Optimal Data Analysis
Vol. 8 (June 13, 2019), 122-129

2155-0182/10/$3.00

Randomized Blocks Designs: Omnibus *vs.* Pairwise Comparison, Fixed *vs.* Relative Optimal Discriminant Threshold, and Raw *vs.* Ipsative *z*-Score Measures

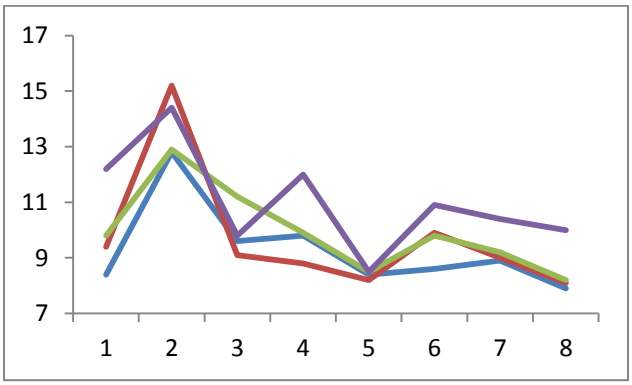
Paul R. Yarnold, Ph.D. and Ariel Linden, Dr.P.H.
Optimal Data Analysis, LLC Linden Consulting Group, LLC

This study extends recent research assessing the use of relative thresholds in matched-pairs designs, for a randomized blocks design in which four treatments are randomly assigned to blood samples drawn from each of eight people (each person treated as a block[1]). Both raw and ipsatively standardized plasma clotting times are compared between treatments.

For applications that involve ordered attributes ("dependent variables"), optimal discriminant (or data) analysis (ODA) compares the values of one or more attributes between two or more groups of observations with respect to a *fixed* discriminant threshold which maximizes the models predictive accuracy, normed *vs*. chance, for the sample.[2-30] However, recent research with matched-pairs designs found that using a *relative* discriminant threshold to evaluate an alternative hypothesis *separately for each of the matched pairs* of subjects can find inter-group differences otherwise too small to be identified vis-à-vis fixed thresholds.[31,32]

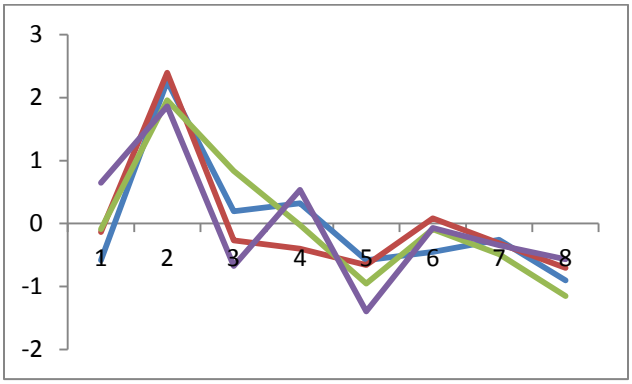
Fleiss gives an example of a randomized blocks experiment in which four treatments are randomly assigned to blood samples which were drawn from each of eight different people (each person is treated as a block), and plasma clotting time is compared between treatments.[1] In Figure 1 (raw minutes) and Figure 2 (ipsative $z_{minutes}$) treatments are color-coded (in the Figures) as **blue**, **red**, **green** and **purple** for treatments 1, 2, 3 and 4, respectively.

Figure 1: Clotting Time in Minutes (Vertical Axis) by Four Different Treatments (Colors), Separately for Each of Eight Patients[1]



2155-0182/10/$3.00

Figure 2: Ipsatively Standardized Clotting Time (Vertical Axis) by Four Different Treatments (Colors), Separately for Each of Eight Patients



Classic Analytic Approaches

Raw Data

For comparison *vs*. ODA methodology, we first replicated the ANOVA described by Fleiss[1] in which a statistically significant *F* value was estimated for the overall model: F=6.62, p<0.025. However, when applying the Sheffe adjustment for multiple hypothesis testing, the only statistically significant treatment effect was between the first and fourth treatments (p<0.037).

