Optimal Data Analysis
Vol. 9 (March 10, 2020), 57-62

2155-0182/10/$3.00

Implementing CTA from Within Stata: Assessing the Quality of the Randomization Process in Randomized Controlled Trials (*Invited*)

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

In randomized controlled trials (RCT) of sufficient size, we expect the treatment and control groups to be balanced on both observed and unobserved characteristics, and any imbalances are considered to be due to chance. CTA can be used to determine whether treatment assignment can be predicted by observed pre-intervention covariates–separately or interacted with other covariates. In this paper, I describe how to assess the quality of the randomization process in RCTs using the new Stata package for implementing CTA.

Prior papers[1-7] introduced the new Stata package called **oda**[8] for implementing ODA from within the Stata environment. In this paper, I introduce the new Stata package called **cta**[9] for implementing CTA from within the Stata environment. This package is a wrapper for the CTA software[10], and therefore 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 to evaluate the quality of the randomization process in RCTs. In simple terms, if the randomization process is implemented correctly, then the treatment and control groups should have identical distributions on all pre-intervention characteristics. It is possible that the groups will differ marginally on some characteristics due to chance alone, and typically those differences are adjusted for in the outcomes analysis. However, if a substantial number of characteristics are imbalanced, then the investigator should be concerned about non-adherence to the treatment assignment protocol, which in turn leads to selection bias.

In **cta** we assess if treatment groups are balanced on observed characteristics–separately as well as part of an interaction with one or

2155-0182/10/$3.00

more other covariates by specifying the treatment assignment indicator as the *class* variable and the pre-intervention covariates as *attributes*. If a model is indeed identified, the investigator must then consider whether this indicates a failure in the randomization process or chance.

Methods

Example 1

Data

This example uses data from a parallel-group, stratified, clinical trial that examined whether a comprehensive, hospital-based, transitional care intervention reduces readmissions for participants with congestive heart failure (CHF) and chronic obstructive pulmonary disease (COPD).[11] The intervention involved nurses implementing motivational interviewing-based health coaching to improve patients' health behaviors, which in turn was expected to empower patients to better manage their own health care and reduce unplanned readmissions.[12-16] This example limits the data to only the CHF cohort, which includes 129 treated patients and 128 controls.

Analytic process

For evaluating the quality of the randomization process in an RCT, **cta** is implemented with the following syntax (see the help file for **cta** for a complete description of the syntax options):

```
cta treat gender age insurance living pam
admits er chfadmits chfdays indexlos
copd cevd pain diab ami renal obes,
pathcta("C:\CTA\")
store("C:\CTA\output") cat(gender
insurance living copd cevd
pain diab ami renal obes) iter(10000)
loo(stable) prune(0.05) enumerate
```

The above syntax is explained as follows: The outcome variable "treat" is the *class* variable; the 16 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 of the 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 retains only the top three maximally accurate attributes) should be conducted. (Yarnold and Soltysik[10] 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 the CTA software (the complete output is stored in the specified directory with the extension ".out"). However, with these data CTA *could not find a model*. Thus, we can conclude that the randomization process was effective in ensuring balance between study groups.

Example 2

Data

The National Supported Work (NWS) Demonstration was a U.S. federally- and privately funded program that aimed to provide work experience for individuals who faced economic and social problems prior to enrollment in the program. Candidates for the experiment were selected on the basis of eligibility criteria, and then were randomly assigned to, or excluded from, the training program. We use the same subset of NSW data used by LaLonde[17]—samples of 297 male treated and 425 control observations. Data were retrieved from: http://users.nber.org/~rdehejia/nswdata2.html.

