Psychological Methods
2004, Vol. 9, No. 4, 403–425

1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.4.403

Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies

Daniel F. McCaffrey, Greg Ridgeway, and Andrew R. Morral
RAND Corporation

Causal effect modeling with naturalistic rather than experimental data is challenging. In observational studies participants in different treatment conditions may also differ on pretreatment characteristics that influence outcomes. Propensity score methods can theoretically eliminate these confounds for all observed covariates, but accurate estimation of propensity scores is impeded by large numbers of covariates, uncertain functional forms for their associations with treatment selection, and other problems. This article demonstrates that boosting, a modern statistical technique, can overcome many of these obstacles. The authors illustrate this approach with a study of adolescent probationers in substance abuse treatment programs. Propensity score weights estimated using boosting eliminate most pretreatment group differences and substantially alter the apparent relative effects of adolescent substance abuse treatment.

Experimental studies offer the most rigorous evidence with which to establish treatment efficacy, but they are not always practical or feasible. Experimental treatment evaluations can be expensive to field and may be too slow to produce answers to pressing questions. In some cases random assignment to treatments is impractical, as with evaluations of the relative effectiveness of hospitals that might be geographically dispersed. Even when randomization to treatment conditions is logistically feasible, ethical concerns are often raised in community treatment settings when conventional wisdom favors one condition (Shadish, Cook, & Campbell, 2002).

Because of the challenges in fielding experimental studies, alternative methods are widely used to study treatment effects. In some circumstances, powerful quasi-experimental methods can be used to provide compelling evidence regarding treatment effects (Shadish et al., 2002; West, Biesanz, & Pitts, 2000). In many evaluation contexts, however, observational data from nonequivalent groups often represent the best available data on the effectiveness of widely used or important interventions. For example, almost all studies of community-based substance abuse treatment use this design (e.g., Gerstein & Johnson, 1999; Hser et al., 2001; Hubbard, Cavanaugh, Craddock, & Rachal, 1985; Sells & Simpson, 1979).

Identifying true, causal effects from observational studies of nonequivalent groups is challenging in part because treatment assignment mechanisms are neither known nor random. For instance, patients and those who refer them select treatments that they believe best suit their needs and resources. Because of these variations in treatment selection, patients entering different care models are likely to exhibit different pretreatment characteristics that may affect outcomes.

To minimize the confounding of treatment effects with pretreatment group differences, researchers frequently use statistical "case-mix adjustment" techniques. Statistical case-mix adjustment attempts to remove selection biases from treatment effect estimates by accounting for observed covariates that are expected to predict both outcomes and the treatment selection. The methods include analysis of covariance (ANCOVA) models (with and without correction for measurement error), gain score models, instrumental variable approaches, and propensity score models. Excellent general discussions of these approaches can be found in Shadish et al. (2002) or West et al. (2000).

Unlike the more common and traditional ANCOVA, propensity score methods account for differences between treatment and control groups by modeling the selection

Daniel F. McCaffrey, Greg Ridgeway, and Andrew R. Morral, Drug Policy Research Center, RAND Public Safety and Justice Program.

Additional materials are on the Web at http://dx.doi.org/10.1037/1082-989X.9.4.403.supp.

Portions of this research reported were supported by Center for Substance Abuse Treatment/Substance Abuse and Metal Health Services Administration Contract 270-97-7011 (Westat prime) and by National Institute on Drug Abuse Grants R01 DA017507-01 (Andrew R. Morral) and R01 DA015697-01 (Daniel F. McCaffrey).

Correspondence concerning this article should be addressed to Daniel F. McCaffrey, RAND, 201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516. E-mail: danielm@rand.org

process. The propensity score is the probability that a study participant is assigned to the treatment of interest rather than to a comparison group given a set of observed characteristics. Rosenbaum and Rubin (1983) showed that, conditional on this score, all observed pretreatment covariates are independent of group assignment and, in large samples, covariates will be distributed equally in both groups and will not confound estimated treatment effects. Rosenbaum and Rubin (1984) suggested stratification or matching on the propensity score when modeling treatment effects. Hirano, Imbens, and Ridder (2003) and Farley et al. (2002) have used propensity scores for weighting observations in treatment effect models.

Although these methods have recently begun to receive widespread use (cf. Connors et al., 1996; Dehejia & Wahba, 1999; Fiebach et al., 1990; Lieberman et al., 1996; Mojtabai & Graff Zivin, 2003; Stone, Obrosky, Singer, Kapoor, & Fine, 1995), the literature contains few proposed methods for the critical step of building propensity score models (Dehejia & Wahba, 1999; Hirano & Imbens, 2001; Rosenbaum & Rubin, 1984). Nearly all examples in the literature use a parametric logistic regression model that assumes covariates are linear and additive on the log-odds scale. The model may also include select interaction or nonlinear terms chosen through forward selection methods. More flexible approaches to modeling dichotomous outcomes have received little attention in the estimation of propensity scores.

In this article, we describe the use of generalized boosted models (GBM), a multivariate nonparametric regression technique, to estimate the propensity score. GBM is a general, automated, data-adaptive modeling algorithm that can estimate the nonlinear relationship between a variable of interest and a large number of covariates. GBM is appealing in the context of case-mix adjustment because it can predict treatment assignment from a large number of pretreatment covariates while also allowing for flexible, nonlinear relationships between the covariates and the propensity score. Other methods are less flexible and require variable selection. Variable selection risks biasing estimates of treatment effects because it omits covariates that are important to treatment selection or misspecifies the functional form of the relationship between covariates and treatment selection.

We demonstrate the use of GBM for estimating the propensity scores in the Adolescent Outcomes Project (AOP). The AOP is a study of the effects of a community-based residential substance abuse treatment program for adolescent probationers using the Phoenix Academy treatment model in comparison with the effects of referral of similar youths to alternative settings. The baseline interview contained dozens of measures of participants' pretreatment demographic characteristics, drug use, criminal history, psychological functioning, and other risk factors. Although youths who entered the Phoenix Academy differed from those in the comparison condition, we show that a GBM-derived propensity score model provides a means of weighting the comparison group that reduces or eliminates most pretreatment group differences. Comparison of outcomes associated with the Phoenix Academy versus the weighted comparison condition provides estimates of the Phoenix Academy treatment effect relative to alternative probation dispositions, with little confounding by the observed pretreatment differences. In addition, we highlight advantages of this method for estimating propensity scores over standard logistic regression approaches.

The Treatment Effect

Following Rosenbaum and Rubin (1983) and others (Holland, 1986; Imbens, 2003), we use counterfactuals to define the treatment effects from an observational study with a treated and an untreated comparison group. Every member in the population has two potential values for any outcome; one in which the individual is assigned to the treatment condition, y_1, and one in which the individual is assigned to a comparison group, y_0. Only one of these values is observed for each individual. The other counterfactual outcome cannot be observed. The treatment effect is $E(y_1) - E(y_0)$, where expectation is over the entire population.

Often, however, the effect of interest is on the treatment effects generalized only to clients like those typically entering a particular program or type of treatment, the so-called average treatment effect on the treated, denoted ATE_1 (Wooldridge, 2001). Let z be an indicator for treatment; $z = 1$ if the individual received treatment and 0 otherwise. Then $E(y_1|z = 1)$ is the average outcome of a treatment participant after receiving treatment, and $E(y_0|z = 1)$ is the average outcome of treatment participants if they had received the comparison condition instead. The treatment effect on the treated is

$$\text{ATE}_1 = E(y_1|z = 1) - E(y_0|z = 1). \quad (1)$$

When treatment effects are not constant, the effect of treatment on the treated can differ from those for the entire population. The treatment effect on the treated is of particular interest in the AOP example, where the Phoenix Academy treatment is expected to be most beneficial for youths with problems like those currently assigned to the Phoenix Academy. In the remainder of this article, we focus on estimating the treatment effect on the treated, though our discussion and results generalize to the treatment effect on the entire population with minor modifications.

The outcomes y_0 are unobserved for every participan receiving treatment, so $E(y_0|z = 1)$ must be estimated from the comparison group. However, as discussed in Rosenbaum and Rubin (1983), the average outcome from the comparison group generally will not yield an unbiased estimate of the $E(y_0|z = 1)$ because of pretreatment dif

ferences between groups. Propensity scores adjust for these differences by accounting for treatment selection.

The Propensity Score and Estimation of Treatment Effects

The propensity score is the probability that a member of the population receives treatment rather than the comparison condition. Conditioning on this quantity can provide an unbiased estimate of treatment effects. That is, if $\mathbf{x}$ denotes a vector of observed pretreatment characteristics, the propensity score, $p(\mathbf{x})$, is equal to $Pr(z = 1|\mathbf{x})$. Rosenbaum and Rubin (1983) showed that, conditional on $p(\mathbf{x})$, the distribution of $\mathbf{x}$ does not depend on z. In other words, conditioning on the propensity score ensures closely matched covariate distributions for all observed pretreatment variables across the treatment and comparison groups, as would be expected with random assignment designs. Rosenbaum and Rubin (1983) also proved that if the joint distribution of y_1 and y_0 is independent of z conditional on $\mathbf{x}$, then they are independent of z conditional on the propensity score. That is, conditional on $p(\mathbf{x})$, the distribution of the observable y_0 for the comparison group equals the distribution for the unobservable y_0 values of the treatment group. The observed values from the comparison group can be used to estimate $E[y_0|z = 1, p(\mathbf{x})]$, which can then be used to estimate ATE_1.

Propensity Score Weighting

The propensity score can be used to weight the observation when estimating the treatment effect (Hirano et al., 2003; Rosenbaum, 1987). To estimate $E(y_0|z = 1)$, let participant i in the comparison sample have weight $w_i = p(\mathbf{x}_i)/[1 - p(\mathbf{x}_i)]$, the odds that a randomly selected participant with features $\mathbf{x}$ would go to the treatment. We observe $y_i = y_{1i}$ if participant i is in the treatment group and $y_i = y_{0i}$ if participant i is in the comparison group. The weighted mean of the observed outcomes for the comparison group is

$$\hat{E}(y_0|z = 1) = \frac{\sum_{i \in C} w_i y_i}{\sum_{i \in C} w_i}, \tag{2}$$

where $i \in C$ denotes the ith observation in the comparison group, and summation is over the set of observations in this group. We use the notation $\hat{E}(y_0|z = 1)$ to denote that Equation 2 provides an estimate of the expected value. This estimate of $E(y_0|z = 1)$ is unbiased by selection into treatment if y_0 is independent of z given $\mathbf{x}$ (i.e., no hidden bias remains). Letting N_T denote the number of individuals in the treatment group and $i \in T$ denote the ith observation in this group, the sample mean for these individuals, $\hat{E}(y_1|z = 1) = \sum_{i \in T} y_i/N_T$, estimates the average outcome under treatment for the treated. The estimated effect of treatment on the treated is $\text{EATE}_1 = \hat{E}(y_1|z = 1) - \hat{E}(y_0|z = 1)$. Wooldridge (2001) and Hirano et al. (2003) discussed other estimators of $E(y_0|z = 1)$ using propensity weights. Appendix A provides a detailed derivation of Equation 2 and a brief discussion of other proposed estimators.

The weighting proposed in Equation 2 is analogous to reweighting procedures used in survey sampling to adjust for observations having unequal probabilities of inclusion in the sample. Heuristically, the observed values of y_0 are treated like a sample with individuals sampled with probability $1 - p(\mathbf{x})$. Thus, the denominator of w_i, $1 - p(\mathbf{x})$, accounts for the oversampling of individuals into the comparison group and weights these individuals back to the entire population of both treated and comparison participants. However, weighting the pooled sample to match the $\mathbf{x}$ distribution of the treatment cases requires weighting each observation by $p(\mathbf{x}_i)$, the numerator in w_i. Those comparison participants that have features uncharacteristic of the treatment population will have $p(\mathbf{x})$ near zero and therefore a weight near zero. Comparison participants with features that are characteristic of the treatment population will have larger $p(\mathbf{x})$ and therefore larger weights.

For a large sample size, the weighted treatment effect estimate will be nearly unbiased provided that several assumptions hold. Foremost, the potential outcome y_0 must be independent of treatment conditional on $\mathbf{x}$. That is, the observed covariates explain all the pre-existing differences between treatment and comparison groups that could affect outcomes. In addition, the stable unit treatment value assumption (Rubin, 1978) must hold. The stable unit treatment value assumption requires that individuals' potential outcomes be unaffected by the treatment assignment of other participants and other factors unrelated to treatment (i.e., no peer effects or treatment contamination can exist). Finally, weighting to estimate treatment effects on the treated requires $p(\mathbf{x}) < 1$ for all participants. This requirement means that no participant can have a 100% chance of being in the treatment condition. If this requirement is met, then any participant in the treatment group has the potential to match a participant in the comparison group.[1]

In general, weighted means have greater sampling vari-

[1] Estimating treatment effects for the whole population, not just the treated, requires both y_0 and y_1 to be independent of treatment assignment, conditional on the observed covariates and $0 < p(\mathbf{x}) < 1$. Furthermore, approximate unbiasedness in large sample when the weights are estimated from the data requires additional assumptions on the propensity score function and its estimator. Hirano, Imbens, and Ridder (2003) provided details on these assumptions and the properties of the estimators.

ance than do unweighted means from a sample of equal size. Kong, Liu, and Wong (1994) captured this increase in variance by computing the effective sample size (ESS) of the weighted comparison group as

$$\text{ESS} = \frac{\left(\sum_{i \in C} w_i\right)^2}{\sum_{i \in C} w_i^2}. \quad (3)$$

The ESS is approximately the number of observations from a simple random sample needed to obtain an estimate, with sampling variation equal to the sampling variation obtained with the weighted comparison observations. Therefore, the ESS will give an estimate of the number of comparison participants that are comparable with the treatment group.

