Tutorial in Biostatistics

Statistics in Medicine

Received 20 January 2012, Accepted 17 January 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5753

A tutorial on propensity score estimation for multiple treatments using generalized boosted models

Daniel F. McCaffrey,[a*†] Beth Ann Griffin,[b] Daniel Almirall,[c] Mary Ellen Slaughter,[a] Rajeev Ramchand[b] and Lane F. Burgette[b]

The use of propensity scores to control for pretreatment imbalances on observed variables in non-randomized or observational studies examining the causal effects of treatments or interventions has become widespread over the past decade. For settings with two conditions of interest such as a treatment and a control, inverse probability of treatment weighted estimation with propensity scores estimated via boosted models has been shown in simulation studies to yield causal effect estimates with desirable properties. There are tools (e.g., the `twang` package in R) and guidance for implementing this method with two treatments. However, there is not such guidance for analyses of three or more treatments. The goals of this paper are twofold: (1) to provide step-by-step guidance for researchers who want to implement propensity score weighting for multiple treatments and (2) to propose the use of generalized boosted models (GBM) for estimation of the necessary propensity score weights. We define the causal quantities that may be of interest to studies of multiple treatments and derive weighted estimators of those quantities. We present a detailed plan for using GBM to estimate propensity scores and using those scores to estimate weights and causal effects. We also provide tools for assessing balance and overlap of pretreatment variables among treatment groups in the context of multiple treatments. A case study examining the effects of three treatment programs for adolescent substance abuse demonstrates the methods. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: causal effects; causal modeling; GBM; inverse probability of treatment weighting; `twang`

1. Introduction

The use of propensity scores to control for pretreatment imbalances on observed variables in non-randomized or observational studies examining the causal effects of treatments or interventions has become widespread over the past decade. Authors have used propensity scores to match [1, 2], stratify (subclassify) [3, 4], or weight [5, 6] the samples from the treatment and control groups so that the distributions (or features of the distributions such as the means) of observed pretreatment characteristics are similar across the treatment and control groups, thereby reducing or eliminating confounding.

Propensity score techniques are advantageous compared with regression-based, covariate-adjustment techniques – which correct for imbalances between groups on pretreatment covariates by controlling for them in regression models for the outcomes – for at least five reasons. First, by summarizing all pretreatment variables to a single score, propensity scores are an important dimension reduction tool for evaluating treatment effects. This characteristic of propensity scores is particularly advantageous over standard adjustment methods when there exists a potentially large number of pretreatment covariates [3]. Second, propensity score methods derive from a formal model for causal inference, the potential

[a]*The RAND Corporation, 4570 Fifth Avenue, Pittsburgh, PA 15213, U.S.A.*
[b]*The RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202-5050, U.S.A.*
[c]*Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson St., Suite 2204, Ann Arbor, MI 48104-2321, U.S.A.*
**Correspondence to: Daniel F. McCaffrey, The RAND Corporation, 4570 Fifth Avenue, Pittsburgh, PA 15213, U.S.A.*
[†]*E-mail: danielm@rand.org*

Copyright © 2013 John Wiley & Sons, Ltd.

outcomes framework, so that causal questions can be well defined and explicitly specified and not conflated with the modeling approach as they are with traditional regression approaches. Third, propensity score methods do not require modeling the mean for the outcome. This can help avoid bias from misspecification of that model [7]. Fourth, propensity score methods avoid extrapolating beyond the observed data unlike parametric regression modeling for outcomes, which extrapolate whenever the treatment and control groups are disparate on pretreatment variables [1]. Lastly, propensity score adjustments can be implemented using only the pretreatment covariates and treatment assignments of study participants without any use of the outcomes. This feature of propensity score adjustments is valuable because it eliminates the potential for the choice of model specification for pretreatment variables to be influenced by its impact on the estimated treatment effect [8].

Most studies that use propensity scores to control for imbalances compare just two treatment groups of interest (e.g., treatment and control). Nonetheless, a number of papers have shown that propensity score methods can be extended to the multiple treatment case with three or more conditions of interest (e.g., treatment A, treatment B, and control; [9–11]). Theoretical work of Imbens [9] and Imai and van Dyk [10] developed the causal models and justification for the use of propensity scores to remove bias in cases with multiple treatment conditions, whereas the work of Robins and colleagues [12] developed the use of marginal structural models and inverse probability of treatment weighting (IPTW) for modeling causal effects from multiple treatments. Other authors have provided more specific guidance on implementing the approaches from Imbens [9] and Imai and van Dyk [10] in practice. For example, Lechner [13] provided a step-by-step matching protocol for multiple treatments, which has been utilized numerous times in the economics literature. A citation search found 76 papers that cited Lechner's paper, most involving economic evaluations. Zanutto *et al.* [14] described how to use stratification (subclassification) on the propensity score in the multiple treatment case, and most recently, Spreeuwenberg *et al.* [15] presented a tutorial for using the multinomial propensity scores as controls in the outcome regression model.

Despite these developments on the use of propensity score matching and stratification for more than two treatments, practical guidance on the use of propensity score weighting when examining multiple treatment conditions has received very limited attention. In particular, there is very limited guidance on how to estimate the propensity scores or the subsequent weights. Moreover, the existing applications have generally relied on parametric estimation of the propensity score via the multinomial, nested, or ordinal logistic regression model for multiple treatments [12–16]. Spreeuwenberg *et al.* [15] provided a step-by-step guide to causal modeling with multiple treatments that suggests multinomial logistic or probit models or ordinal logistic models be used to estimate the propensity scores and gives guidance on when one model may be preferable to another. However, the paper offers no guidance on variable selection or propensity score model tuning within this parametric framework. Zanutto and colleagues [14] also suggested using ordinal logistic regression to estimate the propensity scores for multiple doses of treatment. They recommend using the iterative approach of Rosenbaum and Rubin [3] and offer the necessary modifications to apply the approach to multiple treatments.

Recent studies of propensity score estimation in the binary case of two treatments show that, in terms of bias reduction and mean squared error (MSE), machine learning methods outperform simple logistic regression models with iterative variable section [17–19]. By extension, machine learning methods may also be advantageous in the multiple treatments setting. One such machine learning technique that has been frequently utilized in the two-treatment case [5, 18, 20] is the generalized boosted model (GBM). GBM estimates the propensity score for the binary treatment indicator using a flexible estimation method that can adjust for a large number of pretreatment covariates. GBM estimation involves an iterative process with multiple regression trees to capture complex and nonlinear relationships between treatment assignment and the pretreatment covariates without over-fitting the data [5, 21–24]. It works with continuous and discrete pretreatment variables and is invariant to monotonic transformations of them. Further, one of the most useful features of GBM for estimating the propensity score is that its iterative estimation procedure can be tuned to find the propensity score model leading to the best balance between treated and control groups, where balance refers to the similarity between different groups on their propensity score weighted distributions of pretreatment covariates.

In light of the potential advantages of boosting in the case of three or more treatment conditions, this paper provides researchers with a tutorial on implementing propensity score weighting using GBM when examining multiple treatments. Building on Frölich [11], we begin by describing a variety of causal effect estimands of potential interest when examining more than two treatment conditions (Section 2). Then, in Section 3, we describe how to estimate the multiple treatment propensity scores using GBMs, and we introduce useful diagnostic criteria for assessing balance. In Section 4, we use data from a large,

Copyright © 2013 John Wiley & Sons, Ltd.

observational study of adolescent substance users to illustrate GBM-based propensity score estimation and evaluation of diagnostic criteria for assessing balance in the context of an outcomes analysis of the relative effectiveness of three different outpatient substance abuse programs for adolescents.

2. Causal effects with multiple treatments

Causal effect estimation begins with an explicit determination of the causal effect 'estimands' that formalize the scientific quantities of interest. The causal effect estimands then become the target of estimation procedures. Multiple treatments allow for a variety of causal estimands. Not all estimands will be of interest for every study; the analyst must choose the estimands of interest depending on the scientific context. Hence, we begin our tutorial with the development of a causal model and a review of the possible causal estimands for multiple treatments as these must be the foundation of causal analysis.

We introduce the causal estimands for multiple treatments in the context of a case study on adolescent treatment effects. The Substance Abuse Mental Health Services Administration's Center for Substance Abuse Treatment (SAMHSA CSAT) collected data on adolescents receiving substance abuse treatment services from a variety of different outpatient community-based treatment settings including (i) traditional programs (community), (ii) programs implementing the evidence-based motivational enhancement therapy plus cognitive behavior therapy (MET/CBT-5) treatment protocol, and (iii) programs embedded in strengthened communities that link various systems involved in the identification, referral, and treatment of alcohol and substance abusing youth (Strengthening Communities for Youth (SCY)). One question of interest is the relative effects of these three alternative approaches for the treatment of substance use on outcomes at 12 months post-intake. For clarity in our presentation, we will specify causal effects in terms of these three treatment approaches (community, MET/CBT-5, and SCY). However, our methodology is not restricted to studies with three treatments; it extends naturally to any number of treatments given sufficient data.

Let M denote the number of treatments being studied with $M = 3$ in our case study. Following the potential outcomes approach for causal inference [25, 26], every youth in the population has a potential outcome that will result if he or she receives services from each of the three alternative treatment programs. For an individual, we denote these potential outcomes as $Y[t]$ for $t = 1, 2, 3$, where $t = 1$ denotes the individual's potential outcome had he or she received the community program, $t = 2$ the MET/CBT-5 program, and $t = 3$ the SCY program. When comparing alternative treatments, the causal effect of interest for an individual is defined as the difference among the potential outcomes for the *same* individual. Thus, possible causal effects of interest might be the relative effectiveness of all possible pairs of treatments: community versus MET/CBT-5; community versus SCY; and MET/CBT-5 versus SCY. For an individual, we denote these pairwise effects by $D[t', t''] = Y[t'] - Y[t'']$ for all $t' \neq t''$. In general, there exist $M(M-1)/2$ individual pairwise effects.

Causal effects for individuals generally cannot be estimated because of the fundamental problem of causal inference: We cannot observe an individual under each of the multiple treatments being compared [26]. Instead, we only observe what happens to an individual under the treatment condition they actually received. The causal effects we consider involve summary statistics of the individual effects across populations (or subpopulations) of interest. We consider two such summaries in this paper:

1. Average treatment effect (ATE): The ATE of treatment t' relative to treatment t'' is the comparison of mean outcomes had the *entire* population been observed under one treatment, t', versus had the entire population been observed under another treatment, t'' [27]. An example of an ATE from our case study is the mean outcome had all youth in our study been treated in the community programs compared with the mean outcome had all youth in the study been treated at the MET/CBT-5 programs. More formally, the ATE for comparing treatment t' and t'' equals $\mathrm{E}(D[t', t'']) = \mathrm{E}(Y[t'] - Y[t'']) = \mathrm{E}(Y[t']) - \mathrm{E}(Y[t''])$, where expectation is over the entire population. If we define μ_t as the mean outcome for the entire population when treated with treatment t, that is, $\mu_t = \mathrm{E}(Y[t])$, then the ATE of treatment t' relative to t'' is $\mu_{t'} - \mu_{t''}$. In our case study, with three treatment groups, there exist three ATEs, one for each pairwise comparison (Table I).
2. Average treatment effect among the treated (ATT): In the multiple treatments setting, the definition of ATT depends on what is meant by 'the treated'. This language, which is borrowed from the treatment versus control literature, requires more careful explanation when extended to the multiple

Copyright © 2013 John Wiley & Sons, Ltd.

Table I. Causal estimands for the effects of multiple treatments.

Effect	ATE	ATT		
		Community cases	MET/CBT-5 cases	SCY cases
MET/CBT-5 vs. community	$\mu_2 - \mu_1$	$\mu_{1,2} - \mu_{1,1}$	$\mu_{2,2} - \mu_{2,1}$	*
SCY vs. community	$\mu_3 - \mu_1$	$\mu_{1,3} - \mu_{1,1}$	*	$\mu_{3,3} - \mu_{3,1}$
SCY vs. MET/CBT-5	$\mu_3 - \mu_2$	*	$\mu_{2,3} - \mu_{2,2}$	$\mu_{3,3} - \mu_{3,2}$

The estimands for cells denoted by * theoretically exist but have limited utility in applications, so we do not consider them in this paper.

treatments settings. The ATT of treatment t'' among those treated with treatment t' (also stated as the ATT of t' relative to t'') is the comparison, among study participants who were treated with t', of their mean outcome when treated with treatment t', as they were, with the mean outcome they would have had if they had instead been treated with treatment t'' [27]. For instance, the mean outcome for youth in the community programs versus the mean outcome for those youth had they instead been treated at the MET/CBT-5 programs is an ATT that is of interest in our case study. More formally, if we let $\mu_{t',t''}$ equal the mean for youth who receive treatment t' had they instead received treatment t'', that is, $\mu_{t',t''} = \mathrm{E}(Y[t''] \mid T = t')$, then the ATT of treatment t' relative to t'' is $\mu_{t',t'} - \mu_{t',t''}$. In our example, with three treatment groups, there exist six ATTs: one for each pair of treatments, t' and t'', and either the population receiving t' or t'' as the one of interest (Table I). The $\mu_{t',t''}$ for $t' \neq t''$ are commonly referred to as the 'counterfactual means' because they estimate the mean outcomes for individual for treatments they did not receive, that is, the counterfactuals.

The ATEs and ATTs can differ when the treatment effects $D[t', t'']$ are not constant across individuals (i.e., there exists treatment effect heterogeneity). For instance, if standard community outpatient care is better for some youth, but MET/CBT-5 is better than community care for others, then the ATE, $\mu_2 - \mu_1$, may differ from the ATTs $\mu_{1,2} - \mu_{1,1}$ and $\mu_{2,2} - \mu_{2,1}$.

