$  
(7)
$$PATT_{w_1|w_1,w_2} = E \left[ \frac{\sum_{t \in w_1} Y_i(t)}{|w_1|} - \frac{\sum_{t \in w_2} Y_i(t)}{|w_2|} | T_i \in w_1 \right].$$  
(8)
In (7) and (8), the expectation is over all units, \( i = 1, \ldots, N \), 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 \( \text{PATE}_{w_1, w_2} = E \left[ \frac{Y_{i(t_1)} + Y_{i(t_2)}}{2} - \frac{Y_{i(t_3)} + Y_{i(t_4)} + Y_{i(t_5)}}{3} \right] \).

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 \( w_3 = \{t_3\} \), \( \text{PATE}_{w_1, w_3} - \text{PATE}_{w_1, w_2} = \text{PATE}_{w_2, w_3} \).

For reference group \( w_1 = \{t_1\} \), there are traditionally \( Z - 1 \) possible pairwise PATT’s, one for each of the treatments which the reference group did not receive (McCaffrey et al., 2013). The PATT’s are also transitive, such that \( \text{PATT}_{w_1|w_1, w_3} - \text{PATT}_{w_1|w_1, w_2} = \text{PATT}_{w_1|w_2, w_3} \). 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, \( \text{PATT}_{w_1|w_1, w_2} - \text{PATT}_{w_2|w_2, w_3} \) is generally not equal to \( \text{PATT}_{w_1|w_1, w_3} \).

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

1.3 The generalized propensity score

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

With a binary treatment, knowing \( e_{t_1, t_2}(X) \) is equivalent to knowing \( 1 - e_{t_1, t_2}(X) \). 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 GPS’s, defined as \( R(X) = (r(t_1, X), \ldots, r(t, X)) \), or a function of \( R(X) \) (Imai and Van Dyk, 2004).

Two individuals with the same \( r(t, X) \) for treatment \( t \) may have differing \( R(X) \)’s. For example, for \( T = \{t_1, t_2, t_3\} \), let \( R(X_i), R(X_j), \) and \( R(X_k) \) 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

\[
R(X_i) = (0.30, 0.60, 0.10),
\]
\[
R(X_j) = (0.30, 0.35, 0.35),
\]
\[
R(X_k) = (0.30, 0.10, 0.60).
\]

Even though \( r(t_1, X_i) = r(t_1, X_j) = r(t_1, X_k) = 0.3 \), because \( r(t_2, X_i) \neq r(t_2, X_j) \neq r(t_2, X_k) \) and \( r(t_3, X_i) \neq \ldots \).
with multiple treatments. In part due to this limitation, Imbens (2000) called individual matching less ‘well-suited’ to multiple treatment settings. Only under the scenario of $R(X_i) = R(X_j) = R(X_k)$ 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).

### 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 conditioning on a vector. This can be done by estimating the assignment mechanism as a function of $X$ using the proportional odds model (McCullagh, 1980), such that

$$
\log \left( \frac{P(T_i < t)}{P(T_i \geq t)} \right) = \theta_t - \beta^T X_i, t = 1, ..., Z - 1.
$$

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

The balancing property of $\beta^T X$ can be used to match or subclassify subjects receiving different levels of an ordinal exposure. Lu et al. (2001) used non-bipartite matching to form matched sets based on a function of $\beta^T X$ and the relative distance between exposure levels. While this method does not specify an exact causal estimand, it was used for generating an overall sense 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 $\beta^T X$ values. After subclassification on $\beta^T X$, the distribution of $X$ 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) recommended to combine regression adjustment with subclassification to obtain 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 SUTVA assumption of a constant treatment assignment. 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 dichotimization. Moreover, dichotomization makes identification of an optimal exposure level impossible. Thus, matching or subclassifications methods which maintain all
exposure levels while balancing on $\beta^T X$ are preferred for causal inference with ordinal exposures (Imai and Van Dyk, 2004).

Finally, inverse probability weighting can be used to estimate causal effects from ordinal treatments; a detailed description is provided in Section 1.5.5.

### 1.5 Nominal treatments

Nominal treatments do not follow a specific order. Thus, it is harder to identify a ‘sensible’ function that reduces $R(X)$ to a scalar. Several methods have been proposed to estimate causal effects with multiple treatments. 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 $PATTs$ between multiple treatments using a series of binary comparisons ($SBC$). $SBC$ implements binary propensity score methods within each of the $2^Z$ 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 $\hat{e}_{(t_1,t_2)}(X)$, estimated using logistic or probit regression, or $\hat{r}(t_1, X) / (\hat{r}(t_1, X) + \hat{r}(t_2, X))$, where $\hat{r}(t_1, X)$ and $\hat{r}(t_2, X)$ 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 $e_{(t_1,t_2)}(X, T = t_1)$ and $e_{(t_1,t_2)}(X, T = t_2)$ be the vector of all binary propensity scores among subjects receiving $t_1$ and $t_2$, respectively. We define $E_{2}(t_1, t_2)$ as the indicator for subject $i$ having a binary propensity score for treatments $t_1$ and $t_2$ within the common support:

$$E_{2i}(t_1, t_2) = \begin{cases} 
1 & \text{if } e_{(t_1,t_2)}(X_i) \in e_{(t_1,t_2)}(X, T = t_1) \cap e_{(t_1,t_2)}(X, T = t_2)
\end{cases}$$

