Using Classification Tree Analysis to Generate Propensity Score Weights

Ariel Linden, DrPH[1,2], Paul R. Yarnold, PhD[3]

[1] President, Linden Consulting Group, LLC, Ann Arbor, Michigan, USA

[2] Research Scientist, Division of General Medicine, Medical School - University of Michigan, Ann Arbor, Michigan, USA

[3] President, Optimal Data Analysis, LLC, Chicago, Illinois, USA

Corresponding Author Information:
Ariel Linden, DrPH
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI USA 48103
Phone: (971) 409-3505
Email: alinden@lindenconsulting.org

Key Words: propensity score, machine learning, classification tree analysis, causal inference

Running Header: classification tree analysis and propensity scores

Acknowledgement: We wish to thank Julia Adler-Milstein for her review and feedback on the manuscript. Dr. Yarnold's work was partly funded by a grant from the National Cancer Institute (1R01CA165609-01A1).

ABSTRACT

Rationale, aims and objectives: Widely-used in the evaluation of non-randomized interventions, propensity scores estimate the probability that an individual will be assigned to the treatment group given the observed characteristics. Machine learning algorithms have been proposed as an alternative to conventional logistic regression-based modelling of propensity scores in order to avoid many limitations of linear methods. In this paper we introduce the use of classification tree analysis (CTA) to generate propensity scores. CTA is a “decision-tree”-like classification model that provides accurate, parsimonious decision rules that are easy to visually display and interpret, reports *P* values derived via permutation tests performed at each node, and evaluates potential model cross-generalizability.

Method: Using empirical data, we identify all statistically valid CTA propensity score models and then use them to compute strata-specific, observation-level propensity score weights that are subsequently applied in outcomes analyses. We compare findings obtained using this framework to the conventional method (logistic regression) and a popular alternative machine learning approach (boosted regression), by evaluating covariate balance using standardized differences, model predictive accuracy, and treatment effect estimates obtained using median regression and a weighted CTA outcomes model.

Results: While all models had some imbalanced covariates, main-effects logistic regression yielded the lowest average standardized difference, whereas CTA yielded the greatest predictive accuracy. Nevertheless, treatment effect estimates were generally consistent across all models.

Conclusions: Assessing standardized differences in means as a test of covariate balance is inappropriate for machine learning algorithms that segment the sample into two or more strata. Because the CTA algorithm identifies all statistically valid propensity score models for a sample, it is most likely to identify a correctly specified propensity score model, and should be considered as an alternative approach to modeling the propensity score.

1. INTRODUCTION

Introduced in 1983, the propensity score joined other widely-used methods (e.g. instrumental variables [1,2]) that explicitly model treatment assignment in order to estimate treatment effects in non-randomized studies. The propensity score is defined as the probability of assignment to the treatment group given the observed characteristics [3]. It has been demonstrated that, in sufficiently large samples, if treatment and control groups have similar distributions of the propensity score they generally have similar distributions of the covariates used to create the propensity score (i.e., they exhibit covariate balance). The observed baseline covariates can thus be considered independent of treatment assignment (as if they were randomized), and therefore will not bias treatment effect estimates [3].

Currently there is no consensus regarding how best to estimate the propensity score. In a survey of the literature, Weitzen et al. [4] reported that propensity score estimation is nearly universally performed via logistic regression, and that there is tremendous inconsistency in how models are estimated. For example, some investigators estimate models in which the variable selection process includes only main effects, while others estimate completely saturated models (including all possible interactions, and squared and cubed terms), while others use automated forward or backward stepwise procedures to select variables for model inclusion.

The fundamental concern with this heterogeneous approach to propensity score estimation is that the resulting propensity score model is likely to be misspecified -- that is, the estimated probability of being in the treatment group may differ substantially from the corresponding true probability [5]. With increasing degrees of misspecification it may become implausible to assume that the propensity score accurately represents the underlying covariate distributions, but rather that individuals are not conditionally exchangeable between study groups. In short, a misspecified propensity score may fail to achieve covariate balance between treatment groups, which will subsequently bias treatment effect estimates -- the greater the imbalance, the stronger the bias [6,7].

To avoid the limitations of conventional statistical methods, several investigators have suggested the use of machine learning algorithms as an alternate approach for estimating the propensity score [8-15]. Machine learning algorithms find the best fitting model through automated processes that search through the data to detect patterns that may include interactions between variables, as well as interactions within subsets of variables. This is in contrast to conventional statistics, where a model is chosen and estimated based on an *a priori* hypothesis about the relationship between the variables, and then statistical tests are performed to evaluate whether the data fit crucial assumptions underlying the validity of the findings [16]. In short, machine learning allows the data to dictate the form of the model, whereas conventional statistics attempts to fit the data to an investigator-specified model.

While there are hundreds of machine learning classification algorithms to choose from [17], the models most often examined in the propensity score literature are classification and regression trees (CART) [8,9,11,12], neural networks [11], and ensemble methods such as boosted regression [10,12,14] and random forests [12]. Studies that have conducted head-to-head comparisons between machine learning algorithms and logistic regression for estimating the propensity score have generally found that machine-learning models outperform logistic regression in terms of reduced bias (i.e. the difference between the estimated effect versus the true effect) in the outcome [10-12].