The choice of estimand depends on the substantive questions a study hopes to address and the population that is the target of the treatment. A study can estimate both ATE and ATT, but one or the other typically is better suited for any particular situation. The ATEs are more likely to be of interest compared with ATTs if every treatment potentially might be offered to every member of the population. So, if traditional community-based outpatient programs, programs providing MET/CBT-5, and programs in strengthened communities would each be appropriate for every youth in the population, then the ATEs might be of particular interest.

Conversely, if the research question focuses on the effectiveness of one treatment program t'' (e.g., strengthened communities) if it were to replace care for youth who typically received an alternative form of care, t', (e.g., MET/CBT-5 or community), then the ATT, $\mu_{t',t'} - \mu_{t',t''}$, would be of interest because it measures the relative effectiveness of programs t' and t'' on the population receiving program t'. Similarly, if the research question is how well is a particular treatment program t' matched to the specific population it targets for its treatment, then the ATTs $\mu_{t',t'} - \mu_{t',t''}$ or $\mu_{t',t'} - \mu_{t',t'''}$ are likely to be of interest because they quantify the effects of treatment t' relative to the alternatives on the population targeted by that program. For example, MET/CBT-5 was developed for marijuana users and has less evidence of efficacy among youth who use other drugs such as cocaine or opiates. In this case, estimating the ATE of MET/CBT-5 for a population that includes many youth who do not use marijuana and/or have serious criminal problems might not be of interest, but ATTs would be.

The advantage of ATT is that each treatment program is evaluated only on cases it treats. This is important because youths and treatments may be aligned so that youths assigned to a treatment are the subset of the population who may fare the best with this treatment. For instance, youth with marijuana problems are assigned to MET/CBT-5, which has been shown to be more effective with these youth than others. Disadvantages of ATTs are that they do not support inferences about the relative effects of programs if they are expanded from their base clientele and they cannot be used to determine if changing which youth are treated by different types of programs would improve outcomes overall. For example, the youth in communities that chose to participate in the community strengthening programs (SCY) might not be the most responsive to this approach to treatment; we might find treatment within a strengthened community is more effective on the entire population than on the population served by this program in our study. We could not learn this using ATT, but we could using ATE.

Copyright © 2013 John Wiley & Sons, Ltd.

3. Method for estimating multiple treatment effects using propensity score weights

In Section 2, we presented different types of causal estimands that may be of interest when faced with more than two treatment groups. After determining which estimand is best suited to answer each of a study's research questions, analysts need to obtain accurate estimates of these quantities. This section first describes the general framework of using weighting to estimate pairwise ATEs or ATTs and then provides details how to use GBM to calculate the weights to be plugged into the estimators developed here.

3.1. Notation and assumptions

To allow for a precise presentation of estimators of ATE or ATT, we need additional notation for the observed data. For the $i = 1, \ldots, n$ individuals in the observed data, let T_i denote the observed treatment status for the ith individual; that is, $T_i = t$ if individual i was observed under treatment t, where $t \in \{1, \ldots, M\}$ and M is the total possible number of treatments as noted earlier. Let Y_i denote the observed (rather than a potential) outcome for individual i. Each individual has M potential outcomes, but only the one corresponding to the observed treatment (T_i) is observed. That is, if for a given unit i, $T_i = t$, then $Y_i = Y_i[t]$.[‡] Let $\mathbf{X}_i$ denote the vector of K *observed* pretreatment covariates. In our example, $\mathbf{X}_i$ includes demographics (race, age, and gender), as well as substance use, criminal activities, and emotional functioning among other variables at baseline and $K = 23$ (see Section 5.1.2 for details).

As discussed in Section 1, estimating the effects in Table I is complicated by the fundamental problem of causal inference and the need to compare different youth receiving different treatments rather than the same youth receiving multiple treatments to estimate the causal estimands.

Difficulties arise because of possible confounding variables, which are informally defined as pretreatment variables directly influencing both observed outcomes Y_i and treatment T_i. In observational studies, because the researchers do not control assignment to the multiple treatment conditions, confounding variables could differ across the groups receiving different treatments. These differences could make the estimation of the causal estimands in Table I impossible without assumptions and methods to remove differences among groups on observed confounding variables. IPTW [12], described later, is one methodology to reduce confounding due to observed variables. IPTW relies on two key assumptions in order to produce unbiased (with large samples) estimates of the ATE or ATTs of Table I. The assumptions cannot be tested with data, so analysts need to fully understand the implications of the assumptions and carefully evaluate their plausibility in the context of their observational study. Hence, we begin our discussion of IPTW by describing the two assumptions and then turn to defining the IPTW estimators.

Condition 1 (Sufficient overlap or positivity)
Let T_i be the random treatment assignment variable defined earlier, then

$$0 < \mathrm{pr}(T_i = t|\mathbf{X}) < 1 \text{ for all } \mathbf{X} \text{ and } t$$

The positivity assumption states that each subject has a non-zero probability of receiving each treatment. It implies that there are no values of pretreatment variables that could occur only among units receiving one of the treatments. In the case study, positivity requires that there are no patterns of values for our 23 covariates, which preclude youth from receiving one treatment or another. Lack of overlap in the true (unknown) distribution of observed pretreatment characteristics between groups receiving different treatments would imply the positivity assumption is violated. Although we cannot prove lack of overlap with the observed data (due to small sample error), if we find sets of covariate values that are clearly disjoint across the three treatment groups, then this may suggest lack of overlap (and therefore a violation of the positivity assumption). We discuss our exploration of lack of overlap for the case study in Section 5.2.

Condition 2 (No unknown or unmeasured confounders assumption or exchangeability)

$$\text{Let } T_i[t] = I(T_i = t), \text{ then } T[t] \perp\!\!\!\perp Y[t] \mid \mathbf{X}.$$

[‡]*The observed outcome is well defined as equal to the potential outcome only when the consistency assumption holds so that the potential outcomes have unique value regardless of how the assignment to a treatment resulted in the observed treatment or other factors surrounding treatment such as the other youth in treatment or the specific provider* [28].

Copyright © 2013 John Wiley & Sons, Ltd.

Imbens [9] showed that condition 2 is sufficient for consistent estimation of μ_t and that it is related to the 'strong ignorability' assumption of [29], which requires that treatment indicators be independent of the entire vector of potential outcomes. Informally, condition 2 assumes that the set of observed pretreatment covariates, $\mathbf{X}_i$, is sufficiently rich such that it includes all variables directly influencing both T_i and Y_i (i.e., there exist no unmeasured or unknown confounders not included in $\mathbf{X}_i$). Again, there is no way to test this assumption with the data because it is an assumption about unobserved variables. Typically, analysts try to include a broad set of covariates in the analysis of causal effects to reduce the potential that a confounding variable is omitted inadvertently. Substantive experts may be consulted to identify variables used in the assignment of treatment and which the literature or their experience suggest also are related to outcomes and could result in confounding. Variables known or strongly believed only to be related to treatment assignment should not be used in propensity score models. Controlling for such variables can reduce the precision of estimates and inflate the bias if there are omitted variables [30]. In our case study, the survey instrument used to collect pretreatment variables was designed to collect data useful for treatment assignment and generally known to be related to post-treatment outcomes. Hence, the observed variables are possible confounders, and we test for their relationship to post-treatment outcomes within treatment groups to identify variables for modeling.

3.2. *Inverse-probability of treatment weighting with multiple treatments*

It is well established that the averages of the observed outcomes among people who receive treatment t' can be biased or inconsistent as estimates of $\mu_{t'}$ or $\mu_{t'',t'}$ for $t'' \neq t'$ if there is confounding, that is, if individuals receiving different treatments are different in terms of their potential outcomes [29]. The problem is that samples receiving the different treatments typically differ in their distributions of pretreatment variables and, therefore, possibly differ in terms of their observed outcomes in ways that are not attributable to treatment. Assuming all the variables with pretreatment differences are observed and that the groups have at least some members with similar covariates, that is, conditions 1 and 2 hold, then we should in principle be able to reweight a treatment sample to make the distribution of covariates match that of any of the other treatment groups. Several authors [11, 12, 31] have shown this intuition holds formally provided the weights equal the reciprocal of the probability that a study participant received the treatment he or she received. This approach to estimating the population means of potential outcomes is a form of IPTW. We use IPTW to estimate μ_t for each value of t and use these estimates to obtain treatment effects. Estimates $\mu_{t',t''}$ for different pairs of treatments also use weighting, but the weights are modified because the population of interest is not the entire population.

3.2.1. Pairwise ATEs. When interest is in estimating the pairwise ATEs for a set of M treatments (e.g., the three pairwise ATEs of Table I: $\mu_2 - \mu_1$, $\mu_3 - \mu_1$, and $\mu_3 - \mu_2$), we need consistent estimates of the population means of the potential outcomes for each of the treatments (μ_1, μ_2, and μ_3) to obtain estimates of the desired causal effect. Let $p_t(\mathbf{X})$ denote the *propensity score*, the probability that an individual with pretreatment characteristics $\mathbf{X}$ receives treatment t ($p_t(\mathbf{X}) = \mathrm{pr}(T[t] = 1 \mid \mathbf{X})$). A consistent estimate of μ_t is given by the weighted mean

$$\widehat{\mu}_t = \frac{\sum_{i=1}^{n} T_i[t] Y_i w_i[t]}{\sum_{i=1}^{n} T_i[t] w_i[t]}, \tag{1}$$

where the weights satisfy

$$w_i[t] = \frac{1}{p_t(\mathbf{X}_i)}, \tag{2}$$

provided conditions 1 and 2 hold [11, 32]. We estimate treatment effects of interest using the formulas for the estimands in Table I with the unknown population means replaced by their estimates. For instance, we estimate the ATE of MET/CBT-5 relative to traditional community-based care, $\mu_2 - \mu_1$, by the $\widehat{\mu}_2 - \widehat{\mu}_1$, and similar replacements would be used to estimate the other two ATEs.

3.2.2. Pairwise ATTs. When interest is in estimating the pairwise ATTs for a set of M treatments for the population receiving one of the treatments t' (e.g., the two pairwise ATTs for community cases in Table I: $\mu_{1,2} - \mu_{1,1}$ and $\mu_{1,3} - \mu_{1,1}$), we need consistent estimates for the mean of the potential outcomes for youth like those who received the treatment t' had they received the other treatment conditions ($\mu_{1,2}$

Copyright © 2013 John Wiley & Sons, Ltd.

and $\mu_{1,3}$ in our example). Under conditions 1 and 2, a consistent estimate of $\mu_{t',t''}$ will be given by the weighted mean

$$\widehat{\mu}_{t',t''} = \frac{\sum_{i=1}^{n} T_i[t'']Y_i w_i[t',t'']}{\sum_{i=1}^{n} T_i[t'']w_i[t',t'']}, \tag{3}$$

where

$$w_i[t',t''] = p_{t'}(\mathbf{X}_i)/p_{t''}(\mathbf{X}_i), \tag{4}$$

and $\mu_{t',t'}$ will be consistently estimated using

$$\widehat{\mu}_{t',t'} = \frac{\sum_{i=1}^{n} T_i[t']Y_i}{\sum_{i=1}^{n} T_i[t']},$$

the unweighted mean of observed outcomes from units assigned to treatment t' [11].

Here, the appropriate weight for individuals receiving treatment t'' is the ratio of the probabilities for receiving t' and t'', $p_{t'}(\mathbf{X})/p_{t''}(\mathbf{X})$. Intuitively, we weight each individual by the reciprocal of their probability of receiving the treatment they received relative to the probability of receiving the target treatment. Individuals with covariate values that are much more common in their own treatment group than in the target group (i.e., $p_{t''}(\mathbf{X})$ is very large relative to $p_{t'}(\mathbf{X})$ or $p_{t'}(\mathbf{X})/p_{t''}(\mathbf{X})$ is small) have small weights because they are relatively too common in their sample but not the targeted sample. Individuals with covariate values that are much more common in the target treatment group than in their own group ($p_{t''}(\mathbf{X})$ is very small relative $p_{t'}(\mathbf{X})$ or $p_{t'}(\mathbf{X})/p_{t''}(\mathbf{X})$ is large) have large weights because there are relatively too few of these types of individuals in their sample, and they are most representative of the target treatment group.

Again, we estimate the treatment effect of interest using the formulas for the ATT estimands in Table I with the unknown population means replaced by their estimates.

We note that estimating pairwise ATEs and ATTs can be implemented easily by fitting weighted regression models in the survey packages of many commonly used software packages, including SAS, STATA, and R, and the online supporting information[§] for this paper contains examples of code for using each of these packages to estimate ATE and ATT. The weighted models must include dummy indicators for $M-1$ of the treatment groups to obtain estimates of pairwise ATTs and ATEs (also called treatment contrasts) along with 95% confidence intervals (and corresponding p-values) for assessing statistical significance. In the next section, we describe multinomial logistic regression for estimating propensity scores and then focus on how to use one particularly robust method (GBM) for estimating the propensity scores and subsequent weights for comparing multiple treatment conditions.

4. Estimating propensity scores and assessing balance

4.1. Estimating multiple propensity scores using multinomial logistic regression

Multinomial logistic regression is a very commonly used approach to modeling the relationship between covariates and outcomes that take on a small number of discrete values, such as assignment to one of three treatment conditions, and has been proposed for estimating propensity scores with multiple treatments [15, 32]. It models the probability that an outcome (e.g., treatment assignment, T_i) equals each of its possible values as a function of a linear combination of the covariates and their products and cross products:

$$P(T_i = t|\mathbf{X}_i) = \frac{e^{\boldsymbol{\beta}_t'\mathbf{X}_i}}{1+\sum_{t'=1}^{M-1} e^{\boldsymbol{\beta}_{t'}'\mathbf{X}_i}}, t = 1,\ldots,M-1,$$
$$P(T_i = M|\mathbf{X}_i) = \frac{1}{1+\sum_{t'=1}^{M-1} e^{\boldsymbol{\beta}_{t'}'\mathbf{X}_i}},$$

where $\boldsymbol{\beta}_t$, $t = 1,\ldots,M$ are unknown and estimated from the data. Given estimates for the unknown parameters, the $P(T_i = t|\mathbf{X}_i)$ can be generated and plugged into the formulas for estimating ATEs or

[§] *Supporting information may be found in the online version of this article.*

Copyright © 2013 John Wiley & Sons, Ltd.

