

# Comparing CTA to Boosted Regression for Estimating the Propensity Score (*Invited*)

Ariel Linden, Dr.P.H.  
Linden Consulting Group, LLC

Boosted regression (BR) has been recommended as a machine learning alternative to logistic regression for estimating the propensity score because of its greater accuracy. Commonly known as multiple additive regression trees, BR is a general, automated, data-adaptive modelling algorithm which can estimate the non-linear relationship between treatment assignment (the outcome variable) and a large number of covariates including multiple level interaction terms. However, BR is a “black-box” approach that provides scant information as to how the estimates are derived, and recent research has shown that BR can identify erroneous relationships between outcome and covariates in fabricated random data. The present paper revisits the BR approach used by Linden, et al. (2010) for estimating the propensity score and compares it to a propensity score CTA model which is generated by using the new Stata package for implementing CTA.

Studies in which participants are randomized to treatment are considered the gold standard for assessing causal inference because randomization putatively ensures that the study groups do not differ systematically in their characteristics, and consequently, treatment effects are assumed to be unbiased.<sup>1</sup> If randomization is infeasible, investigators rely on statistical techniques which model treatment assignment in order to control for threats to validity<sup>2,3</sup> which may compromise causal interpretation of the results.<sup>4-8</sup>

Boosted regression (BR) has been recommended as a machine learning alternative to logistic regression for estimating the propensity score because of its greater accuracy.<sup>9</sup> BR, com-

monly referred to as multiple additive regression trees, is a general, automated, data-adaptive modelling algorithm which can estimate the non-linear relationship between the treatment assignment (the outcome variable) and a large number of covariates including multiple level interaction terms.<sup>10</sup> However BR is a “black-box” approach that provides limited information as to how the estimates are derived. Moreover, recent work as shown that BR can identify erroneous relationships between outcome and covariates in fabricated random data.<sup>11</sup>

Herein I employ CTA to generate a propensity score model which was originally estimated using BR in the evaluation of a moti-

vational interviewing-based health coaching intervention for a chronically ill group of participants compared with non-participants.<sup>12</sup> The intervention focused on employing behavioral change to help patients actively engage in their own health care management.<sup>13-17</sup>

## Methods

### Data

A total of 106 individuals with a chronic illness who had participated in the health coaching program and taken the survey twice were compared to 230 individuals with a chronic illness who chose not to participate but also took the survey twice. And, a total of 60 independent variables (attributes) were available for inclusion in the model including: age; gender; job category; self-management level of nine lifestyle and health behaviors; presence and count of up to 10 chronic illnesses; body mass index (BMI); and summary scores from three health-related survey instruments (see Linden et al.<sup>12</sup> for details).

### Analyses

Linden et al.<sup>12</sup> estimated used BR to estimate the propensity score in which the binary treatment indicator was regressed on all 60 covariates. For comparison, the authors also estimated the propensity score using standard logistic regression (LR).

Boosted LR provided superior predictive accuracy across the entire range of propensity scores vs. standard LR (area under the curve [AUC] was 97.4% and 87.4% for boosted regression and standard LR respectively). The boosted regression created 384 regression trees in the iterative process, suggesting that a large number of multiple-level interaction terms were identified. The four independent variables that explained the most variation in the propensity score (either alone or as a component of one or more interaction terms) were: baseline self-efficacy (15%), body mass index (10%), age

(7%), and patient activation measure scores (7%). All remaining attributes explained less than 5% of the overall variation in propensity scores. Results reported for the BR model were for the training sample alone, and therefore we do not know if this extremely high accuracy rate (AUC = 97.4%) was a function of erroneous relationships, or was real.<sup>11</sup>

First, I re-estimated the BR model and perform leave-one-out (LOO) analysis to derive the AUC on the hold-out sample using the Stata package **looclass**.<sup>18</sup> The resulting AUC for the hold-out sample was 73.51%, which is significantly (24.5%) lower than the 97.4% obtained for the training sample ( $P < 0.0001$ ).

Next, I use the new Stata package called **cta**, which implements CTA within the Stata environment to generate a CTA propensity score model using the same 60 attributes.<sup>19</sup> As **cta** is a wrapper for CTA software<sup>20</sup>, the CTA64.exe file (available at <https://odajournal.com/resources/>) must be on the computer for **cta** to work. To download the **cta** package, at the Stata command line type: “ssc install cta” without the quotation marks.