Fitting the Propensity Score Model

In practice, propensity scores are unknown and must be estimated from the data. Accurate treatment effect estimates require that the propensity score model accounts for all covariates related to both treatment selection and outcomes and has the correct functional form. As demonstrated by Drake (1993), propensity score model misspecification can substantially bias the estimated treatment effect.

Provided the necessary assumptions are met, Equation 2 can yield estimated treatment effects that are unbiased in large samples (i.e., converge in probability to the true treatment effects) even when the propensity scores and resulting weights are estimated from data. The required assumptions include those given above and that the propensity score model is sufficiently flexible to describe the relationship between pretreatment characteristics and treatment assignment correctly or with minimal approximation error. Curiously, even if $p(\mathbf{x})$ is known, in many cases using an estimate of $p(\mathbf{x})$ produces better estimates of the treatment effect (discussed in Hirano et al., 2003; Rosenbaum, 1987). As discussed by Rosenbaum (1987) this phenomenon occurs because weighting by the true propensity scores "compensates only for the systematic differences" between groups, whereas weighting by estimated propensity scores "[corrects] for both systematic and chance imbalances" (p. 391).

Because neither the covariates used in selection of treatment nor the functional form of the propensity score models are known, estimation of the propensity scores involves model selection, choosing variables to include and adaptively determining the functional form. Adaptive methods add terms and variables to the model according to criteria such as statistical significance or reduction of prediction error. A key criterion for propensity score model selection is how well the treatment and comparison group covariate distributions match after controlling for the propensity score estimated with the model.

Current methods for estimating propensity scores almost exclusively use parametric linear logistic regression with selected interactions and polynomial terms. The approach used by Dehejia and Wahba (1999) is similar to the method originally proposed by Rosenbaum and Rubin (1984) and is typical of the methods currently used in practice (e.g., see the Mojtabai & Graff Zivin, 2003, study of the efficacy of substance abuse treatment for adults). Conditional on a set of covariates, Dehejia and Wahba fit a model with main effects and then stratified the data by propensity scores, testing for differences in the means and standard deviations between groups within strata. If any differences were statistically significant, then higher order polynomial terms and interactions were added to the model. The process continued until no significant differences remained. Rosenbaum and Rubin (1984) used graphical displays of the distributions of test statistics rather than formal significance tests to select a model with sufficient balance. Selection of covariates often occurs before any estimation of propensity scores.

Hirano and Imbens (2001) selected propensity score models without directly considering the balance of the covariates after weighting as a criterion. Their method combines propensity score weighting and linear regression modeling to adjust for covariates. They first developed a model for the propensity scores. They started with a predetermined set of predictors that included pretreatment covariates (X_1, X_2, . . . , X_k), selected higher order values of the covariates (X_1^2, . . . , X_k^2, and higher order polynomials), and selected interactions between covariates (X_1X_2, . . . , X_1X_k, X_2X_3, . . . , and higher order interactions, if appropriate). To choose predictors from this set to be included in the propensity score model, they repeatedly fit separate logistic regression models, each one predicting treatment assignment using only one of the predictors. They included in the final propensity score model every variable found to have a bivariate association with treatment assignment with a t statistic that exceeded a prespecified limit, t_{prop}.

Having selected a model for propensity scores, Hirano and Imbens (2001) built a model for the outcome that adjusts for covariates and is weighted by the propensity scores. The authors used a method that is analogous to their approach for building the propensity score model. They selected as covariates for the outcomes model every covariate found in a bivariate linear regression to predict the outcome with a t statistic that exceeded a prespecified cutoff, t_{reg}. Thereafter, the final model used propensity weights from the logistic regression model, the selected covariates from the linear regression models, and a treatment indicator to predict outcomes. The coefficient for the treatment indicator was their estimate of the treatment effect.

The authors did not provide specific guidelines for selecting values for t_{prop} and t_{reg}. Instead, they considered a range of values for both t_{prop} and t_{reg} and a range of estimated treatment effects. Similarly, they did not suggest procedures for identifying interactions among variables and other terms to be included in the initial set of predictor variables.

All approaches begin by selecting covariates for the models. In general, all available variables thought to be related to outcomes from empirical or theoretical research and which differ across groups should be included in the propensity score model. With a large number of covariates, however, this approach can quickly exhaust the available degrees of freedom in traditional regression approaches, so modeling is restricted to a subset of the available covariates. As noted in West et al. (2000), there are numerous strategies for selecting covariates. Reichardt, Minton, and Schellenger (1980), proposed limiting analyses to variables of theoretical importance to treatment selection and to those previously demonstrated to predict outcomes. The success of this approach is contingent upon the strength of available theories of treatment selection and the sophistication or earlier empirical analyses of these effects. Alternatively, Rosenbaum (2002) proposed including in propensity score models all pretreatment covariates on which group differences met a low threshold for significance ($|t| > 1.5$). Other rules of thumb exist; for example, Rosenbaum (2002) suggested including covariates unassociated with treatment assignment but related to the outcome. West et al. (2000) noted that pretest scores from the outcome of interest met these criteria.

In other settings, empirical variable selection and forward stepwise procedures are known to produce models that perform poorly with high mean-squared prediction error. We next discuss a modern regression approach, boosting, that offers a flexible and powerful data-adaptive method that can model the effects of large numbers of covariates without greatly reducing the precision of the estimate.

GBMs for Propensity Scores

In the following sections, we describe generalized boosted regression for estimating propensity scores. We begin with an overview of the statistical methods of boosting and generalized boosted regression. We then discuss the application of these methods to propensity score estimation.

Generalized Boosted Modeling

Overview. Boosting is a general, automated, data-adaptive algorithm that can be used with a large number of pretreatment covariates to fit a nonlinear surface and predict treatment assignment. Friedman (2001) and Madigan and Ridgeway (2004) have shown that boosting outperforms alternative methods in terms of prediction error. Many variants of boosting have appeared in machine learning and statistics literature, including the original AdaBoost algorithm (Freund & Schapire, 1997), GBMs (Ridgeway, 1999), LogitBoost (Friedman, Hastie, & Tibshirani, 2000), and the gradient boosting machine (Friedman, 2001). Boosting is particularly effective when the model involves a large set of covariates (Bühlmann & Yu, 2003). We used GBMs because, unlike most other implementations of boosting, this method is tuned to produce models yielding well-calibrated probability estimates. That is, GBM probability estimates match the empirical probabilities of treatment.

GBMs add together many simple functions to estimate a smooth function of a large number of covariates. Each individual simple function lacks smoothness and is a poor approximation to the function of interest, but together they can approximate a smooth function just like a sequence of line segments can approximate a smooth curve. In our implementation of GBMs, each simple function is a regression tree with limited depth.

Regression trees: Basic ideas. A regression tree uses the following recursive algorithm to estimate a function describing the relationship between a multivariate set of independent variables and treatment assignment. Starting with the complete dataset, the tree-fitting algorithm first partitions the dataset into two regions on the basis of the values of a single input variable. For example, if age and sex are covariates, the tree might split the dataset into two partitions, one with observations of people younger than 18 years and the other with observations of people older than or equal to 18 years. Or the tree might split the dataset into males and females. Splits can occur between any pair of observed values of any of the covariates. Within a region defined by the splits, the estimated function equals the sample mean of the output variable for all observations with values for their covariates that are elements of the region. Among all the possible splits, the algorithm selects the one that minimizes prediction error. Appendix B describes this selection more precisely.

The algorithm then further divides each of these partitions into two new partitions. The dataset is now partitioned into four groups, defined by the combination of two splits. Going back to the example with age and sex as covariates, the tree might split the group of people younger than 18 years into people younger than 15 years and people older than or equal to 15 but less than 18 years. It might also split the group of people older than or equal to 18 years into males and females. The dataset is now partitioned into youths younger than 15 years old, youths 15 years or older but younger than 18 years, males 18 years or older, and females 18 years or older. Splitting continues recursively until the tree includes the allowable number of splits. The number of splits determines the complexity of the tree, with each additional split allowing for additional interactions

between variables. Details on regression trees can be found in Breiman, Friedman, Olshen, and Stone (1984).

The GBM algorithm. GBM is an algorithm for iteratively forming a collection of simple regression tree models to add together to estimate the propensity score. Specifically, to simplify computations, GBM models the log-odds of treatment assignment, $g(\mathbf{x}) = \log\{p(\mathbf{x})/[1 - p(\mathbf{x})]\}$, rather than directly modeling propensity scores. The algorithm initially sets $g(\mathbf{x})$ to $\log[\bar{z}/(1 - \bar{z})]$, the constant baseline log-odds of assignment to the treatment, where $\bar{z}$ is the average treatment assignment indicator for the entire sample. The next step of the algorithm searches for a small adjustment, $h(\mathbf{x})$, to add to this initial estimate to improve the fit of the model to the data. Fit is measured by the Bernoulli log-likelihood of Equation 4, with larger values implying better fit:[2]

$$\ell(g) = \sum_{i=1}^{N} z_i g(\mathbf{x}_i) - \log\{1 + \exp[g(\mathbf{x}_i)]\}. \quad (4)$$

Analytically, Equation 4 will yield relatively large values when there is agreement between the $g(\mathbf{x}_i)$ and z_i, such that $z_i = 0$ when $g(\mathbf{x}_i)$ is negative and $z_i = 1$ when $g(\mathbf{x}_i)$ is positive. If the algorithm finds an adjustment that can improve the propensity score model's fit to the data, then $g(\mathbf{x})$, the current model for the log-odds, becomes $g(\mathbf{x}) + h(\mathbf{x})$. The boosting procedure iterates, each time selecting a model adjustment that when added to $g(\mathbf{x})$ offers an increase in the log-likelihood.

Technically, $h(\mathbf{x})$ can be of any form, but we selected $h(\mathbf{x})$ to be a regression tree that models the residuals from the current fit (i.e., the tree models $r_i = z_i - 1/\{1 + \exp[-\hat{g}(\mathbf{x}_i)]\}$ as a function of the covariates, where $1/\{1 + \exp[-\hat{g}(\mathbf{x})]\}$ is the estimate of the propensity score). Appendix B discusses the motivation for choosing $h(\mathbf{x})$ as a regression tree fit to the residuals. Briefly, using regression trees to model the residuals is equivalent to estimating the derivative of the log-likelihood function. Hence, following standard numerical algorithms for function optimization, GBM is an algorithm for finding the maximum-likelihood estimate of the function $g(\mathbf{x})$. As discussed in the next subsection, using trees at this stage is a key factor affecting the flexibility and robustness of the method.

To further reduce prediction error in GBM, Friedman (2002) introduced a stochastic component into the boosting algorithm. At each iteration, GBM selects a different random subsample of the data and uses only that subsample to estimate h. Empirical evidence suggests that subsampling 50% of the observations at each iteration can actually decrease bias and variance in the resulting model fit (Friedman, 2002).

The number of iterations determines the model's complexity and must be determined from the data. With each iteration of the algorithm, the model becomes more complex, fitting additional features of the data. When the number of iterations becomes sufficiently large, the model can predict the responses without error but no longer provides a meaningful estimate of the propensity score. The number of iterations typically is determined by stopping rules that attempt to choose the number of iterations that maximizes the predictive performance on an independent dataset rather than on the same data used to fit the model (Friedman, 2001). We discuss stopping rules for propensity score estimation below.

Advantages of the Boosted Logistic Regression Model

Because the final GBM model is a sum of regression trees, it inherits many of their advantageous properties for estimating propensity scores. Trees are computationally fast to fit (Breiman et al., 1984). Trees handle continuous, nominal, ordinal, and missing independent variables. They can capture nonlinear effects and interaction terms. Trees are also invariant to one-to-one transformations of the independent variables. In other words, whether we use age, log(age), or age^2 as a participant's attribute, we get exactly the same propensity score adjustments. Another important attribute of trees for estimating propensity scores is their ability to adaptively use a large number of covariates even if most are correlated with one another or are unrelated to the treatment assignment.