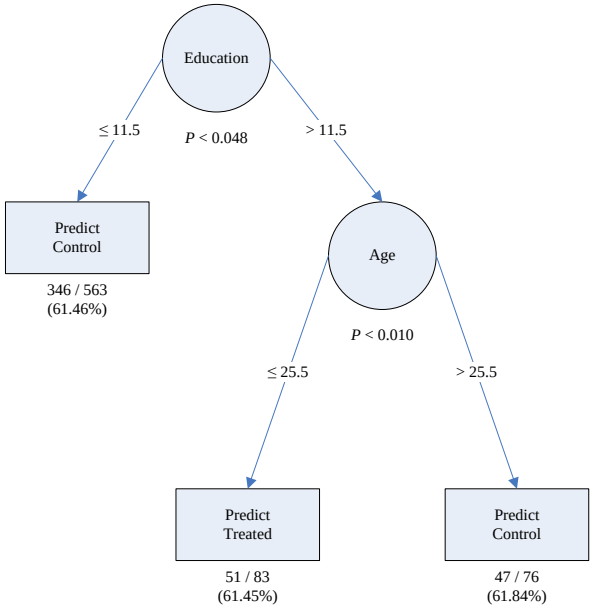
2155-0182/10/$3.00

The following **cta** syntax was used:

```
cta treat age education black married
nodegree re74 re75 hispanic u74 u75,
pathcta("C:\CTA\")
store("C:\CTA\output") cat( black
married nodegree hispanic u74 u75)
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 of the 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 retains only the top three maximally accurate attributes) is conducted.

CTA produced the same results for the unpruned, pruned and enumerated models (this does not normally occur). As seen in the Figure, CTA could discriminate between treated and control units based on education and age, and an interaction of the two. The overall ESS was 9.64% (a weak effect)[18] but the relationships were statistically significant. In summary, an investigator reviewing these results would conclude that the randomization process was not followed correctly in all cases, as those selected for the intervention were predicted to have more than a high-school education, or otherwise were under 25.5 years of age.



Discussion

This paper demonstrates how the new Stata package **cta** can be used to assess the quality of the randomization process in RCTs. CTA allows the investigator to identify possible sources of heterogeneity between groups where there should be none. 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]

2155-0182/10/$3.00

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.[20-38]

References

[1]Linden A (2020). Implementing ODA from Within Stata: An Application to Data From a Randomized Controlled Trial (*Invited*). *Optimal Data Analysis*, 9, 9-13.

[2]Linden A (2020). Implementing ODA from Within Stata: Implementing ODA from Within Stata: An Application to Estimating Treatment Effects using Observational Data (*Invited*). *Optimal Data Analysis*, 9, 14-20.

[3]Linden A (2020). Implementing ODA from Within Stata: An Application to Dose-Response Relationships (*Invited*). *Optimal Data Analysis*, 9, 26-32.

[4]Linden A (2020). Implementing ODA from Within Stata: Assessing Covariate Balance in Observational Studies (*Invited*). *Optimal Data Analysis*, 9, 33-38.

[5]Linden A (2020). Implementing ODA from Within Stata: Evaluating Treatment Effects for Survival (Time-to-Event) Outcomes (*Invited*). *Optimal Data Analysis*, 9, 39-44.

[6]Linden A (2020). Implementing ODA from Within Stata: Evaluating Treatment Effects in Multiple-Group Interrupted Time Series Analysis (Invited). *Optimal Data Analysis*, 9, 45-50.

[7]Linden A (2020). Implementing ODA from Within Stata: Identifying Structural Breaks in Single-Group Interrupted Time Series Designs (Invited). *Optimal Data Analysis*, 9, 51-56.

[8]Linden A (2020). ODA: Stata module for conducting Optimal Discriminant Analysis. *Statistical Software Components S458728, Boston College Department of Economics*.

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

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

[11]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.

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

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

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

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

[16]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.

[17]LaLonde RJ (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604-620.

2155-0182/10/$3.00

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

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

[27]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.

[28]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.

[29]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.

[30]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.

[31]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.

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

[33]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.

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

2155-0182/10/$3.00

[35]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.

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

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

[38]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.