ATTs. Maximum likelihood is the standard approach to estimating the coefficients, and all of the commonly used statistical software packages include routines for fitting these models. This is an extension of logistic regression for binary or dichotomous variables when $M = 2$.

The challenge for propensity score estimation is choosing the correct set of interactions and polynomial terms among the covariates to capture any nonlinearities in their relationship to treatment assignment. There is no standard method for model selection in the context of estimating propensity scores for IPTW for multiple treatments. There is not even a standard method for estimating propensity scores for weighting with two treatments. For example, Foster [16] used multinomial regression to estimate propensity scores for different doses of treatment but did not describe the use of any model selection procedure. For two treatments, the most common approach to variable selection is an iterative approach first suggested in [3] and used in [4], which is similar to the approach used by Zanutto and colleagues described later [14]. But this approach is tailored to stratification and is not specifically tuned for weighted estimation.

Zanutto and colleagues [14] presented a method for selecting the form of propensity score model in the context of multiple treatment for stratification. A similar approach could be used for weighting. The algorithm for this approach involves six steps:

1. Fit a simple model with only main effects for all the proposed covariates.
2. For $T = 1, \ldots, M$, find the areas of common support for $\widehat{\beta}_t'\mathbf{X}_i$ among all treatment groups, retain observations that are in the area of common support for all T.
3. Test for balance on the covariates.
4. Add to the model polynomial and cross product terms for covariates that do not balance.
5. Fit the new model.
6. Repeat steps 3–5 until all the covariates are balanced.

Step 2 of the algorithm is meant to ensure sufficient overlap among the treatment groups so that Condition 1 holds. To test for balance in step 3, Zanutto *et al.* [14] and Spreeuwenberg *et al.* [15] suggested testing whether treatment assignment, modeled with $M-1$ indicator variables for the M treatment conditions, is a significant predictor of each covariate controlling for propensity score strata. If the $M-1$ degree-of-freedom F-test for treatment assignment is significant, then the covariate is not balanced. The approach is similar to the testing method in [3] for two treatments. In the context of using propensity scores to create strata for profiling multiple treatment programs, Huang and colleagues [33] used the same procedure for testing for balance and updating their model.

The significance test needs to be modified for estimating propensity scores for weighting. Rather than controlling for propensity score strata in the model for each covariate, the data must be weighted. The test would then entail – for each pretreatment variable – running a one-way analysis of variance on treatment assignment with each observation weighted by its IPTW and using the $M - 1$ degree-of-freedom F-test for treatment as the test for balance. The F-test would need to account for weighting. Again if the test rejects the null hypothesis of no group differences, then the groups are not balanced; otherwise, there is no evidence of imbalance. An advantage of testing is that it checks for balance on all treatments. However, the significance test could fail to find group differences significant when the covariates are imbalanced because the test statistic is imprecise because of weighting. For these reasons, we and other authors do not recommend using significance test to assess balance. Instead, we suggest the statistics described in Sections 4.2.2 and 4.2.3.

Step 4 is likely to prove challenging in practice. With multiple covariates and a model for each treatment level, there are many possible interaction and polynomial terms to add to the model any time covariates do not balance. With even a moderate number of variables and treatments, it will be very time consuming and probably prohibitive for analysts to try all variations of possible polynomial and interactions terms to add to the model. Obtaining good balance will rely on the analyst's ability to identify candidate terms to test models.

Beyond the challenges of implementing this algorithm, parametric logistic regression has been shown in simulation experiments to yield causal effects estimates with high MSE [17–19, 31, 34] because linearity assumptions of logistic regression can lead to very small probabilities and extremely large weights. Large errors can result even if the model is correctly specified [31], and logistic regression can be particularly problematic if the model is misspecified [17–19, 31, 34]. The possibility of extreme weights also exists in the multiple treatments setting with multinomial regression. GBM can mitigate both these challenges to using multinomial regression. It has automated variable selection [5] and, in simulations, has proven to provide more stable weights than parametric models [17, 18].

Copyright © 2013 John Wiley & Sons, Ltd.

4.2. *Estimating multiple propensity scores using GBM*

We begin with a brief description of GBM and how it has been used in the binary treatment case and then turn to applying the methods to more than two treatment conditions.

GBM fits a piecewise constant model to predict a dichotomous outcome (e.g., a binary treatment indicator). The model consists of many simple regression trees [35] iteratively combined to create an overall piecewise constant function. The iterative fitting algorithm begins with a single simple regression tree, and at each new iteration, another tree is added. The new tree is chosen to provide the best fit to the residuals of the model from the previous iteration. This chosen tree also provides the greatest increase to the log likelihood for the data. When combining trees, the predictions from each tree are shrunken by a scalar less than one to improve the smoothness of the resulting piecewise constant model and the overall fit. Each iteration increases the likelihood, and with enough iterations, the model is sufficiently flexible to perfectly fit the data. The actual algorithm typically considers many iterations of adding trees, which may overfit the data. To avoid this, GBM selects an intermediate iteration (or number of trees) for the final model so as to minimize an external criterion such as out-of-sample prediction error or – in the case of propensity score estimation – imbalance on the pretreatment covariates across the 'treatment' and 'control' groups [5].

4.2.1. Using GBM to obtain propensity score weights for binary treatments. As noted earlier, we can utilize GBM to estimate propensity score weights in the binary treatment setting. The key is to use GBM in the iterative fashion described earlier with the optimal iteration (number of trees) for estimating the propensity scores set to be the one that minimizes a 'stopping rule' criterion based on the difference between the weighted distributions of the pretreatment covariates in the two treatment conditions. Several authors [5, 17] have found that among a variety of propensity score estimation methods, GBM used in this fashion provides estimated weights that yield the best balance of the pretreatment variables and estimated treatment effects with the smallest MSE in the binary treatment case.

In practice, various stopping rules have been utilized to select the optimal iteration of GBM for use in estimating propensity score weights, including rules based on summary statistics (i.e., maximum or mean) of absolute standardized bias (SB, also referred to as the absolute standardized mean difference) or the Kolmogorov–Smirnov (KS) statistic, which compare the means or the distributions of the covariates between treatment groups. The summaries are computed across the pretreatment covariates used in the GBM.

For each covariate, the standardized bias (absolute standardized mean difference) equals the absolute value of the difference between the weighted mean for treatment group and the weighted mean for the control group divided by unweighted standard deviation of the pooled sample for ATE or divided by the unweighted standard deviation of the treatment group for ATT. For ATE, for covariate k ($k = 1, \ldots, K$),

$$\mathrm{SB}_k = |\bar{X}_{k1} - \bar{X}_{k0}|/\widehat{\sigma}_k,$$

where $\bar{X}_{kt}$ is the weighted mean of the covariate for treatment ($t = 1$) or control ($t = 0$) and $\widehat{\sigma}_k$ is the standard deviation of the covariate for the pooled sample. For ATT,

$$\mathrm{SB}_k = |\bar{X}_{k1} - \bar{X}_{k0}|/\widehat{\sigma}_{k1},$$

where the weights equal one for members of the treatment group and $\widehat{\sigma}_{k1}$ is the standard deviation of the covariate for the treatment group.

Standardized bias in the binary treatment setting has been shown to be advantageous for assessing balance because it provides analysts with a way to assess the size of the difference between treatment groups in the distributions of the variables that is consistent across pretreatment covariates. Further, because the level of imbalance is placed on the same scale for all covariates, it allows rank-ordering covariates by the amount of imbalance; this allows consumers of the study results to see which observed covariates are most and least imbalanced. Generally, standardized mean differences of less than 0.20 are considered small, 0.40 are considered moderate, and 0.60 are considered large [36]. Some authors consider any standardized difference greater than 0.25 as evidence of imbalance and a potential source of bias, whereas other authors may wish to use more a more conservative cutoff [7, 37, 38]. In this paper, we focus on differences larger than 0.20 as problematic, but analysts can choose their own cutoff value as deemed appropriate for their given setting.

Copyright © 2013 John Wiley & Sons, Ltd.

The KS statistic depends on the weighted empirical distribution functions for the treatment and control samples. For covariate k, these are defined as $\mathrm{EDF}_{tk}(x) = \sum_{i=1}^{n} w_i[t]T_i[t]I(X_{ik} \leqslant x)/\sum_{i=1}^{n} w_i[t]T_i[t]$, for $t = 0$ or 1 and $I(X_{ik} \leqslant x)$ equal one if $X_{ik} \leqslant x$ and zero otherwise. The KS statistic for this covariate is

$$\mathrm{KS}_k = \sup_x |\,\mathrm{EDF}_{1k}(x) - \mathrm{EDF}_{0k}(x)\,|\,.$$

For ATT, the weights equal one for members of the treatment group.

The KS is valuable because it compares the entire distribution rather than just the mean. However, unlike standardized bias, the distribution of the KS statistic depends on the sample size so there can be no universal objective guideline for what constitutes a large or small value. For modest to large sample sizes, we tend to consider KS statistics greater than 0.10 as indications of imbalance. The KS statistic is very useful for comparing among models because even if there are no absolute standards, the relative sizes of the KS statistics for weights from different model fits can rank the models.

The Toolkit for Weighting and Analysis of Nonequivalent Groups (`twang`) is an R package that implements propensity score estimation via GBM using one (or all) of the four different stopping rules for selecting the optimal GBM iteration described earlier (e.g., mean standardized bias, maximum standardized bias, mean KS, or maximum KS across the pretreatment covariates).

4.2.2. Using GBM to estimate ATEs of multiple treatments. Given that GBM has proven to be effective in studies of two treatments, we propose an extension of the method for estimating propensity score weights when there are more than two treatments. Specifically, when interest lies in estimating the ATEs, we propose using GBM in the following fashion to obtain weights. First, create dummy indicators for each of the M treatment programs, denoted $T_i[t]$ as earlier (e.g., receiving community treatment program versus not receiving community treatment program). Then, fit separate GBMs to each dummy treatment indicator and obtain the estimated propensity score for the given treatment in question (e.g., the probability of receiving the community treatment program). Finally, use the estimated propensity scores from each of the GBM fits to compute the ATE weights (Equation (2)) needed to estimate treatment effects.

The general intuition behind this approach is for any given treatment, $T_i = t$, estimating ATE weights for this treatment group only requires knowing the probability that each case assigned to this group received that assignment rather than one to *any* other treatment. The weights do not differentiate among the alternative treatments. Hence, for each treatment group, estimating the ATE weight is like the binary treatment case. Because GBM has been established to yield good ATE weights in the binary treatment setting, it should in principle be able to do the same for each of the separate fits for the $T_i[t]$. Fitting GBM one treatment at a time should produce, for individuals assigned to that particular treatment group, propensity scores and corresponding inverse probability of treatment weights, which balance the pretreatment characteristics between the group and the entire population. Critically, for each treatment indicator, we propose that the estimated propensity score, $p_t(\boldsymbol{X}_i)$, be computed from the iteration of the GBM fit, which yields the 'best balance' between units with $T_i = t$ ($T_i[t] = 1$) and the pooled sample from all treatments. We use this as the balance criterion because ATEs are average effects across the entire population. The entire population is the key target group that each treatment group should match when weighted, as opposed to, say, those with $T_i[t] = 1$ versus those with $T_i[t] = 0$.

Balance can be defined by any of the stopping rules described earlier (mean standardized bias, maximum standardized bias, mean KS, or maximum KS). However, the balance metrics are modified for ATE with multiple treatments. For covariate k ($k = 1, \ldots, K$) and treatment t ($t = 1, \ldots, M$),

$$\mathrm{PSB}_{tk} = |\bar{X}_{kt} - \bar{X}_{kp}|/\widehat{\sigma}_{kp} \tag{5}$$

where PSB_{tk} is the 'population' standardized bias, $\bar{X}_{kt} = \left(\sum_{i=1}^{n} T_i[t]X_{ki}/\widehat{p}_t(\mathbf{X}_i)\right) / \left(\sum_{i=1}^{n} T_i[t]/\widehat{p}_t(\mathbf{X}_i)\right)$ is the propensity score weighted mean of the covariate, $\widehat{p}_t(\mathbf{X}_i)$ is the estimated propensity score for the treatment obtained from our GBM fits described earlier, and $\bar{X}_{kp}$ and $\widehat{\sigma}_{kp}$ denote the unweighted mean and standard deviation of the covariate for the pooled sample across all treatments. It allows us to directly assess how similar each treatment group is to the population in terms of covariate means both before and after weighting. Hence, PSB_{tk} is an appropriate criterion for assessing balance when one has more than two treatment conditions. An alternative might be to compute standardized bias for every possible two-group comparison and summarize balance across those standardized biases. We find such computations to be cumbersome and less direct than standardizing with respect to the entire population to which we are trying to draw inferences with ATEs.

Copyright © 2013 John Wiley & Sons, Ltd.