The boosting framework overcomes many of the known shortcomings of traditional regression tree approaches, which use only a single tree. For instance, large tree models can produce highly variable estimates when modeling with many covariates. When using a single tree to predict treatment assignment, the model is unable to capture main effects and lacks smoothness, often resulting in relatively poor estimates of the probability of assignment to the treatment group. However, GBM consists of a linear combination of many trees, combined in the boosting framework in such a way that this combination can capture main effects, can produce a smooth fit, and can often outperform single regression tree models (Friedman et al., 2000; Friedman, 2001).

Estimating main effects and interactions with GBM. The following example demonstrates how GBM combines trees to achieve smooth functions unattainable by the single discrete partitioning of the space provided by one large tree model. If we allow each tree to have only one split, then

[2] The log-likelihood is the log of the joint probability of the observed vector of treatment assignments given the function $g(\mathbf{x})$ provides the true log-odds of treatment assignment. Equation 4 is the standard log-likelihood used for fitting linear logistic regression models when $g(\mathbf{x})$ is linear.

each additional tree is necessarily a function of only one variable. The estimate, $\hat{g}(\mathbf{x})$, may therefore look like the following:

$$\hat{g}(\mathbf{x}) = g_0 + g_1(\text{age}) + g_2(\text{drug use}) + g_3(\text{drug use}) + g_4(\text{male}) + g_5(\text{age}) + \ldots. \quad (5)$$

The first term, g_0, is the log of the odds of the baseline rate. The first regression tree, $g_1(\text{age})$, has a single split on age, g_2 is a tree that splits on a drug use index, g_3 again splits on drug use, g_4 splits on sex, and g_5 also splits on age. Because the algorithm adds terms sequentially, categorical variables, such as the male indicator, could appear in Equation 5 multiple times even though there are limited ways to split such variables. After g_5 makes an adjustment for age, GBM may find at a later iteration that splitting on male provides the best-fitting model for the current residual. Grouping together those trees that split on the same variable (e.g., g_1 and g_5), we see that allowing only a single split per tree is equivalent to fitting an additive model. In Equation 6 $g_1^*(\text{age})$ simply refers to the sum of all those trees that split on age.

$$\hat{g}(\mathbf{x}) = g_0 + g_1^*(\text{age}) + g_2^*(\text{drug use}) + g_3^*(\text{male}). \quad (6)$$

The terms in Equation 6 can approximate many curves, including linear or quadratic terms, as well as curves that are not well approximated by low-order polynomials. If we allow each tree to have two splits then Equation 5 may take the following form:

$$g(\mathbf{x}) = g_0 + g_1(\text{age, drug use}) + g_2(\text{drug use}) + g_3(\text{male, age}) + g_4(\text{drug use}) + \ldots. \quad (7)$$

Collecting the trees as we did in Equation 6, we see that the algorithm fits an additive model with two-way interactions.

Using GBM to Estimate Propensity Scores

The desirable properties of GBM make it a natural tool for estimating propensity scores. Resulting propensity scores can then be used with Equation 2 to produce estimates of ATE_1, Equation 1, which we denote as EATE_{GBM}. GBM-based propensity scores can also be used for stratification or matching (Rosenbaum & Rubin, 1984, 1985). In this article, we focus on propensity score weighting. Zador, Judkins and Das (2001) also explored boosting to estimate propensity scores using MART rather than the GBM implementation presented here.

When estimating propensity scores, we suggest using all available covariates when fitting GBM. The algorithm will adaptively choose the variables to include in the prediction model. Our experience has shown that, even with large numbers of predictors, GBM can produce models that balance the covariate distributions across the groups and provide good mean-square prediction errors, even when applied to independent validation samples.

We have experimented extensively with the various tuning parameters involved in GBM and offer here our recommendations based on our experiences. Future research may produce further refinements or modifications to the methods. We allow a maximum of four splits for each simple tree used in the model, allowing for all four-way interaction between all covariates to be considered for optimizing the likelihood function at each iteration. This choice represents a compromise between identification of the correct functional form for the model and precise estimation of the model. In practice, we have found that higher order interactions offered no additional improvement in prediction error. Generally, we expect that unless samples are very large, it is unlikely that estimated five-way or higher order interactions would improve the predictive accuracy of the model. As discussed in Appendix B, the model also requires specification of a shrinkage parameter. We suggest using a value .0005, a relatively small shrinkage, ensuring a smooth fit. We also suggest subsampling 50% of the dataset for the regression tree fitting at each iteration.

We suggest stopping the algorithm at the number of iterations that minimizes the average standardized absolute mean difference (ASAM) in the covariates. To calculate the ASAM, for each covariate we calculate the absolute value of the difference between the mean for the treatment group and the weighted mean for the control group, divided by the standard deviation for treatment group. These are the standardized absolute mean differences, and we average these across covariates to obtain the ASAM, our measure of balance between the groups. To make the effect size comparisons comparable across alternative weightings, the denominator of the ASAM uses the standard deviation of only the treatment group, which is unaffected by the propensity weights. In our experience, the ASAM always initially decreases with each additional iteration and reaches a minimum, following which the ASAM increases with additional iterations. Thus, we suggest stopping when ASAM is minimized.[3]

We fit GBM using the generalized boosted modeling package developed at the RAND Corporation (Ridgeway, 2004) for the R statistical environment.[4] Both R and the GBM package are freely available. Details on how to obtain and use GBM and code for estimating propensity score

[3] There is no guarantee that the ASAM will have global minimum value. If a minimum is not obtained, other estimation methods might be required.

[4] R is a full-featured, freely available language and environment for statistical computing and graphics (Ihaka & Gentleman, 1996). R's general syntax and approach to statistical computing is the same as S-plus. However, the two packages have some different functions for programming and conducting statistical analyses.

using GBM can be found at http://dx.doi.org/10.1037/1082-989X.9.4.403.supp. To find the iteration that minimizes the ASAM, we run the GBM algorithm for a large number of iterations (e.g., 20,000 in our example). R's optimize function efficiently selects the number of iterations that minimizes the ASAM.

Estimating the Variance in $EATE_{GBM}$

The variance in $EATE_{GBM}$ depends on (a) the variance in the GBM estimates of the propensity scores, (b) the variability in covariates across samples, and (c) the variance in the outcomes within groups. Frequently, variance calculations for propensity score-based treatment effect estimates ignore the uncertainty in the propensity score model itself (e.g., Hirano & Imbens, 2001). Ignoring the model uncertainty results in easily computed variance estimates and has been shown to be an upper bound for the actual sampling variability of the estimated treatment effect for the observed sample. However, we are primarily interested in knowing whether the treatment will work on future participants, participants from the same population that would undergo the same treatment assignment process. The simple variance calculations can underestimate the variance of estimates of ATE_1 when either logistic regression or GBM is used to estimate propensity scores. Variance formulas for GBM do not exist, so sample re-use methods, such as the bootstrap or jackknife (Efron & Tibshirani, 1993), are natural alternatives for estimating the variance of $EATE_{GBM}$.

For the leave-one-out jackknife estimate of the variance of $EATE_{GBM}$, we deleted observation i from the data to obtain a jackknife sample and re-estimate the $EATE_{GBM}$ on this sample to obtain the jackknife replicate $EATE_{GBM(i)}$. We repeated this for all the observations in the data and calculated the average of the jackknife replicates, $EATE_{GBM(\cdot)}$. The variance estimate is given by

$$\hat{V}(EATE_{GBM}) = \frac{N_C + N_T - 1}{N_C + N_T} \times \sum (EATE_{GBM(i)} - EATE_{GBM(\cdot)})^2. \quad (8)$$

Details on the jackknife are found in Chapter 11 of Efron and Tibshirani (1993).

Sensitivity Analysis for Hidden Bias

Hidden bias results when individuals with the same values on observed covariates have different probabilities of treatment assignment. For example, if treatment assignment depends in part on an unobserved covariate then two individuals with the same values of the observed covariates but different values of the unobserved covariates will have different probabilities of treatment assignment. Of particular concern is the possibility that individuals in the treatment group will have greater than assumed probability of treatment and that the error in the propensity score model will be correlated with the outcome variable. Hidden bias cannot be estimated from observed data but can be explored through sensitivity analysis by adapting methods suggested by Rosenbaum (2002, Chapter 4). Rosenbaum's methods apply to matching and stratification by propensity scores but extend naturally to our weighted estimator.

The presence of hidden bias means that individuals with the same values for the covariate vector $\mathbf{x}$ have different odds of treatment. That is, in the presence of hidden bias, there exists for every individual in the sample an unobservable random variable, a_i, such that odds of treatment are not $w_i = \exp[g(\mathbf{x}_i)]$, as we assumed but $a_i w_i$. We say that the strength of the hidden bias is $G > 1$, if all of the random variables a_i are between $1/G$ and G. If the strength of the hidden bias is $G = 2$, for example, then for any individual in the sample, the odds of treatment could be twice as large or half as large as we assumed or anywhere in between. Larger values of G correspond to greater hidden bias, greater possible errors in weights, and greater possible bias in our estimated treatment effects.

We conduct a sensitivity analysis for hidden bias by assuming that hidden bias of a given strength, G, exists and measure changes in the estimated treatment effect. We repeat this at increasing values of G. If inferences about the treatment effect remain unchanged as G becomes large, then we have added confidence that our results would not change even if we could obtain and account for additional variables. Or in other words, differences between the treatment and comparison groups in unobserved variables would need to be large before they could undermine our estimated treatment effects.

Hidden bias depends not only on how much the weights will vary but also on the correlation between the values of a and the values of the outcome of interest. The absolute value of hidden bias increases with the absolute value of this correlation. Therefore, in our sensitivity analysis, we maximized this correlation by finding a set of values of a for the comparison cases to maximize and to minimize the estimated weighted comparison group mean. By doing this, we bound the possible hidden bias for a given value of G, the possible error in the odds of treatment for any individual in the population.

For the outcome y, sensitivity analysis follows these steps:

1. Pick a value of G near 1;
2. Find an N_C-vector of a values to maximize $S = \sum_{i=1}^{N_C} a_i w_i y_i / \sum_{i=1}^{N_C} a_i w_i$ subject to the constraint that $1/G \leq a_i \leq G$;

3. Repeat Step 2, finding a vector of a values to minimize S;

4. Repeat Steps 1 to 3 with increasing values of G.[5]

Choices for the range in G include increasing G until either the maximum or minimum equals the mean for the treatment group, so that the estimated treatment effect is zero or increasing G to where inferences about G change.

Our approach is analogous to Rosenbaum's (2002) use of upper and lower bounds on p values to quantify the errors from a hidden bias of G. Our approach is also similar to bounding used by Shadish, Hu, Glaser, Kownacki, and Wong (1998) when exploring bias from attrition. If small values of G result in large discrepancies between the bounds and the estimated effect, then the estimate is highly susceptible to hidden bias and should be interpreted with caution. If the bounds are close to the estimate even for large values of G, then we can have confidence in estimated effects.

Case Study: The AOP

The AOP is a study comparing the outcomes of 449 youthful offenders under the supervision of the Los Angeles Department of Probation (LADP): 175 received treatment at the Phoenix Academy following referral by the LADP (treatment); 274 received alternative services including treatment at other residential programs (comparison). Phoenix Academy is a 150-bed substance abuse treatment program providing long-term residential care for adolescents under 18 years old, which uses a modified therapeutic community approach (Jaycox, Marshall, & Morral, 2002; Morral, Jaycox, Smith, Becker, & Ebener, 2003; Morral, McCaffrey, & Ridgeway, 2004). Therapeutic community treatment is an experiential treatment approach that uses counseling, encounter groups, and mutual self-help to foster behavior change (De Leon & Dietch, 1985; Jainchill, 1997). Youth in the comparison group received the standard services that treatment youth would have received had they not gone to Phoenix Academy.

Successful data collection strengthened the AOP. Study follow-up retention was excellent. At the 12-month assessments, more than 90% of the baseline sample ($N = 449$) were located and successfully interviewed. See Morral et al. (2003) for additional details on recruiting, including eligibility criteria. The principal data collection instrument at each of the four assessments was a version of the Global Appraisal of Individual Needs (GAIN; Dennis, 1998), an extensive instrument that collects detailed data on substance use and related risk factors.

Pretreatment Risk Factors

Research on risk factors for poor psychosocial outcomes of youths in substance abuse treatment has revealed a wide range of pretreatment characteristics that may be associated with treatment outcomes (e.g., Catalano, Hawkins, Wells, & Miller, 1991; Orlando, Chan, & Morral, 2003; Williams, Chang, & Addiction Centre Research Group, 2000). This case study included 41 pretreatment variables from the GAIN that a review of the literature suggested could influence treatment assignment and treatment outcomes. These included demographic characteristics, lifetime and recent drug use, criminal histories, drug problems, treatment readiness indices, psychological functioning indices, measures of home and social environment, school and work performance measures, and other variables listed in Table 1. Although this is only a subset of all of the variables available in the GAIN, 41 covariates are many more than propensity score models typically include.

