Optimal Data Analysis
Vol. 8 (May 30, 2019), 114-121

2155-0182/10/$3.00

Using Fixed and Relative Optimal Discriminant Thresholds in Randomized Blocks (Matched-Pairs) Designs

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

Optimal discriminant analysis (ODA) is often used to compare values of one (or more) attributes between two (or more) groups of observations with respect to a *fixed* discriminant threshold that maximizes accuracy normed against chance for the sample.[1-29] However, a recent study using a matched-pairs design found that using a *relative* discriminant threshold to assess an (exploratory or confirmatory) *a priori* hypothesis *separately for each pair* of observations can identify inter-group differences which otherwise are too subtle to be identified by using fixed thresholds.[30] The present investigation replicates the finding regarding efficacy of relative thresholds for matched-pairs designs, this time for a randomized blocks design consisting of two patient groups (one group assigned to take an antidepressant drug, the other group assigned to take a placebo) between which a numerical measure of depression was compared.[31] Several recommendations are made concerning use of improved modern optimal statistical alternatives for this class of experimental design.

Fleiss presents an example of a randomized blocks (matched-pairs) experiment comparing scores on a depression measure for 60 patients formed into 30 pairs matched on gender, age (within one decade for females, or two decades for males), and time of study entry (within one month).[31] Data are given by pair in Table 1: the first value in each row is score on the Hamilton depression scale (higher scores indicate worse depression) for the patient in the *Imipramine* condition; the second value is the depression score for the paired patient in the *Placebo* condition; and the third value is the difference between these scores—a negative difference indicates the patient in the Placebo condition has a higher depression score (is more depressed) than the patient in the Imipramine condition). In the Table, the pairs were first sorted in order of increasing depression score for the patient in the Imipramine condition; for tied scores, data were sorted a second time by increasing depression score for the patient in the Placebo condition.

2155-0182/10/$3.00

Table 1: Hamilton Depression Scale Scores of 60 Patients in 30 Matched Pairs[31]

Group			Relative Threshold	
Imipramine	Placebo	I-P	I≤P	I<P
3	3	0	1	0
3	8	-5	1	1
3	9	-6	1	1
4	3	1	0	0
4	5	1	1	1
4	6	1	1	1
4	7	1	1	1
4	7	1	1	1
5	2	3	0	0
5	6	-1	1	1
5	8	-3	1	1
5	11	-6	1	1
6	4	2	0	0
6	8	-2	1	1
6	8	-2	1	1
6	9	-3	1	1
6	11	-5	1	1
6	12	-6	1	1
7	5	2	0	0
7	7	0	1	0
7	10	-3	1	1
7	10	-3	1	1
8	8	0	1	0
8	9	-1	1	1
8	11	-3	1	1
9	7	2	0	0
10	5	5	0	0
10	10	0	1	0
11	9	2	0	0
12	9	3	0	0

Classic Analytic Approaches

There are three common classic approaches to statistical analysis for designs in which there are matched pairs or grouped (blocked) data. The first approach is using the paired *t*-test which requires the usual assumptions of linear models—a sufficient sample size and normal distribution of the data. Using a paired *t*-test we derived a mean difference of -1.267 between groups (95% CI: -2.36, -0.17), $p \leq 0.025$.

The second approach is a non-parametric Wilcoxon signed-rank test[32] which tests equality of matched pairs of observations (the null hypothesis is that the distributions are the same). Using this test we obtain $p \leq 0.027$.

A third approach, an extension of the Wilcoxon signed-rank test, is the Hodges-Lehmann treatment effects estimator with distribution-free confidence intervals.[33] This method entails estimating the average difference in outcomes (x-y) for all n(n+1)/2 possible pairs and then deriving the overall median of all averages (the Hodges-Lehmann estimator). Using this approach, we derive a median difference estimate of -1.5 (95% CI: -2.5, 0).