A similar modification is used for the KS statistic for assessing balance with multiple treatments. The populations KS statistics (PKS) equals

$$\mathrm{PKS}_{tk} = \sup_x |\, \mathrm{EDF}_{tk}(x) - \mathrm{EDF}_{pk}(x) \,| \tag{6}$$

where $\mathrm{EDF}_{pk}(x) = \sum_{i=1}^{n} I(X_{ik} \leqslant x)/n$, the unweighted empirical distribution function for the pooled sample across all treatments. Again, the stopping rules use the maximum or the minimum of either PSB_{tk} or PKS_{tk} across the pretreatment covariates to measure the balance of a treatment group relative to the population and to choose the optimal GBM iteration.

In our case study example, the procedure for estimating ATE weights proceeds as follows. We fit three separate binary GBMs. The first would estimate $\widehat{p}_1(\mathbf{X}_i)$, the probability that each youth received community care given his or her pretreatment covariates. The second estimates $\widehat{p}_2(\mathbf{X}_i)$, the probability that each youth received MET/CBT-5 given his or her pretreatment covariates, and the third estimates $\widehat{p}_3(\mathbf{X}_i)$, the probability that each youth received an SCY program given his or her pretreatment covariates.

For ATE weights, our method fits M separate models or probability estimates. It is not fitting a multinomial model to the treatment group assignments. Consequently, for a given youth, the sum of the $\widehat{p}_t(\mathbf{X}_i)$, $t = 1, \ldots, M$, ($\widehat{p}_1(\mathbf{X}_i)$, $\widehat{p}_2(\mathbf{X}_i)$, and $\widehat{p}_3(\mathbf{X}_i)$ in our example) may not equal one as would be the case if we were to jointly estimate these multiple treatment propensity scores as is carried out with multinomial logistic regression models. The fact that the estimated propensity scores do not sum to one is not a problem for estimating weighted ATEs, because for each youth, only $\widehat{p}_t(\mathbf{X}_i)$, the probability of being assigned the treatment he or she received, is used for weighting his or her data. For each treatment, the primary goal is to weight the group that received that treatment to match the entire sample and yield an estimate of μ_t that is essentially unconfounded by the observed pretreatment covariates so these estimated means can be used to estimate treatment effects. We can achieve this goal by estimating the probability that $T[t] = 1$ ignoring the probability of assignment to other treatments provided the estimated probabilities yield weights that balance the group receiving treatment t and the entire sample. This is particularly important because obtaining weights via GBM has proven to be more successful for the task of balancing two groups than the parametric logistic model.

It is possible to use GBM to estimate propensity scores for multiple treatments in which the estimated probabilities satisfy $\sum_t \widehat{p}_t(\mathbf{X}_i) = 1$. The estimation procedure uses the following steps: (1) choose one treatment group as a holdout, say $t = M$; (2) for treatment group $T = 1$, subset the data to youth with $T = 1$ or $T = M$; (3) use this subsample of treatment groups with $T = 1$ and M to fit a GBM model to obtain estimates, $\tilde{p}_1(\mathbf{X}_i)$ of the $P(T = 1 \mid \mathbf{X}_i)$; (4) calculate the odds ratio $\mathrm{OR}_{i1} = \tilde{p}_1(\mathbf{X}_i)/(1-\tilde{p}_1(\mathbf{X}_i)$; (5) repeat steps 2–4 for the remaining treatments other than the holdout to obtain OR_{it} for $t = 2$ to $M - 1$; (6) set $\mathrm{OR}_{iM} = 1$; and (7) set $\widehat{p}_t(\mathbf{X}_i) = \mathrm{OR}_{it}/\sum_{j=1}^{M} \mathrm{OR}_{ij}$. These probabilities will sum to one by construction. The choice of using $t = M$ was arbitrary. Any of the treatment groups could be chosen as the holdout. However, the estimates could be sensitive to the choice of holdout – changing the holdout could yield different estimated probabilities. Another possible shortcoming to this method is that the GBM model is fit using the subsample of treatment groups t and M so that the model is not tuned for balancing group t to the entire population by default. Hence, we would need to test the balance of the desired samples separately from the automated fitting in routines such as the `twang` package in R . We prefer the alternative approach described earlier, even though it does not yield probabilities that sum to one, because that approach directly uses the powerful nonparametric modeling capabilities of GBM to obtain weights with good balancing properties and yields only one estimate rather than one for each possible holdout group.

4.2.3. Using GBM to estimate ATTs of multiple treatments. When estimating the ATTs for a treatment t', the goal is to estimate $\mu_{t',t''}$ by the weighted mean of the outcomes from the group receiving treatment t'' weighted to match the group receiving treatment t'. To achieve this goal requires balance between the weighted distributions of covariates for the $T = t''$ group and the distribution for the $T = t'$ group. Following Equation (4), we could set $w_i = \widehat{p}''_t(\mathbf{X}_i)/\widehat{p}'_t(\mathbf{X}_i)$ with the probabilities estimated using the algorithm for estimating weights for ATE. However, those weights were chosen to assure balance between the $T = t''$ group and the entire sample and the $T = t'$ group and the entire sample. The weights were not chosen to balance the weighted $T = t''$ group and the unweighted $T = t'$ group. We have found that tuning the GBM fits specifically for the estimation goal (i.e., tuning for ATE when estimating ATE or tuning for ATT when estimating for ATT) yields better balance and subsequent treatment effect estimates with smaller MSE. Hence, we do not recommend using ratios of probabilities

Copyright © 2013 John Wiley & Sons, Ltd.

estimated for estimating ATEs to estimate ATT. Rather, when estimating weights for ATTs, our method is to fit a GBM to the treatment indicator for $T = t'$ using only the subsample with $T = t''$ and $T = t'$ using the standard stopping rules for estimating ATT with a binary treatment. We then assign individuals in treatment group t'' the ATT weights that result from this binary fit. We repeat this procedure for all $t'' \neq t'$. This approach for estimation with more than two treatments directly takes advantage of the excellent balancing properties that GBM has been shown to have for estimating ATT weights for two treatment groups [17–19]. Because our target population is a particular treatment group (here t'), it is important to obtain weights that make every other treatment condition look like t'. To achieve this, we must find the weights that make the covariate distributions for each other treatment group match the distributions for the t' group. It is not necessary to have all conditions in the same estimation model; instead, a pointed comparison between t' and t'' for each t'' is sufficient and should yield ATT weights with relatively superior balance for each pairwise comparison. Intuitively, estimating ATT only involves data from two groups so we can repeatedly use the tools for estimation with two groups to obtain the desired causal estimates.

4.3. Assessing balance with multiple treatments

Although our proposed method for estimating propensity score weights for multiple treatments via GBM checks the balance of groups when fitting the GBM model for each treatment group, it is also important to have useful diagnostic criteria for assessing overall balance across the multiple groups. An overall summary of balance across groups will be particularly useful for describing the results of the causal analyses in presentations and publications. For both ATE and ATT estimations, we create overall summary measures of balance by taking the maximum of the balance metrics for each treatment. Hence, we summarize balance with $\text{PSB}_k = \max_t \text{PSB}_{tk}$ and $\text{PKS}_k = \max_t \text{PKS}_{tk}$ for ATE and $\text{TSB}_k = \max_t \text{SB}_{tk}$ and $\text{TKS}_k = \max_t \text{KS}_{tk}$ for ATT. We can then summarize metrics PSB_k, PKS_k, TSB_k, and TKS_k across covariates using either the maximum or the average by using before and after weighting plots like those shown in Figures 1–3 in Section 5.

4.4. Doubly robust estimation

Successfully reweighting the sample yields values of all the summary statistics (PSB_{tk} and PKS_{tk} or TSB_{tk} and TKS_{tk}), which are small for all covariates. For both sets of standard bias statistics (PSB_{tk} and TSB_{tk}), we consider values of less than 0.20 to be small, 0.40 to be moderate, and 0.60 to be large as with traditional standardized bias measures from the binary treatment setting. Covariates that remain imbalanced after weighting could potentially confound estimated treatment effects and pose a challenge for which no single solution will be best in all settings.

Sometimes, imbalance indicates separation between the groups on some dimension of the covariates, and restricting the samples to more comparable subsets can improve balance and yield accurate treatment effect estimates for a well-defined subset of the population. For instance, sometimes, one treatment group contains only a subset of the individuals found in other groups, and by conducting ATT rather than ATE analyses, balance may be more easily obtained. To illustrate using our case study, because MET/CBT-5 was specifically targeted for marijuana users rather than youth using other substances such as cocaine or opiates, it may be the case that if the community and SCY groups include many youth who use substances other than marijuana, it may be difficult to weight the MET/CBT-5 sample to match the entire population. Instead, focusing in on ATT for effects of MET/CBT-5 sample might result in better balance and unconfounded treatment effects.

When imbalances remain after weighting, another commonly used approach models the outcomes including the imbalanced covariates in the model. Model parameters including treatment effects are obtained by fitting a weighted regression. This approach is most advisable when the remaining imbalances are modest because the idea is that when the differences in the groups are modest, the model can control for them and any errors in the model will tend to be small. Estimating treatment effects through weighted regression on treatment indicators and covariates is a form of 'doubly robust' estimation, which combines a model for the outcome with weighting to obtain an estimator that yields consistent estimates of the treatment effect if either the model for the outcome or the propensity score model is correct but not necessarily both [31]. There are many different estimators that have the doubly robust property, one of which is weighted linear regression [31], and they can be used even if all the covariates balance to provide protection against possible errors in the propensity score model. Doubly robust estimation can

Copyright © 2013 John Wiley & Sons, Ltd.

also be more efficient than the simple weighted estimator [39]. Hence, some analysts might always combine weighting and modeling. However, concerns may arise when there are many covariates to include in the weighted regression model. It is likely that such cases will require variable selection, thereby undermining one of the key features of the propensity score approach: separating the modeling to control of confounding from the estimation of the treatment effect. For this reason, some analysts might prefer weighting alone.

Some papers have clearly shown that doubly robust estimation yield more accurate treatment effect estimates [39], but these have not considered multiple covariates with differing relationships with the outcome. We prefer to use weighted models alone or to control for pretreatment covariates that remain imbalanced after weighting in the regression model for the outcome(s). We also use doubly robust estimation, but we identify key covariates substantively (e.g., baseline values of the outcomes). Further, we prefer to control for these covariates in models for the outcome without additional variable selection. This is akin to the analysis of randomized trial data where a set of pretreatment covariates are specified in the design of the trial (before data are collected) and adjusted for in the model for the primary outcomes without additional variable selection.

4.5. *Effective sample sizes*

In general, weighted treatment effect estimates such as those defined in Equations (1) and (3) can have greater sampling variance than unweighted means from a sample of equal size. The effective sample size (ESS) of the weighted treatment group is a conservative way to capture the impact of this increase in variance on precision and power. Specifically, $\text{ESS}_t = \left(\sum_{i=1}^{N} T_i[t]w_i\right)^2 / \sum_{i=1}^{N} T_i[t]w_i^2$ for treatment t where $w_i = 1/\widehat{p}_t(X_i)$ if one is estimating ATEs and $w_i = W_i[t', t'']$ if one is estimating ATTs. The ESS_t provides a useful measure of the disparity in the weights for a treatment group's sample and the potential loss in precision from weighting. The ratio of the ESS_t to the number of observations in a treatment group sample equals the loss in precision because of weighting, if the outcome is uncorrelated with the weights [40]. If the outcome is correlated with the weights, then there may be no loss in precision because of weighting, or the loss may be much smaller than suggested by the ratio of the ESS_t to the sample size [40]. Regardless, very small values of the ESS_t relative to the sample size indicate a great disparity in the weights for the treatment group sample and typically that a small number of units receive very high weights relative to the majority of the units in the sample. Such a disparity in the weights could lead to unstable estimates dominated by a few cases and signal weak overlap among the groups. Furthermore, if two alternative GBM fits yield essentially equal balance but one yields a larger ESS_t than the other, then the fit yielding the larger ESS_t is preferred. Hence, ESS_t can be very useful for choosing among alternative models and assessing the overall quality of a model, even if it provides a possibly conservative picture of the loss in precision because of weighting.

5. Illustrative data example

The approach described earlier for estimating multiple propensity scores using GBM and assessing balance for different types of causal estimands is implemented in the `twang` package in R. We now illustrate our approach for estimating the different causal estimands (pairwise ATEs and ATTs) using this package and the CSAT's observational data on three different outpatient treatment approaches: community-based care, MET/CBT-5, and SCY. We first provide additional details about each treatment approach and the sample from each program. Then, we describe the observed pretreatment variables available for removing differences among the groups and the outcome of interest, which measures substance use frequency. We estimate the three propensity score models and finally estimate all pairwise ATE and ATT effects.

5.1. *Data and measures*

5.1.1. Study samples. We collected data for our example under three distinct discretionary grant programs administered by the SAMHSA's CSAT. Data from the three programs can be readily merged because each had a 12-month follow-up and utilized the same survey instrument to measure pretreatment covariates and outcomes.

The community sample in our analysis comes from the Adolescent Treatment Model (ATM) program, which was an observational study of 10 'exemplary' community-based care treatment programs funded

Copyright © 2013 John Wiley & Sons, Ltd.

in 1998–1999 to collect detailed survey information on clients at intake and 90-day intervals for 1 year post-intake to study the relative effectiveness of various models of adolescent substance abuse treatment [41]. Extensive details about the programs can be found in [42]. We only use data from the three outpatient programs of the ATM for our community sample.

The MET/CBT-5 sample comes from the Effective Adolescent Treatment (EAT) program, which was designed to disseminate the evidence-supported treatment to community-based treatment programs starting in 2003. The EAT included 37 community-based adolescent substance abuse treatment sites, all of which delivered MET/CBT-5. Details about the treatment and the EAT sites can be found in [43–45].