One can argue that using a parametric model to conduct a treatment effects analysis with four treatments and a sample size of eight is inappropriate given that the assumptions required of such a model are easily violated (e.g., sample size, normality of the data, etc.). As a more appropriate alternative, we perform a Skillings-Mack test[33]—a generalization of the Friedman non-parametric test which is a suitable alternative to repeated measures ANOVA for small samples and possibly non-normally distributed data. The overall model was statistically significant (p<0.0001), and four of the pairwise comparisons between treatments achieved significant *p* values (at the conventional p<0.05 level) derived using 1,000 simulations (1 *vs*. 3: p<0.0001; 1 *vs*. 4: p<0.0001; 2 *vs*. 4: p<0.0070; 3 *vs*. 4: p<0.0190).

Ipsative Standard Data

The replication of Fleiss' ANOVA using the standardized data was not statistically significant (p=1.00). Similarly, the Skillings-Mack test using the standardized data was not statistically significant (p<0.79).

ODA Fixed Discriminant Threshold

Raw Data

Raw clotting times of all four groups (Figure 1) were compared by omnibus four-class-category ODA analysis. Training analysis identified a model with a moderate effect size (ESS=37.5) which wasn't statistically significant (p<0.28).

All six possible pairwise comparisons of raw clotting times among the four groups were conducted by two-class-category ODA: none were statistically significant (all p>0.08).

Ipsative Standard Data

Ipsatively standardized clotting times of all four groups (Figure 2) were compared by omnibus four-class-category ODA analysis. Training analysis identified a model with a moderate effect size (ESS=25.0) which wasn't statistically significant (p<0.99).

All six possible pairwise comparisons of ipsatively standardized clotting times among the four groups were conducted using ODA: none were statistically significant (all p>0.66).

ODA Relative Discriminant Threshold

Raw Data

No multiclass omnibus relative threshold model yet exists.[31,32] Pairwise comparisons of corresponding raw clotting times (Figure 1) between pairs of treatments assessed if time for treatment X is *strictly less than* time for treatment Y, for each subject considered separately. Results showed treatment 1 had 100% lower clotting time when considered on a pairwise basis *vs*.

2155-0182/10/$3.00

treatments 3 or 4: for each, binomial $p<0.00391$. Treatment 2 had 82.5% lower clotting time *vs*. treatment 4 (binomial $p<0.032$). Finally, under the strictly-less-than ("always lower") criterion the pairwise comparison of treatments 3 and 4 was not statistically significant (75% difference in clotting times, $p<0.11$), however under the *less-than-or-equal* ("never higher") criterion the difference was statistically significant (82.5% difference in clotting times, $p<0.032$). Other pairwise comparisons were not statistically significant.

Ipsative Standard Data

Pairwise comparisons of corresponding ipsative *z*-scores (Figure 2) within pairs of treatments assessed if the *z*-score for treatment X is strictly less than the *z*-score for treatment Y (there were no equivalent *z*-scores), considered separately for each subject in the sample: none were statistically significant (all $p>0.21$).

Comments

As occurred herein, different statistical methods may generate consistent as well as inconsistent analytic conclusions.[34,35]

For example highly inconsistent findings were obtained herein for *omnibus* comparisons of *raw* clotting times (Figure 1): Skillings-Mack identified a training analysis effect that yielded lowest Type I error ($p<0.0001$); ANOVA found a training analysis effect which met the criterion for per-comparison (not for experimentwise[35]) statistical significance ($p<0.025$); and ODA using a fixed discriminant threshold found no effect ($p<0.28$). No multiclass optimal omnibus relative discriminant threshold methodology is yet available: a novometric procedure to identify and evaluate all possible omnibus models is thus needed.[2,30,36-38]

Inconsistent findings also emerged for *pairwise* comparisons of *raw* clotting times. ANOVA with a Sheffe post-estimation adjustment found a statistically significant ($p<0.037$) treatment effect only in the comparison between treatments 1 and 4; Skillings-Mack identified significant pairwise differences between treatments 1 *vs*. 3 and 1 *vs*. 4 (p's<0.0001), 2 *vs*. 4 ($p<0.007$), and 3 *vs*. 4 ($p<0.019$); ODA using a *fixed* discriminant threshold found no statistically significant pairwise differences (p's>0.08); ODA using a *strictly-less-than relative* discriminant threshold identified significant pairwise differences between treatments 1 *vs*. 3 and 1 *vs*. 4 (p's<0.00391), and 2 *vs*. 4 ($p<0.032$); and ODA using a *less-than-or-equal relative* discriminant threshold identified a pairwise difference between treatments 3 *vs*. 4 ($p<032$).