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

$$PATT_{E_{2i}(t_1|t_1,t_2)} = E[Y_i(t_1) - Y_i(t_2)|T_i = t_1, E_{2i}(t_1, t_2) = 1].$$

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 \( e_{t_1,t_2}(X) \) and \( e_{t_1,t_3}(X) \), respectively; (2) Using 1:1 matching, pairs of units receiving \( t_1 \) or \( t_2 \) are matched using \( \hat{e}_{t_1,t_2}(X) \) and pairs of units receiving \( t_1 \) or \( t_3 \) are matched using \( \hat{e}_{t_1,t_3}(X) \); (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 was no match for the reference unit to a unit receiving \( t_3 \).

Let \( E_{3i} \) be the indicator for having two pairwise binary PS’s within their respective common supports, such that
\[
 E_{3i} = \begin{cases} 
 1 & \text{if } E_{2i}(t_1, t_2) = 1 \text{ and } E_{2i}(t_1, t_3) = 1 \\
 0 & \text{if otherwise}. 
\end{cases}
\]

CRM attempts to estimate the following treatment effects:
\[
\begin{align*}
PATT_{E_{3}(t_1|t_1, t_2)} & = E[Y_i(t_1) - Y_i(t_2)|T_i = t_1, E_{3i} = 1] \\
PATT_{E_{3}(t_1|t_1, t_3)} & = E[Y_i(t_1) - Y_i(t_3)|T_i = t_1, E_{3i} = 1] \\
PATT_{E_{3}(t_2|t_2, t_3)} & = E[Y_i(t_2) - Y_i(t_3)|T_i = t_1, E_{3i} = 1]
\end{align*}
\]

The next section will explain the possible issues that arise from such procedure.

1.5.3 Binary PS applications to multiple treatments  The following hypothetical example with \( Z = 3 \) illustrates issues with the implantation of binary PS tools, as in SBC and CRM, when there are multiple treatments.

Let \( X_i = (x_{1i}, x_{2i}) \) be a vector of covariates for subject \( i \), and we will assume that \( X_i|T_i = t \sim N(\mu_t, \mathbb{I}) \), where \( \mu_t \) is a 2 x 1 mean vector and \( \mathbb{I} \) is the 2 x 2 identity matrix. Without loss of generality we will 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, $Z$, and orthogonal to it, $W$ (Rubin and Thomas, 1992a). Matching on the true or estimated propensity score does not introduce any bias in $W$ when $X_i | T_i$ follows a multivariate normal distribution. In addition, after matching, $W$ will have the same expected second moment (Rubin and Thomas, 1992b). Specifically, when matching treatment 1 to treatment 2 with $a = 2$, $Z_{12} = (\frac{1}{2})' X_1 / \sqrt{2} = \sqrt{2} X_2$ and $W_{12} = X_1$. After matching, Rubin and Thomas (1992b) showed that

$$E(Z_{m_{12}}) = 2 - \Omega(N_{t_2}, n_{t_2}) \approx 2 - 2\pi \log \left( \frac{N_{t_2}}{n_{t_2}} \right)$$

$$E(Z_{m_{12}}) = 0 + \Omega(N_{t_1}, n_{t_1}) \approx 2 + 2\pi \log \left( \frac{N_{t_1}}{n_{t_1}} \right)$$

where $Z_{m_{12}}$ and $Z_{m_{21}}$ 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 $a = 2$, $\mu_m = E(Z_{m_{12}}) = E(Z_{m_{21}})$ 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 $W_{13} = X_2$.

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 $\mu_m$ and another with mean $\mu_m$. Note that $W_{12}$ and $W_{12}$ are independent and have similar means (Rubin and Thomas, 1992b). Similarly, $W_{13}$ and $W_{13}$ are independent and have similar means. Lastly, $W_{12}$ is independent from $W_{13}$. 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 $W_{12}$ that has mean that is close to zero and group 3 will have $W_{13}$ that is 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 0.10). Table 1 shows the median covariate values among those receiving each treatment, using only the subjects that remain after matching.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>0.71 (0.56, 0.82)</td>
<td>0.72 (0.57, 0.83)</td>
</tr>
<tr>
<td>$t_2$</td>
<td>0.70 (0.56, 0.84)</td>
<td>0.01 (-0.16, 0.20)</td>
</tr>
<tr>
<td>$t_3$</td>
<td>0.01 (-0.19, 0.18)</td>
<td>0.72 (0.58, 0.85)</td>
</tr>
</tbody>
</table>

2.5th, 97.5th percentiles shown in parenthesis

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 $X_1$ and $X_2$ among those matched, with one ellipse for each treatment. The triangles represent the pre-matched sample mean for those at each treatment, while the ‘+’ signs are the mean covariate values among those matched. After matching, the covariate spaces of those receiving each treatment are unique, and there is limited overlap between subjects receiving $t_2$ and $t_3$.