In this paper, we introduce classification tree analysis (CTA) [18,19] and assess whether it offers a superior alternative to logistic regression and boosted regression for estimating propensity scores. CTA is a "decision-tree"-like classification model that provides accurate, parsimonious decision rules that are easy to visually display and interpret, while reporting *P* values derived via permutation tests performed at each node -- making this approach particularly attractive to investigators coming from statistics-based disciplines as compared to other machine learning approaches. In our proposed approach, once a CTA model is generated, strata-specific propensity score weights are computed for all observations in the sample. These weights are then applied in the subsequent outcomes analysis. We illustrate the implementation of the CTA-weighting framework and compare it to weighting approaches using propensity scores derived from conventional logistic regression as well as from boosted logistic regression that is presently the most popular machine learning approach for estimating the propensity score [10].

The paper is organized as follows. In the Methods section we provide a brief introduction to CTA, and describe the data source and analytic framework employed in the current study. The Results section reports and compares the results of the logistic regression, boosted regression and CTA-weighting framework. The Discussion section describes the specific advantages of the CTA-weighting framework for estimating the propensity score and evaluating treatment effects compared with logistic regression and other machine learning approaches, and discusses how CTA can be applied more broadly within the causal inferential framework.

2. METHODS

2.1 A brief introduction to Classification Tree Analysis

In its simplest form, CTA is an optimal discriminant analysis (ODA) model [20]. ODA is a machine-learning algorithm that finds the cutpoint(s) on an ordered attribute (variable) that maximally discriminates between two or more classes (e.g. treatment groups) [21]. The optimal cutpoint is determined by iterating through each value on the attribute and calculating the effect strength for sensitivity [ESS], which is the mean sensitivity amongst the classes, standardized to a 0 - 100% scale where 0 represents the discriminatory accuracy expected by chance, and 100% represents perfect discrimination. By definition, the maximally accurate predictive model uses the "optimal" cutpoint achieving the highest ESS. This model is further subjected to a non-parametric permutation test to assess the statistical validity of that cutpoint. Finally, reproducibility and generalizability of the model are assessed using cross-validation methods [18,22,23].

CTA models use one or more attributes to classify a sample of observations into two or more subgroups that are represented as model endpoints (these are called "terminal nodes" in alternative decision-tree methods). Subgroups are known as "sample strata" because the CTA model stratifies the sample into subgroups of observations that -- with respect to model attributes -- are homogeneous within and heterogeneous between strata [18]. The initial "hierarchically-optimal" CTA algorithm involves chained ODA models in which the initial ("root") node represents the attribute achieving the highest ESS value for the entire sample, and additional nodes yielding greatest ESS are iteratively added at every step on all model branches [24]. In contrast, the enumerated-optimal CTA algorithm explicitly evaluates all possible combinations of the first three nodes, which dominate the solution [25]. The most robust globally-optimal (GO) CTA algorithm explicitly evaluates all possible solutions (called the descendant family), and identifies the model reflecting the best combination of ESS and parsimony (i.e. the model yielding highest ESS using the fewest strata). The software that implements ODA and CTA models provides users with a vast array of options for controlling the modeling process, and a comprehensive description can be found elsewhere [18].

2.2 Data

We use data from a primary care-based medical home pilot program that invited patients to enroll if they had a chronic illness or were predicted to have high costs in the following year [26]. The goal of the program was to lower healthcare costs for program participants by providing intensified primary care [27]. The retrospectively collected data consist of observations for 374 program participants and 1,628 non-participants. Eleven pre-intervention characteristics were available; these included demographic variables (age and gender), health services utilization (primary care visits, other outpatient visits, laboratory tests, radiology tests, prescriptions filled, hospitalizations, emergency department visits, and home-health visits), and total medical costs. The outcome was total medical costs in the program year (see [26] for a more comprehensive description).

2.3 Estimating the propensity score

This study compared three different modelling approaches for estimating propensity scores. Fundamental characteristics of each approach are described in this section, and corresponding methods for computing propensity scores are described in the next section.

The first approach, which is the most commonly used in practice, involves estimating a logistic regression model to predict program participation status using the eleven pre-intervention covariates described above, all entered as main effects. We also estimate a fully saturated logistic regression model which includes the eleven main effects, all possible interactions (including squared terms), and cubed terms for continuous variables. The fully saturated model represents the extreme use of logistic regression for estimating the propensity score, in which every possible relationship between the covariates and outcome (treatment assignment) is explored.

The second approach uses a popular machine learning algorithm called boosted logistic regression for estimating the propensity score [10]. Boosted regression is a procedure in a family of machine learning classifiers called ensemble methods, which combines a large number of relatively simple models (e.g. decision trees) adaptively to optimize predictive performance. Boosting follows a sequential process in which decision trees are fitted iteratively to random subsets of the data, gradually increasing emphasis on observations modelled poorly by the existing collection of trees. The final boosted model is a linear combination of many trees (usually hundreds to thousands) that can be thought of as a regression model where each term is a tree [28]. Here we apply the boosted approach to estimating propensity scores as described by McCaffrey et al [2004], and we implement it in Stata using the user-written program `BOOST` [29], setting the maximum iterations at 20,000, the shrinkage factor to 0.0005, the percentage of data to be used as training data at 80%, the fraction of training observations to be used to fit an individual tree at 50%, and allow up to seven interactions to be assessed. All eleven pre-intervention covariates were used to predict program participation.

