Optimal Data Analysis
Vol. 9 (March 12, 2020), 63-67

2155-0182/10/$3.00

Implementing CTA from Within Stata: Characterizing Participation in Observational Studies (*Invited*)

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

In contrast to randomized studies where individuals have no control over their treatment assignment, participants in observational studies self-select into the treatment arm and are therefore likely to differ in their characteristics from those who elect not to participate. These differences may explain part, or all, of the difference in the observed outcome. In this paper, I describe how the new Stata package for implementing CTA can identify patterns in the data that distinguish study participants from non-participants, revealing potentially complex relationships among individual characteristics that may bias the outcome analysis.

A prior paper[1] introduced the new Stata package called **cta**[2] for implementing CTA from within the Stata environment. This package is a wrapper for the CTA software[3], thus the CTA64.exe file must be loaded on the computer for the **cta** package to work (CTA software is available at https://odajournal.com/resources/). To download the **cta** package, at the Stata command line type: "ssc install cta" (without the quotation marks).

This paper demonstrates how the **cta** package can be used in observational studies to identify patterns in the data that distinguish study participants from non-participants, revealing potentially complex relationships among individual characteristics that may bias the outcome analysis. Arming investigators with this information can serve two purposes. First, from an administrative perspective, the results of a CTA analysis as applied to study participation could help identify new candidates for enrollment who may most benefit from the intervention.[4] Second, the resulting CTA model identifies the sources of confounding that should be adjusted for when evaluating treatment effects.[5-9]

In **cta** we assess selection bias by simply specifying the treatment assignment indicator as the *class* variable and all the observed pre-intervention covariates as *attributes*.

Methods

Data

This paper uses data from a prior evaluation of a health plan–based program intended to reduce 30-day readmission rates for patients hospitalized with one or more chronic illnesses. The intervention was modeled after that described in Linden and Butterworth,[10] which focused on behavioral change to help patients actively engage in their own health care, which in turn was expected to reduce the likelihood of readmission.[11-15] This subset of the retrospectively collected data consists of observations for 1398 participants and 7957 nonparticipants.

Ten pre-intervention characteristics available for every observation included demographic variables ([Age] and [Gender]), health services use in the 12 months prior to the index hospitalization (office visits [Office], emergency department [ED] visits, hospitalizations [Admits]), length of stay for the index [Index] hospitalization, indicator variables for whether the patient had congestive heart failure [CHF] and/or chronic obstructive pulmonary disease [COPD], the patient's Charlson comorbidity index score [CCI],[16] and a diagnosis-based risk adjustment score [Riskscore]. The outcome was the number of days post-discharge from the index hospitalization: patients were classified as censored if they were lost to follow-up prior to 30 days, or if they did not experience a readmission within 30 days.[17]

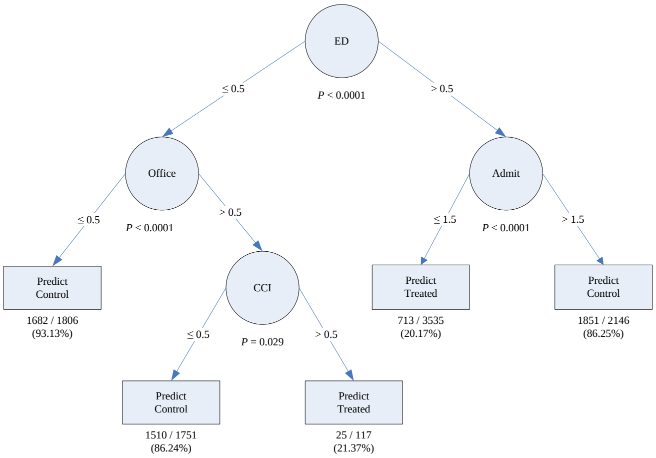
Analytic process

For evaluating self-selection in an observational study, **cta** is implemented with the following syntax (see the help file for **cta** for a complete description of the syntax options):

```
cta treat age gender riskscore admits er
office index cci chf copd , pathcta("C:\
CTA\") store("C:\CTA") cat( gender chf
copd) iter(10000) loo(stable) prune(0.05)
enumerate
```

The above syntax is explained as follows: The outcome variable “treat” is the *class* variable; the 10 variables listed until the comma are covariates specified as the *attributes*; the directory path where the CTA64.exe file is located on my computer is “C:\CTA\”; the directory path where the output and other files generated during the analysis should be stored is "C:\CTA\output"; the *cat*() option indicates which attributes are categorical; the number of iterations (repetitions) for computing a permutation *P*-value is 10,000; leave-one-out analysis is used and attributes are only retained if they are stable; the tree is pruned with a *P*-value of 0.05 used as the cutpoint for inclusion; and an enumerated model (which enumerates the first three nodes) is conducted. (Yarnold and Soltysik[3] provide a complete description of the CTA modeling process and interpretation of results).