All three approaches were computed using Stata (StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC). The Hodges-Lehmann model was estimated using the user-written module ALIGNEDPAIRS.[34]

ODA Fixed Discriminant Threshold

Hamilton scale scores (Table 1) were compared between Imipramine *vs*. Placebo patient groups via exploratory ODA using fixed discriminant threshold values.[35] Analysis identified a training model (if score≤6 predict group=Imipramine, otherwise predict group=Placebo) which yielded marginally significant ($p<0.090$) moderate effect strength (ESS=30.0).

A directional fixed-threshold ODA was also conducted to test the *a priori* confirmatory hypothesis that patients in the Imipramine group would have lower Hamilton depression scores than patients in the Placebo group. The same model (and associated ESS) identified in non-directional training analysis also emerged in confirmatory analysis, however the effect was statistically significant ($p<0.043$). One-sample (leave-one-out, or LOO) cross-generalizability analysis revealed that while model sensitivity in

2155-0182/10/$3.00

predicting Imipramine patients was unchanged in training and LOO analysis, LOO sensitivity for Placebo patients fell (from 70% in training) to 56.7%, yielding a relatively weak (ESS=16.7) statistically insignificant (p<0.16) effect.

Thus, assessed via a fixed discriminant threshold applied to the entire sample, scores on the Hamilton Scale were unable to reproducibly discriminate Imipramine *vs*. Placebo patients. This is due to misclassification of Imipramine group patients having scores at or near (i.e., 5-7) the model threshold (6 points)—which is half of the maximum observed score (12 points, Table 1). Only *if* it is hypothesized that Imipramine patients had lower depression scores than Placebo patients, *and* if only training results are of interest, *then* the moderate training effect (ESS= 30) was statistically significant (p<0.043).

ODA Relative Discriminant Threshold

Two directional hypotheses and corresponding relative threshold criteria used to evaluate each pairwise comparison are: (1) the Hamilton scale score of the Imipramine patient is less than or equal to the Hamilton scale score of the Placebo patient; and (2) the Hamilton scale score of the Imipramine patient is strictly less than the Hamilton scale score of the Placebo patient.

As seen in Table 1, by the first criterion 22 of 30 Imipramine patients had a Hamilton scale score which was less than or equal to the corresponding score of the paired Placebo patient. If it is assumed that Hamilton scale score is a uniform random variable then p(success)= 0.50 for each pair, and the binomial probability of 22 successes in 30 trials is p<0.00545. By the second criterion, 18 of 30 Imipramine patients had a Hamilton scale score which was strictly less than the corresponding score of the paired Placebo patient (p<0.0806).

Comments

As is often the case, using alternative statistical methods produced different analytic conclusions.[36] Considering legacy methods first, the parametric paired t-test found a statistically significant difference in mean depression scores between Imipramine *vs*. Placebo groups—but Table 1 indicates markedly skewed data for the former group thereby invalidating the assumed normality and calling into question the validity of the obtained p-value. While non-parametric Wilcoxon signed-rank test[37,38] found depression scores of Imipramine patients had significantly lower ranks *vs*. matched Placebo patients, distribution-free confidence intervals for the Hodges-Lehmann treatment effects estimator indicated this effect overlapped zero—thereby indicating a non-significant difference.

Next consider the findings obtained via ODA. First, for a *fixed* discriminant threshold a marginal training effect emerged in *exploratory* analysis, and a significant training effect in *confirmatory* analysis—which became insignificant in *LOO* analysis due to misclassification of Imipramine patients with depression scores near the threshold. Thus, while the confirmatory model explains results obtained for the *present* sample, comparable predictive accuracy is *not* expected by applying the *identical* threshold to classify an *independent* random sample. Second, for a *relative* threshold, a statistically significant effect was obtained for the hypothesis that depression scores of Imipramine patients were less than or equal to (never greater than) depression scores of matched Placebo patients, and a marginally significant effect emerged if it was hypothesized that depression scores of Imipramine patients were strictly lower (always less) than depression scores of matched Placebo patients.