The third approach uses the GO-CTA algorithm [18]. For any given dataset, multiple propensity score models having 90% power to test a non-directional hypothesis with experimentwise *P* <0.05 may be generated depending on the subset of covariates and interactions included. We utilize the GO-CTA approach to identify and select the optimal model in the family of all statistically valid CTA models that exist for the sample, evaluating all eleven pre-intervention covariates for inclusion. Point estimates and exact discrete 95% confidence intervals (CIs) are computed for ESS and D (ESS normed for parsimony) for every model in the family, for model performance as well as for chance: if model and chance 95% CIs overlap then the model is judged to be statistically invalid. The GO model is defined as the CTA model within the family of models which has the smallest D statistic. Generalizability of model performance is estimated presently using leave-one-out (LOO) cross-validation. We constrained all CTA models to yield identical predictive accuracy in training and LOO analysis [18,30]. Once all the CTA models were generated, weights were computed for individuals in all end-point strata.

2.4 Generating propensity score weights

Reflecting conventional practice, the inverse probability of treatment weight (IPTW) was computed for each individual in the sample [31]. The IPTW is based on the conditional probability of an individual receiving his/her own treatment: $IPTW_i = (Z_i / p_i) + ([1 - Z_i] / [1 - p_i])$. In this approach an individual i in the treatment group ($Z = 1$) receives a weight equal to the inverse of the estimated propensity score p, and an individual in the control group ($Z = 0$) receives a weight equal to the inverse of 1 minus p. The IPTW weights the treated and control groups to reflect the characteristics of the combined sample in order to estimate the average treatment effect [32,33].

In contrast, for CTA models a stratified weight is generated for each individual based on both their actual treatment assignment and their specific stratum (model endpoint): observations have identical weights if they are classified into the same endpoint and they have the same actual treatment assignment (i.e. treated or non-treated). CTA model-based stratified weights are computed using the following formula:

$$\frac{n_s \times \Pr(Z = z)}{n_{Z\,=\,z,s}} \tag{1}$$

where n_s is the total number of individuals in a given stratum s, $\Pr(Z = z)$ is the estimated probability of assignment to treatment group z (i.e., the proportion of individuals actually receiving treatment z in the sample), and $n_{z\,=\,z,s}$ is the total number of individuals in stratum s who were actually assigned to treatment z. Thus, the weight is proportional to the ratio of the number of individuals in a given stratum relative to the number of individuals within that stratum who do (not) receive treatment. Taken together, the stratification reduces bias in the observed covariates used to create the propensity score, and the weighting standardizes each treatment group to the target population. We developed this stratified weighting approach for the CTA models to ensure that weights conform exactly to the underlying geometry and findings of the CTA model. Although stratified weighting has been shown to produce less bias than IPTW when the propensity score is misspecified [34], we apply IPTW in the comparison models to be consistent with other (prior) studies using machine learning for generating propensity score weights [12].

2.5 Estimating treatment effects

For all propensity score weighted models, we estimated treatment effects using two approaches. In the first approach we estimate treatment effects using quantile (median) regression, in which the outcome variable (medical costs in the program year) is regressed on the treatment indicator, the weights specified as sampling weights, and standard errors and confidence intervals computed via a bootstrap procedure with 2,000 repetitions [35]. Quantile regression is used because medical costs are highly skewed and contain several outliers.

In the second approach, we estimate treatment effects using another CTA model (other than the initial model that generated the propensity scores). Here, medical costs are specified as the attribute, treatment assignment is specified as the class variable, and the weights are used for adjustment. Exact *P* values were estimated using 25,000 Monte Carlo experiments, and LOO analysis (single-case jackknife analysis) was performed to assess potential cross-generalizability of the model in correctly classifying individuals outside of the sample used for model estimation [23,36].

2.6 Performance metrics

We use several methods for assessing the performance across the propensity score estimation models. First, we use the absolute standardized difference statistic for assessing whether weighting on the propensity score successfully balanced the covariates [37]:

$$SD = \frac{\left|\overline{X}_T - \overline{X}_C\right|}{\sqrt{\frac{(S_T)^2 + (S_C)^2}{2}}} \tag{2}$$

where the numerator is the absolute difference in means between the treatment and control groups (denoted as *T* and *C*, respectively) and the denominator is a 50:50 pooled standard deviation [38]. While there is currently no universally-recognized cut-off point as to what is considered the upper limit of balance, Normand et al. [39] suggest that a standardized difference of less than 0.10 is indicative of good balance.

We use ESS to assess the accuracy of fit amongst the various outcome models. The ESS statistic is a chance-corrected (0 = the level of accuracy expected by chance) and maximumcorrected (100 = perfect prediction) index of predictive accuracy. The formula for computing ESS for binary case classification is [40]:

ESS = [(Mean Percent Accuracy in Classification –50)]/ 50 x 100% (3),

where

Mean Percent Accuracy in Classification = (sensitivity + specificity)/2 x 100 (4).

Based on simulation studies, Yarnold and Soltysik [40] consider ESS values less than 25% to indicate a relatively weak, 25% to 50% to indicate a moderate, 50% to 75% to indicate a relatively strong, and 75% or greater to indicate a strong effect. Using ESS, an investigator may directly compare the performance among the various propensity score and outcome models, regardless of structural features of the analyses, such as sample size and the measurement metric.