The SCY group comes from the same named program (Strengthening Communities for Youth), which implemented procedures meant to improve community collaboration. Specifically, in 2001–2002, CSAT initiated its SCY program to fund 12 communities to 'strengthen their drug and alcohol identification, referral, and treatment systems for youth' (SAMHSA/CSAT 2001 GFA No. TI 01-004); see [46] for more details. The current analysis is restricted to the eight SCY communities that collected data at baseline and 12 months post-intake using the Global Appraisal of Individual Needs (GAIN; see details in the following discussion) without extensive modifications and were willing to share their data for studies of the SCY program.

Providers in each of the three types of programs offer outpatient community-based care but with different practices and under different conditions: The ATM group (community) represents standard practice, the EAT group (MET/CBT-5) represents practices that are translating evidence-based protocols in community settings, and the SCY group represents practices within an enhanced community setting with strengthened ties among its various drug and alcohol identification, referral, and treatment systems. For all three programs, clients were interviewed using the same biopsychosocial instrument (GAIN [47]) at intake and for a year thereafter.

5.1.2. Pretreatment measures. All pretreatment measures come from the GAIN. The GAIN has eight sections assessing background and demographic characteristics (baseline only), substance use, physical health, risk behaviors, mental health, environment, legal, and educational/vocational problem areas. It includes over 100 symptom counts, change scores, and service utilization indices, most with Cronbach's alphas over 0.85 [47]. A test–retest analysis of key GAIN substance use and substance use problem indices in a sample of 210 adolescents indicated good reliability over a 90-day interval ($r > 0.72$ for each index) [46]. We include a total of $K = 23$ pretreatment measures (Table III), which have been shown in previous analyses to explain at least 1% of the variance in frequently used outcomes for evaluating substance abuse treatment programs [43]. Missing values on the 23 pretreatment variables were low (mean $= 0.73\%$ and max $= 4.85\%$) and controlled for in our propensity score models, which include missing value indicators as well as the observed values of each pre-treatment variable (details in following section).

5.1.3. Outcome. The Substance Frequency Scale is an 8-item scale that assesses the average proportion of alcohol and other drug using days in the past 90 taking into account heavy use and problem days. Higher scores indicate increased frequency of substance use in terms of days used, days staying high most of the day, and days causing problems. Inter-item reliability was good ($\alpha = 0.80$) in the ATM dataset.

5.2. Estimating causal effects

Table II lists the steps we use when estimating the causal effects of multiple treatments using weighting. The following subsections highlight implementation of each step using the community, MET/CBT-5, and SCY treatment programs. As noted earlier, $t = 1$ denotes community, $t = 2$ denotes MET/CBT-5, and $t = 3$ denotes SCY.

Step 1. Estimating propensity scores

When interested in ATE

Suppose we are interested in estimating the pairwise ATEs. In this case with three treatments, there are three possible ATEs of interest (MET/CBT-5 vs. Community, $\mu_2 - \mu_1$; SCY vs. Community, $\mu_3 - \mu_1$; and SCY vs. MET/CBT-5, $\mu_3 - \mu_2$). To estimate all three ATEs, the first step of the analysis fits three binary GBMs, one for each treatment program indicator (community, MET/CBT-5, or SCY). Each

Copyright © 2013 John Wiley & Sons, Ltd.

Table II. Steps in estimating effects of multiple treatments.

1. Estimate propensity scores for each treatment by running `mnps()` function in `twang`
2. Assess balance using desired balance metric(s)
3. Assess overlap among treatment samples
4. Estimate the targeted population mean for each treatment
   - 4.1 Set weights equal to the reciprocal of the propensity scores for ATE or the ratio of treatment to target treatment propensity scores for ATT
   - 4.2 Calculate the weighted mean for each treatment
5. Estimate causal treatment effects
   - 5.1 If all covariates are balanced, use difference in weighted means to estimate treatment effects
   - 5.2 If some covariates remain imbalanced, estimate effects via a doubly robust procedure such as fitting a weighted regression model that includes unbalance covariates

GBM uses an ATE stopping rule that selects the iteration that yields optimal balance defined by PSB_{tk} and/or PKS_{tk}. We use the mean and maximum standardized bias stopping rules when exploring fits for this analysis.

When interested in ATT

To illustrate what happens when estimating ATT weights, we begin by assuming that it is of interest to estimate ATT for youth like those who were in the MET/CBT-5 condition. In this case, there are two ATTs of potential interest (MET/CBT-5 vs. Community, $\mu_{2,2} - \mu_{2,1}$; and MET/CBT-5 vs. SCY $\mu_{2,2} - \mu_{2,3}$ for youth like those receiving MET/CBT-5). Therefore, the analysis begins by creating two subsamples from the overall dataset: One that contains youth in the MET/CBT-5 and community groups, and the other that contains youth in the MET/CBT-5 and SCY groups. To estimate the weights for estimating the ATT of MET/CBT-5 relative to community among youth like those who received MET/CBT-5 ($\mu_{2,2} - \mu_{2,1}$), we fit a binary GBM propensity score model for the MET/CBT-5 treatment indicator to the pooled sample with youth from the MET/CBT-5 and community samples. We use an ATT stopping rule that selects the iteration of GBM, which yields optimal balance defined by TSB_{tk} and/or TKS_{tk}. From this model, we obtain estimates of the probability that each youth in the subsample received MET/CBT-5 rather than community care. The ATT weights equal one for youth in the MET/CBT-5 sample, and it equals the ratio of the propensity score to one minus the propensity score (the odds of receiving MET/CBT-5 rather than community care) for youth in the community sample. These are the same weights we would use if there were only two treatments. In essence, we are estimating the pairwise treatment effects for multiple treatments by repeatedly using the methods for causal effects for cases with just two treatments.

We then repeat the GBM estimation procedure of using the sample with youth from the MET/CBT-5 and SCY groups to estimate the ATT of MET/CBT-5 relatively; SCY for youth like those receiving MET/CBT-5, $\mu_{2,2} - \mu_{2,3}$. Again, we use the binary GBM propensity score model for the MET/CBT-5 treatment to estimate the probability of receiving MET/CBT-5 rather than SCY. The weights are one for youth in the MET/CBT-5 sample and the odds of receiving MET/CBT-5 rather than SCY for the youth in the SCY sample.

We use a similar process to estimate ATT for youth like those receiving community care and ATT for youth like those receiving SCY. For expository purposes, we estimate all the pairwise ATT estimates using each of the three treatments to define the target population. Applications might usually focus on only a single target population.

We used the `mnps()` function in `twang` to estimate both the ATE and ATT weights for our case study. The function automates the propensity score and weight estimation process by running the GBM fitting algorithm for many iterations and selecting the iteration to minimize the user-specified stopping rule. It also produces weights from the selected model and repeats all the steps for all the treatment groups.

Step 2. Assess balance using desired balance metric(s)

Using ATE weights

Table III shows for each pretreatment variable the unweighted and ATE weighted means for each treatment program and the unweighted overall population mean and standard deviation. Cells marked with

Copyright © 2013 John Wiley & Sons, Ltd.

Table III. Means for treatment groups (unweighted and ATE weighted) and the population (pooled sample, unweighted).

	Unweighted means MET/			Weighted means MET/			Population	
Pretreatment covariate	Comm.	CTB-5	SCY	Comm.	CTB-5	SCY	Mean	SD
Demographics (race/ethnicity)								
Percent non-Hispanic white	0.67*	0.51	0.35*	0.59*	0.48	0.46	0.48	0.50
Percent African American (nonH)	0.14	0.08*	0.32*	0.12	0.14	0.17	0.16	0.37
Percent Hispanic	0.10*	0.24	0.2	0.2	0.22	0.22	0.21	0.41
Percent other	0.09	0.17	0.14	0.1	0.16	0.16	0.15	0.36
Substance use								
In recovery	0.20	0.25	0.18	0.19	0.23	0.22	0.23	0.42
Substance frequency scale	0.14	0.11	0.15	0.14	0.12	0.13	0.13	0.14
Subs prob scale (past month)	3.11	2.71	3.36	3.15	2.89	2.95	2.96	3.56
Subs prob scale (past year)	7.93*	6.3	7.09	7.82*	6.55	6.66	6.73	4.39
Subs dep scale (past month)	1.02	0.89	1.2	0.94	0.96	1.02	1.01	1.66
Subs dep scale (past year)	3.10*	2.33	2.69	3.15*	2.44	2.48	2.54	2.29
Days drunk/high (past 90 days)	13.78	11.42	18.44	15.2	13.33	13.87	14.06	23.13
Criminal activities								
Internal mental distress scale	7.72	7.82	9.33	8.35	8.13	8.51	8.3	8.71
Problem orientation scale	0.99	0.69	1	1.09	0.76	0.82	0.82	1.59
Emotional problems scale	0.28*	0.21	0.24	0.25	0.22	0.23	0.23	0.19
Behavioral complexity scale	12.26*	9.81	10.89	11.8	10.28	10.26	10.44	7.99
Previous mental health trt	0.45	0.36	0.41	0.42	0.38	0.41	0.39	0.49
Mental health trt scale	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.05
Environmental risk								
Social risk scale	13.84	12.31*	13.39	13.89	12.57	12.61	13.46	4.55
Crime environment scale	0.24*	0.05	0.13	0.13	0.07	0.09	0.1	0.23
Environmental risk scale	31.43*	35.47	37.23	33.97	35.79	35.72	35.75	9.41
Living in house/apartment	0.85	0.88	0.82	0.88	0.87	0.86	0.86	0.35
Living in jail/correctional facility	0.11	0.05	0.06	0.06	0.05	0.05	0.06	0.24
Other living situation	0.04	0.06	0.13	0.05	0.08	0.08	0.08	0.27
Training activity scale	0.55*	0.66	0.57*	0.61	0.63	0.62	0.64	0.26
N or ESS (weighted)	444	2459	1351	101.17	2113.64	1025.14	4254	

Cells marked with an * denote pretreatment covariates for which PSB_{tk} is greater than 0.20 within a given program.

an asterisk denote pretreatment covariates for which PSB_{tk} is greater than 0.20 within a given program. Table III also highlights the ESSs for the community, MET/CBT-5, and SCY programs after the ATE weights were applied.

Community. Before weighting, youth in the community program were very different from the population; PSB_{1k} was greater than 0.20 for 13 of the 24 pretreatment covariates. Specifically, youth in the community program had higher mean behavioral complexity, crime violence, emotional problems, illegal activities, controlled environment, criminal justice system, and substance frequency, problems, and dependence scales than the population, yet lower mean environmental risk and training activity scales than the population average. They also had higher mean number of days out of the past 90 spent in an institutionalized setting, a greater percentage of whites, and a lower percentage of Hispanics than the population. After weighting, the values of PSB_{1k} were generally attenuated for all pretreatment covariates and only five values greater than 0.20. Specifically, youth in the community program still had higher mean illegal activities and controlled environment scales, a greater percentage of whites, and lower mean environmental risk scales than the population average after weighting.

As expected, the community sample took a large hit with respect to ESS after weighting (originally $N = 444$, whereas ESS after ATE weighting $= 101.17$). This results from the fact that the distribution of pretreatment covariates in the community was very different from the overall population.

Copyright © 2013 John Wiley & Sons, Ltd.

MET/CBT-5. Youth in the MET/CBT-5 program only had two pretreatment covariates with PSB_{2k} greater than 0.20 prior weighting. Specifically, MET/CBT-5 youth had lower mean social risk scale than the population average and a lower percentage of African Americans. Weighting removed both these differences. Also, because the MET/CBT-5 sample was so similar to the overall population, the weights were not highly differential, and the ESS had only a minor reduction compared with the unweighted sample: $N = 2459$ versus the ESS after ATE weighting equal to 2113.64.

SCY. The pretreatment variables of youth in the SCY also were very similar to those of the overall population prior to weighting. Only the percentage of African Americans and whites and training activity scale were imbalanced prior to weighting. Weighting removed these differences. As with the MET/CBT-5, the similarity of the means of pretreatment variables for SCY and overall samples prior to weighting resulted in an ESS that differed very little from the actual sample size: $N = 1351$ versus the ESS after ATE weighting equal to 1025.14.

Using ATT weights

Figures 1–3 contain a series of plots for assessing the balance between groups on pretreatment variables before and after ATT weighting. For each treatment, there are two rows of panels, one for comparing it with each of the other two treatments. For instance, Figure 1 is ATT when community is the target population. In this figure, the top row compares the community and MET/CBT-5 samples before and after the MET/CBT-5 sample is weighted to match the distribution of pretreatment variables in the community sample. The left panel shows the results using the mean standardized bias stopping rule, and the right panel shows the results using the maximum standardized bias stopping rule. The bottom row compares the community and the SCY samples when the SCY sample is weighted to match the community sample. Each point in a panel represents the SB_k or effect size difference between samples for a pretreatment variable. The left column of points in a panel are the SB_k for the unweighted samples, and the right column of points are the values when the comparison treatment is weighted to match the community sample. Lines connect the values for the same pretreatment variable before and after weighting. A closed red circle indicates a covariate for which the difference between treatment group means is statistically significant. An open red circle and red line identify a variable for which the standardized bias or effect size increases with weighting; a blue line identifies a variable for which balance improves with weighting. Figures 2 and 3 show analogous results for ATT weights when the 'treated' population is MET/CBT-5 or SCY, respectively. A corresponding table with the ATT weighted means for each comparison is available in the Supporting information.

Community. In both rows of Figure 1, the samples are very different prior to weighting with many solid red dots greater than 0.20. The youth in both the MET/CBT-5 and SCY samples differ from the youth in the community sample on many pretreatment variables, and this is reflected in the panels of the figure. After weighting, all of the large differences between the means of the pretreatment variables between community and weighted MET/CBT-5 samples are greatly reduced, but five remain greater than 0.20. This holds regardless of the stopping rule used to select the GBM for the propensity scores. The balance is nearly invariant to the stopping rule. There are also large differences between the community and SCY samples prior to weighting, although there are fewer of these than for MET/CBT-5. Weighting again reduces these greatly, but three remain above 0.20.