![95% quantiles of bivariate distribution](image)

Fig 3: 95% quantiles of bivariate $X_1$ and $X_2$ distribution among subjects matched for $Z = 3$. Means of pre and post-matched cohort depicted by symbols.

1.5.4 Resulting causal estimands when using SBC and CRM  

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). Those ‘eligible’ for all treatments simultaneously, however, are more representative of the clinical trial we are hoping to replicate, in which each subject has a realistic chance of receiving all treatments.

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

\[
\begin{align*}
    r(t, X)^{(low)} &= \max \left( \min (r(t, X|T = t_1)), ..., \min (r(t, X|T = t_Z)) \right) \\
    r(t, X)^{(high)} &= \min \left( \max (r(t, X|T = t_1)), ..., \max (r(t, X|T = t_Z)) \right)
\end{align*}
\]

where $r(t, X|T = \ell)$ is the treatment assignment probability for $t$ among those who received treatment $\ell$. This is a
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 \( r(t, X) \notin (r(t, X)^{\text{low}}, r(t, X)^{\text{high}}) \forall t \in T \) may have \( X \) 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 GPS’s are not disproportionately impacted by those dropped (adapted from the binary treatment scenario in Imbens and Rubin (2015)).

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

\[
E_{4i} = \begin{cases} 
1 & \text{if } r(t, X_i) \in (r(t, X)^{\text{low}}, r(t, X)^{\text{high}}) \forall t \in T \\
0 & \text{if otherwise}
\end{cases}
\]

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, \( \text{PATT}'s \) among subjects eligible for all treatments are defined as follows.

\[
\text{PATT}_{E_{4i}(t_1|t_1,t_2)} = E[Y_i(t_1) - Y_i(t_2)|T_i = t_1, E_{4i} = 1]
\]

\[
\text{PATT}_{E_{4i}(t_1|t_1,t_3)} = E[Y_i(t_1) - Y_i(t_3)|T_i = t_1, E_{4i} = 1]
\]

\[
\text{PATT}_{E_{4i}(t_1|t_1,t_Z)} = E[Y_i(t_1) - Y_i(t_Z)|T_i = t_1, E_{4i} = 1]
\]

There are two benefits to our definition of eligibility. First, all estimands in (13) are transitive; \( \text{PATT}_{E_{4i}(t_1|t_1,t_2)} \) and \( \text{PATT}_{E_{4i}(t_1|t_1,t_3)} \), 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 \( r(t, X)^{\text{low}} < r(t, X) < r(t, X)^{\text{high}} \forall t \), extrapolation to subjects that did not received a specific treatment is reduced.

1.5.5 Inverse probability weighting (IPW) Another common approach for estimating causal effects with multiple treatments uses the inverse probability of treatment assignment as weights (Imbens, 2000; Feng et al., 2011; McCaffrey et al., 2013). For IPW, a relaxed version of the assumption of a regular treatment assignment is adopted. Let \( I_i(t) = \{1 \text{ if } T_i = t, 0 \text{ otherwise} \} \). IPW requires that \( \forall t \in T \), \( P(I_i(t) = 1|Y_i(t), X_i) = P(I_i(t) = 1|X_i) \), which allows the estimation of average outcomes within subpopulations defined by pre-treatment status and is referred to as weak unconfoundedness (Imbens, 2000). Imbens (2000) acknowledges that the contrast between weak unconfoundedness and strong unconfoundedness is ‘not very different.’
Feng et al. (2011) implemented IPW to estimate \( PATE \)'s between each pair of treatments, such that to contrast \( t_1, t_2 \in \mathcal{T} \),

\[
\begin{align*}
PATE_{t_1, t_2} &= E[Y_{i}(t_1)] - E[Y_{i}(t_2)] \\
E[Y_{i}(t_1)] &= \left( \sum_{i=1}^{N} \frac{I(T_i = t_1)Y_i}{r(t_1, X_i)} \right) \left( \sum_{i=1}^{N} \frac{I(T_i = t_1)}{r(t_1, X_i)} \right)^{-1} \\
E[Y_{i}(t_2)] &= \left( \sum_{i=1}^{N} \frac{I(T_i = t_2)Y_i}{r(t_2, X_i)} \right) \left( \sum_{i=1}^{N} \frac{I(T_i = t_2)}{r(t_2, X_i)} \right)^{-1}.
\end{align*}
\]

When using IPW, extreme weights that are close to 0 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. As in the binary treatment setting, it is possible to trim or eliminate subjects with extreme weights. However, Kilpatrick et al. (2012) found that inclusion of high weights was necessary to maintain covariate balance in the multiple treatment setting, and that removal of extreme weights could result in bias in an unknown direction.