While ESS compares the predictive accuracy of every given model versus chance, different models may achieve the same level of normed accuracy using different numbers of sample strata. Because model complexity increases as the number of sample strata increases, the D (for "distance") statistic standardizes model ESS for parsimony. The formula for computing D for binary case classification is [18]:

D = 100/(ESS/2)-2 (5),

where the resulting value gives the number of additional effects of identical strength (i.e., ESS) observed for the model that are needed to obtain a theoretically ideal model having perfect accuracy using the minimum number of strata possible for the sample: if accuracy is perfect then D = 0 [41].

Finally, we assess the generalizability (external validity) of the models using LOO cross-validation. We conduct these analyses to assess how well the model predicts treatment assignment to new study participants who may have somewhat different characteristics than those in the original sample. The ESS of the cross-validated model is compared to those of the original model using the entire data-set. The model is considered generalizable if the accuracy measures remain consistent with those of the original model. Current practice guidelines recommend constraining CTA models to have identical ESS in training (total sample) and LOO analysis as a means of inhibiting overfitting and maximizing cross-generalizability [18,42].

2.6 Analytic software

Stata 14.1 (StataCorp., College Station, TX, USA) was used to perform logistic regression and boosted logistic regression for estimating the propensity score and quantile regression for estimating treatment effects (outcome model). We estimated the two logistic regression models (main effects only, and fully saturated) using a user-written command for Stata, `LOOCLASS` [43], which performs LOO and produces several classification measures. We estimated a boosted logistic regression implementing the user-written program `BOOST` [29], within a modified wrapper program of `LOOCLASS` to provide the LOO estimates for the boosted model. Standardized differences were computed using a user-written command for Stata, `COVBAL` [44]. GO-CTA was conducted to generate and assess the accuracy of propensity score models, and to model outcomes, and was performed using CTA software [18,25].

3. RESULTS

Table 1 presents the observed pre-intervention characteristics of the participants and non-participants in the pilot study [26]. Continuous variables are summarized by the mean and standard deviation, and categorical variables are presented as number and percent. For balance measures, we report the standardized difference, for which perfect balance is zero and the conventional *P* value, where variables with values ≤ 0.05 may be considered imbalanced. It is clear that the participant group differed markedly from the non-participant group on every characteristic. On average, participants were older, were less likely to be female, and had higher utilization and costs than non-participants. All standardized differences exceeded the recommended value of 0.10, and all *P* values were ≤ 0.05. Thus, it is readily apparent that this non-randomized study exhibits substantial selection bias.

Table 2 summarizes the structure (number of strata, smallest strata N) and performance (ESS, D) of all CTA models that emerged for discriminating between study participants and non-participants. Disqualified models either failed to achieve the minimum denominator criterion (N≥34) specified in power analysis (steps 1-4), or had 95% CIs for D (steps 13-14) lower than for the globally-optimal model in the family (step 12). The descendant family (DF) thus consisted of the eight models in steps 5-12: note that all eight models had ESS 95% CIs that overlapped and reflected relatively strong normed predictive accuracy, and all eight models had chance ESS 95% CIs that overlapped and reflected relatively weak normed predictive accuracy. However, model 12 is unambiguously identified as the GO model since its 95% CI for D lay below corresponding 95% CIs for all other models in the DF [18,41].

Although all eight models in the DF may be used to construct propensity scores, for this exposition we limit our analysis to the four models illustrated in Appendix Figures 1-4, respectively. Results are reported for the least complex two-strata GO model [CTA-2] (Table 2, step 12); the most accurate (highest ESS) next-least complex four-strata model [CTA-4] (step 9); the sole intermediate-complexity six-strata model [CTA-6] (step 8); and the nine-strata model [CTA-9] -- the first and most complex member of the DF, offering greatest stratification granularity (step 5). All of these models had overlapping 95% CIs for ESS, and all correctly classified at least 3 of 4 non-participants, and 4 of 5 program participants.

Table 3 presents standardized differences of all covariates and the average absolute standardized difference for each of the seven (logistic-main effects, logistic-saturated, boosted, and four CTA) weighted propensity score models. The main effects only logistic regression model with IPTW achieves the lowest average standardized difference amongst the models, but is far from ideal in achieving covariate balance. For example, age remains substantially unbalanced between participants and non-participants, and to a lesser degree so do the number of prescriptions filled, emergency department visits, and other outpatient visits. Interestingly, the saturated logistic regression model performed worse in achieving covariate balance than the main effects only model. This may be due to the very large number of covariates used in the estimation model (166) relative to the number of observations (2002), resulting in data patterns known as complete or quasi-complete separation [45]. All other models performed substantially worse than logistic regression with IPTW in achieving covariate balance.

Table 4 presents treatment effect estimates using median regression for the seven weighted propensity score models and also for a naïve estimate, which is simply a regression of the outcome on the treatment indicator without adjustment for confounding. All models show that the median costs of the participants in the program year were higher than the median costs of non-participants. The treatment effect estimates for the seven weighted models span a relatively narrow range between $738 and $1536, and all models except for the saturated logistic model (*P* <0.094) achieve statistical significance (*P* <0.0001).

Table 5 presents treatment effect estimates using weighted CTA outcome models for the seven weighted propensity score models and also for the naïve estimate. For every analysis, the first row of data are for the training (full sample) analysis, and the second row for leave-one-out (LOO) one-sample jackknife analysis. For all models, observations having costs less than or equal to the tabled threshold value (each threshold value is computed using the indicated model) are predicted to be from the non-participant group, and observations having costs that are greater than the tabled threshold are predicted to be from the participant group.

