Optimal Data Analysis
Vol. 5 (May 31, 2016), 41-52

2155-0182/10/$3.00

Using Machine Learning to Model Dose-Response Relationships via ODA: Eliminating Response Variable Baseline Variation by Ipsative Standardization

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

A maximum-accuracy machine-learning method for predicting dose of exposure based on distribution of the response variable was recently introduced. Herein we demonstrate the advantages of eliminating baseline variation in the response variable via transformation by ipsative standardization. Using data measuring forearm blood flow responses to intra-arterial administration of Isoproterenol, findings obtained using optimized discriminant analysis and a general estimating equation are compared separately for black and white males (and pooled data) using raw versus ipsatively standardized blood flow data. Findings using raw versus ipsatively standardized forearm blood-flow response data were incongruous. The standardized responses of blacks and whites were indistinguishable through 150 ng/min doses; responses of blacks were elevated at 300 ng/min dose, but at 400 ng/min dose responses regressed to 150 ng/min-dose-levels; while at 400 ng/min dose whites had the greatest response levels observed in the study. Using raw data there was no evidence of inter-method statistical conclusion agreement; baseline variability resulted in failure to statistically confirm numerous inter-dose responses; and the dose-response model yielded moderate predictive accuracy. Using standardized data there was significant evidence of inter-method statistical conclusion agreement; eliminating baseline variability yielded more findings of statistically reliable inter-dose responses; and the dose-response model yielded relatively strong predictive accuracy. This study adds to a growing literature demonstrating that ipsative standardization of the response variable studied in single-case or multiple-observation “repeated measures” designs yields generalizable models that generate the most accurate predictions (normed against chance) that are analytically possible for the sample data.

Understanding the precise nature of an association that may exist between the *dose* of any administered substance (conceptualized as constituting an experimental manipulation in statistics and in systems engineering), and the resulting measured *response* (operationalized using process or "throughput" variables, and outcome variables), is by definition crucial in establishing practice guidelines, demonstrating safety, adherence, and efficacy for evidence-based health care practice.

Recently a machine learning approach was introduced as an alternative to conventional linear models for modeling dose-response (D-R) relationships. Linden, Yarnold, and Nallamothu[1] demonstrated that the "optimal data analysis" or ODA algorithm (also known as the "optimizing" or "maximum-accuracy" algorithm) readily performs the usual functions of conventional parametric statistical methods, such as testing an omnibus model for statistical significance and then conducting any needed *post-hoc* follow-up pairwise comparisons between doses.[2,3] However Linden et al.[1] also emphasized three key, unique advantages that ODA offers in the area of maximum-accuracy D-R research.

First, for every unique empirical application, ODA identifies the exact, non-parametric, assumption-free, pruned, cross-validated "customized" model that employs minimum complexity (maximum parsimony) to explicitly maximize predictive accuracy for the given application. Widely-considered aspects or features of model predictive accuracy are captured in common metrics such as specificity or positive predictive value, for example. The overall ("omnibus") predictive accuracy of any statistical classification model is summarized using an intuitive effect strength index that is called "ESS", and ranges from 0 (the level of predictive accuracy expected by chance) to 100 (perfect, errorless prediction) for every unique empirical application. The ESS statistic facilitates direct comparison of omnibus strength of competing models.[3]

Second, it is straightforward and simple to use an ODA model to make maximum-accuracy point-predictions at the individual subject ("observation") level. This is important, for example, in studies of an individual's adherence to a treatment regimen.

Finally, findings in D-R (and conceptually parallel) experiments can be applied in observational studies using analytic weights for covariate adjustment (available for all ODA models)—for example, in after-market drug studies, research investigating exposure to environmental hazards, or multivalued interventions in which self-selection is likely to bias the outcome.[4-6]

While ODA clearly offers unique capabilities for modeling D-R relationships, the small, imbalanced data sets that are ubiquitous in D-R research often provide weak statistical power -- a limitation that reduces the effectiveness of all statistical methods. And, while not influencing the validity of statistical conclusions reached on the basis of analysis conducted using ODA (for which no distributional assumptions are made), small D-R samples often fail distributional assumptions that are the foundation that support the validity of alternative analytic approaches and their associated estimated *P* values.