Outcomes

For this demonstration we selected two drug use outcomes measured 12-months after pretreatment interviews, days of alcohol use in the previous 90 days, and days of marijuana use in the previous 90 days. Single items on the GAIN measured both outcomes. We estimated the treatment effect relative to baseline values using the change in days of alcohol use and change in days of marijuana use. For these outcomes, a negative value of the population average treatment effect indicated that Phoenix Academy is more effective than the comparison sites in reducing substance use, whereas a positive value indicated the opposite.

Statistical Methods

We used GBM to estimate propensity scores, tuning the model so that the treatment and weighted comparison group are well-matched on the pretreatment covariates by minimizing the ASAM. Using the tuning parameter settings described previously in the section, *Using GBM to Estimate Propensity Scores,* we ran GBM for 20,000 iterations before searching for the number of iterations that minimized the ASAM for the 41 pretreatment covariates.

Logistic regression model for comparison. For comparison, we also estimated propensity scores using the more common logistic regression approach described above. Using conventional procedures, we first modeled propensity using the subset of our 41 pretreatment variables with significant ($p < .05$, two-tailed test) bivariate relationships with treatment assignment. Because some analysts have recommended relaxing significance requirements for this variable selection, we also developed a logistic regression

[5] The constrained optimizations of Steps 2 and 3 can be replaced by unconstrained optimizations by reparameterizing the problem with $a_i = 1/\{[1 + \exp(-\delta_i)]G\} + G/[1 + \exp(\delta_i)]$ and solving for values of the δs to optimize S.

Table 1
Pretreatment Characteristics and Group Difference Effect Sizes (d) Between Phoenix Academy (PA) and Comparison Condition (COMP) on All Baseline Covariates Before and After Propensity Score Weighting

Covariate	Unweighted					Propensity score weighted[a]			
	PA		COMP			COMP			
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*d*	*p*
Demographics									
Age (years)	15.82	0.91	15.31	1.28	0.56	15.76	1.11	0.07	0.58
Race (%)									
African American	8.57	28.07	18.61	38.99	−0.36	12.23	32.76	−0.13	0.26
Latino/Hispanic	60.00	49.13	52.19	50.04	0.16	55.24	49.73	0.10	0.43
White	20.57	40.54	13.14	33.84	0.18	18.62	38.92	0.05	0.72
Female (%)	18.29	38.77	9.12	28.85	0.24	8.17	27.39	0.26	0.01
School/work participation									
Last grade completed	9.04	1.25	8.59	1.36	0.36	8.85	1.28	0.15	0.20
Recency of last school/training[b]	5.41	1.16	5.28	1.36	0.11	5.22	1.41	0.16	0.21
Recency of paid work[b]	1.32	1.45	1.13	1.32	0.13	1.16	1.35	0.11	0.34
Current drug use/problems									
Days of alcohol/drug use (in past 90 days)[c]	6.19	3.17	3.77	3.59	0.76	5.59	3.35	0.19	0.10
Days drunk/high (in past 90 days)[c]	4.18	3.43	2.51	3.17	0.49	3.91	3.38	0.08	0.51
Substance involvement (recent)[c]	0.86	0.81	0.40	0.66	0.56	0.77	0.82	0.10	0.44
Substance use intensity index[c]	7.61	4.36	4.59	4.72	0.69	6.94	4.75	0.16	0.19
Substance problem index (in past month)[c]	1.61	1.32	0.70	1.08	0.68	1.39	1.28	0.16	0.17
Current withdrawal index	1.53	0.28	1.49	0.24	0.16	1.51	0.26	0.08	0.47
Self-reported treatment need for (%)									
Alcohol	4.57	20.95	5.47	22.79	−0.04	7.89	26.95	−0.16	0.29
Marijuana	32.00	46.78	6.93	25.45	0.54	21.10	40.81	0.23	0.08
Other drugs	27.43	44.74	12.04	32.61	0.34	18.45	38.79	0.20	0.07
Drug use history									
Age of first use	12.55	1.84	11.97	3.13	0.32	12.38	2.48	0.09	0.49
Substance problem index (lifetime)[c]	3.05	0.78	2.22	1.31	1.07	2.89	0.99	0.21	0.08
Substance involvement (lifetime)[c]	2.15	0.60	1.73	0.71	0.69	2.03	0.62	0.18	0.11
Substance disorder level (%)									
Physical dependence	60.00	49.13	37.23	48.43	0.46	57.51	49.43	0.05	0.67
Dependence	10.29	30.46	6.20	24.17	0.13	7.66	26.60	0.09	0.46
Abuse	23.43	42.48	27.01	44.48	−0.08	22.95	42.05	0.01	0.92
Use	6.29	24.34	29.56	45.72	−0.96	11.88	32.35	−0.23	0.05
Prior drug treatments[c]	0.98	1.77	0.52	1.19	0.26	0.91	1.45	0.04	0.72
Smoking recency[b]	2.93	1.64	2.25	1.64	0.42	2.85	1.64	0.05	0.68
Injection drug use recency[b]	1.85	2.13	1.14	1.76	0.33	1.31	1.87	0.25	0.03
Criminal history									
Lifetime arrests[c]	1.82	0.87	1.85	1.55	−0.04	1.92	1.22	−0.11	0.42
Arrest recency[b]	2.88	1.16	2.81	1.15	0.06	2.94	1.22	−0.05	0.69
Arrests (in past 90 days)[c]	0.76	0.62	0.70	0.60	0.10	0.72	0.59	0.06	0.58
Crime recency[b]	2.54	1.51	2.58	1.44	−0.03	2.59	1.43	−0.03	0.79
Crime days (in past 90 days)[c]	4.26	3.47	3.20	3.24	0.31	4.07	3.38	0.05	0.65
Property crimes (in past 90 days)[c]	1.90	2.48	1.65	2.77	0.10	1.78	2.85	0.05	0.70
Violent crimes (in past 90 days)[c]	0.98	1.51	1.08	1.57	−0.06	1.03	1.39	−0.03	0.76
Drug crimes (in past 90 days)[c]	1.60	2.89	1.12	2.56	0.17	1.50	2.75	0.03	0.76
Days in a controlled environment[c]	2.33	3.24	2.85	3.66	−0.16	2.35	3.27	−0.01	0.95
Treatment readiness									
Treatment resistance index	1.11	1.01	0.97	0.98	0.13	1.10	0.98	0.01	0.93
Treatment motivation index	2.52	1.32	1.35	1.38	0.89	2.22	1.46	0.23	0.07
Self-efficacy index	3.38	1.19	3.42	1.23	−0.03	3.34	1.25	0.03	0.79
Problem orientation index	1.38	1.91	0.50	1.27	0.46	1.00	1.67	0.20	0.08

Table 1 (*continued*)

Covariate	Unweighted					Propensity score weighted[a]			
	PA		COMP			COMP			
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*d*	*p*
Social environment									
Environmental risk index	30.61	9.76	28.94	11.06	0.17	31.09	10.61	−0.05	0.67
General social support	6.31	2.06	6.45	2.11	−0.06	6.27	2.25	0.02	0.89
Health and mental health									
Past-year health[d]	1.87	1.12	1.55	1.10	0.29	1.71	1.09	0.15	0.20
Psychological distress recency[b]	1.14	1.58	1.30	1.77	−0.10	1.31	1.70	−0.10	0.38
Somatic symptoms index	1.14	1.34	0.92	1.15	0.17	0.99	1.14	0.12	0.26
Depressive symptoms index	2.39	1.88	2.05	1.88	0.18	2.26	1.87	0.07	0.56
Anxiety symptoms index	2.82	2.58	2.61	2.52	0.08	2.64	2.35	0.07	0.50
Complex behavior index	12.84	8.53	12.11	9.69	0.09	13.00	9.19	−0.02	0.88
Average absolute effect size					0.31			0.11	

Note. Sample sizes: PA, $n = 75$; COMP, $n = 274$. Effective sample size after weighting: COMP, $n = 107.5$. Effect sizes (*d*) calculated as the difference between group means divided by the standard deviation for PA. The standard deviation for PA is unaffected by propensity score weighting and allows for comparison pre- and postweighting.

[a] PA cases are not weighted, so only values for COMP change with weighting.

[b] Recency scale spans 0 (*never*) to 6 (*past two days*).

[c] Past 90-day frequency and count variables with a range greater than 15 are square root transformed to reduce variable skew.

[d] Past-year health scale ranged from 0 (*Excellent*) to 4 (*Poor*).

model of propensity scores using a $p < .20$ variable inclusion criterion. The second step fit a main effects logistic regression model using the selected variables and calculated propensity score weights from the resulting predicted probabilities. The third step tested for significant differences ($p > .05$, two-tailed test) between treatment means and the weighted comparison group means of all covariates used for fitting the model. The fourth step identified the variable with the largest absolute effect size for these group differences and interacted this variable with all the other covariates and itself (e.g., the procedure used in Mojtabai & Graff Zivin, 2003). These interaction terms were included in the set of covariates and Steps 2–4 repeated until no differences were significant at Step 3. The final models provided the propensity scores for estimating treatment effects, which we called $\text{EATE}_{\text{logit},.05}$ and $\text{EATE}_{\text{logit},.20}$.

Standard error estimation. The standard error of each treatment effect estimator was estimated using the leave-one-out jackknife. For each estimator the entire estimation process was replicated with each jackknife replicate sample. Thus, our variance estimators accounted for variability of the adaptive model selection methods used in each method.

Results

Estimated Propensity Scores and Weights

Following the procedures we discussed above, we let the algorithm iterate until the comparison group weights derived from it minimized between group differences on the 41 pretreatment characteristics. The resulting model had a deviance $R^2 = .521$ (Hosmer & Lemeshow, 1989, p. 148). We can decompose the overall improvement in the model's log-likelihood, shown in Equation 4, into components attributable to each of the 41 covariates, as a measure of the relative influence of each variable (Friedman, 2001). About 30% of the increase in model likelihood is due to four covariates: treatment motivation index, substance use intensity index, complex behavior index, and substance problem index (past month). Three of these four variables are related to substance use, which is reassuring because Phoenix Academy is the only disposition specifically designated as a substance use treatment program. We can probe the marginal contribution of each of these factors using partial dependence plots (Friedman, 2001). These plots illustrate the nonlinear relationships between each covariate and the log-odds that a youth is assigned to Phoenix Academy, conditional on the effects of the other covariates (see Figure 1). This figure shows that after accounting for the influence of other covariates, youths are more likely to belong to the Phoenix Academy condition if they (a) have a treatment motivation index score of 2 or greater, (b) report more recent drug and alcohol use (on the substance use intensity index), (c) have scores below 16 (the 60th percentile) on the complex behavior index, and (d) report more recent problems associated with substance use. The signals that the GBM detects are consistent with Phoenix Academy admission practices and goals.

Of the four most important variables, the ones that does not directly concern substance use is the complex behavior index, a count of problem behaviors associated with atten-

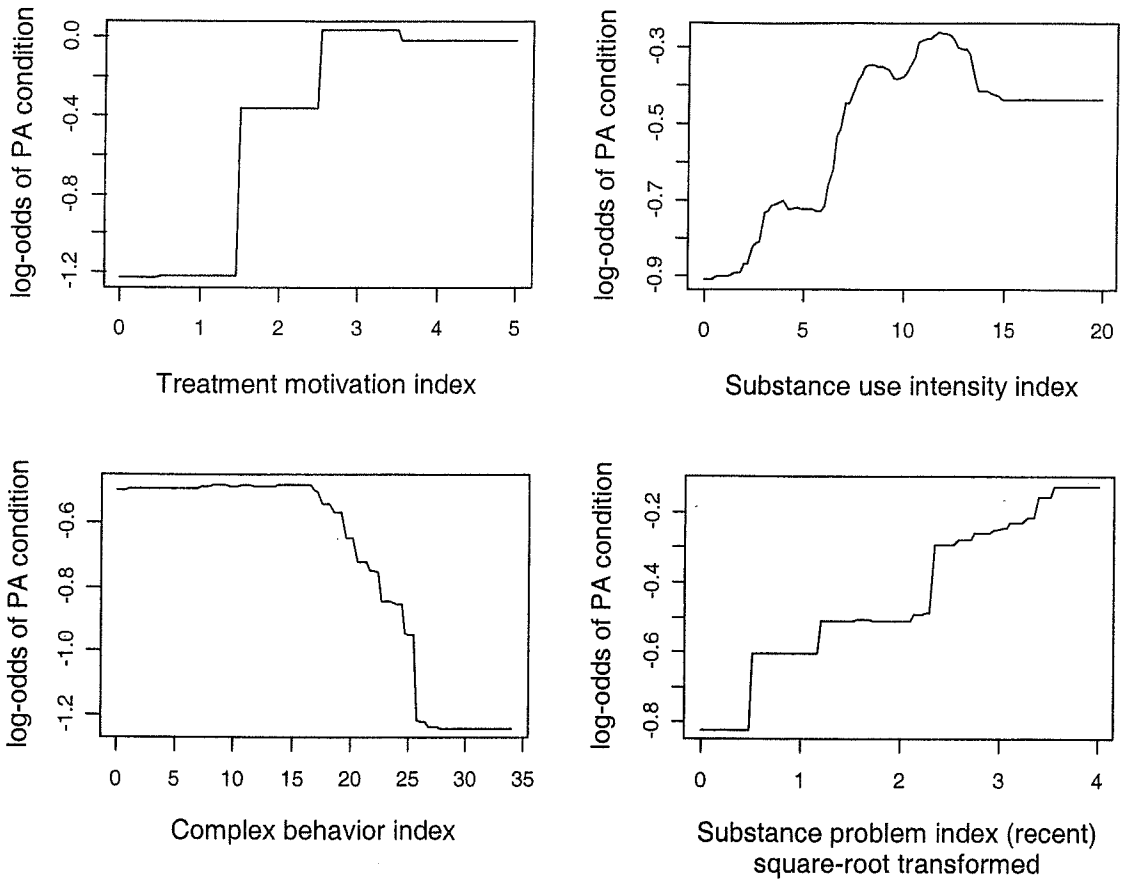

Figure 1. Partial dependence plots for the four variables accounting for the majority of generalized boosted model propensity model improvement. These plots illustrate the nonlinear relationships between covariates and the log odds of the probability a youth is assigned to Phoenix Academy (PA), marginalizing with respect to the other 40 covariates in the model. Figures are plotted on the entire range of each covariate.

tion deficits, hyperactivity, and conduct disorder. The importance of this variable in the model is interesting and highlights an advantage of the GBM methods were used. Specifically, Table 1, discussed in detail later, shows that in a univariate analysis, the behavioral complexity index (Variable 41) does not appear to distinguish group membership well and therefore would likely be excluded from case-mix adjustment models that can accept only a small number of covariances. Because it proves to be influential in determining group membership, it suggests that dropping variables on the basis of mean comparisons alone can be counterproductive. Even though the mean behavioral complexity indexes in the treatment and comparison groups are very similar, after accounting for other covariates subject with higher scores on the behavioral complexity index are much more likely to be in the comparison group.