For example, the cutpoint in the unweighted naïve model indicates that non-participants were predicted to have medical costs ≤ $2664 while participants were predicted to have costs > $2664. The accuracy (and LOO cross-generalizability) of these predictions is represented by the respective sensitivities, overall ESS, and permutation *P* values. In the case of the naïve estimate, the full sample sensitivity of the non-participant group was 68.2%, indicating that 68.2% of non-participants we accurately predicted to have costs ≤ $2664. Similarly, the full sample sensitivity of the participant group was 82.6%, indicating that 82.6% of participants were accurately predicted to have costs > \$2664. The ESS for the naïve model was 50.8%, indicative of relatively strong overall classification accuracy [40]. Furthermore, the exact P<0.0001 for the naïve model indicates that the participant group had statistically higher cost than then non-participant group. Finally, the model is generalizable, as indicated by LOO values that are nearly identical to those of the full sample.

All weighted models were statistically significant (exact P<0.0001). Full sample weighted ESS (WESS) values ranged between 28.5% and 37.3%, and LOO WESS values ranged between 26.3% and 36.6%, indicative of moderate classification accuracy and cross-generalizability [40]. Taken together the findings of the CTA outcomes analyses were qualitatively similar to findings derived via median regression. That is, the participant group had statistically higher costs than the non-participant group, across all models. We close with concluding comments.

4. DISCUSSION

Given that main-effects logistic regression generated propensity scores weights that yielded the lowest mean standardized difference measure of covariate balance, and produced median-regression-based treatment effect estimates that were consistent with estimates of all the other models, one may question the value of using alternative approaches to generate propensity scores. However, in using empirical data where the true treatment effect is never known, we highlight challenges investigators face when developing propensity score models using logistic regression to derive the best (i.e. least biased) estimate.

First, neither the main effects nor the fully saturated logistic regression models generated propensity scores that yielded good covariate balance, indicated by standardized differences for several covariates that were substantially higher than the recommended upper bound of 0.10 [39]. If in fact it is possible to attain a correctly specified logistic regression model for the present sample, then it lies somewhere between these extreme (main effects only, versus completely saturated) specifications. However, a correctly specified logistic regression model is unlikely to be discovered by using a manual variable selection approach.

Second, as an increasing number of variables, interactions, and polynomial terms are added to the model, violations of statistical assumptions underlying the validity of the model estimates become increasingly likely. This underscores a clear advantage of using automated machine learning algorithms, which require no statistical assumptions in selecting model terms, as an alternative to logistic regression for generating propensity scores.

Third, as expected, varying the specifications for estimating the logistic regression model yielded qualitatively different findings. The main effects logistic regression model produced estimated treatment effects consistent in magnitude and statistical significance to estimates of all weighted models except for the saturated logistic regression model, which produced an estimated treatment effect that was substantially lower than that obtained by all other models, and was not statistically significant (Table 4). This finding supports conducting sensitivity analysis to assess the consistency of findings obtained by different models (or specifications) as a standard practice in the propensity score modeling process, in order to increase confidence in the validity of the analytic results [46]. By design, the CTA framework conducts such a sensitivity analysis for the propensity score models. In the present study, analysis identified all 14 potential CTA-based propensity score models that exist for the study data, of which eight met all statistical validity criteria (for exposition we proceeded with four of these eight models). All CTA models produced overlapping outcomes, cutpoints, ESS, and *P* values for all weighted models, and exhibited consistency in the degree of generalizability of the estimates. In achieving similar outcomes under different propensity score model specifications, we gain confidence that the analytic approach produces valid results.

Our empirical results also reveal that the standard approach to assess covariate balance as an indicator of comparability between study groups is problematic. None of the machine learning-based models (nor the saturated or boosted logistic models) achieved covariate balance using the criterion of an average standardized difference <0.10. This indicates that the standardized difference is not an appropriate metric for assessing comparability between study groups when such models are implemented. The standardized difference measures the difference in the means of two (assumingly normal) distributions. However, machine learning algorithms rarely deal with entire distributions of a variable, but rather subsets -- and interactions between subsets -- of available variables. Therefore, metrics based on distributional assumptions of the entire variable are not relevant to machine learning models.

On the other hand, CTA models by design provide results in a decision-tree-like format that allows for direct inspection of balance, with all individuals that end in the same stratum (terminal node) comparable on all the attributes that define that terminal node. Concomitantly, this format also indicates the degree of overlap between study groups in these covariate patterns. More specifically, any terminal node that contains 100% of observations from a single study group has no counterfactual and thus causal inferences cannot be made about the effects of the intervention on that subset of observations. In such cases, all observations with no counterfactual within a given terminal node may be dropped from the analysis, and the CTA model should be re-estimated. While this methodology is applicable to CTA and classification and regression trees (CART) algorithms that provide results in a decision-tree or decision-rules format, it is not clear how best to assess covariate balance when using "black-box" algorithms (e.g. boosted regression, random forests, support vector machines, etc.).

An important issue associated with the use of machine learning tools for generating propensity score models is the choice and number of variables determined by the model. The recommended approach for estimating the propensity score is to "be liberal in terms of including variables that may be associated with treatment assignment and/or the outcomes" [46]. However, classification algorithms are specifically designed to exclude variables that do not contribute to predictive accuracy. Indeed, CTA explicitly maximizes ESS so forcing additional variables into a CTA model will reduce ESS and/or D. Moreover, many "black-box" machine learning algorithms do not report the number or identify of variables included in the model. Taken together, it is clear that recommendations for estimating propensity score models could be improved by including the application of machine learning techniques.

