Optimal Data Analysis
Vol. 10 (April 25, 2021), 50-57

2155-0182/10/$3.00

Using ODA to Estimate Propensity-Weight-Adjusted Treatment Effects for Multi-Valued Treatments

Paul R. Yarnold, Ph.D., Fred B. Bryant, Ph.D., Ariel Linden, Dr.P.H.
Optimal Data Analysis, LLC, Loyola University Chicago, Linden Consulting Group, LLC

We demonstrate the use of optimal data analysis to obtain a hierarchically optimal classification tree-based propensity score model for an application with three (treatment) groups, and to assess outcome differences between treatment groups after weighting observations by propensity scores to reduce threats to causal inference.

Studies in which the participants are randomized to treatment conditions are considered the gold standard for assessing causal inference, because randomization putatively ensures that the study groups do not systematically differ with respect to their characteristics, and consequently that estimated treatment effects may be assumed to be unbiased.[1] When randomization is infeasible, investigators rely on statistical techniques to model treatment assignment[2-8] in order to control for threats to validity which may compromise causal interpretation of the results.[9-13]

A common approach used to estimate treatment effects in studies involving multiple treatment arms is to first estimate propensity scores using multinomial logistic regression, then compute the inverse probability of treatment weights, and finally use treatment weights in a weighted regression to evaluate outcomes.[8]

A recent paper demonstrated performing this process using two machine learning algorithms—boosted regression to compute the probabilities of treatment, and optimal data analysis[14-18] (ODA) to evaluate the weighted outcomes, thereby avoiding all assumptions required by conventional parametric methods.[19]

A second paper demonstrated use of the random forest algorithm to compute the probabilities of treatment, and ODA to evaluate the weighted outcomes.[20] In both studies (using boosted regression and random forests) ODA identified no difference in propensity score-weighted outcomes between class categories.

The present paper demonstrates how to perform this process using the ODA algorithm to compute the probabilities of treatment and to evaluate the weighted outcomes, thus yielding *maximum possible predictive accuracy* while requiring no distributional assumptions.[17,18]

2155-0182/10/$3.00

Methods

Data

Data for our empirical example are taken from a disease management program for patients with congestive heart failure (CHF), which was implemented in a large health plan located in the western USA.[6,8] Individuals with CHF were contacted and invited to enroll in the program. Those agreeing to participate received one of the following interventions: (1) periodic telephone calls (CALL), from a nurse to discuss self-management behaviors (n=654), or (2) remote tele-monitoring (RTM), entailing daily electronic transmission of the participant's disease-related symptoms to a database followed by a call from the nurse if symptoms appeared to indicate the onset of an acute exacerbation (n=705). Assignment to either intervention arm was conducted by the program nurse and based largely on subjective assessment of the patient's psychosocial needs, past levels of health care utilization, and the patient's preferred level of contact.[21-23] The primary goal of the intervention was to reduce avoidable hospitalizations.[24] Patients with CHF, but not participating in the program, received their usual medical care and were assigned as CONTROL (n=6612) patients in this study (see [6], and [8] for a comprehensive description).

Ordinal pre-intervention variables (or "attributes" in the ODA paradigm) available to discriminate the CALL, RTM and CONTROL groups (a three-category "class" variable) included age; number of prescriptions; number of admissions; number of ER visits; number of office visits; and days in hospital. Binary "yes *vs*. no" pre-intervention variables included diabetes without complications; diabetes with complications; mild liver disease; moderate to severe liver disease; cancer; metastatic cancer; chronic obstructive pulmonary disease; rheumatoid arthritis; congestive heart failure; cerebral vascular disease; peripheral vascular disease; renal disease; HIV/AIDS; and dementia.

The OUTCOME variable was number of hospitalizations in the year after the individual entered the study.

Analytic Process

Data analysis reported herein was conducted via MegaODA software, which is available at no cost for individual *non-commercial* use.[17,25-32] Additional interfaces for MegaODA software are available for the R suite[33] and for Stata[34] (to download the **oda** package, at the Stata command line type: *ssc install oda*).