1.5.6 Matching for multiple treatments Recently, attempts have been made to group several subjects together who have similar \( R(X) \), 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' GPS's 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 often yielded lower standardized covariate bias when compared to CRM and SBC, although in some settings, these results were reversed.

As currently constructed, only \( t_1 \) can be used as the reference treatment using WithinTrio, where \( n_{t_1} = \min \{ n_{t_1}, n_{t_2}, n_{t_3} \} \). Because all subjects receiving \( t_1 \) are matched, there is 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. PATTs generalizable to those receiving treatment \( t_2 \) or \( t_3 \) cannot be estimated using WithinTrio. Lastly, WithinTrio cannot currently handle scenarios involving more than three treatment types. The computational requirements of this approach also grow exponentially with an increased number of units. Widespread public software for this algorithm's implementation remains unavailable.

Tu, Jiao and Koh (2012) examined a clustering algorithm to bin units into subclasses based on their \( \hat{R}(X)'s \) using simulations. The authors showed that \( K - \text{means} \) clustering (KMC, Johnson et al. (1992)) on the logit transformation of the GPS vector,  

\[
\log \left( R(X) \right) = \left( \log \left( \frac{r(t_1, X)}{1-r(t_1, X)} \right), ..., \log \left( \frac{r(t_Z, X)}{1-r(t_Z, X)} \right) \right),
\]

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 $\textit{PATEs}$ or $\textit{PATTs}$. One possible issue with clustering on $R(X)$ is that some subclasses may not include units from all treatment groups, which will require extrapolation to that subgroup. There are no known implementations of $KMC$ to estimate causal effects for a nominal exposure with real data.

2. MATCHING ON A VECTOR OF GENERALIZED PROPENSITY SCORES

To address the limitations of current strategies for multiple treatments effect estimation, we propose a new algorithm which can match subjects with similar $R(X)$. The proposed algorithm, vector matching ($VM$), forms matched sets for estimating (13)-(14) among subjects eligible for all treatments.

2.1 Vector Matching

With multiple treatments and GPS’s, we propose subclassifying subjects using $KMC$, and matching subjects only if they appear in the same subclass. Matching within a subclass ensures that units matched to one another are nearly perfect matches on one GPS and roughly similar on other treatment assignment probabilities. Matching within each subclass is performed with replacement, because it has been shown to yield lower bias in comparison to matching without replacement with binary treatment (Abadie and Imbens, 2006). Moreover, matching with replacement allows estimation of $\textit{PATT}$’s which are generalizable to each treatment group, and not just the group with the smallest sample size. We now explicate and summarize the procedure for a reference treatment $t \in T = \{t_1,...,t_Z\}$.

1. Estimate $R(X_i), i = 1,..., N$ using, for example, a multinomial logistic model.
2. Drop units outside the common support (e.g., those with $E_4 = 0$), and re-fit the model once.
3. $\forall t' \neq t$
   - (a) Cluster units receiving all other treatments besides $t$ and $t'$ using KMC on the logit transform of the remaining estimated GPS scores, $\tilde{R}_{t,t'}(X) = (r(t, X) \ \forall \ell \neq t, t')$. This forms $K$ strata of subjects, with similar $Z - 2$ GPS scores in each $k \in K$.
     - Example: with $Z = 5$, $T = \{t_1,...,t_5\}$, reference treatment $t_1$ and letting $t' = t_2$, $VM$ would use KMC on $\logit(r(t_3, X_i), r(t_4, X_i), r(t_5, X_i))$
   - (b) Within each strata $k \in K$, match those receiving $t$ to those receiving $t'$ on $\logit(r(t, X_i))$ within each $k$. Matching is performed with replacement using a caliper of 0.05.
     - 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 $n_{t_1,E_4=1}$ triplets can be generated using $VM$, where $n_{t_1,E_4=1}$ is the number of subjects receiving $t_1$ with $E_{4i} = 1$. For $Z = 3$, $VM$ reduces to:
1. Match those receiving $t_1$ to those receiving $t_2$ on $\logit(r(t_1, X_i))$ within $K$–means strata of $\logit(r(t_3, X_i))$
2. Match those receiving $t_1$ to those receiving $t_3$ on $\logit(r(t_1, X_i))$ within $K$–means strata of $\logit(r(t_2, X_i))$
3. Extract the subjects receiving $t_1$ who were matched to both subjects receiving $t_2$ and $t_3$

We implemented $VM$ by matching on $\logit(r(t, X))$ as well as $r(t, X)$ within strata estimated using KMC. The logit transformation produced smaller biases, which parallels findings observed with binary treatment (Rosenbaum and Rubin, 1985). The in strata matching procedure is implemented using the Matching (Sekhon, 2008) package in R statistical software (R Core Team, 2014).

### 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 $X$ using simulations. SBC is not included in the analysis because it cannot be used to contrast three or more treatments simultaneously.