CTA methodology holds several advantages over conventional logistic regression for estimating propensity score models, such as using an automated process for optimizing variable selection, being unencumbered by the assumptions required of parametric models, and insensitivity to skewed data and outliers [18]. Moreover, the built-in sensitivity analysis for GO-CTA is more likely to consistently identify a correctly specified propensity score model than when using logistic regression. Additionally, while this paper has demonstrated the implementation of the CTA framework to generate propensity score weights for pretest-posttest studies with a binary treatment, the approach can be extended to any study design that may utilize propensity score weights (see for example [48-52]).

CTA methodology also carries advantages over other machine learning algorithms for estimating propensity score models. In contrast to the more computationally-intensive machine learning techniques typically favored for generating propensity scores, CTA models offer transparency in the computational approach, interpretable formulae, and straightforward visual displays of the final model [53]. Moreover, the GO-CTA algorithm identifies all statistically valid propensity score models for a sample, which vary in terms of predictive accuracy (ESS) as well as parsimony (number of strata) [18]. As a general rule, a simpler model is always preferred over a more complex model, assuming both have the same classification accuracy. Finally, CTA includes permutation tests, adjusted for multiple comparisons, to ensure that the final model meets rigorous statistical assumptions, and can use multiple methods to assess potential cross-generalizability [18]. Thus, one may consider CTA as an “all-in-one” classification algorithm that combines the synergies of machine learning and conventional statistics. That is, the machine learning component ensures that the final model achieves maximum accuracy (as measured by cross-validated ESS), and the permutation tests, performed at each node, ensure that the model’s discriminatory ability has met accepted levels of statistical significance.

The primary limitation of the CTA framework -- as is the case with every approach used to evaluate non-randomized studies -- is the models are generated using only the available data. No matter how sophisticated the algorithm, unobservable factors such as unmeasured motivation to change health behaviors may confound the outcomes in healthcare interventions [54,55].

5. CONCLUSION

In summary, this paper introduced a novel machine learning framework for generating propensity score weights to evaluate treatment effects in observational studies. This framework offers many advantages over both logistic regression as well as other machine learning algorithms, such as explicit maximization of accuracy, parsimony, sensitivity, statistical robustness, and transparency. Because the CTA algorithm identifies all statistically valid propensity score models for a sample, it is most likely to identify a correctly specified propensity score model, and should be considered as an alternative approach to modeling the propensity score.

REFERENCES

1. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996;91:444– 455.
2. Linden A, Adams J. Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice* 2006;12: 148-154.
3. Rosenbaum PR, Rubin DB. The central role of propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
4. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety* 2004;13:841e53.
5. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American Journal of Epidemiology* 2006;163:1149-56.
6. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993;49:1231– 1236.
7. Smith J, Todd P. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* 2001;91:112–18.
8. Cook EF, Goldman L. Asymmetric stratification. An outline for an efficient method for controlling confounding in cohort studies. *American Journal of Epidemiology* 1988;127:626-639.

9. Barosi G, Ambrosetti A, Centra A, Falcone A, Finelli C, Foa P, et al. Splenectomy and risk of blast transformation in myelofibrosis with myeloid metaplasia. Italian Cooperative Study Group on Myeloid with Myeloid Metaplasia. *Blood* 1998;91:3630-3636.

10. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 2004;9:403-425.

11. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety* 2008;17:546-555.

12. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; 29(3): 337–346.

13. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* 2010; 63(8): 826–833.

14. Wyss R, Ellis AR, Brookhart MA, Girman CJ, Funk MJ, LoCasale R, Stürmer T. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology* 2014;180:645-55.

15. Neugebauer R, Schmittdiel JA, van der Laan MJ. A Case Study of the Impact of Data-Adaptive Versus Model-Based Estimation of the Propensity Scores on Causal Inferences

from Three Inverse Probability Weighting Estimators. *The International Journal of Biostatistics* 2016;12:131-55.

16. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science* 2001;16:199–231.
17. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 2014;15:3133–3181.
18. Yarnold PR, Soltysik RC. *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books, 2016. DOI: 10.13140/RG.2.1.1368.3286
19. Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine* 1997;16:1451-1463.
20. Yarnold PR, Soltysik R.C. Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences* 1991;22:739-752.
21. Linden A, Yarnold PR. Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice* 2016;22:875-885.
22. Linden A, Yarnold PR, Nallamothu BK. Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice* 2016;22:860-867.

23. Linden A, Yarnold PR. Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice* 2016;22:855-859.

24. Yarnold PR. Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement* 1996;56:656-667.

25. Soltysik RC, Yarnold PR. Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis* 2010;1:144-160.

26. Linden A. Identifying spin in health management evaluations. *Journal of Evaluation in Clinical Practice.* 2011;17:1223-1230.

27. Linden A, Adler-Milstein J. Medicare disease management in policy context. *Health Care Finance Review* 2008;29:1-11.

28. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *Journal of Animal Ecology* 2008;**77:**802–813.

29. Schonlau M. Boosted regression (boosting): An introductory tutorial and a Stata plugin. *Stata Journal* 2005;5:330–354.