Which analytic finding should be used? The answer to this question depends upon the conceptual orientation of the researcher. If one justifies on a theoretical basis that *means* are the appropriate moment with which to compare the response distributions of the groups, then t-test should be reported (presently the validity of the p-value is suspect). If one's theoretical orientation indicates that *rankings* are the appropriate

2155-0182/10/$3.00

moment to compare response distributions of the groups then the Wilcoxon signed-rank test using the Hodges-Lehmann treatment effects estimator should be reported (presently the confidence interval for the group comparison overlaps zero). And, if one's theoretical orientation is that the appropriate way to compare response distributions of groups is to compare the entire response distributions—identify the model that explicitly maximizes the effect strength normed *vs*. chance, obtain exact *p*-values while making no distributional assumptions, and estimate the cross-generalizability of the model—then ODA with fixed (for robust effects) or relative (for subtle, pairwise effects) should be reported.

The study data are a good pilot sample. Given failure of the training effect to reproduce in LOO analysis for the fixed threshold model, and failure of the "strictly-less-than" relative threshold model to reach statistical significance, it is natural to wonder how to improve the study and obtain a statistically significant, moderate (or stronger) effect strength in cross-generalizability analysis (ODA routinely employs LOO reproducibility analysis, which isn't available for the legacy methods). Achieving statistical significance is easily done (and thus is basically meaningless) by using a larger sample—the size is computed based on the LOO results obtained presently (the first axiom of novometric theory is the sample must provide adequate statistical power to test the alternative hypothesis[39,40]). Increasing the model validity effect strength is more challenging (and thus theoretically and/or translationally meaningful), requiring improving the measurement precision of the attribute, particularly in the response scale region close to the value of the fixed threshold[41], and/or replacing the singular depression score with a battery of measures offering greater theoretical clarity, and measurement granularity and precision.[42-44]

Another, arguably the most influential aspect of the analysis in need of consideration, is the method used in the matching process—or blocking, as Fleiss[31] states. Given that the outcome analysis may be biased due to confounding, it is imperative that the matching/blocking process eliminates confounding of observed covariates. In small samples with few covariates, matching directly on the available covariates may suffice. However, as the number of covariates increases, and as their distribution differs between the treatment and control conditions, methods which stratify and weight individuals into blocks of the propensity score to adjust for observed confounding should be considered.[45]

These techniques are incorporated within the ODA, CTA, and novometric frameworks.[46-48] For example, in any given substantive area of scientific application, optimal ("maximum-accuracy") methods are available to assist researchers to obtain a clear understanding of the factors which must be considered as being potential threats to causal inference: to identify variables characterizing participation in both discretionary treatment[49] and observational[26] research (i.e., to identify possible confounding variables), and to identify structural breaks in single[50,51] and multiple-group[52] interrupted time series analysis, dose-response studies[53,54] and in research investigating mediating processes.[55] 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[46] and randomized trials[56] to remove otherwise undetected threats to causal inference. Globally-optimal (i.e., 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.[57-60]

The present findings further illustrate why we strongly advocate using the 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

2155-0182/10/$3.00

concerning how health care interventions and policy changes are evaluated.[61,62]

References

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

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

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

[4]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.

[5]Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995).Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine*, *10*, 601-606.

[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]Lyons JS (1997). The evolving role of outcomes in managed health care. *Journal of Child and Family Studies*, *6*, 1-8.

[8]Ostrander R, Weinfurt KP, Yarnold PR, August G (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology*, *66*, 660-672.

[9]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.

[10]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.

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

[12]Grammer LC, Zeiss CR, Yarnold PR, Shaughnessy MA (2002). Human leucocyte antigens (HLA) and trimellitic anhydride (TMA) immunologic lung disease. *Respiratory Medicine, 18*, 473-477.

[13]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.

[14]Stalans LJ, Yarnold PR, Seng M, Olson DE, Repp M. (2004). Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis. *Law & Human Behavior*, *28*, 253-271.

[15]Taft CT, Pless AP, Stalans LJ, Koenen KC, King LA, King DW (2005). Risk factors for partner violence among a national sample of combat veterans. *Journal of Consulting and Clinical Psychology, 73*, 151-159.