In stark contrast to inconsistent findings obtained between methods in analysis of raw clotting time data, a *uniformly consistent* finding of *no effect* emerged in analyses of *ipsatively standardized* data vis-à-vis ANOVA ($p=1.00$), Skillings-Mack ($p<0.79$); ODA using a *fixed* discriminant threshold in *omnibus* ($p<0.99$) or *pairwise* (p's>0.66) comparisons; and ODA using *relative* discriminant thresholds in pairwise analyses (p's>0.21).

What explanation underlies the different findings which were obtained by analyzing raw *vs*. ipsatively standardized data? Quite simply, this is a classical form of Simpson's Paradox (i.e., paradoxical confounding) whereby pooling raw clotting times of different people ignored existing baseline interpersonal differences and yielded invalid and inconsistent results—which vanish once interpersonal baseline differences in mean and variance are eliminated using separate (ipsative) standardizations.[39] The present study is yet another example of a ubiquitous problem permeating all quantitative sciences—the use of inappropriate measurement metrics[39-47] and/or suboptimal statistical methods[48-51] may fail to identify true underlying effects and instead find confounded effects—which in reality simply *don't* exist.

Two fundamental weaknesses underlie the experimental methodology used presently. One weakness is the inadequate sample size: the

2155-0182/10/$3.00

first axiom of novometric statistical theory is the sample must yield adequate statistical power to test the *a priori* alternative hypothesis.[36-38] The second design weakness lies in the sample of observations itself, in light of the evidence of paradoxical confounding attributable to baseline differences. Because outcome analysis may be biased by confounding, it is imperative to use an appropriate matching/blocking process to eliminate possible confounding which is attributable to observed covariates that exert influence on the outcome. For example, in studies with small (sufficiently large) samples and few covariates, matching directly on the covariates may suffice. However, as the number of covariates increases and the distributions differ increasingly between the treatment and control conditions, methods to stratify and weight individuals into blocks (of the propensity score) and thereby to adjust for observed confounding should be considered.[52]

These techniques are incorporated in the ODA, CTA and novometric frameworks.[53-55] In any substantive analytic application, optimal ("maximum-accuracy") methods are available to assist researchers to obtain a clear understanding of the factors which must be considered to be potential threats to causal inference: to identify variables characterizing participation in both discretionary treatment[56] and observational[27] research (to identify possible confounding variables), and to find structural breaks in single[57,58] and multiple-group[59] interrupted time series analysis, dose-response studies[47,60] and in research investigating mediating processes.[61] Optimal ODA and CTA methods are available to assist researchers to identify and correct (in real time) covariate interactions which exist in data from matched[53] and randomized trials[62] to remove otherwise undetected threats to causal inference. Globally-optimal ("novometric") analysis is available to identify all statistically unique propensity score models that maximize classification accuracy and vary as a function of complexity, which exist within a sample: this makes model misspecification impossible, and is used in both time-to-event and single-case precision forecasting.[63-66] The present findings further illustrate why we strongly advocate using ODA, CTA and novometric frameworks to draw causal inferences about treatment effects in observational data and in data from randomized controlled trials. Clearly, changes are needed in guidelines concerning how health care interventions and policy changes are evaluated.[67,68]

Finally, several new tantalizing research directions are indicated by our recent studies of the use of relative discriminant thresholds in maximum-accuracy models. First, an enhanced incarnation of novometric CTA is needed, that combines fixed and relative thresholds—always using the fixed threshold first. And, application of these methods should be extended to research designs involving dyads, couples, and "self *vs*. other" configurations, as well as to designs that involve temporal comparisons.

References

[1]Fleiss JL (1986). *The design and analysis of clinical experiments*. New York, NY: Wiley (pp. 125-129).

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

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

[4]Yarnold PR, Martin GJ, Soltysik RC, Nightingale SD (1993). Androgyny predicts empathy for trainees in medicine. *Perceptual and Motor Skills*, *77*, 576-578.

[5]Yarnold PR, Soltysik RC, Martin GJ (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, *13*, 1015-1021.

2155-0182/10/$3.00

[6]Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement, 56*, 656-667.

[7]Collinge W, Yarnold PR, Raskin E (1998). Use of mind/body self-healing practice predicts positive health transition in chronic fatigue syndrome: A controlled study. *Subtle Energies & Energy Medicine, 9*, 171-190.

[8]Feinglass J, Yarnold PR, Martin GJ, McCarthy WJ. (1998). A classification tree analysis of selection for discretionary treatment. *Medical Care, 36*, 740-747.