30. Yarnold PR, Linden A. Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis* 2016;5:65-73.

31. Linden A, Adams JL. Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice* 2010;16:175-179

32. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–560.

33. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987;82:387–394.

34. Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics* 2010;35:499-531.

35. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to the bootstrap technique. *Disease Management and Health Outcomes* 2005;13: 159-167.

36. Linden A, Adams J, Roberts N. The generalizability of disease management program results: getting from here to there. *Managed Care Interface* 2004;17:38-45.

37. Flury BK, Reidwyl H. Standard distance in univariate and multivariate analysis. *The American Statistician* 1986;40:249–251.

38. Linden A, Samuels SJ. Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice* 2013;19:968–975.

39. Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology* 2001;54:387–398.

40. Yarnold PR, Soltysik RC. *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books, 2005.

41. Yarnold PR, Linden A. Theoretical aspects of the D statistic. *Optimal Data Analysis* 2016;5:171-174.

42. Yarnold PR. Determining jackknife ESS for a CTA model with chaotic instability. *Optimal Data Analysis* 2016;5:11-14.

43. Linden A. LOOCLASS: Stata module for generating classification statistics of Leave-One-Out cross-validation for binary outcomes. Statistical Software Components s458032, Boston College Department of Economics, 2015. Downloadable from http://ideas.repec.org/c/boc/bocode/s458032.html [Accessed on 12 January 2017].

44. Linden A. COVBAL: Stata module for generating covariate balance statistics. Statistical Software Components s458188, Boston College Department of Economics, 2016. http://ideas.repec.org/c/boc/bocode/s458188.html [Accessed on 12 January 2017].

45. Allison PD. Convergence failures in logistic regression. *SAS Global Forum* 2008;360:1-11.

46. Linden A, Adams J, Roberts N. Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice* 2006;12:140-147.

47. Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science* 2010;25:1–21.

48. Linden A, Adams JL. Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice* 2010;16:180-185.

49. Linden A, Adams JL. Applying a propensity-score based weighting model to interrupted time series data: improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice.* 2011;17:1231-1238.

50. Linden A, Adams JL. Combining the regression-discontinuity design and propensity-score based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice.* 2012;18(2):317-325.

51. Linden A. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice* 2014;20(6):1065-1071.

52. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management* 2004;7:180-190.

53. Linden A, Yarnold PR. Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice* 2016;22:839-847.

54. Linden A, Roberts N. Disease management interventions: What's in the black box? *Disease Management* 2004;7:275-291.

55. Linden A, Butterworth S, Roberts N. Disease management interventions II: What else is in the black box? *Disease Management* 2006;9:73-85.

Table 1: Baseline (12 months) characteristics of program participants and non-participants (Linden [2011]).

	Participants (*N*=374)	Non-Participants (*N*=1628)	Standardized difference	*P*-value[a]
Demographic characteristics				
Age	54.9 (6.71)	43.4 (11.99)	1.177	<0.001
Female	211 (56.4%)	807 (49.6%)	0.137	0.017
Utilization and Cost				
Primary care visits	11.3 (7.30)	4.6 (4.35)	1.110	<0.001
Other outpatient visits	18.0 (16.65)	7.2 (10.61)	0.772	<0.001
Laboratory tests	6.1 (5.27)	2.4 (3.31)	0.844	<0.001
Radiology tests	3.2 (4.46)	1.3 (2.48)	0.524	<0.001
Prescriptions filled	40.6 (29.96)	11.9 (17.14)	1.174	<0.001
Hospitalizations	0.2 (0.52)	0.1 (0.29)	0.403	<0.001
Emergency department visits	0.4 (1.03)	0.2 (0.50)	0.287	<0.001
Home-health visits	0.1 (0.88)	0.0 (0.38)	0.108	0.012
Total costs	8236 (9830)	3047 (5817)	0.643	<0.001

[a] A two-tailed t-test for independent samples was used for continuous variables, and a Chi-square test was used for dichotomous variables. Continuous variables are reported as mean (standard deviation) and dichotomous variables are reported as *N* (percent).

Table 2: All CTA models discriminating between study participants and non-participants

Step	Strata	Smallest Strata N	ESS for Model (95% CI)	ESS for Chance (95% CI)	D (95% CI)
1	16	6	68.02 (62.93 – 72.83)	1.68 (0.29 – 4.97)	7.52 (5.97 – 9.43)
2	15	16	67.40 (62.31 – 72.35)	1.64 (0.00 – 4.93)	7.26 (5.73 – 9.07)
3	11	24	67.30 (62.52 – 71.93)	1.75 (0.22 – 5.16)	5.34 (4.29 – 6.59)
4	14	29	67.07 (61.80 – 72.03)	1.65 (0.33 – 4.94)	6.87 (5.44 – 8.65)
5	9	48	66.81 (62.16 – 71.45)	1.71 (0.07 – 5.33)	4.47 (3.60 – 5.48)
6	8	55	66.14 (61.07 – 70.98)	1.70 (0.05 – 5.21)	3.97 (3.20 – 5.10)
7	7	84	65.07 (59.89 – 70.08)	1.68 (0.03 – 4.96)	3.76 (2.99 – 4.69)
8	6	107	63.53 (58.60 – 68.31)	1.90 (0.07 – 5.33)	3.44 (2.78 – 4.24)
9	4	187	62.25 (57.34 – 66.94)	1.87 (0.23 – 5.36)	2.43 (1.98 – 2.98)
10	4	222	58.67 (54.39 – 62.84)	1.83 (0.19 – 5.45)	2.82 (2.37 – 3.35)
11	4	244	58.05 (53.09 – 62.91)	1.82 (0.18 – 5.44)	2.89 (2.36 – 3.53)
12	2	675	57.84 (52.40 – 63.14)	1.94 (0.30 – 5.23)	1.46 (1.17 – 1.82)
13	2	823	45.46 (39.77 – 51.18)	1.89 (0.08 – 5.67)	2.40 (1.91 – 3.03)
14	2	984	6.85 (0.16 – 13.33)	1.92 (0.27 – 5.20)	27.20 (13.0 – 1,248)