As expected, when using ATT weights that aim to make youth in the MET/CBT-5 and SCY programs look like community youth, we see a large reduction in ESS. For both programs, the ESS is 85% lower than the unweighted sample sizes for these two programs (originally $N = 2459$ and 1351 for MET/CBT-5 and SCY, respectively; ESS after ATT community weighting = 379.96 and 199.17, respectively).

MET/CBT-5. The top row of Figure 2 compares the MET/CBT-5 and community samples before and after the community sample is weighted to match the MET/CBT-5 sample. The samples differ substantially on many variables. The SB_k are larger in this comparison than when estimating ATT for the community sample because the variance of the pretreatment variables is smaller in the MET/CBT-5 sample than in the community sample yielding denominators of the standardized biases that are smaller for ATT on MET/CBT-5. Weighting nearly removes all the difference: Only two remain greater than 0.20. The great reduction in SB_k that results from weighting suggests that a subset of the community

Copyright © 2013 John Wiley & Sons, Ltd.

sample is similar to the MET/CBT-5 sample and, when they are up-weighted relative to the other youth in the community sample, the two groups are similar in terms of the pretreatment variables. However, this subsample is small. When weighting community youth to look like MET/CBT-5 youth, we have a 90% drop in the ESS after weighting: ESS in community is 42.19 after weighting, and the sample size is 444. As shown in the bottom row of Figure 2, the MET/CBT-5 and SCY samples are fairly similar prior to weighting with the largest standardized bias being just greater than 0.4. Weighting makes the differences very small. Again, all the results are relatively invariant to the stopping rule. The loss in ESS for the SCY group is about 43% (ESS of 769 for the sample of 1351 youth), which is large given the relative similarity of the pretreatment variables in the SCY and MET/CBT-5 samples priors to weighting.

SCY. Figure 3 shows the results for weighting community and MET/CBT-5 samples to match the SCY sample. As noted previously, the community sample differs from the SCY sample, and this is again reflected in the large values for some of the SB_k in the left-hand side columns of the two panels in the top row of the figure. Weighting reduces all the large differences, but the three variables with largest differences prior to weighting, race, illegal activities, and the environmental risk scale have the largest differences after weighting. The $SB_k > 0.2$ for each of these variables. The SCY sample has fewer white youth with low environmental risks and high rates of criminal activities than the community sample and weighting cannot fully correct these differences. Conversely, the MET/CBT-5 sample is similar to the SCY sample before weighting, and with weighting, the MET/CBT-5 sample matches the SCY sample very well with no SB_k greater than 0.15. There is a very large loss in ESS for the community group when weighted to match the SCY sample and 55% loss for the MET/CBT-5 sample.

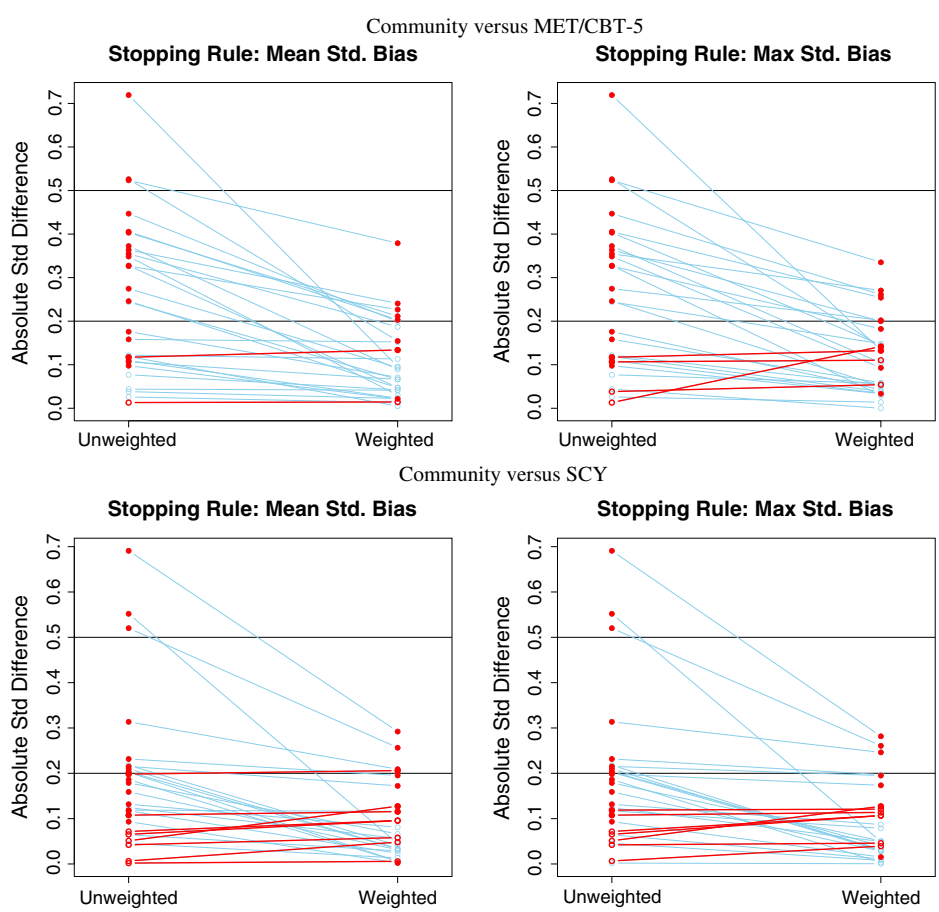

Figure 1. Effect size plots for assessing the balance of pretreatment variables on youth like those receiving Community care for estimating pairwise ATT effects for Community treatment.

Copyright © 2013 John Wiley & Sons, Ltd.

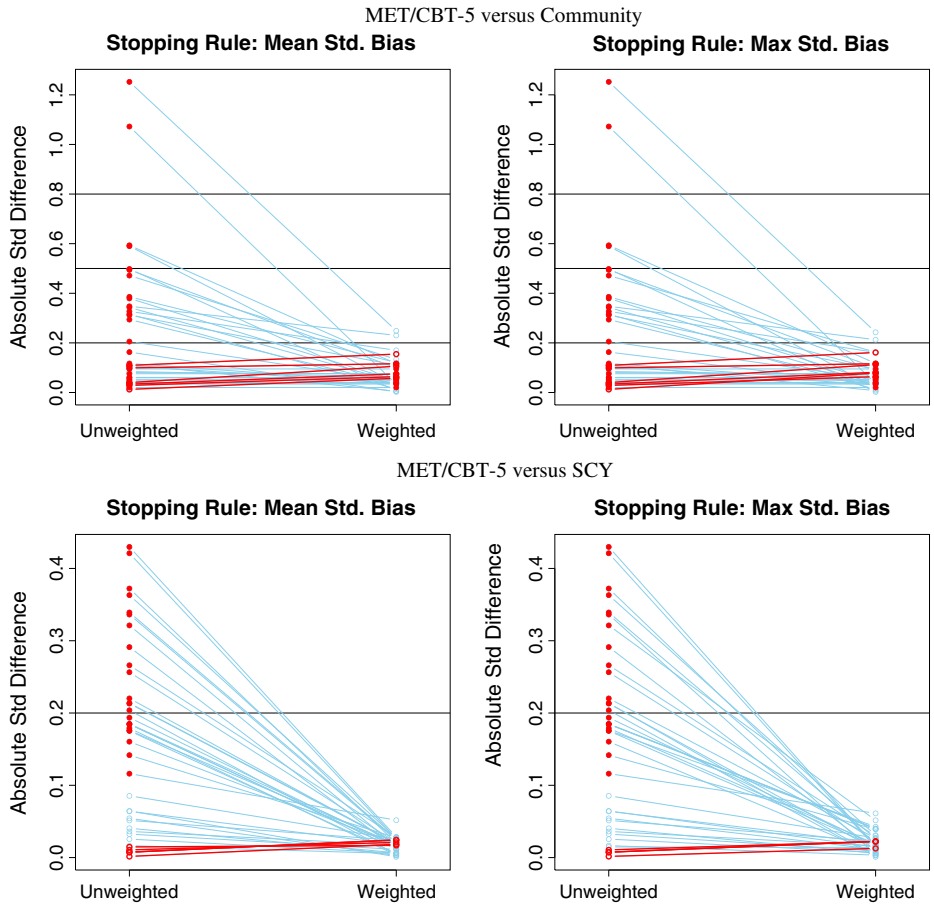

Figure 2. Effect size plots for assessing the balance of pretreatment variables on youth like those receiving MET/CBT-5 for estimating pairwise ATT effects for MET/CBT-5.

Step 3. Assessing overlap

As discussed in Section 3.1, weighted treatment group means yield consistent estimates of the population means of the potential outcomes μ_t, for $t = 1, 2$, or 3, provided there is overlap among the groups such that every youth could have received each treatment and there are no values of the pretreatment variables that occur only in one of the treatment conditions. It is difficult to visually assess overlap in high dimensions. Consequently, a common approach to assessing overlap compares the distributions of the estimated propensity scores across the treatments. We have found that box plots work well for comparing the distributions of propensity scores. To test the overlap for ATE, we estimate the propensity score model for each treatment, community, MET/CBT-5, and SCY, and calculate the propensity scores for every youth in the sample regardless of the youth's actual assignment. We then plot the distributions of estimated propensity scores using a separate box plot for the youth receiving each type of treatment. For example, we first estimate the propensity score model for community treatment using GBM. We then evaluate these propensity scores for youth in community, MET/CBT-5, and SCY samples, and plot the distributions of values for each of the three groups using side-by-side box plots. We repeat these steps for the other two treatments. Figure 4 presents the three sets of three box plots of the distributions of propensity scores with scores for community in the top panel, MET/CBT-5 in the center panel, and SCY in the bottom panel.

There are no specific rules for what constitutes sufficient overlap, but substantial overlap of the box plots is desirable (i.e., the panels for MET/CBT-5 and SCY), and great separation of the plots (the panel for community) is typically considered insufficient. However, separation in the propensity score box plots does not necessarily imply that weighting cannot balance the pretreatment variables, and we have even found examples where tuning the GBM model to achieve greater balance in the covariates resulted greater separation of the propensity score box plots.

Copyright © 2013 John Wiley & Sons, Ltd.

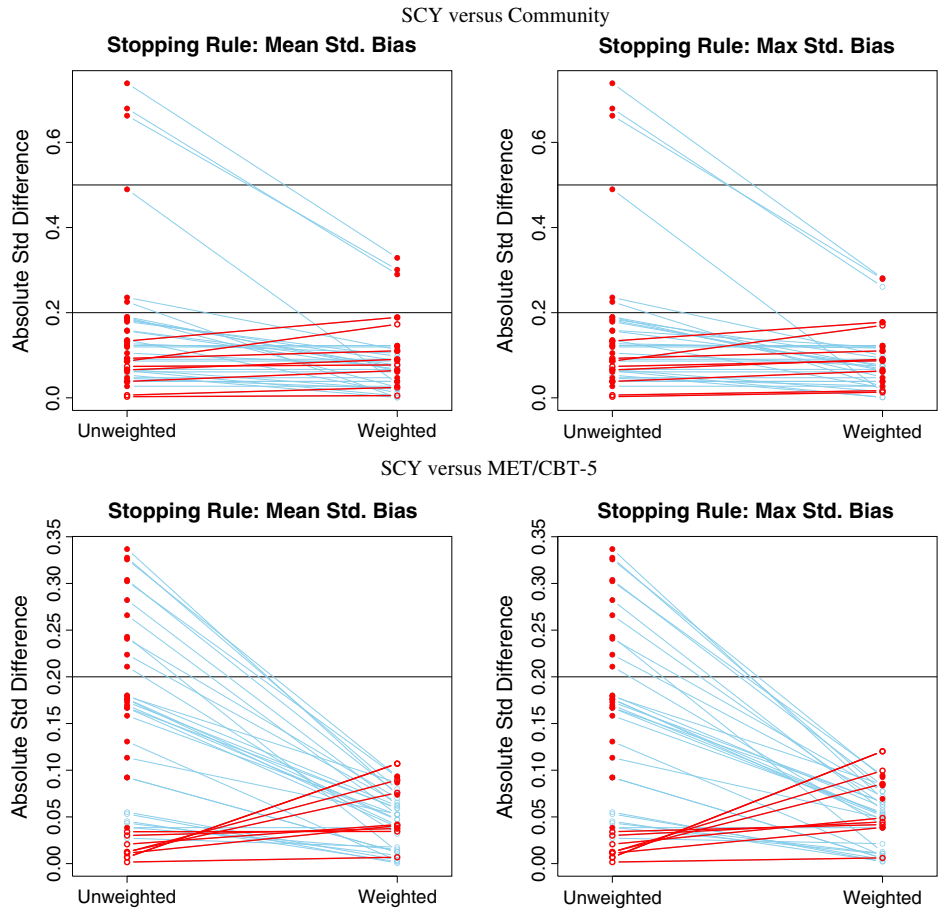

Figure 3. Effect size plots for assessing the balance of pretreatment variables on youth like those receiving SCY for estimating pairwise ATT effects for SCY.

Generally, we use a combination of the overlap plot and the balance table to assess whether the groups are sufficiently similar to support causal estimation of the treatment estimands. The lack of overlap and poor balance suggests that the groups may be sufficiently distinct so as not to support causal estimation. For example, the overlap between the community and other groups is poor (see the top panel of Figure 4) and, as discussed previously in step 2, imbalance between the community group and the overall population remains on several variables including race and measures of substance use problems and criminal activity. White youth with high rates of criminal activity and substance use problems and symptoms of dependence the year before treatment intake appear in the community sample but not in the other samples. For these youths, we might not be able to use the other samples to estimate their counterfactual outcomes following MET/CBT-5 or SCY instead of community treatment. We might need to consider focusing only on ATT for the other groups or subsetting the population in other ways. As the figures show, we have good overlap when focusing on MET/CBT-5 or SCY. For the purposes of the tutorial, we will continue with the estimation of the pairwise ATE and ATT effects, but we need to be cautious in our interpretation of the ATE estimates.