#### 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 $X$ was generated for $N = n_{t_1} + n_{t_2} + n_{t_3}$ subjects receiving one of three treatments, $T \in \{t_1, t_2, t_3\}$, 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 Appendix 5. The values of $X$ were generated from multivariate symmetric distributions such that

\begin{align*}
T_i &= t_1, i = 1, \ldots, n_{t_1} \\
T_i &= t_2, i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1} \\
T_i &= t_3, i = n_{t_1} + \gamma n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1} + \gamma^2 n_{t_1} \\
\end{align*}

(16) \hspace{1cm} X_i | \{T_i = t_1\} \sim f(\mu_1, \Sigma_1), i = 1, \ldots, n_{t_1} \\
(17) \hspace{1cm} X_i | \{T_i = t_2\} \sim f(\mu_2, \Sigma_2), i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1} \\
(18) \hspace{1cm} X_i | \{T_i = t_3\} \sim f(\mu_3, \Sigma_3), i = n_{t_1} + \gamma n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1} + \gamma^2 n_{t_1} \\
(19) \hspace{1cm} \mu_1 = ((b, 0, 0), \ldots, (b, 0, 0))^T, \mu_2 = ((0, b, 0), \ldots, (0, b, 0))^T, \text{ and } \mu_3 = ((0, 0, b), \ldots, (0, 0, b))^T \\
(20) \hspace{1cm} \Sigma_1 = \begin{pmatrix} 1 & \tau & \ldots & \tau \\ \tau & 1 & \ldots & \tau \\ \ldots & \ldots & \ldots & \ldots \\ \tau & \tau & \ldots & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} \sigma_2 & \tau & \ldots & \tau \\ \tau & \sigma_2 & \ldots & \tau \\ \ldots & \ldots & \ldots & \ldots \\ \tau & \tau & \ldots & \sigma_2 \end{pmatrix}, \text{ and } \Sigma_3 = \begin{pmatrix} \sigma_3 & \tau & \ldots & \tau \\ \tau & \sigma_3 & \ldots & \tau \\ \ldots & \ldots & \ldots & \ldots \\ \tau & \tau & \ldots & \sigma_3 \end{pmatrix}
**Table 2**

**Simulation factors**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels of factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{t_1}$</td>
<td>{500, 2000}</td>
</tr>
<tr>
<td>$\gamma = \frac{n_{t_2}}{n_{t_1}} = \frac{n_{t_3}}{n_{t_2}}$</td>
<td>{1, 2}</td>
</tr>
<tr>
<td>$f$</td>
<td>{tr, Normal}</td>
</tr>
<tr>
<td>$b$</td>
<td>$B = \frac{b}{\sqrt{1+\frac{\sigma_2^2}{\sigma_3^2}}} \text{ takes levels } {0, 0.25, 0.50, 0.75, 1.00}$</td>
</tr>
<tr>
<td>$\tau$</td>
<td>{0, 0.25}</td>
</tr>
<tr>
<td>$\sigma_2^2$</td>
<td>{0.5, 1, 2}</td>
</tr>
<tr>
<td>$\sigma_3^2$</td>
<td>{0.5, 1, 2}</td>
</tr>
<tr>
<td>$P$</td>
<td>{3, 6}</td>
</tr>
</tbody>
</table>

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

\[
B = \frac{b}{\sqrt{1+\frac{\sigma_2^2}{\sigma_3^2}}}
\]

in order to evaluate the reduction in initial bias somewhat independently of the variance ratios $\sigma_2^2$ and $\sigma_3^2$.

Due to the small number of eligible subjects remaining when $P = 6$ and $n_{t_1} = 500$, these simulations are discarded, leaving 1080 simulation factors. For each simulation condition, 200 data sets are generated, and on each data set, $VM$ (using $K = 5$ strata), $CRM$, and $IPW$ are used to identify matched and weighted sets. For $CRM$, we used a caliper of 0.05 (Austin, 2011).

### 3.2 Simulation metrics

While several metrics have been proposed for evaluating the success of matching with binary treatments (see Austin, Grootendorst and Anderson (2007); Austin (2009), for example), 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 $\psi_i$ 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

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

For $IPW$, $\psi_i = \frac{1}{r(t, X_i)}$ is each subject’s weight, where $r(t, X)$ is estimated using multinomial logistic regression, and $n_{trip}$ is simply the number of subjects receiving each treatment $t$.