In the ODA paradigm, the accuracy that is obtained by a model is corrected to eliminate the effect of chance. This is accomplished by computing the effect strength for sensitivity (ESS) index on which 0 represents the classification accuracy expected by chance, 100 indicates perfect (errorless) classification, and values less than 0 indicate accuracy which is less than what is expected by chance.[18,30]

The first step of the analysis employs MegaODA to conduct hierarchically-optimal (HO) classification tree analysis (CTA) to identify the model which most accurately classified the three class categories.[35,36] First, all attributes which have three or more levels are used to identify the first node of the HO-CTA model.[37-40] Second, for all attributes having fewer levels (e.g., a binary attribute) than the class variable (e.g., a multicategorical attribute), the roles of the class variable and attribute are switched.[41] This will cause one endpoint in the developing tree model to combine two of the three class categories. Accordingly, deeper in the tree, additional attributes may be identified that disentangle comingled class categories—if such inclusion yields a tree model with greater ESS. Third, when a tree model cannot be further grown then it is pruned to identify the structure that yields maximum ESS.[42,43]

In the second step of the analysis the CTA model having maximum ESS is used to generate propensity score weights. This process

2155-0182/10/$3.00

is performed by first computing propensity scores for every observation in the sample (i.e., the probability of being in the treatment category endpoint), and then weights are computed based on those propensity scores.[44,45]

In the third (final) step of the analysis, the propensity score weight is specified as a weight in the ODA software and the outcome analysis is conducted.

Results

ODA was used to evaluate the *unadjusted effect* of the intervention, comparing the outcome of all three participant groups *without* propensity-score weighting. The optimal model was:

if outcome=0 then class=CONTROL;
if 0<outcome≤4 then class=CALL;
if outcome>4 then class=RTM.

This model had $p<0.0001$, and ESS= 5.73%, indicating a weak effect.[17] The model was stable in leave-one-out (LOO) jackknife cross-generalizability analysis, $p<0.0001$.[46,47] Classification performance of this model in training and LOO analyses is shown in Table 1.

Table 1: Unadjusted Classification Performance

Actual	*Predicted* Class			
Class	Control	Call	RTM	Sensitivity
Control	5356	1210	46	81.0%
Call	457	189	8	28.9%
RTM	528	166	11	1.6%
PV	84.5%	12.1%	16.9%	

Note: PV=predictive value

HO-CTA was then used to identify the best model predicting observation actual class assignments. A single-attribute model involving age emerged as the strongest model possible for the present data (ESS=17.18%, a relatively weak effect[17]), $p<0.0001$. This model was stable in LOO analysis, $p<0.0001$. All other CTA models using ordinal attributes had ESS<11%, which eroded if additional attributes were included in the model. No "inverted" CTA models using binary attributes achieved ESS exceeding 10%. The single-attribute model yielding maximum ESS for these data was:

if age<60 then class=CONTROL;
if 59<age≤64 then class=CALL;
if age>64 then class=RTM.

This model had $p<0.0001$, and ESS= 17.18%, indicating a weak effect. The model was stable in LOO analysis, $p<0.0001$. Classification performance of this model in training and LOO analyses is shown in Table 2. These results were used to create propensity score weights.

Table 2: Classification Performance of HO-CTA Model Predicting Actual Class

Actual	*Predicted* Class			
Class	Control	Call	RTM	Sensitivity
Control	2807	1139	2666	42.4%
Call	208	141	305	21.6%
RTM	117	92	496	70.4%
PV	89.6%	10.3%	14.3%	

Note: PV=predictive value

Finally, ODA was used to evaluate the propensity-score *adjusted effect* of the intervention, comparing the outcome of all three participant groups *with* propensity-score weighting. The optimal model was:

if outcome≤1.5 then class=CONTROL;
if 1.5<outcome≤6.5 then class=CALL;
if outcome>6.5 then class=RTM.

This model had $p<0.0027$, and weighted ESS=2.19%, indicating a weak effect. The model was stable in LOO analysis, $p<0.0022$.

2155-0182/10/$3.00

Classification performance of this model in training and LOO analyses is shown in Table 3.

Table 3: Propensity-Score Adjusted Classification Performance

Actual	*Predicted* Class			
Class	Control	Call	RTM	wSensitivity
Control	6168	429	15	92.3%
Call	576	77	1	12.0%
RTM	632	72	1	0.1%
wPV	89.6%	9.5%	2.6%	

Note: wSensitivity=weighted sensitivity, wPV=weighted predictive value

Discussion

A common approach for estimating treatment effects in studies with multiple treatment arms involves estimating propensity scores vis-à-vis multinomial logistic regression, then computing the inverse probability of treatment weights, and finally using the treatment weights in weighted regression to evaluate outcomes. For data used in this study neither boosted regression nor random forest approaches were able to distinguish outcomes between treatment modalities.[19,20] In contrast, we found a statistically significant difference in outcomes between treatment groups using propensity score weights developed using a multicategorical ODA model.