Figure 2 shows the distribution of propensity scores for the treatment and comparison groups. Naturally, the treatment group tends to have fairly high propensity scores. A small number of comparison observations also have high propensity scores, but most have scores of less than .30. This leads to generally small observation weights for most comparison participants, as shown in Figure 3, and a few youths with weights exceeding 1.0 or 1.5. None of the weights are excessive. The largest accounts for slightly more than 2% of the total sum of the weights. The variability of the weights reduced the ESS of the comparison group, calculated as in Equation 3, from 274 before weighting to 107.5. This implies that the weighting effectively filtered out 167 comparison participants that were incomparable with the treatment participants.

Ideally, we would like to see greater overlap between treatment and comparison propensity scores, which yield a larger ESS for the same amount of bias reduction. With less overlap, treatment effect estimates will have larger variances, and there is some danger that propensity score weighting will not succeed in producing a comparison

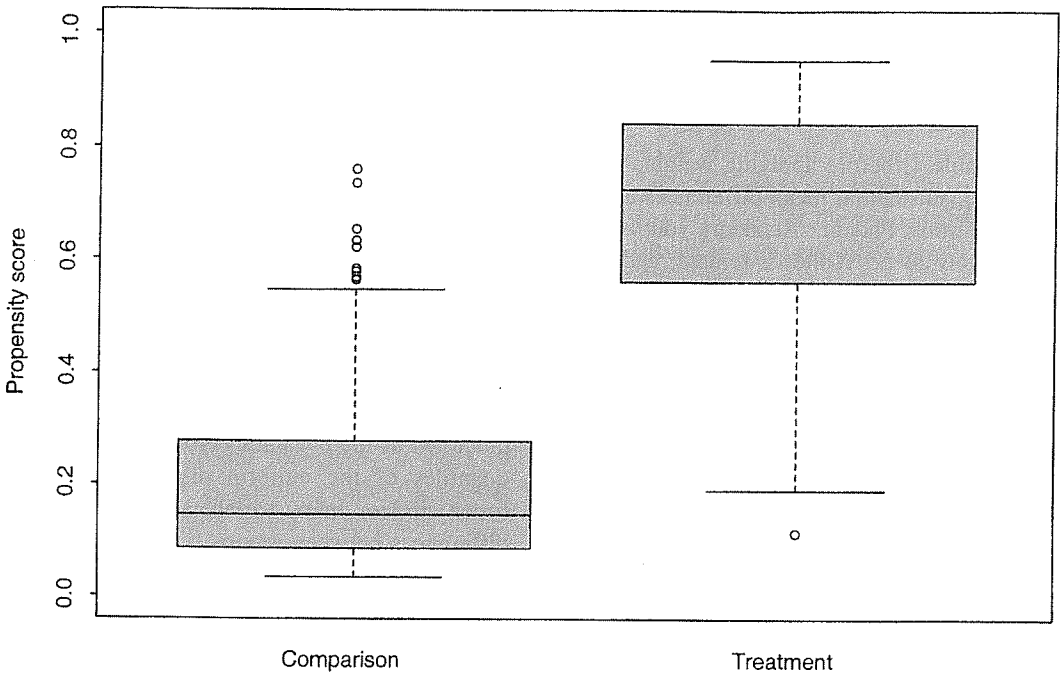

Figure 2. Boxplots of the generalized boosted model propensity scores for the comparison and treatment observations. The boxes mark the first and third quartiles of the propensity scores with solid lines drawn at the medians. The dashed lines extending from the boxes indicate the medians plus and minus 1.5 times the interquartile range. Propensity scores more extreme than that are indicated with open circles.

group with covariate distributions well-matched to the treatment group. However, nonlinearities in GBM imply that distances between propensity scores do not equate to distances between the covariates on the covariate scale or bias in the treatment effects. As discussed below, the covariates in the present example are well-balanced after weighting. We have found that less disparity between groups in the distribution of the propensity scores does not correspond to better balance in means for the covariates. In fact, for the AOP dataset, comparisons of GBM solutions using different numbers of iterations revealed that fits resulting in treatment and comparison groups having more disparate estimated

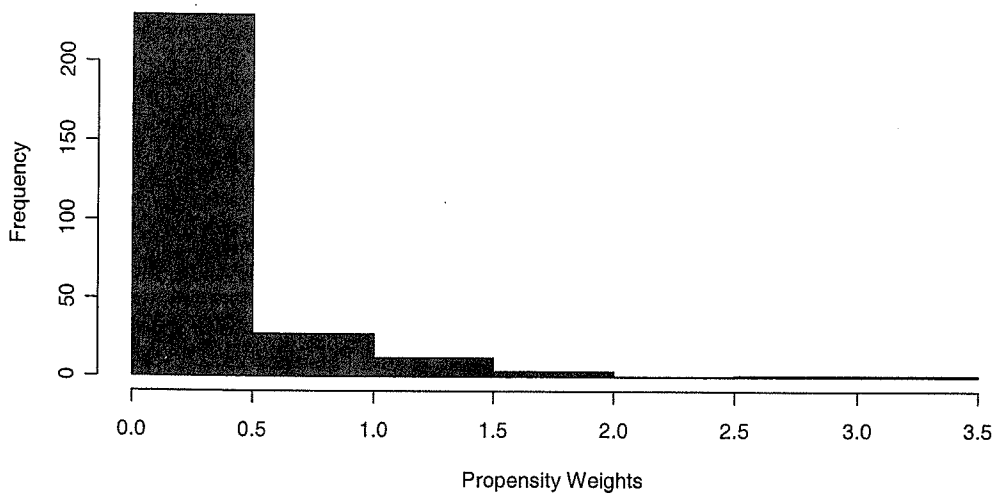

Figure 3. Distribution of the generalized boosted model propensity score weights for the comparison group.

propensity scores often resulted in better balance on the covariates.

Weighting the Comparison Condition

As displayed in Table 1, before applying weights to the comparison condition, substantial mean differences are observed between conditions. Phoenix Academy youths were older; were more likely to be White or female; and were more involved with drugs, alcohol, and drug crimes. Comparison condition youths were more involved with violent crimes and reported better health and a greater number of family members that have been in jail or prison for significant lengths of time. Not surprisingly, Phoenix Academy youths had higher treatment readiness scores, reported more drug use and related problems and more recent (non-drug-related) crime, and were more likely to report needing treatment for marijuana and other drug use. As a rule of thumb an effect size of .2 is considered small, .5 medium, and .8 large when considering likely substantive importance (Cohen, 1988). Across the 41 pretreatment variables used in the model, the unweighted mean absolute effect size is .307. Moreover, 10 variables have effect sizes greater than .5, with the substance problem index (Variable 15) having an effect size greater than 1.0.

After weights derived from the propensity scores we applied, differences between groups diminished substa tially, with the average absolute effect size dropping 65% .107. No variable has an effect size over .3, and only variables have effect sizes larger than .2. Phoenix Acaden youths remain somewhat more likely to be female and report recent injection drug use.

Figure 4 illustrates that after weighting, differences b tween groups on the 41 pretreatment characteristics we close to those we would expect had cases been random assigned to treatment and comparison groups. p values fro independent tests in which the null hypothesis is true ha a uniform distribution. Figure 4 shows a plot comparing tl quantiles of the p values before and after weighting to tl quantiles of the uniform distribution. Before weightir (open circles), many variables have statistically significa differences between groups (i.e., with p values near zerc After weighting (closed circles), the p values follow the 4 line, the cumulative distribution of a uniform variable on [1], as would be expected in test for covariate differences a random experiment. In a random experiment, the nu hypothesis of no difference in covariate means betwe treatment and comparison groups is true. The p value is tl probability that a test statistic would exceed the observe

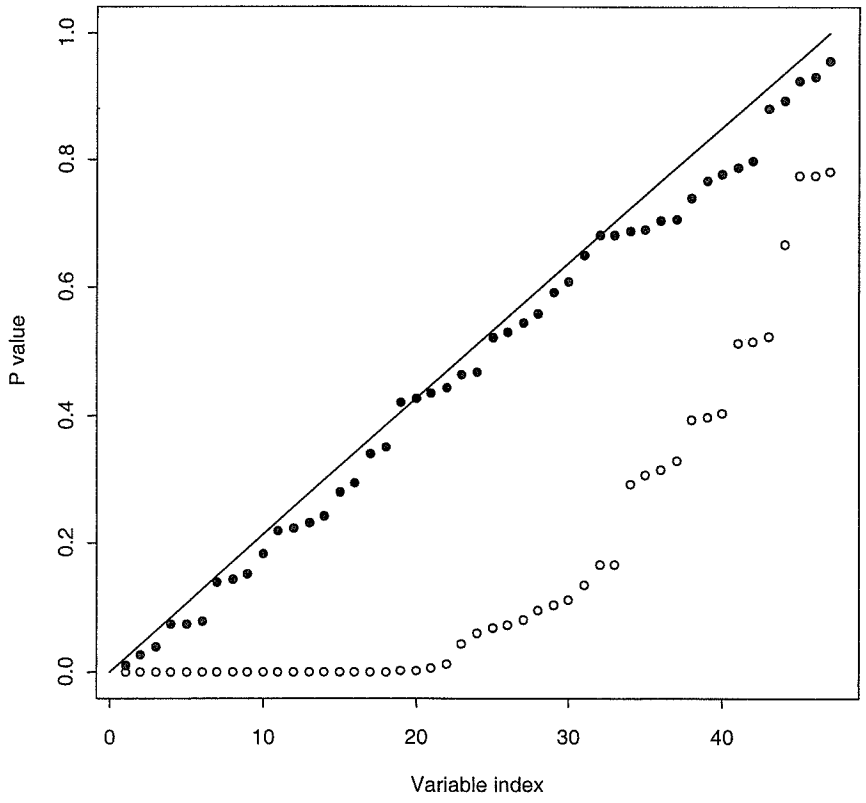

Figure 4. Plot comparing the quantiles of the uniform distribution to the quantiles of the p values for t tests of group differences on 41 baseline covariates, with categorical variables dummy coded. Open circles are the p values prior to weighting. Solid circles are p values after weighting.

test statistic under the null hypothesis. p values are random variables on the interval 0 to 1, and by definition, when the null hypothesis is true, the probability that a p value is less than any proportion (say .05) is that same proportion. Thus, the p values for testing group mean differences among the covariates for a true experiment would follow a uniform distribution on the interval [0, 1]. For this figure, categorical variables were dummy coded, resulting in 47 significance tests.

Outcome Analyses

The first two rows of Table 2 show the estimated treatment effects and 95% confidence intervals. An unweighted analysis would declare significant reductions in days of marijuana use, with Phoenix Academy youths decreasing use by 10 days more than those in the comparison group. The GBM-based propensity score adjustment, on the other hand, indicate that the difference between groups was only about 6 days, and this was not statistically significant. The loss of statistical significance for this comparison did not appear to be attributable to the small increase in the size of the confidence interval resulting from GBM-based weighting. Unadjusted, the groups had similar change in alcohol use. Weighting indicated that youths attending the Phoenix Academy had a smaller decrease, but this difference was not significant. A full longitudinal analysis of the outcomes for this dataset is in Morral et al. (in press).