[9]Yarnold PR, Michelson EA, Thompson DA, Adams SL (1998). Predicting patient satisfaction: A study of two emergency departments. *Journal of Behavioral Medicine, 21*, 545-563.

[10]Kanter AS, Spencer DC, Steinberg MH, Soltysik RC, Yarnold PR, Graham NM (1999). Supplemental vitamin B and progression to AIDS and death in black South African patients infected with HIV. *Journal of Acquired Immune Deficiency Syndrome, 21*, 252-253.

[11]Kucera CM, Greenberger PA, Yarnold PR, Choy AC, Levenson T. (1999). An attempted prospective testing of an asthma severity index and a quality of life survey for 1 year in ambulatory patients with asthma. *Allergy and Asthma Proceedings, 20*, 29-38.

[12]Mueser KT, Yarnold PR, Rosenberg SD, Drake RE, Swett C, Miles KM, Hill D. (2000). Substance use disorder in hospitalized severely mentally ill psychiatric patients: Prevalence, correlates, and subgroups. *Schizophrenia Bulletin, 26*, 179-193.

[13]Collinge W, Yarnold PR (2001). Transformational breath work in medical illness: Clinical applications and evidence of immunoenhancement. *Subtle Energies & Energy Medicine, 12*, 139-156.

[14]Arozullah AM, Parada J, Bennett CL, Deloria-Knoll M, Chmiel JS, Phan L, Yarnold PR (2003). A rapid staging system for predicting mortality from HIV-associated community-acquired pneumonia. *Chest, 123*, 1151-1160.

[15]Donenberg GR, Bryant FB, Emerson E, Wilson HW, Pasch KE (2003). Tracing the roots of early sexual debut among adolescents in psychiatric care. *Journal of the American Academy of Psychiatry, 42*, 594-608.

[16]Layden BL, Minadeo N, Suhy J, Abukhdeir AM, Metreger T, Foley K, Borge G, Crayton JW, Bryant FB, Mota de Freitas D. (2004). Biochemical and psychiatric predictors of Li^+ response and toxicity in Li^+-treated bipolar patients. *Bipolar Disorders, 6*, 53-61.

[17]Coakley RM, Holmbeck GN, Bryant FB (2006). Constructing a prospective model of psychosocial adaptation in young adolescents with spina bifida: An application of optimal data analysis. *Journal of Pediatric Psychology, 31*, 1084-1099.

[18]Arozullah AM, Lee SD, Khan T, Kurup S, Ryan J, Bonner M, Soltysik RC, Yarnold PR (2006). The roles of low literacy and social support in predicting the preventability of hospital admission. *Journal of General Internal Medicine, 21*, 140-145.

[19]Pachman LM, Abbott K, Sinacore JM, Amoruso L, Dyer A, Lipton R, Ilowite N, Hom C, Cawkwell G, White A, Rivas-Chacon R, Kimura Y, Ray L, Ramsey-Goldman R (2006). Duration of illness is an important variable for untreated children with juvenile dermatomyositis. *The Journal of Pediatrics, 148*, 247-253.

2155-0182/10/$3.00

[20]Nebeker JR, Yarnold PR, Soltysik RC, Sauer BC, Sims SA, Samore MH, Rupper RW, Swanson KM, Savitz LA, Shinogle J, Xu W (2007). Developing indicators of inpatient adverse drug events through non-linear analysis using administrative data. *Medical Care, 45*, S81-S88.

[21]Stalans LJ, Hacker R, Talbot ME (2010). Comparing nonviolent, other-violent, and domestic batterer sex offenders: Predictive accuracy of risk assessments on sexual recidivism. *Criminal Justice and Behavior, 37*, 613-628.

[22]Snowden JA, Leon SC, Bryant FB, Lyons JS (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis. *Journal of Child and Adolescent Psychology, 36*, 8-18.

[23]Rhode P, Stice E, Gau JM (2012). Effects of three depression prevention interventions on risk for depressive disorder onset in the context of depression risk factors. *Prevention Science, 13*, 584-593.

[24]Pape TLB, Guernon A, Lundgren S, Patil V, Herrold AA, Smith B, Blahnik M, Picon LM, Harton B, Peterson M, Mallinson T, Hoffman M. (2013). Predicting levels of independence with expressing needs and ideas 1 year after severe brain injury. *Rehabilitation Psychology, 58*, 253-262.