An important ubiquitous limitation of research is the use of low-fidelity binary measures such as gender (male, female).[48] A problem with binary measures treated as a class variable, or as an attribute, is their use may not be *statistically motivated*. For example, comparison between, or measure of, incidence of "black" *vs*. "white" cancer patients creates paradoxical confounding if different subgroups of either or both categories exist. That is, combining low- and high-risk groups can create composite groups which don't represent any of the observations in the combined sample.[49]

An important irreconcilable limitation of logistic regression analysis[50] and least-squares regression analysis[51] is inability to use multicategorical variables as attribute or class variable. In the present study several ordinal measures were discretized to allow their analysis using multinomial regression. For example, a three-level measure of diabetes was deconstructed into two binary attributes: diabetes without complications (0=no; 1=yes), and diabetes with complications (0=no; 1=yes). In the present article this coding was used to facilitate direct comparison between boosted regression and ODA.[19] However, for ODA these two variables could be integrated into a single three-category scale (0=no diabetes; 1=diabetes without complications; 2= diabetes with complications) which is more parsimonious and potentially more informative: (a) complicated diabetes might be separated from the combination of no diabetes and mild diabetes; (b) no diabetes might be separated from the combination of diabetes with and without complications; or (c) no diabetes might be separated from diabetes without complications, which might be separated from diabetes with complications. Note that this limitation of regression analyses applies regardless of whether the variable serves as a dependent (attribute) or independent (class) variable.

Parallel arguments may be made for the use of ordinal (e.g., Likert-type) variables as class variables, or as attributes.[52] Taken to the limit this line of consideration suggests that the greater the fidelity of the variables which are used in research, the more powerful the effects which may be discovered. In this regard, novometric statistical analysis empowers maximum-accuracy analysis of ratio-level measures.[53-58]

CTA analysis of multicategorical class variables with many levels are computationally demanding, particularly when a phenomenon under investigation has a functional relationship

2155-0182/10/$3.00

which involves parabolic[59] or more complex non-linear shapes. While novometric analysis of such phenomena lies outside the capability of present computing machinery, the advent of quantum computing architecture offers the promise of rendering such analyses viable.[60]

Finally, although the propensity-score weighted estimated effect was statistically significant and stable in LOO cross-generalizability analysis, the first Axiom of novometric analysis maintains that the 95% exact discrete confidence interval for the model lies outside the corresponding confidence interval for chance, based on sample data.[61,62] While this methodology is available[32] in the R suite for binary class variables, it is not yet available (but is currently being constructed) for multicategorical class variables.

References

[1]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.

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

[3]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.

[4]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.

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

[6]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.

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

[8]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.

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

[10]Linden A (2007). Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes*, *15*, 7-12.

[11]Linden A (2013). Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology*, *13*, 1-7.

[12]Yarnold PR, Bryant FB, Soltysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression *away* from the mean. *Optimal Data Analysis*, *2*, 19-25.

[13]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.

2155-0182/10/$3.00

[14]Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, *22*, 739-752.

[15]Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: One-tailed hypotheses. *Educational and Psychological Measurement*, *54*, 646-653.

[16]Carmony L, Yarnold PR, Naeymi-Rad F (1998). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research*, *74*, 223-238.

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

[18]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[19]Linden A, Yarnold PR (2021). Implementing ODA from within Stata: Combining boosted regression and ODA to estimate treatment effects for multi-valued treatments. *Optimal Data Analysis*, *10*, 18-23.

[20]Linden A (2021). Executing ODA from within Stata: Combining random forests and ODA to estimate treatment effects for multi-valued treatments (*Invited*). *Optimal Data Analysis*, *10*, 12-17.

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

[22]Biuso TJ, Butterworth S, Linden A (2007). A conceptual framework for targeting prediabetes with lifestyle, clinical and behavioral management interventions. *Disease Management*, *10*, 6-15.

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

[24]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.

[25]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, *2*, 194-197.

[26]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the Wheat. *Optimal Data Analysis*, *2*, 202-205.

[27]Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, *2*, 220-221.

[28]Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software (*Invited*). *Optimal Data Analysis, 2*, 2-6.