Alternative Propensity Score Estimators

We compared the GBM-based method to the two logistic regression-based methods discussed above. Figure 5 contains a scatterplot of the propensity scores for the comparison group estimated by GBM versus the propensity scores for that group estimated by logistic regression with the .05 inclusion criteria for main effects. The correlation was moderately high (.82); however, the propensity scores from logistic regression were more dispersed than were the GBM estimates. Moreover, the logistic regression estimates tended to be greater than the corresponding GBM estimates, except at low values of both. In one case, the logistic regression estimate for the propensity score was very close to one, resulting in a very large weight that greatly exceeded any of the GBM weights. The plot for logistic regression with the .20 inclusion criteria was very similar.

Table 2 also compares the treatment effect estimates of our recommended GBM methods for estimating propensity scores with the two alternatives. Because we do not know the true effect sizes, we cannot say which estimator is best. Rather, we compare the estimators on three characteristics that should relate to bias and variance in the estimated treatment effect: (a) prediction error in the propensity score model, (b) balance between groups on the means of the covariates, and (c) variability in the estimated treatment effect.

The GBM model for propensity scores had smaller prediction error than that of the logistic alternatives. To avoid bias in our estimate of prediction error that results from estimating the error metrics with the same data used to fit the models, we used the jackknife replicate samples to estimate prediction error. Each jackknife replicate sample predicted treatment assignment for the held-out observation. We measured prediction error for this observation using the deviance metric,

$$\text{Deviance}_i = -2(z_i \log p^{[i]}(\mathbf{x}) + (1 - z_i)\log[1 - p^{[i]}(\mathbf{x})]), \tag{9}$$

where $p^{[i]}(\mathbf{x})$ is the estimated propensity score from the i^{th} model fit, with observation i left out. The deviance is -2

Table 2

Treatment Effect Estimates and Their Properties Using Unadjusted Sample Means and Three Alternative Propensity Score Weighting Methods: Generalized Boosted Models (GBM), and Two Logistic Regression Models

	Treatment effect estimation method							
	Unadjusted		GBM		Logit (0.05)		Logit (0.20)	
Summary statistic	*M*	CI	*M*	CI	*M*	CI	*M*	CI
Estimated treatment effect								
Marijuana	−11.8	−19.7, −3.8	−5.9	−16.2, 4.3	−1.9	−12.7, 8.8	−5.2	−24.4, 14.1
Alcohol	−1.2	−5.5, 3.0	2.8	−3.6, 9.3	1.5	−10.2, 13.3	3.1	−10.5, 16.7
Deviance			466.4		539.2		511.4	
Average standardized absolute mean difference (ASAM)		0.31		0.11		0.14		0.20
Standard error treatment effect (marijuana)		4.0		5.2		6.6		11.8
Standard error treatment effect (alcohol)		2.2		3.3		7.2		8.3

Note. Deviance is a measure of prediction error based on each observation's contribution to the log-likelihood (see Equation 9). The ASAM measures average effect sizes for group differences on pretreatment covariates after any propensity score weights are applied. Standard errors were estimated using the jackknife procedure.

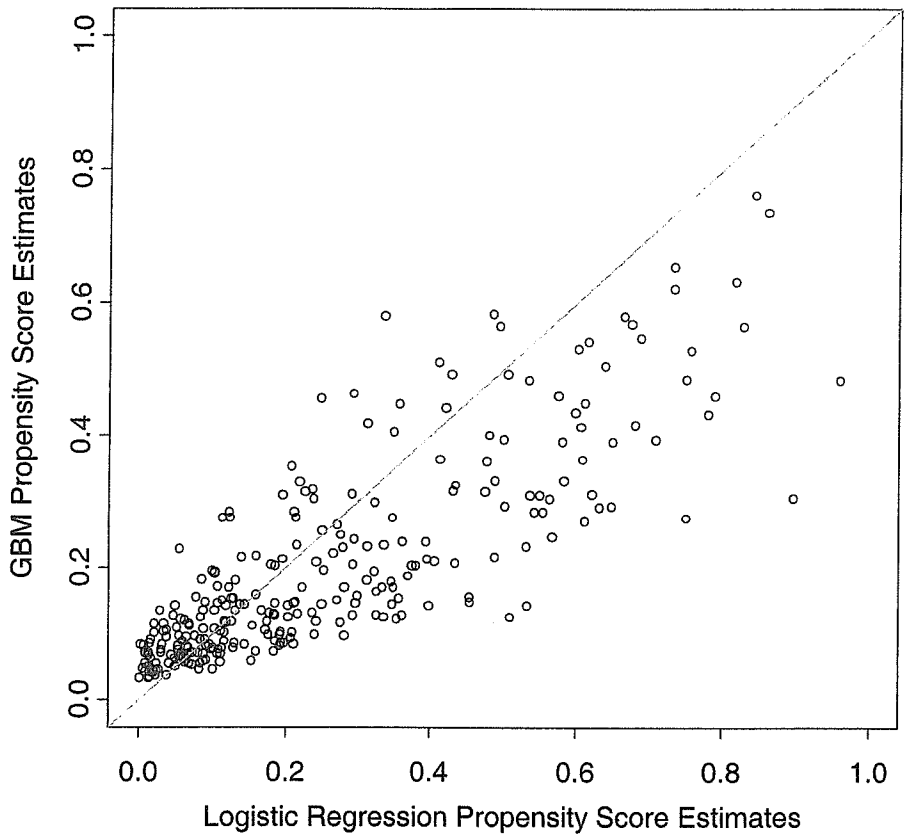

Figure 5. Scatterplot of propensity scores estimated by the generalized boosted model (GBM) versus logistic regression for all observations in the comparison group of the Adolescent Outcomes Project study. Logistic regression used a 0.05 inclusion criterion for selecting main effects.

times the contribution of observation i to the log-likelihood. A large deviance implies that the observed value of z_i is unlikely, given the estimated $p^{[i]}(\mathbf{x})$, which signals a poor fit. We totaled these deviances across all observations to calculate the prediction error for a model. The prediction errors for the logistic regression-based method using the .05 and .20 criteria were, respectively, 16% and 10% larger than those for the GBM method. This indicated that GBM provided the more accurate estimates of $p(\mathbf{x})$.

Although accuracy in estimating $p(\mathbf{x})$ offers some validation, GBM also balanced the covariates better than the logistic regression approaches, offering additional evidence that it is more capable of removing bias in baseline differences between the two groups. The average absolute effect size was .11 for the GBM model compared with .14 and .20 for the .05 and the .20 logistic methods, respectively. Although not shown in the table, GBM resulted in no between-group effect sizes of .3 or greater on the pretreatment covariates, whereas the .05 and the .20 logistic models resulted in propensity scores allowing, respectively, 4 and 9 such large differences after weighting.

The GBM model yielded effect size estimates with substantially smaller standard errors than did the logistic me ods. The standard errors of $\text{EATE}_{\text{logit},.05}$ and $\text{EATE}_{\text{logi}}$ for alcohol were 2.2 and 2.5 times larger than that of EATE_{GBM}, and for marijuana the ratios were 1.3 and 2

The different methods provided substantially diffe estimates of the treatment effects. All the effects are posi for alcohol, ranging from 1.5 to 3.1 days, but each estin has a confidence interval that includes zero. The eff were all negative for marijuana but, again, all confide intervals included zero. Although the estimated treatn effects were sensitive to the choice of propensity sc method, the effects appeared weak and within the e bands of the methods. Data with stronger effects m show more consistency across methods.

Sensitivity Analysis

We conducted a sensitivity analysis for estimating possible effects of hidden bias on our treatment ef estimate concerning change in days of marijuana use. S cifically, we examined how the treatment effect estin might change if hidden bias with a magnitude between (

2 and $G = 4$ were present, and we identified the smallest value of G for which the possible treatment effect estimate bounds included a null treatment effect. Ideally, the goal of the sensitivity analysis was to lend confidence to our primary treatment effect estimate by demonstrating that even if covariates that exert medium to large influence on the odds of treatment assignment are unobserved, our treatment effect estimates would not be dramatically altered.

As a benchmark against which to consider the appropriate range of values for G, we studied the effect on our sample weights had we omitted from our propensity score model the variable found to be the most important predictor of treatment assignment—the treatment motivation index. Treating our original propensity score estimates as the true scores, we find that omitting the best predictor of assignment has the effect of increasing some case weights by a factor as large as 2.5 and reducing others by half. In other words, had treatment motivation index been the sole source of hidden bias, the strength of the hidden bias would have a value of G between 2.0 and 2.5. Therefore, in the sensitivity analysis we present in this article, we took values of 2.0 to represent a moderate hidden bias effect and also considered the possible effects of hidden bias twice as large ($G = 4$). As discussed previously, this sensitivity analysis presented a worst-case scenario in which it assumed that the observations in the comparison group with the most extreme values of the outcome were also those with the greatest weight misspecification.

Table 3 presents the results. For four values of G, we estimated the upper and lower bound on the estimated treatment effect that would result from a hidden bias. If hidden bias was large, so that weights could be as much as 4 times too large or small, we might have estimated treatment effects as large as 28.06 (implying treatment actually was less effective than the comparison) to −29.87 (implying that treatment was substantially more effective than the comparison). With G as small as 1.24, the estimated treatment effect could be zero given the observed outcomes and possible effect of the hidden bias. Thus, the estimates in this example are sensitive to hidden bias. In other samples with larger treatment effects and larger sample size, such analyses can potentially show that treatment effects are robust to possible hidden bias.

Table 3
Sensitivity Analysis for Estimated Treatment Effect on Change in Marijuana Use

G	Bounds on treatment effect estimate	
	Maximum	Minimum
1.24	0.00	−11.32
2.00	13.78	−20.58
3.00	23.19	−26.52
4.00	28.06	−29.87

Note. G is a constant representing the strength of hidden bias. Larger values of G correspond to greater possible hidden bias. Maximum and minimum possible treatment effect given hidden bias changes odds of treatment assignment by no more than G and no less than $1/G$.

Discussion

Propensity scores estimated by GBM provide an appealing method for removing the confounding effects of observed covariates on treatment effects estimated with data from nonequivalent groups. GBM offers an adaptive, automated method for estimating propensity scores, which accommodates data with many pretreatment variables, various types of covariates (continuous, nominal, or ordinal), and missing values. Because it is nonparametric model, it reduces the chance of model misspecification errors that have been shown to bias estimates of treatment effects in case-mix adjusted analyses (Drake, 1993). Creating propensity scores and associated weights can purge estimates of treatment effects of the confounding effects of many pretreatment differences in groups. The same weights can also be used to assess the treatment effect for several different outcomes, thus making complex modeling for each outcome variable unnecessary.

The AOP example demonstrates the advantage of GBM for propensity scores. The relationship between pretreatment variables and treatment assignment was distinctly nonlinear as shown in Figure 1. Alternative methods for estimating the propensity scores, such as linear logistic regression, would not capture these nonlinearities, even if the model included low-order polynomial terms. Also, the GBM was fit to 41 correlated variables that were both discrete and continuous. Modeling these variables using a variable-selection method (such as stepwise deletion) and logistic regression has been shown to produce unstable estimates in other context (Breiman, 1996) and resulted in highly variable treatment effect estimates in our example.

Even though the treatment and comparison groups differed considerably at baseline, weighting balanced the group means on nearly all of the 41 variables in the model. These adjustments were critical. Unadjusted group means suggested that the Phoenix Academy reduced marijuana use more than the comparison condition did. However, the difference after weighting was much smaller and was not statistically significant. Nevertheless, weighting did not greatly inflate standard errors so that bias reduction was achieved with minimal gain in variance.

Weighting by propensity scores did not remove group differences in every pretreatment variable in the dataset. However, remaining differences were small and were distributed, as we might expect, with random assignment in a controlled experiment. We used change scores to account for some of the residual difference across groups. To adjust for imbalance that remains after weighting, one might also

use a weighted ANCOVA to estimate treatment effects rather than differences in weighted means. Whereas linear covariate adjustment can be very problematic when group differences are as large as they are prior to weighting, linear adjustments combined with propensity score adjustment can be more effective than propensity score adjustment alone (Huppler-Hullsiek & Louis, 2002; Rosenbaum, 2002).

Covariates often are measured with error. For example, youths may under- or overreport their level of drug use in the months preceding treatment intake. Under the assumption that treatment assignment depends on the observed error-prone covariate (e.g., placement of probationers depends on self-reported drug use), measurement error would not matter. However, selection might depend on the precisely measured covariates (e.g., the true value of drug use rather than self-report), independent of measurement error, in which case the assumption of no measurement error underlying the propensity score method is violated. Sensitivity analyses can be used to explore the possible bias due to measurement error. On the other hand, if the propensity score model results in good balance across groups for the error-prone measures, it might also reduce the confounding effects of the error-free measures. Indeed, using meta-analysis, Shadish and Ragdale (1996) and Heinsman and Shadish (1996) have found that nonequivalent control group studies in which groups have been matched on important pretreatment covariates produce reasonably good estimates of the treatment effects observed in related studies using randomized experimental designs.