Notes: Strata is the number of model endpoints (terminal nodes); smallest strata N is the number of observations in the endpoint with the smallest number of observations among all endpoints in the model; ESS is a measure of normed predictive accuracy (0=accuracy expected by chance; 100=perfect accuracy); exact 95% confidence intervals for model and chance ESS are computed using 10,000 bootstrap and Monte Carlo iterations, respectively; and the D statistic indicates the number of additional effects with equivalent ESS needed to obtain a theoretically ideal model with perfect accuracy and maximum possible parsimony for the application [Y&L 2016; Y&S, 2016].

Table 3: Absolute standardized differences of baseline covariates, and average standardized difference, for all propensity score models.

	Absolute Standardized differences						
Characteristic	Logistic (main effects)	Logistic (saturated)	Boosted	CTA-2	CTA-4	CTA-6	CTA-9
Age	0.397	0.603	0.863	0.980	0.626	0.757	0.743
Female	0.027	0.068	0.148	0.184	0.079	0.133	0.193
Primary care visits	0.094	0.239	0.501	0.829	0.337	0.364	0.237
Other outpatient visits	0.126	0.170	0.297	0.659	0.347	0.443	0.424
Laboratory tests	0.017	0.212	0.389	0.588	0.295	0.129	0.176
Radiology tests	0.006	0.146	0.207	0.407	0.159	0.194	0.159
Prescriptions filled	0.154	0.232	0.545	0.301	0.564	0.45	0.171
Hospitalizations	0.037	0.022	0.113	0.244	0.152	0.211	0.061
Emergency department visits	0.141	0.139	0.145	0.213	0.147	0.038	0.15
Home-health visits	0.015	0.025	0.031	0.031	0.024	0.001	0.001
Total costs	0.039	0.123	0.253	0.417	0.274	0.342	0.190
Average standardized difference	0.096	0.180	0.317	0.441	0.273	0.278	0.228

Notes: Inverse probability of treatment weights (IPTW) were used with logistic and boosted logistic regression models, and stratified weights were used with classification tree analysis (CTA) models.

Table 4: Treatment effect estimates using quantile (median) regression as the outcome model. CIs were computed using a bootstrap procedure with 2,000 repetitions

Model	Participants	Non-participants	Difference	95% CI	*P*-value
Naïve	4,819	1,799	3,020	[2758,3282]	< 0.0001
Logistic (main)	3,518	2,346	1,172	[651,1693]	< 0.0001
Logistic (saturated)	2,841	2,103	738	[-126,1602]	0.094
Boosted	3,480	2,000	1,480	[886,2074]	< 0.0001
CTA-2	3,554	2,018	1,536	[1103,1969]	< 0.0001
CTA-4	3,407	2,042	1,365	[669,2061]	< 0.0001
CTA-6	3,310	2,083	1,227	[685,1769]	< 0.0001
CTA-9	3,084	2,111	973	[430,1516]	< 0.0001

Table 5: Treatment effect estimates using CTA as the outcome model

Model	Cutpoint Predicting Non-participants	Sensitivity* (Non-participants)	Sensitivity* (Participants)	WESS*
Naïve	≤ 2664	68.2	82.6	50.8
LOO		68.2	82.4	50.5
Logistic (main)	≤ 1470	32.1	96.5	28.5
LOO		32.0	94.2	26.3
Boosted	≤ 1730	43.3	92.6	35.9
LOO		43.3	92.3	35.5
Logistic (sat)	≤ 1470	35.2	96.8	32.0
LOO		35.1	91.3	26.4
CTA-2	≤ 2425	59.0	78.0	37.0
LOO		59.0	73.4	32.4
CTA-4	≤ 1980	48.9	88.4	37.3
LOO		48.9	87.8	36.6
CTA-6	≤ 1740	42.0	92.3	34.2
LOO		42.0	87.0	28.9
CTA-9	≤ 1953	46.6	89.4	36.1
LOO		46.6	88.0	34.6

Notes: * All estimates are weighted with the exception of the naïve model. WESS is weighted ESS: 0=weighted ESS expected by chance, 100=perfect prediction. For every analysis, the first row of data are for the training (full sample) analysis, and the second row of data are for the leave-one-out (LOO) one-sample jackknife analysis. For all models, observations having costs less than or equal to the tabled threshold value (computed by the ODA algorithm) are predicted to be from the non-participant group (coded as 0), and observations having costs that are greater than the tabled threshold are predicted to be from the participant group (coded as 1). Exact *P* <0.0001 for all tabled ESS values. The D statistic is not needed to further norm ESS for parsimony, because all of the Tabled models had two terminal nodes (endpoints).