The following syntax generated the CTA model (see the help file for **cta** for a complete description of the syntax options):

```
cta coached cov1-cov60, pathcta("C:\CTA\  
store("C:\CTA\output") loo(stable)  
iter(10000) prune(0.05) enumerate
```

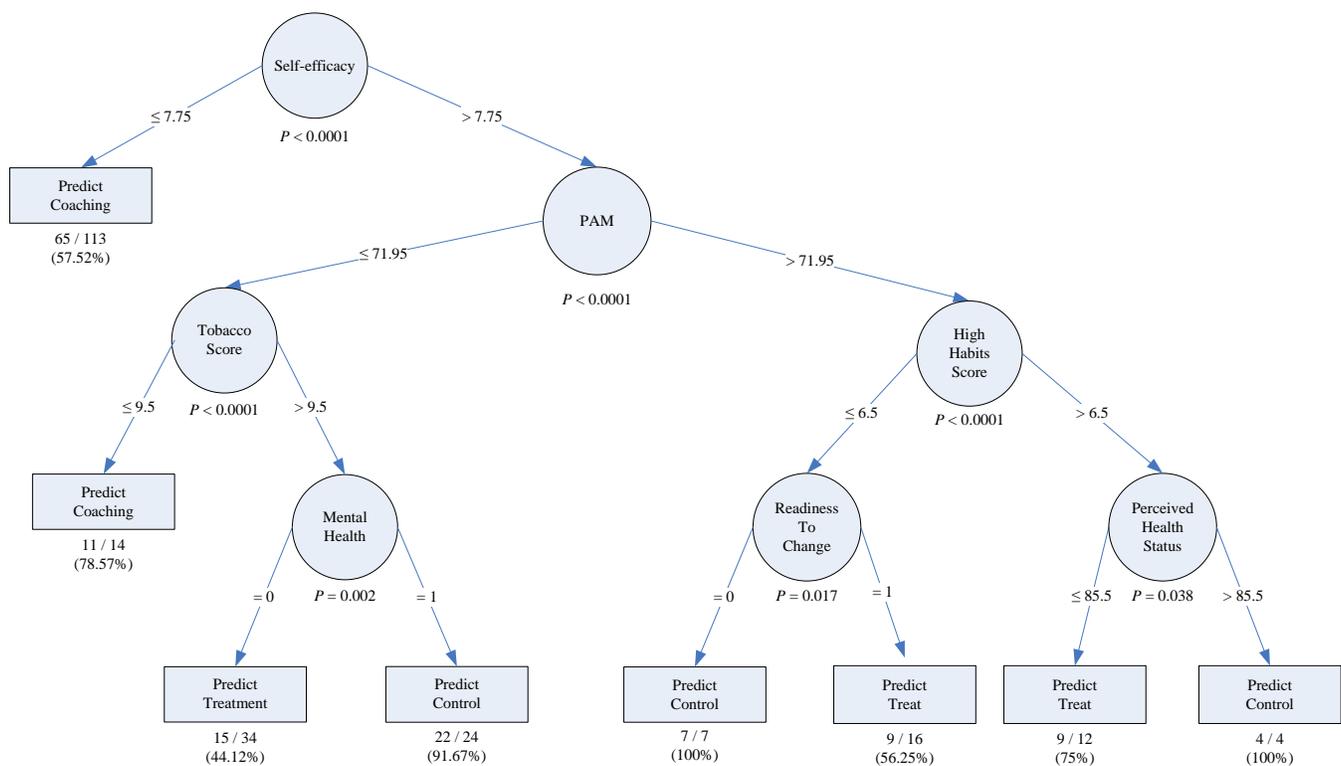
The above syntax is explained as follows: The outcome variable “coached” is the *class* variable; the 60 variables listed before the comma (cov1-cov60) are covariates specified as *attributes*; the directory path for the CTA64.exe file on my computer is “C:\CTA\”; the directory path where output and other files generated during analysis are stored is “C:\CTA\output”; the number of iterations (repetitions) for computing a permutation *P*-value is 10,000; leave-one-out (LOO) cross-generalizability analysis is used to identify and retain attributes having the same classification performance in LOO and

training (total sample) analysis; the tree is pruned with experimentwise  $P < 0.05$  used as the cutpoint for inclusion; and a model enumerating the first three nodes was used (Yarnold and Soltysik<sup>20</sup> describe the CTA modeling process as well as the interpretation of CTA results). **cta** produces an extract of the total output produced by CTA software: the complete output is stored in the specified directory with the extension “.out”.

Below I present a diagram of the enumerated model, which achieved overall ESS of 61.72 (a relatively strong effect)<sup>21</sup>—which is equivalent to an AUC of 81%. While this value is lower than that achieved by the BR model for

the *training* sample, it is actually more accurate because it ensures that the model is stable in *LOO*. In other words, the comparison between the CTA and BR models should be based on the AUCs for the validity (*LOO*) analysis, which is 73.51% and 81%, for the BR (*LOO*) and CTA, respectively. Therefore, the CTA model is inherently more accurate.

While an explanation of the variables in the diagram are beyond the scope of this paper (see Linden et al.<sup>12</sup> for a comprehensive discussion), it is important to convey the transparency of the CTA approach—as opposed to the BR model which operates as a black-box.