The AOP study design attempted to create similar treatment and comparison groups by restricting the comparison group to probationers eligible for the same dispositions from the same criminal justice system during the same time period. This design controlled for many of the selection issues associated with referrals to the Phoenix Academy of Los Angeles. However, even with this careful design, the selection of youths for the Phoenix Academy resulted in nonequivalent groups prior to adjustment. Even with our powerful methods of adjusting for many covariates, we cannot guarantee that selection bias does not exist. Studies should try to design equivalent groups so that balance is easier to achieve, the ESS is large, and the likelihood of hidden biases seems more remote. However, the AOP demonstrates that this is not always possible, and results should be presented with appropriate caveats.

Although GBM offers many advantages over other modeling approaches, the analyst must still tune the model. We have found that models with four levels generally fit as well as more complex models, but this restricts the models to no more than four factor interactions. Shrinkage also affects the fit. Values smaller than .0005 can provide better models at a cost of additional computation and a decreasing marginal improvement in performance.

Our weighted estimator, as given in Equation 2, differs from similar estimators suggested in Wooldridge (2001) and Hirano et al. (2003). Wooldridge suggested using N_T rather than the sum of the weights for the comparison group in the denominator in Equation 2. Hirano and colleagues suggested using the sum of the probabilities for the entire sample in the denominator of Equation 2. All three denominators are estimates of N_T. When the average of estimated probabilities nearly equals the overall probability of being in the treatment group, then all three estimators will provide similar results. When the propensity score model is poorly calibrated and estimated probabilities deviate from the true probabilities of treatment assignment, like the logistic regression models in the AOP example, the sum of the weights or the probabilities will differ from N_T, and the three alternative estimators may differ. In particular, the numerator is also sensitive to the sum of the weights, so the Wooldridge and Hirano et al. estimators can produce treatment effect estimates that vary with the scale of the weights. Our estimator is invariant to the scale of the weights and is more robust to poor calibration in the propensity score. However, the smallest bias or variance an estimator yields is likely to depend on the weights and the correlation between the weights and the outcomes. Additional research is necessary to determine whether one estimator has a smaller mean squared error than the others.

As noted above, propensity score weighted estimates of the treatment effect provide approximately unbiased estimates of the population average treatment effect for the treated provided the appropriate assumptions hold. These analytic results assume large sample sizes approaching infinity. We know of no published studies on the properties of these estimators for small or moderate sized samples. A very limited simulation study that we conducted with a simple treatment assignment function suggested that estimates can be approximately unbiased for small samples and that the rate at which bias decreases with sample size depends on the overlap in the distributions of pretreatment variables between treatment and control groups. The rate of decrease increases with the overlap between the groups. However, these simulation results are very limited, and additional research on the small sample properties of the estimators is necessary.

The goal of case-mix adjustment should be to derive treatment effect estimates with minimum bias and variance. More research is needed on optimizing propensity score models in this way. Current approaches, including our GBM method, adaptively add terms to these models until they satisfy a data-dependent criterion. For instance, the common logistic regression approaches add terms until no significant pretreatment differences remain between groups after conditioning on propensity scores. For GBM, we suggest allowing the algorithm to iterate until the ASAM is minimized. In both cases, emphasis is placed on removing bias resulting from covariate differences. However, these adap

tive procedures might achieve bias reduction by creating highly variable weights and small ESSs.

Improved methods for optimizing propensity score models might exchange worse balance on covariates for substantial reductions in variance. In the AOP example, alternative stopping rules for the GBM model resulted in models with the ASAM about 40% larger than the models we present and jackknife standard errors for the alcohol and marijuana treatment effects that were about 8 and 15% smaller than those presented in this article. The alternative GBM models used a stopping rule that resulted in a less complex model with fewer terms. Similarly restricting the logistic regression models to exclude interactions also greatly reduced the standard error of the estimated treatment effects.

Despite the need for these refinements, we believe the approach to case-mix adjustment presented in this article represents a substantial improvement over the present alternatives available to researchers.

References

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics, 24,* 2350–2383.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Bühlmann, P., & Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association, 98,* 324–339.

Catalano, R. F., Hawkins, J. D., Wells, E. A., & Miller, J. (1991). Evaluation of the effectiveness of adolescent drug abuse treatment, assessing the risk for relapse, and promising approaches for relapse prevention. *The International Journal of the Addictions, 25,* 1085–1140.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *The Journal of the American Medical Association, 276,* 889–897.

Dehejia, J. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 98,* 1053–1062.

De Leon, G., & Deitch, D. (1985). Treatment of the adolescent substance abuser in a therapeutic community. In A. S. Friedman & G. M. Beschner (Eds.), *Treatment services for adolescent substance abusers.* Rockville, MD: National Institute on Drug Abuse.

Dennis, M. L. (1998). *Global Appraisal of Individual Needs (GAIN) manual: Administration, scoring and interpretation.* Retrieved April 28, 2004, from http://www.chestnut.org/LI/gain/index.html

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49,* 1231–1236.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Farley, D. O., Short, P. F., Elliott, M. N., Kanouse, D. E., Brown, J. A., & Hays, R. D. (2002). Effects of CAHPS health plan performance information on plan choices by New Jersey Medicaid beneficiaries. *Health Services Research, 37,* 985–1007.

Fiebach, N. H., Cook, E. F., Lee, T. H., Brand, D. A., Rouan, G. W., Weisberg, M., & Goldman, L. (1990). Outcomes in patients with myocardial infarction who are initially admitted to stepdown units: Data from the multicenter chest pain study. *American Journal of Medicine, 89,* 15–20.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55,* 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29,* 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis, 38,* 367–378.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics, 28,* 337–374.

Gerstein, D. R., & Johnson, R. A. (1999). *Adolescents and young adults in the National Treatment Improvement Evaluation Study.* (National Evaluation Data Services Report). Rockville, MD: Center for Substance Abuse Treatment.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1,* 154–169.

Hirano, K., & Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2,* 259–278.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71,* 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

Hser, Y. I., Grella, C. E., Hubbard, R. L., Hsieh, S. C., Fletcher, B. W., Brown, B. S., & Anglin, M. D. (2001). An evaluation of drug treatments for adolescents in 4 United States cities. *Archives of General Psychiatry, 58,* 689–695.

Hubbard, R. L., Cavanaugh, E. R., Craddock, S. G., & Rachal, J. V. (1985). Characteristics, behaviors, and outcomes for youth in the TOPS. In A. S. Friedman & G. M. Beschner (Eds.), *Treatment services for adolescent substance abusers* (DHHS Publication No. ADM 85-1342, pp. 49–65). Washington, DC: U.S. Government Printing Office.

Huppler-Hullsiek, K., & Louis, T. A. (2002). Propensity score

modeling strategies for the causal analysis of observational data. *Biostatistics, 2,* 1–15.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5,* 299–314.

Imbens, G. (2003). *Nonparametric estimation of average treatment effects under exogeneity: A Review* (National Bureau of Economic Research, Technical Report No. T0294). Retrieved April 28, 2004, from http://www.nber.org/papers/t0294

Jainchill, N. (1997). Therapeutic communities for adolescents: The same and not the same. In G. De Leon (Ed.), *Community as method: Therapeutic communities for special populations and special settings* (pp. 161–178). Westport, CT: Praeger.

Jaycox, L. H., Marshall, G. N., & Morral, A. R. (2002). Phoenix Academy at Lake View Terrace, CA: Clinical manual and program description of an adolescent therapeutic community. Retrieved April 28, 2004, from http://www.chestnut.org/LI/bookstore/index.html

Kong, A., Liu, J., & Wong, W. (1994). Sequential imputation and Bayesian missing data problems. *Journal of the American Statistical Association, 89,* 278–288.

Lieberman, E., Lang, J. M., Cohen, A., D'Agostino, R., Jr., Datta, S., & Frigoletto, F. D., Jr. (1996). Association of epidural analgesia with caesareans in nulliparous women. *Obstetrics and Gynecology, 88,* 993–1000.

Madigan, D., & Ridgeway, G. (2004). Discussion of "Least angle regression" by Efron et al. *Annals of Statistics, 32,* 465–469.

Morral, A. R., Jaycox, L. H., Smith, W., Becker, K., & Ebener, P. (2003). An evaluation of substance abuse treatment services for juvenile probationers at Phoenix Academy of Lake View Terrace. In S. Stevens & A. R. Morral (Eds.), *Adolescent substance abuse treatment in the United States: Exemplary models from a national evaluation study* (pp. 213–234). New York: Haworth Press.

Morral, A. R., McCaffrey, D. F., & Ridgeway, G. (2004). Effectiveness of community-based treatment for substance abusing adolescents: 12-month outcomes from a case-control evaluation of a Phoenix academy. *Psychology of Addictive Behaviors, 18,* 257–268.

Mojtabai, R., & Graff Zivin, J. (2003). Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: A propensity score analysis. *Health Services Research, 38,* 233–259.

Orlando, M., Chan, K., & Morral, A. R. (2003). Retention of court-referred youths in residential treatment programs: Client characteristics and treatment process effects. *American Journal of Drug and Alcohol Abuse, 29,* 337–357.

Reichardt, C. S., Minton, B. A., & Schellenger, J. D. (1980). The analysis of covariance (ANCOVA) and the assessment of treatment effects. In *Prevention Evaluation Research Monograph II—Outcome* (Section D4, chap. VIII). Denver: University of Denver, Department of Psychology.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics, 31,* 172–181.

Ridgeway, G. (2004). *GBM 1.1-2 package manual.* Retrieved April 28, 2004, from http://cran.r-project.org/doc/packages/gbm.pdf

Rosenbaum, P. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82,* 387–394.

Rosenbaum, P. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

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

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39,* 33–38.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 7,* 34–58.

Sells, S. B., & Simpson, D. D. (1979). Evaluation of treatment outcome for youths in the Drug Abuse Reporting Program (DARP): A follow-up study. In G. M. Beschner & A. S. Friedman (Eds.), *Youth drug abuse* (pp. 571–628). Lexington, MA: Lexington Books.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R., & Wong, S. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3,* 3–22.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64,* 1290–1305.

Stone, R. A., Obrosky, S., Singer, D. E., Kapoor, W. N., & Fine, M. J. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. *Medical Care, 33,* AS56–AS66.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–88). New York: Cambridge University Press.

Williams, R. J., Chang, S. Y., & Addiction Centre Research Group. (2000). A comprehensive and comparative review of adolescent substance abuse treatment outcome. *Clinical Psychology: Science and Practice, 7,* 138–166.

Wooldridge, J. (2001). *Econometric analysis of cross section and panel data.* Cambridge: MIT Press.

Zador, P., Judkins, D., & Das, B. (2001). Experiments with MART to automate model building in survey research: Applications to the National Survey of Parents and Youth. In *Proceedings of the annual meeting of the American Statistical Association* [CD-ROM]. Alexandria, VA: American Statistical Association.