Steps 4 and 5. Estimating means and causal treatment effects

This section combines steps 4 and 5 in Table II to illustrate how one would estimate ATE and ATT causal estimands.

Pairwise ATEs

The estimate of the population mean had all youth received community care (μ_1) equals the weighted mean of youth in the community care sample where the weight for each youth is the reciprocal of the estimated propensity for receiving community care. We call this estimate $\widehat{\mu}_1$. Only the outcomes of youth

Copyright © 2013 John Wiley & Sons, Ltd.

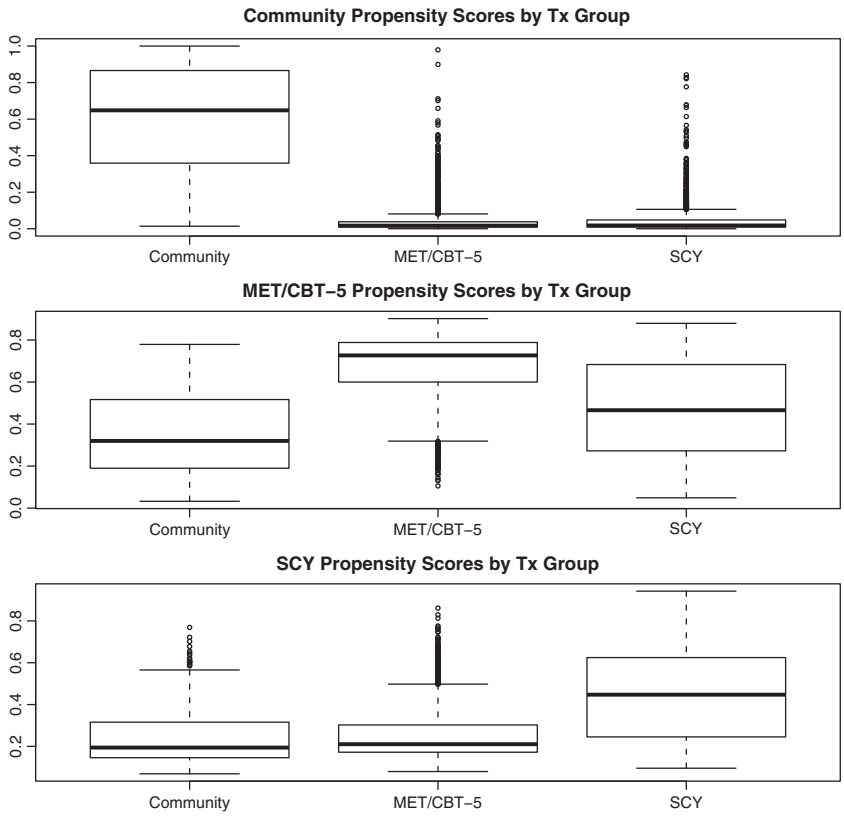

Figure 4. Overlap assessment. Each panel presents box plots by treatment group of the estimated propensity scores for one of the treatments, $\text{pr}(T[t] = 1 \mid \mathbf{X})$ for ever youth in the sample. The top panel presents Community ($t = 1$), the middle panel presents MET/CBT-5 ($t = 2$), and the bottom panel presents SCY ($t = 3$).

in the community sample are used to estimate this mean because only these youth received community care. Their weights depend only on their propensity scores for receiving community care. Similarly, the estimate of the population mean had all youth received MET/CBT-5 equals the weighted mean of youth in the MET/CBT-5 sample where the weight for each youth equals the reciprocal of the estimated propensity scores for receiving MET/CBT-5, and the estimate of the population mean had all youth received SCY equals the weighted mean of youth in the SCY sample where the weight for each youth equals the reciprocal of the estimated propensity score for receiving SCY. We call these estimates $\widehat{\mu}_2$ and $\widehat{\mu}_3$, respectively. We estimate the pairwise ATE by differences of the estimated population means: the average treatment effect of community care relative to MET/CBT-5 equals $\widehat{\mu}_1 - \widehat{\mu}_2$; the average treatment effect of community care relative to SCY equals $\widehat{\mu}_1 - \widehat{\mu}_3$; and the average treatment effect of MET/CBT-5 relative to SCY equals $\widehat{\mu}_2 - \widehat{\mu}_3$.

Alternatively, we can pool data from three samples and estimate the ATE and population means through a weighted linear regression of substance use frequency on indicator variables for two of the three treatment groups. For each youth, the weight is the reciprocal of the propensity for the treatment the youth received: The weights for youth in the community care sample equal the reciprocal of their propensity scores for receiving community care; the weights for youth in the MET/CBT-5 sample equal the reciprocal of their propensity scores for receiving MET/CBT-5; and the weights for youth in the SCY sample equal the reciprocal of their propensity scores for receiving SCY. There is one weight variable, and its value depends on each youth's treatment status. If we include indicators for MET/CBT-5 and SCY, then the coefficients on the indicator variables estimate the ATE of MET/CBT-5 relative to community ($\mu_2 - \mu_1$) and SCY relative to community ($\mu_3 - \mu_1$). The estimate of the ATE of MET/CBT-5 relative to SCY ($\mu_2 - \mu_3$) is estimated by the difference of the coefficient for the indicators (the coefficient for the MET/CBT-5 indicator minus the coefficient for the SCY indicator). When fitting the weighted regression, analysts should use a software package that accounts for weighting when estimating standard errors such as the `survey` package in R or the SURVEYREG procedure in SAS or by specifying the weights as analysis weights in STATA. The intercept is the estimate of the population mean for community, and

Copyright © 2013 John Wiley & Sons, Ltd.

Table IV. Treatment group means and pairwise ATEs and ATTs for substance frequency scale (×100) before and after weighting.

	Community	MET/CBT-5	SCY	Difference (95% CI)
Unweighted				
MET/CBT-5 vs. Community	11.4	6.7		−4.7 (−5.9, −3.5)
SCY vs. Community	11.4		7.7	−3.8 (−5.1, −2.4)
SCY vs. MET/CBT-5		6.7	7.7	1.0 (0.1, 1.8)
ATE weighted				
MET/CBT-5 vs. Community	8.7	7.5		−1.1 (−3.0, 0.8)
SCY vs. Community	8.7		7.7	−1.0 (−3.0, 1.0)
SCY vs. MET/CBT-5		7.5	7.7	0.1 (−0.8, 1.0)
ATT weighted to match community sample				
Community vs. MET/CBT-5	11.5	8.6		2.9 (0.7, 5.2)
Community vs. SCY	11.5		8.8	2.8 (0.4, 5.1)
ATT weighted to match MET/CTB-5 sample				
MET/CBT-5 vs. Community	8.5	6.8		−1.7 (−4.5, 1.1)
MET/CBT-5 vs. SCY		6.8	7.4	−0.6 (−1.5, 0.3)
ATT weighted to match SCY sample				
SCY vs. Community	9.0		7.6	−1.4 (−3.5, 0.7)
SCY vs. MET/CBT-5		8.3	7.6	−0.7 (−1.8, 0.5)

the intercept plus the coefficient on the dummy indicator variables for each of the other treatments equal the estimates of their population means.

Although estimating weighted means or fitting a weighted regression yields the same ATE estimates, the regression approach has two advantages. First, it provides standard errors that can be used for confidence intervals and hypothesis testing. Second, the regression approach naturally extends to doubly robust estimation in the form of weighted regression that includes treatment indicators and covariates. In this doubly robust approach, the treatment effects still can be estimated as the difference between pairs of coefficients on the treatment group indicator variables. To obtain estimates of the population means from a model that controls for covariates, we need to estimate the marginal means [48]. To do this, we center the covariates at their unweighted overall sample means and then use the same combinations of coefficients used in the model without covariates to estimate the various population means.[¶]

The Supporting information provides a sample code for estimating the pairwise ATE and population means from models with and without additional covariates. Table IV shows the unweighted and weighted regression results for the substance frequency scale at 12 months.

MET/CBT-5 versus Community. There are significant differences between the unweighted outcomes for youth in both MET/CBT-5 and youth receiving the traditional community care, with youth receiving community care having more frequent drug use at the follow-up than the youth in the alternative sample. However, as previously discussed, there are substantial pretreatment differences among groups, so we should not interpret these differences as causal effects. When we weight the sample to control for pretreatment differences and estimate the effects of treatment for the pooled sample, we find that the estimated effects are much closer to zero than the raw differences and they are not statistically significant.

SCY versus Community. Again, there were significant differences between these two groups before weighting, but the differences are much smaller and not significant when weighted.

SCY versus MET/CBT-5. The substance use frequency at follow-up for the youth from the MET/CBT-5 and SCY sample is very similar before weighting. Weighting has little effect on the difference between these two groups because the two samples are very similar to the population on pretreatment variables prior to weighting.

[¶]*For dichotomous outcomes modeled by logistic regression, centering the covariates is not sufficient to yield marginal means, and the methods of Graubard and Korn* [48] *must be used to estimate the population means.*

Copyright © 2013 John Wiley & Sons, Ltd.

Pairwise ATTs

To estimate the ATT of MET/CBT-5 relative to community care for youth like those receiving community care, we first subset the data to the MET/CBT-5 and community samples. We estimate the mean for youth like those receiving community care had they received MET/CBT-5, $\mu_{1,2}$, by the ATT weighted mean of the MET/CBT-5 sample. The mean for youth like those receiving community care had they received community care, $\mu_{1,1}$, equals the unweighted mean for the community care sample. We can again estimate the means and treatment effect using a weighted regression of substance use frequency on the community indicator from the combined community and MET/CBT-5 samples. There is one weight variable equal to one for youth in the community sample and equal to the ATT weight for youth in the MET/CBT-5 sample. The coefficient on the community indicator equals the estimated ATT ($\mu_{1,1}-\mu_{1,2}$). The means for the groups are estimated by the intercept (MET/CBT-5) and the sum of the intercept and coefficient on the community indicator (community).

As with ATE, we can include covariates in the model for the outcome in our weighted regression to obtain doubly robust estimates of the ATT. To obtain an estimates of $\mu_{1,2}$, we center the covariates at the mean of the community sample and use the intercept to estimate the counterfactual mean just as we did in the model without covariates. We can use the unweighted mean for the community sample to estimate $\mu_{1,1}$.

We can repeat this approach using the SCY sample rather than the MET/CBT-5 sample to estimate the relative effects of SCY and community care for youth like those who receive community care. We can use a similar approach to estimate the ATTs for youth like those receiving MET/CBT-5 and SCY.

We suggest fitting two separate models, one for each pairwise ATT estimate. However, analysts can pool datasets used to estimate the two sets of ATT weights, taking care not to duplicate the community sample, and fit a single model with indicator variables for MET/CBT-5 and SCY. The coefficients on the indicator variables for MET/CBT-5 and SCY estimate $\mu_{1,2} - \mu_{1,1}$ and $\mu_{1,3} - \mu_{1,3}$ and will equal the estimates from the two separate models (except for the change in sign on the estimates) when no covariates are included in the model. When the model includes covariates, fitting models separately for each comparison and pooling may yield different treatment effect estimates because of differences in the estimates of the coefficients on the covariates.

Table IV contains all the pairwise ATT estimates.[‖]

Community. The results in Table IV indicate that among youth like those who are treated in the community setting, community treatment is less effective than MET/CBT-5 or SCY. The estimated effects of community versus MET/CBT-5 or SCY for youth like those in the community sample are −0.029 (CI: −0.052 to −0.007) and −0.028 (CI: −0.051 to −0.004), respectively, after weighting and controlling for unbalanced pretreatment variables for each comparison. Both estimates are statistically significant. This could suggest that community treatment is poorly suited for the youth who typically receive treatment in this setting. However, the groups are very different before weighting, a few sizeable differences remain after weighting, and the overlap between the groups is weak. Hence, it is possible that the adjustments have not fully removed the confounding of the estimated causal effects by pretreatment variables.

MET/CBT-5. Among youth who received MET/CBT-5, this treatment approach does not appear to have beneficial effects relative to either of the other treatments. The estimate of the mean substance use frequency for youth who received MET/CBT-5 had they instead received the community treatment (i.e., the ATT weighted mean for the community sample) is 8.5, which is notably higher than the estimated 6.8 for these youths outcomes following MET/CBT-5. However, the difference is imprecise and not significant in large part because the effective community sample size is so low after ATT weighting of the community youth to match the distribution of background variables of youth receiving MET/CBT-5. The weighted mean of the outcomes from the SCY sample is very close to the mean for the MET/CBT-5 sample and not significant.

[‖]*Because of very small amounts of missing data on some of the pretreatment variables used for the modeling and for predicting means, the mean for the community care for ATT with community as the target does not equal the unweighted mean for the entire community sample and likewise for MET/CBT-5 and SCY.*

Copyright © 2013 John Wiley & Sons, Ltd.

SCY. The results of the pairwise ATT analyses targeting youth in the SCY sample parallel those of ATT for MET/CBT-5. There are notable differences between the weighted community mean and the SCY mean, but the difference is not significant. The estimated effect of SCY relative to MET/CBT-5 on the SCY sample is very small, again, indicating no differences in these two types of treatments.