[25]Belknap SM, Moore H, Lanzotti SA, Yarnold PR, Getz,M, Deitrick DL, Peterson A, Akeson J, Maurer T, Soltysik RC, Storm J (2008). Application of software design principles and debugging methods to an analgesia prescription reduces risk of severe injury from medical use of opioids. *Clinical Pharmacology and Therapeutics, 84*, 385-392.

[26]Lavigne JV, Dahl KP, Gouze KR, LeBailly S A, Hopkins J (2015). Multi-domain predictors of oppositional defiant disorder symptoms in preschool children: Cross-informant differences. *Child Psychiatry & Human Development, 46*, 308-319.

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

[28]Kiguradze T, Temps WH, Yarnold PR, Cashy J, Brannigan RE, Nardone B, Micali G, West DP, Belknap SM (2017). Persistent erectile dysfunction in men exposed to the 5α-reductase inhibitors, finasteride, or dutasteride. *PeerJ*, 5: e3020 https://doi.org/10.7717/peerj.3020

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

[30]Yarnold PR, Linden A (2019). Novometric stepwise CTA analysis discriminating three class categories using two ordered attributes. *Optimal Data Analysis, 8*, 68-71.

[31]Yarnold PR (2019). Fixed *vs*. relative optimal discriminant thresholds: Pairwise comparisons of raters' ratings for a sample. *Optimal Data Analysis, 8*, 103-106.

[32]Yarnold PR, Linden A (2019). Using fixed and relative optimal discriminant thresholds in randomized blocks (matched-pairs) designs. *Optimal Data Analysis, 8*, 114-121.

[33]Skillings JH, Mack GA (1981). On the use of a Friedman-type statistic in balanced and unbalanced block designs. *Technometrics, 23*, 171-177.

[34]Yarnold PR (2019). ODA *vs*. χ^2, *r*, and τ: Trauma exposure in childhood and duration of participation in eating-disorder treatment program. *Optimal Data Analysis, 8*, 72-75.

2155-0182/10/$3.00

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

[36]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis, 3*, 55-77.

[37]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis, 3*, 78-84.

[38]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[39]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement, 56*, 430-442.

[40]Bryant FB, Siegel EKB (2010). Junk science, test validity, and the Uniform Guidelines for Personnel Selection Procedures: The case of *Melendez v. Illinois Bell*. *Optimal Data Analysis, 1*, 176-198.

[41]Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis, 1*, 67-100.

[42]Yarnold PR, Soltysik RC, Collinge W (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis, 2*, 37-42.

[43]Yarnold PR, Soltysik RC (2013). Ipsative transformations are *essential* in the analysis of serial data. *Optimal Data Analysis, 2*, 94-97.

[44]Yarnold PR (2013). Comparing attributes measured with "identical" Likert-type scales in single-case designs with UniODA. *Optimal Data Analysis, 2*, 148-153.

[45]Yarnold PR (2013). Comparing responses to dichotomous attributes in single-case designs. *Optimal Data Analysis, 2*, 154-156.

[46]Yarnold PR (2013). Ascertaining an individual patient's *symptom dominance hierarchy*: Analysis of raw longitudinal data induces Simpson's Paradox. *Optimal Data Analysis, 2*, 159-171.

[47]Yarnold PR, Linden A. (2016). Using machine learning to model dose-response relationships: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis, 5*, 41-52.

[48]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.

[49]Linden A, Yarnold PR (2019). Effect of sample size on discovery of relationships in random data by classification algorithms. *Optimal Data Analysis, 8*, 76-80.

[50]Linden A, Yarnold PR (2019). Multi-layer perceptron neural net model identifies effect in random data. *Optimal Data Analysis, 8*, 94-96.

[51]Linden A, Bryant FB, Yarnold PR (2019). Logistic discriminant analysis and structural equation modeling both identify effects in random data. *Optimal Data Analysis, 8*, 97-102.

[52]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.

2155-0182/10/$3.00

[53]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.

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

[55]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.

[56]Feinglass J, Yarnold PR, Martin GJ, McCarthy WJ (1998). A classification tree analysis of selection for discretionary treatment. *Medical Care*, *36*, 740-747.

[57]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.

[58]Linden A, Yarnold PR (2018). The Australian Gun Buyback Program and rate of suicide by Firearm. *Optimal Data Analysis*, *7*, 28-35.

[59]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.

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

[61]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.

[62]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.

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

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

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

[66]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.

[67]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.

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

Author Notes

No conflict of interest was reported.