Appendix A

Derivation of the Treatment Effect of the Treated Estimator

This section discusses an importance sampling style derivation of the average treatment effect on the treated, $E(y_1|z = 1) - E(y_0|z = 1)$. Whereas the first expectation is trivial to estimate, the second one is not directly estimable.

$$E(y_0|z = 1) = \int y_0 f(y_0|z = 1)dy_0$$

$$= \int\int y_0 f(y_0, \mathbf{x}|z = 1)d\mathbf{x}dy_0 \quad \text{(A1)}$$

The second equality in Equation A1 introduces the pretreatment measures, which can be high dimensional. Although we do not have a sample from $f(y_0, \mathbf{x}|z = 1)$ for the treatment group, we do have a sample from $f(y_0, \mathbf{x}|z = 0)$—all those participants assigned to the comparison group. We can multiply and divide the integrand in Equation A1 by $f(y_0, \mathbf{x}|z = 0)$ to get closer to an expression we can estimate from our data.

$$E(y_0|z = 1) = \int\int y_0 \frac{f(y_0, \mathbf{x}|z = 1)}{f(y_0, \mathbf{x}|z = 0)} f(y_0, \mathbf{x}|z = 0)d\mathbf{x}dy_0$$

$$= E\left(y_0 \frac{f(y_0, \mathbf{x}|z = 1)}{f(y_0, \mathbf{x}|z = 0)}\middle| z = 0\right) \quad \text{(A2)}$$

The expectation in Equation A2 is over the distribution of the comparison group, but the expectand is not directly observed in the data. We can derive weights by applying Bayes' Theorem to the numerator and denominator in Equation A2.

$$E(y_0|z = 1) = \int\int y_0 \frac{f(z = 1|y_0, \mathbf{x}) f(y_0, \mathbf{x}) f(z = 0)}{f(z = 0|y_0, \mathbf{x}) f(y_0, \mathbf{x}) f(z = 1)} \times f(y_0, \mathbf{x}|z = 0)d\mathbf{x}dy_0 \quad \text{(A3)}$$

$$= \frac{f(z = 0)}{f(z = 1)} \int\int y_0 \frac{f(z = 1|y_0, \mathbf{x})}{f(z = 0|y_0, \mathbf{x})} \times f(y_0, \mathbf{x}|z = 0)d\mathbf{x}dy_0 \quad \text{(A4)}$$

The expression $f(z = 1|y_0, \mathbf{x})$ is the probability that a participant with pretreatment variables equal to $\mathbf{x}$ and outcome in the comparison condition equal y_0 is assigned to the control group. We cannot assess this probability from the data without an assumption. Following Rosenbaum and Rubin (1983), we assume that treatment assignment is independent of the outcome given $\mathbf{x}$ so that $f(z = 1|y_0, \mathbf{x}) = f(z = 1|\mathbf{x}) = p(\mathbf{x})$ and $f(z = 0|y_0, \mathbf{x}) = f(z = 0|\mathbf{x}) = 1 - p(\mathbf{x})$. In practice, if $\mathbf{x}$ contains all the factors involved in deciding assignment to the treatment program then this assumption is met.

$$E(y_0|z = 1) = \frac{f(z = 0)}{f(z = 1)} \int\int y_0 \frac{f(z = 1|\mathbf{x})}{f(z = 0|\mathbf{x})} f(y_0, \mathbf{x}|z = 0)d\mathbf{x}dy_0$$

$$= \frac{1 - P(z = 1)}{P(z = 1)} \int\int y_0 \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} f(y_0, \mathbf{x}|z = 0)d\mathbf{x}dy_0 \quad \text{(A5)}$$

Note that this requires that $p(\mathbf{x})$ must be strictly less than 1 for all $\mathbf{x}$. Because we have a sample from $f(y_0, \mathbf{x}|z = 0)$, we can estimate the integral in Equation A5 with the sample average,

$$\hat{E}(y_0|z = 1) = \frac{1 - P(z = 1)}{P(z = 1)} \frac{\sum_{i=1}^{N} y_i w_i (1 - z_i)}{\sum_{i=1}^{N} 1 - z_i}, \quad \text{(A6)}$$

where $w(\mathbf{x}) = p(\mathbf{x})/[1 - p(\mathbf{x})]$, the odds of being in the treatment group. Wooldridge (2001, p. 615) used the fraction of treatment participants in the sample to estimate $P(z = 1)$, whereas Hirano et al. (2003) used the sample average of the observed propensity scores. We noted that

$$1 = \frac{1 - P(z = 1)}{P(z = 1)} \int\int \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} f(y_0, \mathbf{x}|z = 0)d\mathbf{x}dy_0$$

$$\approx \frac{1 - P(z = 1)}{P(z = 1)} \frac{\sum_{i=1}^{N} w_i (1 - z_i)}{\sum_{i=1}^{N} 1 - z_i}. \quad \text{(A7)}$$

Dividing Equation A6 by the 1 estimated in Equation A7 produces the estimator for the average treatment effect of the treated having the form of a weighted average.

$$\hat{E}(y_0|z = 1) = \frac{\sum_{i=1}^{N} y_i w_i (1 - z_i)}{\sum_{i=1}^{N} w_i (1 - z_i)}. \quad \text{(A8)}$$

(*Appendixes continue on next page*)

Appendix B

Details of the Boosting Algorithm

Let z_i be the treatment indicator for participant i. The likelihood principle implies that to get the best estimates of $p(\mathbf{x})$ we should examine the expected Bernoulli log-likelihood function,

$$\mathrm{E}(\ell(p)) = \mathrm{E}(z \log p(\mathbf{x}) + (1 - z)\log[1 - p(\mathbf{x})]|\mathbf{x}). \quad \text{(B1)}$$

Equation B1 implies that we will evaluate any choice for $p(\mathbf{x})$ by how well on average it assigns large probabilities when $z = 1$ and small probabilities when $z = 0$. The true $Pr(z = 1|\mathbf{x})$ maximizes Equation B1. It is important to note that this expectation is with respect to all future participants that might undergo the same treatment assignment process. Conventional practice uses the logistic transform of $p(\mathbf{x})$ to simplify some of the analysis.

$$p(\mathbf{x}) = \frac{1}{1 + \exp[-g(\mathbf{x})]}. \quad \text{(B2)}$$

The logistic transform ensures that, regardless of the value of $g(\mathbf{x})$, $p(\mathbf{x})$ will always be in [0, 1]. If we substitute Equation B2 into Equation B1 then we have the log-likelihood in terms of the regression function $g(\mathbf{x})$, $\ell(g)$, as shown in Equation B3.

$$\mathrm{E}[\ell(g)] = \mathrm{E}(zg(\mathbf{x}) - \log\{1 + \exp[g(\mathbf{x})]\}|\mathbf{x}). \quad \text{(B3)}$$

If we restrict $g(\mathbf{x})$ to be a linear combination of $\mathbf{x}$ and maximize an estimate of the expected log-likelihood with the sample participants, we have exactly a linear logistic regression. However, we allow g to be a member of a flexible family of functions and use boosting to choose the function.

Boosting is a numerical method useful for finding functions that maximize expressions such as Equation B3 from data. The algorithm works as follows. Assume that we have an initial estimate of the function that maximizes Equation B3, which we will call $\hat{g}(\mathbf{x})$. Commonly the overall sample log-odds, $\hat{g}(\mathbf{x}) = \log[\bar{y}/(1 - \bar{y})]$, provides the initial estimate. We would like to improve upon this initial estimate by adding a small adjustment to it. That is, we want to find an $h(\mathbf{x})$ such that

$$\mathrm{E}(\ell(\hat{g} + \lambda h)) > \mathrm{E}\ell(\hat{g}). \quad \text{(B4)}$$

The new improvement offers an increase in the expected log-likelihood and, therefore, we can update our current guess as

$$\hat{g}(\mathbf{x}) \leftarrow \hat{g}(\mathbf{x}) + \lambda h(\mathbf{x}) \quad \text{(B5)}$$

for some step size λ. The remaining problem is how to find an $h(\mathbf{x})$ that satisfies Equation B4.

The derivative of Equation B3, with respect to $g(\mathbf{x})$, indicates the local "direction" to move $g(\mathbf{x})$ for the greatest increase in the expected log-likelihood. Friedman (2001) suggested that such a derivative, therefore, is a reasonable adjustment to our current g.

$$h(\mathbf{x}) = \frac{\partial}{\partial g(\mathbf{x})}\mathrm{E}(L(g)) = \mathrm{E}\left(z - \frac{1}{1 + \exp[-g(\mathbf{x})]}\middle|\mathbf{x}\right)$$
$$= \mathrm{E}(z - p(\mathbf{x})|\mathbf{x}). \quad \text{(B6}$$

The best direction in which we should adjust $\hat{g}(\mathbf{x})$ is a kind c residual, the difference between the treatment indicator and th probability of assignment to the treatment. We cannot comput Equation B6 directly (doing so would require knowledge of th $Pr(z_i = 1|\mathbf{x})$), but we can estimate it with our sample using flexible least squares regression procedure. We use a regressio tree algorithm (Breiman et al., 1984) to estimate $h(\mathbf{x})$ to yield nonparametric and robust prediction model. The regression tre predicts these residuals, $z - p(\mathbf{x})$, from $\mathbf{x}$ using a piecewis constant function, selecting the splits to minimize the mea squared residuals. After fitting a regression tree to the residuals, w can update our estimate for $\hat{g}(\mathbf{x})$, as in Equation B5.

Given a regression tree estimate for h, the update expression i Equation B5 indicates that we simply need to do a line search fc the λ that offers the greatest increase in the log-likelihood. Friec man (2001) suggested a computational shortcut when using re gression trees to estimate Equation B6. The tree is a piecewis constant model. It partitions the participants according to the features into regions, $T_1, T_2, \ldots, T_K$. Within each region, th residuals are relatively homogeneous and the tree estimates $h(\mathbf{}$ for all $\mathbf{x}$s in the region with a constant equal to the mean of th region's residuals. Rather than using the average of the region residuals to estimate the value of $h(\mathbf{x})$ in a node and then pickin a value of λ so that Equation B4 holds, Friedman (2001) suggeste solving for the best update separately for each region. The tree wi do the work of partitioning the participants and defining th regions. Then the optimal adjustment to $\hat{g}(\mathbf{x})$ can be found sep: rately for each region. That is, for all observations that fall into th kth partition, $\mathbf{x} \in T_k$,

$$h(\mathbf{x}) = \arg\max_{\theta} \sum_{x_i \in T_k} z_i[\hat{g}(\mathbf{x}_i) + \theta] - \log\{1 + \exp[\hat{g}(\mathbf{x}_i) + \theta]\}$$

$$\approx \frac{\sum_{x_i \in T_k} z_i - p(\mathbf{x}_i)}{\sum_{x_i \in T_k} p(\mathbf{x}_i)[1 - p(\mathbf{x}_i)]}, \quad \text{(B'}$$

where $\arg\max_\theta$ denotes the value of θ that minimizes the sum. T avoid expensive computation, the estimate in the last line Equation B7 is based on maximizing a second-order Taylor a proximation of the first line. It is very stable as long as not estimated probabilities are 0 or 1 in any terminal node. In tho cases, we set $h(\mathbf{x}) = 0$ for that region.

Combining all of the pieces yields the following boosting alg rithm for fitting a nonlinear logistic regression model to the trea ment assignment data.

1. Initialize $\hat{g}_0(\mathbf{x}) = \dfrac{log\ \bar{z}}{1 - \bar{z}}$
2. For m in $1, \ldots, M$ do

a. Let $r_i = z_i - 1/1 + \exp[-\hat{g}_{m-1}(\mathbf{x}_i)]$.
b. Construct a tree structured predictor of r_i to partition the features into terminal nodes $T_1, \ldots, T_k$.
c. Compute the updates for each terminal node

$$\theta_k = \frac{\sum_{x_i \in T_k} z_i - p(\mathbf{x}_i)}{\sum_{x_i \in T_k} p(\mathbf{x}_i)[1 - p(\mathbf{x}_i)]}.$$

d. Update the logistic regression model as

$$\hat{g}_m(\mathbf{x}) \leftarrow \hat{g}_{m-1}(\mathbf{x}) + \theta_{k(\mathbf{x})},$$

where $k(\mathbf{x})$ determines to which terminal node an observation with features $\mathbf{x}$ belongs.

The algorithm begins with an initial naive guess for $\hat{g}(\mathbf{x})$, the log of the baseline odds of assignment to the treatment group (Step 1). Then Steps 2a–c find a piecewise constant function that offers an improvement in the observed logistic log-likelihood. Last, Step 2d incorporates this new adjustment.

The variability of $\hat{g}(\mathbf{x})$ can also be reduced by modifying Step 2d of the algorithm using a shrinkage coefficient,

$$g_m(\mathbf{x}) \leftarrow g_{m-1}(\mathbf{x}) + \alpha \cdot \theta_{k(\mathbf{x})}, \tag{B8}$$

where $\alpha \in (0, 1]$. Smaller values for α allow the algorithm to make smaller, finer adjustments rather than large, perhaps overconfident changes. Similar strategies exist for many parametric optimization procedures. Smaller values of α will certainly increase the number of iterations needed to produce good propensity score estimates. However, empirical evidence shows smaller αs result in better model fits. Our strategy is to make α as small as possible so that the marginal improvement in log-likelihood for a small α is negligible. This shrinkage strategy reduces the variance without necessarily increasing the bias.

Because each iteration produces a new estimate of g that increases an estimate of the logistic log-likelihood, as additional iterations are performed, eventually $\hat{g}(\mathbf{x})$ will "overfit" the data. Because balance of the pretreatment characteristics is our primary goal, we iterate until we achieve the best matching measured using the average absolute effect size.

Received January 9, 2003
Revision received April 30, 2004
Accepted May 11, 2004 ■

Correction to Venter et al. (2002)

The article "Power in Randomized Group Comparisons: The Value of Adding a Single Intermediate Time Point to a Traditional Pretest-Posttest Design" (*Psychological Methods,* 2002, Vol. 7, No. 2, pp. 194–209) contained two errors on p. 202. Appendix B correctly shows that Equation 17 should read as

$$\beta_{Y_1Y_0} > 0.1667 - 0.1667\rho_{00}$$

instead of as

$$\beta_{Y_1Y_0} > 0.1667 + 0.1667\rho_{00}.$$

Consequently, Equation 18 should read as

$$\rho_{Y_1Y_0}\frac{\sigma_{Y_1}}{\sigma_{Y_0}} > 0.1667 - 0.1667\rho_{00}$$

instead of as

$$\rho_{Y_1Y_0}\frac{\sigma_{Y_1}}{\sigma_{Y_0}} > 0.1667 + 0.1667\rho_{00}.$$

Corresponding threshold values should read as 0 when $\rho_{00} = 1.00$ and 0.08 when $\rho_{00} = 0.50$.