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

\[
SB_{p12} = \frac{\bar{X}_{p1} - \bar{X}_{p2}}{\delta_{p1}}
\]

and $\delta_{p1}$ is the standard deviation of $X_p$ 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

\[
Max2SB_p = max(|SB_{p12}|, |SB_{p13}|, |SB_{p23}|).
\]

(24)

For all of the matching algorithms and at each \( p \), \( \delta_{p1} \), the standard deviation of \( X_p \) 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, \( |SB_{p12}| + |SB_{p13}| + |SB_{p23}| / 3 \), 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 we are interested in generalizing to. Simulations with \( %Matched \approx 1 \) 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 their covariates’ distributions 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 each of the matching algorithms, \( Max2SB_p \forall p \) and \( %Matched \) are obtained, and averaged across 200 replications. For simplicity, we summarize \( Max2SB_p \forall p \) by averaging over \( p \), such that

\[
Max2SB = \sum_{p=1,...,P} Max2SB_p / P.
\]

### 3.3 Determinants of matching performance

Figure 4 shows boxplots of \( Max2SB \) and \( %Matched \) across each of the simulation factors. \( Max2SB \) was calculated for \( VM \), \( CRM \), \( IPW \), 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 4, \( Max2SB \) exceeds a cutoff of 0.20 in 25% of the combinations when using \( IPW \), compared to 19% when using \( CRM \) and to 3% when using \( VM \). There are 16 simulation configurations for which \( IPW \) yields a \( Max2SB \) greater than 1.5.

\( VM \) matched at least 85% of eligible reference subjects in a matched triplet in 97% of the configurations, while only 30% of the configurations for \( CRM \) reached the 85% cutoff. \( VM \) matched at least 95% of the eligible reference group subjects on more than 75% 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 \( Max2SB \) as well as \( %Matched \) (as in Rubin (1979), Cangul et al. (2009)). Because \( %Matched \) was highly skewed, we used Box-Cox power transformations (Sakia, 1992) to make this metric approximately normally
Initial covariate bias $B$ drives the highest proportion of variation in $Max2SB$, 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 $\sigma_2$ and $\sigma_3$. While $\gamma$, 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 $\gamma$; 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 primary ones. Tables 4 to 7 show $Max2SB$ based on different biases ($B$), distributions of $X$ ($f$), number of parameters ($P$), number of subjects receiving $t_1$ ($n_{t_1}$), and the ratio of units receiving $t_2$ to those receiving $t_1$ ($\gamma$).

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

Fig 4: $Max2SB$ for pre-matched cohort and by matching algorithms (left), and $%Matched$ for $VM$ and $CRM$ distributed.
Table 3
ANOVA for VM, CRM and IPW Max2SB: most influential factors

<table>
<thead>
<tr>
<th>Variable</th>
<th>MSE</th>
<th>Variable</th>
<th>MSE</th>
<th>Variable</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>137612</td>
<td>B</td>
<td>154397</td>
<td>B</td>
<td>41354</td>
</tr>
<tr>
<td>p</td>
<td>3092</td>
<td>f</td>
<td>115220</td>
<td>γ</td>
<td>4089</td>
</tr>
<tr>
<td>γ</td>
<td>1063</td>
<td>B*f</td>
<td>42628</td>
<td>σ_t</td>
<td>2271</td>
</tr>
<tr>
<td>σ_t</td>
<td>792</td>
<td>σ_s</td>
<td>22903</td>
<td>σ_s</td>
<td>469</td>
</tr>
<tr>
<td>σ_s</td>
<td>680</td>
<td>p</td>
<td>14746</td>
<td>B:*σ_t</td>
<td>243</td>
</tr>
<tr>
<td>τ</td>
<td>517</td>
<td>n_t1</td>
<td>11707</td>
<td>B:*γ</td>
<td>233</td>
</tr>
<tr>
<td>B*p</td>
<td>490</td>
<td>σ_t</td>
<td>10137</td>
<td>B*σ_s</td>
<td>147</td>
</tr>
<tr>
<td>B*σ_t</td>
<td>182</td>
<td>B*σ_s</td>
<td>5280</td>
<td>p</td>
<td>64</td>
</tr>
<tr>
<td>B*σ_s</td>
<td>158</td>
<td>B*p</td>
<td>3838</td>
<td>B*n_t1</td>
<td>48</td>
</tr>
<tr>
<td>σ_s*p</td>
<td>142</td>
<td>B*n_t1</td>
<td>3776</td>
<td>f</td>
<td>28</td>
</tr>
</tbody>
</table>

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’ using IPW vary substantially (e.g., Table 7). While $γ$ is not a major determinant of Max2SB for IPW, VM and CRM perform better in settings with $γ = 2$ (Tables 5 and 7).

Table 8 shows the %Matched for different values of $n_t1$ and $γ$, averaging over $P$, $f$, $τ$, $σ_2$ and $σ_3$. For low bias and with a larger number of controls ($γ = 2$), 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 $γ = 1$, $B = 1$, and $n_t1 = 1000$, for example, CRM matches only 36% of eligible subjects on average, compared to 90% 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 the covariates bias’ 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.

As in Tu, Jiao and Koh (2012), we also split subjects into strata using KMC, performing an analysis within each subclass. While the authors used $K = 5$ clusters, we also used KMC of find the largest number of clusters possible such that each cluster contained at least ten subjects receiving each treatment. Compared to $K = 5$, these additional clusters yielded lower bias, where bias was measured using a weighted average of within subclass bias. However, VM yielded lower bias in 88% of simulation configurations for $K = 5$, and in 74% of settings with KMC using the larger number of clusters. Pairwise bias using KMC was greater than 0.20 in 58% of configurations with $K = 5$ and in 14% of configurations with a larger cluster numbers.