The **cta** package produces an extract of the total output produced by CTA software (the complete output is stored in the specified directory with the extension “.out”). Here we include a diagram of the pruned model, which achieved overall ESS of 16.17 (a weak effect)—slightly less than the enumerated model (ESS=18.35), but more parsimonious (3 *vs*. 5 endpoints).



Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

In reviewing this diagram, it is evident that those patients predicted to participate in the intervention follow a different pathway than those predicted to serve as controls. That is, a patient is predicted to participate in the intervention if (1) they had less than one ED visit, more than one office visit and a CCI level of at least 1 in the past year, or if (2) they had at least one ED visit and less than two hospital admissions in the prior year. While the accuracy of these predictions is low (21.37% and 20.17% for pathway 1 and 2, respectively), these pathways were statistically significant ($P < 0.0001$).

Discussion

This paper demonstrates how the new Stata package **cta** can be used to determine whether selection bias is a concern in the study, and if so, which variables (and interactions between variables) are the sources of selection bias. CTA provides accurate, parsimonious decision rules that are easy to visually display and interpret, while reporting *P* values derived via permutation tests at every node, in addition to corresponding partial ESS statistics. CTA is also insensitive to skewed data or outliers, and has the ability to handle any variable metric including categorical, Likert-type integer, and real number measurement scales. Moreover, CTA also has the distinct ability to ascertain where optimal (maximum-accuracy) cutpoints are on each variable, which in turn, facilitates the use of measures of predictive accuracy. Moreover, CTA can perform cross-validation using LOO which allows for assessing the cross-generalizability of the model to potentially new study participants or non-participants.[19]

Finally, the findings continue to support our recommendation to employ the ODA and CTA frameworks to evaluate the efficacy of health-improvement interventions and policy initiatives.[20-37]

References

[1] Linden A (2020). Implementing CTA from Within Stata: Implementing CTA from Within Stata: Assessing the Quality of the Randomization Process in Randomized Controlled Trials (*Invited*). *Optimal Data Analysis*, 9, 57-62.

[2] Linden A. (2020). CTA: Stata module for conducting Classification Tree Analysis. *Statistical Software Components S458729, Boston College Department of Economics.*

[3] Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[4] Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice, 22,* 839-847.

[5] Linden A, Adams J (2006). Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12, 148-154.

[6] 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.

[7] 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.

[8] 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.

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

[9]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.

[10]Linden A, Butterworth SW (2014). A comprehensive hospital-based Intervention to reduce readmissions for chronically ill patients: A randomized controlled trial. *American Journal of Managed Care,* 20, 783-792.

[11]Linden A, Roberts N (2004). Disease management interventions: What's in the black box? *Disease Management,* 7, 275-291.

[12]Linden A, Butterworth S, Roberts N (2006). Disease management interventions II: what else is in the black box? *Disease Management*, 9, 73-85.

[13]Biuso TJ, Butterworth S, Linden A (2007). Targeting prediabetes with lifestyle, clinical and behavioral management interventions. *Disease Management*, 7, 6-15.

[14]Linden A, Adler-Milstein J (2008). Medicare disease management in policy context. *Health Care Finance Review*, 29, 1-11.

[15]Linden A, Roberts N (2005). A Users guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care,* 11, 81-90.

[16]Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Disease*, 40, 373-383.

[17]Linden A, Adams J, Roberts N (2004). Evaluating disease management program effectiveness: an introduction to survival analysis. *Disease Management*, 7, 180-190.

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

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

[20]Linden A, Adams J, Roberts N (October, 2003). *Evaluation methods in disease management: determining program effectiveness*. Position Paper for the Disease Management Association of America (DMAA).

[21]Linden A, Yarnold PR, Nallomothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*, *22,* 860-867.

[22]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, 22, 171-174.

[23]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.

[24]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, 22, 848-854.

[25]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, *23,* 703-712.

[26]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.

Copyright 2020 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

[27]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.

[28]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.

[29]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.

[30]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.

[31]Linden A, Yarnold PR (2017). Modeling time-to-event (survival) data using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1299-1308.

[32]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.

[33]Yarnold PR, Linden A (2017). Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis*, *6*, 43-46.

[34]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.

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

[36]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials. *Optimal Data Analysis*, *7,* 46-49.

[37]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials: Application to survival outcomes. *Optimal Data Analysis*, *7*, 50-53.

Author Notes

No conflict of interest was reported.