[29]Yarnold PR (2018). Visualizing application and summarizing accuracy of ODA models. *Optimal Data Analysis*, *7*, 85-89.

[30]Yarnold PR (2013). Creating a data set with SAS™ and maximizing ESS of a multiple regression analysis model for a Likert-type dependent variable using UniODA™ and MegaODA™ software. *Optimal Data Analysis*, *2*, 191-193.

[31]Rhodes NJ (2020). Statistical power analysis in ODA, CTA and Novometrics (Invited). *Optimal Data Analysis*, *9*, 21-25.

2155-0182/10/$3.00

[32]Rhodes JN, Yarnold PR (2020). Generating novometric confidence intervals in R: Bootstrap analyses to compare model and chance ESS. *Optimal Data Analysis*, *9*, 172-177.

[33]Rhodes NJ, Yarnold PR. 2020. ODA: a package and R-interface for the MegaODA software suite. R package version 1.0.1.3. Available: https://github.com/njrhodes/ODA

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

[35]Yarnold PR, Bryant FB (2015). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis*, *4*, 36-53.

[36]Yarnold, P.R. (2013). Initial use of hierarchically optimal classification tree analysis in medical research. *Optimal Data Analysis*, 2, 7-18.

[37]Yarnold PR, Linden A (2021). Implementing ODA from within Stata: Exploratory hypothesis, three-category class variable, continuous attribute. *Optimal Data Analysis*, *10*, 3-9.

[38]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, multicategorical class variable and attribute. *Optimal Data Analysis*, *9*, 162-166.

[39]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, multicategorical class variable and attribute. *Optimal Data Analysis*, *9*, 157-161.

[40]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Nondirectional hypothesis, multicategorical class variable, multicategorical attribute. *Optimal Data Analysis*, *9*, 152-156.

[41]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.

[42]Yarnold PR, Soltysik RC (2010). Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis*, *1*, 23-29.

[43]Yarnold PR (2019). Maximizing classification accuracy of CART® recursive partitioning tree models using optimal pruning. *Optimal Data Analysis*, *8*, 26-29.

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

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

[46]Yarnold PR (2016). Using UniODA to determine the ESS of a CTA model in LOO analysis. *Optimal Data Analysis*, *5*, 3-10.

[47]Yarnold PR (2016). Determining jackknife ESS for a CTA model with chaotic instability. *Optimal Data Analysis*, *5*, 11-14.

[48]Yarnold PR (2018). Minimize usage of binary measurement scales in rigorous classical research. *Optimal Data Analysis*, *7*, 3-9.

[49]Yarnold PR (2020). What is novometric data analysis? *Optimal Data Analysis*, *9*, 195-206.

[50]Wright RE (1995). Logistic regression (pp. 217-244). In: *Reading and understanding multivariate statistics*. APA Books, Washington D.C.

2155-0182/10/$3.00

[51]Yarnold PR (2014). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis*, *3*, 107-109.

[52]Licht MH (1995). Multiple regression and correlation (pp. 19-64). In: *Reading and understanding multivariate statistics*. APA Books, Washington D.C.

[53]Yarnold PR (2016). Novometric *vs*. ODA reliability analysis *vs*. polychoric correlation with relaxed distributional assumptions: Inter-rater reliability of independent ratings of plant health. *Optimal Data Analysis*, *5*, 179-183.

[54]Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. APA Books, Washington D.C, pp. 1-18.

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

[56]Yarnold PR (2013). Analyzing categorical attributes having many response categories. *Optimal Data Analysis*, *2*, 172-176.

[57]Yarnold PR, Soltysik RC (2019). Confirming the efficacy of weighting in optimal Markov analysis: Modeling serial symptom ratings. *Optimal Data Analysis*, *8*, 53-55.

[58]Yarnold PR (2019). Optimal Markov model relating *two* time-lagged outcomes. *Optimal Data Analysis*, *8*, 61-63.

[59]Yarnold PR (2015). An example of nonlinear UniODA. *Optimal Data Analysis*, *4*, 124-128.

[60]https://phys.org/news/2021-04-algorithms-boundaries-quantum-future.html

[61]Yarnold PR (2018). Comparing exact discrete 95% CIs for model *vs*. chance ESS to evaluate statistical significance. *Optimal Data Analysis*, *7*, 82-84.

[62]Yarnold PR (2020). What is novometric data analysis? *Optimal Data Analysis*, *9*, 195-206.

Author Notes

No conflicts of interest were reported.