[16]Coakley RM, Holmbeck GN, Bryant FB (2006). Constructing a prospective model of psychosocial adaptation in young adolescents

2155-0182/10/$3.00

with spina bifida: An application of optimal data analysis. *Journal of Pediatric Psychology, 31*, 1084-1099.

[17]Bennett CL, Nebeker JR, Yarnold PR, Tigue CC, Dorr DA, McKoy JM, Edwards BJ, Hurdle JF, West DP, Lau DT, Angelotta C, Weitzman SA, Belknap SM, Djulbegovic B, Tallman MS, Kuzel TM, Benson AB, Evens A, Trifilio SM, Courtney DM, Raisch DW (2007). Evaluation of serious adverse drug reactions: A proactive pharmacovigilance program (RADAR) versus safety activities conducted by the Food and Drug Administration and pharmaceutical manufacturers. *Archives of Internal Medicine, 167*, 1041-1049.

[18]Kyriacou DM, Yarnold PR, Soltysik RC, Wunderink RG, Schmitt BP, Parada JP, Adams JG (2008). Derivation of a triage algorithm for chest radiography of community-acquired pneumonia in the emergency department. *Academic Emergency Medicine*, *15*, 40-44.

[19]McKoy JM, Bennett CL, Scheetz MH, Differding V, Scarsi KK, Yarnold PR, Sutton S, Chandler K, Palella F, Johnson S, Obadina E, Raisch DW, Parada JP (2009). Hepatotoxicity associated with long-course versus short-course nevirapine use as HIV-prophylaxis among non-HIV infected individuals, HIV-infected pregnant women and their offspring: An analysis from the Research on Adverse Drug events And Reports (RADAR) project. *Drug Safety, 32*, 147-158.

[20]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.

[21]Albuquerque K, Giangreco D, Morrison C, Siddiqui M, Sinacore J, Potkul R, Roeske J (2011). Radiation-related predictors of hematologic toxicity after concurrent chemoradiation for cervical cancer and implications for bone marrow-sparing pelvic IMRT. *International Journal of Radiation Oncology * Biology * Physics, 79*, 1043-1047.

[22]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.

[23]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.

[24]Herrold AA, Pape TLB, Guernon A, Mallinson T, Collins E, Jordan N (2014). Prescribing multiple neurostimulatns during rehabilitation for severe brain injury. *The Scientific World Journal*, Article ID 964578, 1-7.

[25]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.

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

[27]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

2155-0182/10/$3.00

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

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

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

[31]Fleiss JL (1986). *The design and analysis of clinical experiments*. New York, NY: Wiley (pp. 121-125).

[32]Wilcoxon F (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80-83.

[33]Hodges JL, Lehmann EL (1963). Estimation of location based on ranks. *Annals of Mathematical Statistics*, *34*, 598-611.

[34]Linden A (2014). ALIGNEDPAIRS: Stata module to perform the aligned ranks test for matched pairs (Hodges-Lehmann). *Statistical Software Components* S457916, Boston College Department of Economics.

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

[36]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.

[37]Yarnold PR (2015). UniODA *vs*. Wilcoxon rank sum test: A small-sample paired experiment. *Optimal Data Analysis*, *4*, 163-164.

[38]Yarnold PR (2015). UniODA *vs*. Wilcoxon rank-sum test: Invariance over monotonic transformations. *Optimal Data Analysis*, *4*, 200-201.

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

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

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

[42]Yarnold PR, Bryant FB (2014). The role of residuals in optimal and suboptimal statistical modeling. *Optimal Data Analysis*, *4*, 9-11.

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

[44]Yarnold PR (2019). More on: "Optimizing suboptimal classification trees: S-PLUS® propensity score model for adjusted comparison of hospitalized *vs*. ambulatory patients with community-acquired pneumonia". *Optimal Data Analysis*, *8*, 56-59.

[45]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.

[46]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.

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

[48]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in

2155-0182/10/$3.00

multivalued treatments. *Journal of Evaluation in Clinical Practice*, *22*, 875-885.

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

[50]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.

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

[52]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.

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

[54]Yarnold, P.R., 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.

[55]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.

[56]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.

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

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

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

[60]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.

[61]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.

[62]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.