A reduced set of simulations using $Z = 5$, which yielded similar conclusions, are detailed in Appendix 5.

4. DISCUSSION & CONCLUSION

The comparisons of multiple treatments falls under the realm of comparative effectiveness research (CER), which has recently earned an increased role in the health care reform debate (Helfand et al., 2011). CER is the direct comparison
of two or more treatments to determine which works best for which patients. The first step in CER should be to formulate the exact causal question (Rubin, 2010). But while estimating causal effects for binary treatment 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).

This manuscript compares current methods for causal effect estimation with multiple treatments with a novel matching algorithm, $VM$. The proposed algorithm can generate sets of subjects which are roughly equivalent on measured covariates. Simulations demonstrated that relative to other commonly used methods, $VM$ generally yielded the lowest bias in the matched sets. In most settings, nearly all eligible subjects that received the reference treatment were matched using $VM$. Under regular assignment mechanism, differences in these units’ outcomes could be contrasted, providing treatment effects generalizable to the population of subjects receiving $t_1$. Further, we found that pairwise procedures produced poor individual and aggregated matched sets when applied to multiple treatments.

We did not explore covariate adjustment for the GPS or a function of the GPS using a regression model (Filardo et al., 2009, 2007; Dearing, McCartney and Taylor, 2009; Spreeuwenberg et al., 2010). Such techniques are subject to possible model misspecification and extrapolation problems as seen in standard regression adjustment for binary treatment (Dehejia and Wahba, 1998, 2002), and simulations have shown that these strategies can perform worse than matching, stratification, or weighting with multiple treatments (Hade and Lu, 2013).

Although our focus was on the design phase of matching for multiple treatments, it is important to consider how matched sets could be used to approximate estimands (13) - (14). Point estimates using $VM$ matches can be obtained by contrasting those matched using a weighted average, with weights proportional to $\psi_i$. Variance terms for each causal effect can be estimated by using a bootstrap approach, as recommended by Hill and Reiter (2006) for a binary treatment. Abadie and Imbens (2006) derived empirical formulas for variance terms under certain matching scenarios with a binary treatment, and extending this approach for multiple treatments is an area for further research.

While we matched for $PATT$s, $VM$ 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 cohort would generalize to the population as a whole. As noted in Abadie and Imbens (2006), 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.
### Table 4
Max2SB, small/equal sample sizes: \( n_{t_1} = 500, n_{t_2} = 500, n_{t_3} = 500 \)

<table>
<thead>
<tr>
<th>( B )</th>
<th>( f = \text{Normal} )</th>
<th>( f = t_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{VM} )</td>
<td>( \text{CRM} )</td>
</tr>
<tr>
<td>0.00</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>0.25</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>0.50</td>
<td>0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>0.75</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>1.00</td>
<td>0.13</td>
<td>0.25</td>
</tr>
</tbody>
</table>

### Table 5
Max2SB, small/unequal sample sizes: \( n_{t_1} = 500, n_{t_2} = 1000, n_{t_3} = 2000 \)

<table>
<thead>
<tr>
<th>( B )</th>
<th>( f = \text{Normal} )</th>
<th>( f = t_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{VM} )</td>
<td>( \text{CRM} )</td>
</tr>
<tr>
<td>0.00</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>0.25</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>0.50</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>0.75</td>
<td>0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>1.00</td>
<td>0.10</td>
<td>0.20</td>
</tr>
</tbody>
</table>

### Table 6
Max2SB, large/equal sample sizes: \( n_{t_1} = 1000, n_{t_2} = 1000, n_{t_3} = 1000 \)

<table>
<thead>
<tr>
<th>( B )</th>
<th>( f = \text{Normal} )</th>
<th>( f = t_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{VM} )</td>
<td>( \text{CRM} )</td>
</tr>
<tr>
<td>0.00</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>0.25</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>0.50</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td>0.75</td>
<td>0.07</td>
<td>0.20</td>
</tr>
<tr>
<td>1.00</td>
<td>0.10</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Table 7

Max2SB, large/equal sample sizes: $n_1 = 1000$, $n_2 = 2000$, $n_3 = 4000$

<table>
<thead>
<tr>
<th>B</th>
<th>VM</th>
<th>CRM</th>
<th>IPW</th>
<th>VM</th>
<th>CRM</th>
<th>IPW</th>
<th>VM</th>
<th>CRM</th>
<th>IPW</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>0.25</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>0.50</td>
<td>0.04</td>
<td>0.08</td>
<td>0.08</td>
<td>0.04</td>
<td>0.07</td>
<td>0.12</td>
<td>0.06</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>0.75</td>
<td>0.05</td>
<td>0.14</td>
<td>0.13</td>
<td>0.05</td>
<td>0.12</td>
<td>0.33</td>
<td>0.10</td>
<td>0.16</td>
<td>0.14</td>
</tr>
<tr>
<td>1.00</td>
<td>0.08</td>
<td>0.21</td>
<td>0.17</td>
<td>0.07</td>
<td>0.19</td>
<td>0.80</td>
<td>0.16</td>
<td>0.21</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table 8