## Discussion

This paper demonstrates how the new Stata package **cta** can be used to generate a propensity score model which captures all of the

covariates and interactions between covariates that predict treatment assignment. In utilizing CTA for this procedure, the modeling process is automated and the resulting model is maximally accurate—and generalizable. Equally important, CTA produces output that is transparent to the

user, allowing for greater confidence in the model and its statistical results. CTA should therefore be considered the preferred approach over other machine learning approaches as well as commonly-used parametric models (e.g., logistic regression). Finally, the findings continue to support the recommendation to use the ODA and CTA frameworks to evaluate the efficacy of health-improvement interventions and policy initiatives.<sup>22-35</sup>

## References

- <sup>1</sup>Linden A, Yarnold PR (2017). Minimizing imbalances on patient characteristics between treatment groups in randomized trials using classification tree analysis. *Journal of Evaluation in Clinical Practice*, 23, 1309-1315. DOI: 10.1111/jep.12792
- <sup>2</sup>Linden A (2007). Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes*, 15, 7-12.
- <sup>3</sup>Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*, 22, 839-847.
- <sup>4</sup>Linden A, Adams J (2006). Evaluating disease management program effectiveness: An introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12, 148-154.
- <sup>5</sup>Linden A, Adams JL (2010). Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice*, 16, 175-179.
- <sup>6</sup>Linden A, Adams JL (2010). Evaluating health management programmes over time: Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice*, 16, 180-185.
- <sup>7</sup>Linden A (2014). Combining propensity score based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice*, 20, 1065-1071.
- <sup>8</sup>Linden A, Uysal SD, Ryan A, Adams JL (2016). Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine*, 35, 534-552.
- <sup>9</sup>McCaffrey D, Ridgeway G, Morral A (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9, 403-425.
- <sup>10</sup>Ridgeway G (1999). The state of boosting. *Computing Science and Statistics*, 31, 172-181.
- <sup>11</sup>Linden A, Yarnold PR (2019). Some machine learning algorithms find relationships between variables when none exist -- CTA doesn't. *Optimal Data Analysis*, 8, 64-67.
- <sup>12</sup>Linden A, Butterworth SW, Prochaska JO (2010). Motivational interviewing-based health coaching as a chronic care intervention. *Journal of Evaluation in Clinical Practice*, 16, 166-174.
- <sup>13</sup>Linden A, Roberts N (2004). Disease management interventions: What's in the black box? *Disease Management*, 7, 275-291.
- <sup>14</sup>Linden A, Butterworth S, Roberts N (2006). Disease management interventions II: What else is in the black box? *Disease Management*, 9, 73-85.
- <sup>15</sup>Biuso TJ, Butterworth S, Linden A (2007). Targeting prediabetes with lifestyle, clinical and behavioral management interventions. *Disease Management*, 7, 6-15.
- <sup>16</sup>Linden A, Adler-Milstein J (2008). Medicare disease management in policy context. *Health Care Finance Review*, 29, 1-11.

- <sup>17</sup>Linden A, Roberts N (2005). A User's guide to the disease management literature: Recommendations for reporting and assessing program outcomes. *American Journal of Managed Care*, 11, 113-120.
- <sup>18</sup>Linden A (2015). 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, revised 24 Jan 2019.
- <sup>19</sup>Linden A. (2020). CTA: Stata module for conducting Classification Tree Analysis. *Statistical Software Components S458729*, Boston College Department of Economics.
- <sup>20</sup>Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286
- <sup>21</sup>Yarnold PR, Soltysik RC. *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books, 2005.
- <sup>22</sup>Linden A, Yarnold PR (2018). Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice*, 24, 353-361.
- <sup>23</sup>Linden A, Adams J, Roberts N (2004). Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management*, 7, 180-190.
- <sup>24</sup>Linden A, Yarnold PR (2018). Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice*, 24, 380-387.
- <sup>25</sup>Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, 23, 703-712.
- <sup>26</sup>Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, 22, 875-885.
- <sup>27</sup>Yarnold PR, Linden A (2017). Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis*, 6, 43-46.
- <sup>28</sup>Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, 22, 848-854.
- <sup>29</sup>Linden A, Yarnold PR (2016). Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*, 22, 855-859.
- <sup>30</sup>Linden A, Yarnold PR, Nallomothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*, 22, 860-867.
- <sup>31</sup>Yarnold PR, Linden A (2016). Using machine learning to model dose-response relationships via ODA: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis*, 5, 41-52.
- <sup>32</sup>Linden A, Yarnold PR (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 22, 868-874.
- <sup>33</sup>Linden A, Yarnold PR (2017). Modeling time-to-event (survival) data using classification tree analysis. *Journal of Evaluation in Clinical Practice*, 23, 1299-1308.

<sup>34</sup>Linden A, Yarnold PR (2018). Using machine learning to evaluate treatment effects in multiple-group interrupted time series analysis. *Journal of Evaluation in Clinical Practice*, 24, 740-744.

<sup>35</sup>Linden A, Yarnold PR (2018). The Australian gun buy-back program and the rate of suicide by firearm. *Optimal Data Analysis*, 7, 28-35.

### **Author Notes**

No conflict of interest was reported.