5.3. *Comment on the use of multinomial logistic regression to estimate weights*

We also used multinomial logistic regression models to estimate propensity scores and weights for the case study data to examine the balance it might provide as an alternative to using GBM. We estimated the parametric model-based propensity score weights using the same set of pretreatment covariates that were used in our GBM models; however, the multinomial model only includes main effects of the pretreatment covariates, which is a common practice. We did not tune the model further (e.g., by trying various two-way or three-way interactions) because it was not clear which interactions would fix the imbalances shown in the tables, particularly for the pairwise ATEs. Tables 1 and 2 in the Supporting information highlight that in general, the multinomial logistic regression without cumbersome manual model selection did not yield superior weights to our GBM approach in this case study. The ATE weights that resulted from fitting the multinomial logistic model were very unstable and yielded very poor balance. Trimming the weights [19] or removing records with outlier weights improved balance, but it remained inferior to the balance obtained using GBM. For ATT, the methods were more comparable, but GBM continued to provide better balance. We note that 10% of the records in the data had missing values on one or more covariates. GBM automatically adds indicators for missing values and includes them in the model. Multinomial logistic regression requires a manual fix, such as imputation, creating missing data indicators, or dropping incomplete records. When applying the parametric approach, we dropped incomplete records for this illustrative analysis.

6. Discussion

This tutorial aimed to provide practical guidance for researchers on how to utilize propensity score weights when estimating causal treatment effects of multiple treatment conditions. In estimating the multiple treatment propensity score weights, a powerful machine learning method, GBM, was used to obtain robust propensity score weights with better balance properties than a simple parametric model (namely the multinomial logistic) did.

As shown in our example, the use of weighting can improve imbalances when interest lies in comparing more than two treatment programs, thereby allowing researchers to make more robust inferences when estimating treatment effects. Before weighting, the community means differ significantly from the other groups. After weighting, there were no differences between the ATE weighted MET/CBT-5 and SCY samples and the population, and there were no pairwise ATE or ATT effects between these two treatment programs. However, the weights were unable to completely remove differences among the distributions of pretreatment variables between the community and other samples, and lingering imbalances still existed after ATE weighting on five pretreatment covariates. In light of this, we utilized a doubly robust modeling strategy to estimate the ATEs, which controlled for not only the propensity score weights but also the five pretreatment variables for which imbalance still existed after weighting. This method yielded no statistically significant pairwise ATE estimates. Similarly, there were remaining imbalances with ATT weighting between the community sample and both of the other samples. Again, we used a form of doubly robust estimation to estimate the pairwise ATT effects, which were significant and suggested that community treatment is less effective than either of the alternatives for youth like those who receive community care. However, the large differences among groups even after weighting could result in residual confounding even with the doubly robust estimation, especially given the weak overlap between the community and other samples.

The causal modeling of the case study has desirable properties. First, the causal estimands were clearly defined prior to any analysis of the data. Then, transparent estimators of the estimands were defined and estimated. The appropriateness of the samples for creating the counterfactuals necessary for causal effect estimation was assessed via multiple graphical and tabular methods using only pretreatment variables. The quality of the controls for confounding were also assessed graphically and through the balance tables. Finally, the effects were estimated using methods that rely on linear models and weighting to reduce the risk of confounding by observed pretreatment variables.

Copyright © 2013 John Wiley & Sons, Ltd.

When the groups are very dissimilar, as they were for the community sample and either of the other groups in the example, weighting cannot always remove all differences among the groups. The inability to achieve balance is particularly likely when the number of pretreatment variables is very large relative to the overall sample size. But the tools presented here and automated in the `twang` package in R make diagnosing such problems very easy. Also, as we saw in our example, when the groups are disparate, weighting can be very inefficient because most of the weight is applied to very few cases and most of the sample receives very little weight, making precise estimation of the causal effects difficult.

Estimation of multiple causal effects is very similar to estimation of the effects with a single binary treatment or a treatment and a control group. Our proposed method estimates multiple causal effects by applying the binary tools multiple times. The balance metrics for ATE need to be adjusted to compare estimates with the pooled sample rather than to the other treatment mean, but for ATT, the methods simply repeat the binary procedures for each pair of treatments. The powerful tools in the `twang` package conduct these steps and are readily available for analysts to use.

The methods discussed in this tutorial do of course have their limitations. In particular, the methods only remove confounding by observed variables, at best. Condition 2, no unknown or unmeasured confounders or exchangeability, must hold for unbiased estimation of the causal estimands. If there are unmeasured variables that predict outcomes and differ among treatment groups, then the estimates can be biased. This limitation, however, is not specific to the methods we present; indeed, all causal modeling strategies that use observational study data must contend with this limitation in one way or another. Also as discussed earlier, the methods cannot always balance the groups on observed pretreatment variables, or doing so can yield very unstable weights and imprecise estimates. Again, this is not a limitation that is specific to the methods we present. In fact, in our experience, using GBM to estimate propensity scores can achieve balance when other methods cannot and with more stable weights.

Finally, methods for estimating the standard errors of the propensity scores estimated by GBM do not exist. Consequently, principled methods for estimating the standard errors of the subsequent weights also do not exist. In this paper, we approximate the standard errors of the ATE and ATT estimates using robust (or so-called 'sandwich') standard errors. Robins *et al.* [12] suggested a similar approach for marginal structural models with propensity scores estimated by parametric approaches. In limited simulation experiments, robust standard errors for treatment effects estimated with GBM-based weights yield conservative confidence intervals that cover the true value more than the nominal percent of time [18]. However, to our knowledge, theory does not exist that guarantees this to be true for GBM, and this is an area for further statistical research.

Future research should more carefully explore various estimation methods for obtaining propensity score weights when there are more than two treatment conditions in order to better understand under which types of selection methods different estimation techniques prevail. From our personal experience, GBM has outperformed the use of multinomial logistic regression when we have tried to balance more than two treatment groups on pretreatment characteristics. Naturally, there will be applications for which the selection mechanism will bring the performance of the two methods (GBM and multinomial logistic regression) more inline. Table 2 in the Supporting information presents one such case where the two methods have similar results (ATT for the community cases). However, even if multinomial logistic regression can match the performance of GBM in some settings, GBM applied using the methods laid out in this tutorial is a tool that belongs in every causal modeler's tool box; it can yield better balance with less variable weights and less manual user inputs than multinomial logistic regression, and it can be implemented using available software.

Causal modeling is challenging. It requires estimating quantities that cannot be directly observed and measured. Causal estimation of multiple treatment adds additional challenges because multiple comparisons are potentially of interest. Weighting provides a method for making these estimates, and the `twang` package and the approach presented in this tutorial make weighting readily available and relatively easy to implement.

Acknowledgements

The development of this article was funded by the National Institute of Drug Abuse grant number 1R01DA015697 (PI: McCaffrey) and was supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and

Copyright © 2013 John Wiley & Sons, Ltd.

Mental Health Services Administration (SAMHSA) contract #270-07-0191 using data provided by the following grantees: Adolescent Treatment Model (Study: ATM: CSAT/SAMHSA contracts #270-98-7047, #270-97-7011, #277-00-6500, #270-2003-00006 and grantees: TI-11894, TI-11874; TI-11892), the Effective Adolescent Treatment (Study: EAT; CSAT/SAMHSA contract #270-2003-00006 and grantees: TI-15413, TI-15433, TI-15447, TI-15461, TI-15467, TI-15475,TI-15478, TI-15479, TI-15481, TI-15483, TI-15486, TI-15511, TI-15514, TI-15545, TI-15562, TI-15670, TI-15671, TI-15672, TI-15674, TI-15678, TI-15682, TI-15686, TI-15415, TI-15421, TI-15438, TI-15446, TI-15458, TI-15466, TI-15469, TI-15485, TI-15489, TI-15524, TI-15527, TI-15577, TI-15584, TI-15586, TI-15677), and the Strengthening Communities-Youth (Study: SCY; CSAT/SAMHSA contracts #277-00-6500, #270-2003-00006 and grantees: TI-13305, TI-13308, TI-13313, TI-13322, TI-13323, TI-13344, TI-13345, TI-13354). The authors thank these grantees and their participants for agreeing to share their data to support this secondary analysis. The opinions about these data are those of the authors and do not reflect official positions of the government or individual grantees.

References

1. Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science* 2010; **25**(1):1–21.
2. Rosenbaum PR. *Observational Studies*. Springer-Verlag Inc: New York, 2002.
3. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
4. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999; **94**:1053–1062.
5. McCaffrey D, Ridgeway G, Morral A. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* December 2004; **9**(4):403–425.
6. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services & Outcomes Research Methodology* 2001; **2**(3-4):259–278.
7. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; **15**:199–236.
8. Rubin DB. Using propensity scores to help design observational studies: application to tobacco litigation. *Health Services & Outcomes Research Methodology* 2001; **2**:169–188.
9. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**(3):706–710.
10. Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* 2004; **99**(467):854–866.
11. Frölich M. Programme evaluations with multiple treatments. *Journal of Economic Surveys* 2004; **18**(2):181–224.
12. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
13. Lechner M. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*. Physica-Verlag: Heidelberg, 2001; 1–18.
14. Zanutto E, Lu B, Hornik R. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* 2005; **30**(1):59–73.
15. Spreeuwenberg MD, Bartak A, Croon MA, Hagenaars JA, Busschbach JJ, Andrea H, Twisk J, Stijnen T. 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. *Medical Care* 2010; **48**(2):166–174.
16. Foster M. Propensity score matching: an illustrative analysis of dose response. *Medical Care* 2003; **41**(10):1183–1192.
17. Harder VS, Stuart EA, Anthony J. Propensity score techniques and the assessment of measured covariate balance to test causal association in psychological research. *Psychological Methods* 2010; **15**(3):234–249.
18. Lee BK, Lessler J, Stuart E. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; **29**(3):337–346.
19. Lee BK, Lessler J, Stuart E. Weight trimming and propensity score weighting. *PLoS ONE* 2011; **6**(3):e18174.
20. Ramchand R, Griffin BA, Harris KM, Morral AR. Using a cross-study design to assess the efficacy of MET/CBT-5 in treating adolescents with cannabis-related disorders. *Journal of Studies on Alcohol and Drugs* 2011; **72**(3): 380–389.
21. Ridgeway G. The state of boosting. *Computing Science and Statistics* 1999; **31**:172–181.
22. Friedman J. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 2001; **29**(5):1189–1232.
23. Friedman J. Stochastic gradient boosting. *Computational Statistics and Data Analysis* 2002; **38**(4):367–378.
24. Ridgeway G. GBM 1.6-3.1 package manual, November 2011. (Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.151.4024) [Accessed on: February 18, 2013].
25. Rubin DB. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Education Psychology* 1974; **66**:688–701.
26. Holland P. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–960.
27. Wooldridge J. *Econometric of Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA, 2002.
28. Cole SR, Frangakis CE. The consistency statement: a definition or an assumption? *Epidemiology* 2009; **20**(1):3–5.
29. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
30. Meyer JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 2011; **174**(11):1213–1222.

Copyright © 2013 John Wiley & Sons, Ltd.

31. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; **22**(4):523 –539.
32. Feng P, Zhou XH, Zou QM, Fan MY, Li XS. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine* 2012; **31**(7):681–697.
33. Huang IC, Frangakis C, Dominici F, Diette GB, Wu AW. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research* 2005; **40**(1):253–278.
34. Westreich D, Lessler J, Jonsson Funk M. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* 2010; **63**(8):826–833.
35. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Chapman & Hall/CRC: Boca Raton, FL, 1984.
36. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2 edn., Lawrence Erlbaum: New Jersey, 1988.
37. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 2008; **171**:481–502.
38. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**:295–313.
39. Neugebauer R, van der Laan M. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* 2005; **129**:405–426.
40. Little RJ, Vartivarian S. Does weighting for nonresponse increase the variance of survey means? *Technical Report*, Mathematica Policy Research, 2005. (Available at http://www.mathematica-mpr.com/publications/pdfs/weightingnonresponse.pdf) [Accessed on: February 18, 2013].
41. Dennis ML, Dawud-Noursi S, Muck RD, McDermeit (Ives) M. The need for developing and evaluating adolescent treatment models. In *Adolescent Substance Abuse Treatment in the United States: Exemplary Models from A National Evaluation Study*. Hawthorn Press: New York, 2003.
42. Stevens S, Morral AR (eds). *Adolescent Substance Abuse Treatment in the United States: Exemplary Models from A National Evaluation Study*. Hawthorn Press: New York, 2003.
43. Hunter S, Ramchand R, Griffin BA, Suttorp MJ, McCaffrey DF, Morral AR. The effectiveness of community-based delivery of an evidence-based treatment for adolescent substance use. *Journal of Substance Abuse Treatment* 2012; **43**(2):211–220.
44. Riley KJ, Rieckmann T, McCarty D. Implementation of MET/CBT5 for adolescents. *The Journal of Behavioral Health Services and Research* 2008; **35**(3):304–314.
45. Dennis M, Godley SH, Diamond G, Tims FM, Babor T, Donaldson J, Liddle H, Titus JC, Kaminer Y, Webb C, Hamilton N, Funk R. The Cannabis Youth Treatment (CYT) study: main findings from two randomized trials. *Journal of Substance Abuse Treatment* 2004; **27**(3):197–213.
46. Dennis M, Ives ML, White MK, Muck RD. The Strengthening Communities for Youth (SCY) initiative: a cluster analysis of the services received, their correlates and how they are associated with outcomes. *Journal of Psychoactive Drugs* 2008; **40**(1):3–16.
47. Dennis ML. Global appraisal of individual needs (GAIN) administration guide for the GAIN and related measures (version 1299). *Technical Report*, Chestnut Health Systems, Bloomington, IL, 1999.
48. Graubard B, Korn E. Predictive margins with survey data. *Biometrics* 1999; **55**(2):652–659.

Copyright © 2013 John Wiley & Sons, Ltd.