%Matched: The percent of eligible subjects receiving $t_1$ who were matched

<table>
<thead>
<tr>
<th>$n_1$ = 500</th>
<th>$n_1$ = 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma = 1$</td>
<td>$\gamma = 2$</td>
</tr>
<tr>
<td>b</td>
<td>VM</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>0.00</td>
<td>0.96</td>
</tr>
<tr>
<td>0.25</td>
<td>0.95</td>
</tr>
<tr>
<td>0.50</td>
<td>0.94</td>
</tr>
<tr>
<td>0.75</td>
<td>0.91</td>
</tr>
<tr>
<td>1.00</td>
<td>0.88</td>
</tr>
</tbody>
</table>
5. APPENDIX

We implement VM and IPW for $Z = 5$, where $X$ 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, $T \in \{t_1, t_2, t_3, t_4, t_5\}$, with $n_t$ the sample size of subjects receiving treatment $t$. Let $1$ be the $5 \times 5$ identify matrix. The values of $X$ were generated from multivariate symmetric distributions such that

\begin{align*}
T_i &= t_1, i = 1, \ldots, n_{t_1} \\
T_i &= t_2, i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1} \\
T_i &= t_3, i = n_{t_1} + \gamma n_{t_1} + 1, \ldots, n_{t_1} + 2 \gamma n_{t_1} \\
T_i &= t_4, i = n_{t_1} + 2 \gamma n_{t_1} + 1, \ldots, n_{t_1} + 2 \gamma n_{t_1} + \gamma^2 n_{t_1} \\
T_i &= t_5, i = n_{t_1} + 2 \gamma n_{t_1} + \gamma^2 n_{t_1} + 1, \ldots, n_{t_1} + 2 \gamma n_{t_1} + 2 \gamma^2 n_{t_1}
\end{align*}

\begin{align*}
X_i | \{T_i = t_1\} &\sim f(\mu_1, \Sigma), i = 1, \ldots, n_{t_1} \\
X_i | \{T_i = t_2\} &\sim f(\mu_2, \Sigma), i = n_{t_1} + 1, \ldots, n_{t_1} + \gamma n_{t_1} \\
X_i | \{T_i = t_3\} &\sim f(\mu_3, \Sigma), i = n_{t_1} + \gamma n_{t_1} + 1, \ldots, n_{t_1} + 2 \gamma n_{t_1} \\
X_i | \{T_i = t_4\} &\sim f(\mu_4, \Sigma), i = n_{t_1} + 2 \gamma n_{t_1} + 1, \ldots, n_{t_1} + 2 \gamma n_{t_1} + \gamma^2 n_{t_1} \\
X_i | \{T_i = t_5\} &\sim f(\mu_5, \Sigma), i = n_{t_1} + 2 \gamma n_{t_1} + \gamma^2 n_{t_1} + 1, \ldots, n_{t_1} + 2 \gamma n_{t_1} + 2 \gamma^2 n_{t_1}
\end{align*}

\begin{align*}
\mu_1 &= (b, 0, 0, 0, 0)^T, \mu_2 = (0, b, 0, 0, 0)^T, \mu_3 = (0, 0, b, 0, 0)^T, \mu_4 = (0, 0, 0, b, 0)^T, \mu_5 = (0, 0, 0, 0, b)^T \\
\Sigma &= 1
\end{align*}

The following design implicitly assumes a regular assignment mechanism that depends on four factors (Table 9).

<p>| Table 9 |</p>
<table>
<thead>
<tr>
<th>Simulation factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
</tr>
<tr>
<td>$n_{t_1}$</td>
</tr>
<tr>
<td>$\frac{\gamma_{12}}{n_{t_1}} = \frac{\gamma_{23}}{n_{t_1}} = \frac{\gamma_{34}}{n_{t_1}} = \frac{\gamma_{45}}{n_{t_1}}$</td>
</tr>
<tr>
<td>$f$</td>
</tr>
<tr>
<td>$b$</td>
</tr>
</tbody>
</table>

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 due to the small number of matches generated.

The $Max2SB$ for sets generated using VM is lower than 0.08 in each of the 20 settings in Table 9, and, on average, VM matched at least 93% of eligible subjects in each setting. As in simulations with $Z = 3$, $Max2SB$ for IPW was near 0 for normally distributed covariates. For $b = 0.75$ and 1.00 and using the $t_7$ distribution, IPW bias was 0.26 and 0.31, respectively.
REFERENCES


Hott, J. R., Brunelle, N. and Myers, J. A. (2012). Division of Pharmacoepidemiology And Pharmacoeconomics.


from a case study in mental health. Medical care 48 166.