The second data issue—the constitution of the training sample and independent validity samples, is the focus of the present study. Simply stated, combining data from observations representing two or more distinct strata – for example, any possible combination of white and black, young and old, sick and healthy, men and women – may produce erroneous results for one or more of the constituent groups.[1] Referred to as paradoxical confounding, this phenomenon occurs in longitudinal studies involving multiple subjects, or in series involving a single unit of observation (e.g., a single-case study involving a patient's longitudinal symptom ratings, or an indexed time-series such as monthly consumer sentiment rating).[3]

Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Sometimes the data of two or more groups simply cannot be combined without inducing confounding -- for example, because response-score-relevant, qualitatively distinct structural differences ("chasms") exist between the groups.[1,3,7] In such circumstances a single model simply cannot be developed—in the sense that no feasible solution exists—to explain an outcome well for both of such distinctly incompatible strata.

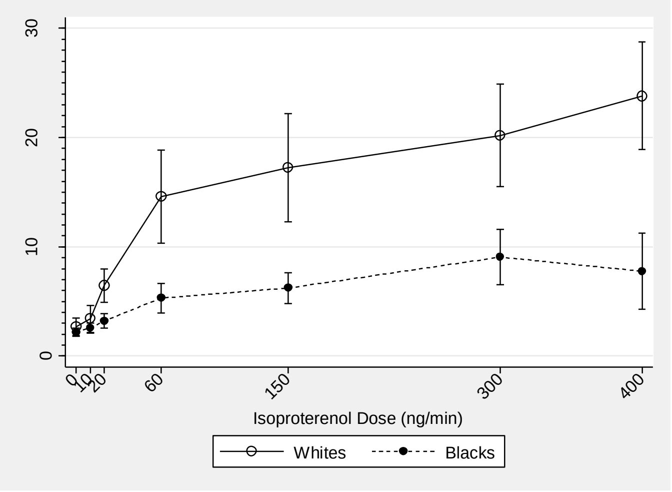
However, sometimes paradoxical confounding arises as a function of measurement artifacts -- specifically in the D-R context, as a function of response values that differ between observations and are known as baseline differences. In serial applications *ipsative standardization* (described ahead) of the response variable is a demonstrated method for eliminating such "nuisance variability" in longitudinal studies involving single cases or multiple observations.[3,7]

The present paper compares the findings of analysis using "ipsatized response data" with prior findings of a parallel analysis using the identical data expressed in their original raw (non-standardized) units of measure. Comparing findings of D-R models developed using general estimating equations (GEE) versus ODA, the paper is organized as follows: we describe data source and analytic methods; present findings using ipsatized response data, and compare them with findings for the raw response data; and discuss advantages of standardized maximum-accuracy machine-learning models of D-R relationships.

Methods

Study data were the same data utilized in a prior analysis[1] of raw (unstandardized) data, drawn from Lang et al. who measured the responses of blood flow in the forearm to the intra-arterial administration of Isoproterenol (in escalating doses of 10 to 400 ng/min) in 9 black and 13 white normotensive men.[8] Analysis of raw data indicated: (1) that forearm blood-flow responses to Isoproterenol were markedly attenuated in normotensive blacks beginning with 20 ng/min doses; and (2) the reduced mean forearm blood-flow responses of blacks in response to doses of 400 ng/min, versus to doses of 300 ng/min, fell within the expected range of the (wide, due to small sample) corresponding confidence intervals.[1,8] This is easily visualized in Figure 1 by examining and comparing means and 95% confidence intervals of the raw forearm blood flow measure, separately by dose and (simultaneously) separately for whites and blacks. Raw data were accessed from a supplement to "*Statistical Modeling for Biomedical Researchers"*.[9]

Figure 1: Forearm Blood Flow Responses (*Raw Units*) to Isoproterenol in Normotensive Blacks (N=9) and Whites (N=12): Values Shown are Means and 95% Confidence Intervals[1,8]



Optimal discriminant analysis (ODA) is a machine learning algorithm that is capable of analyzing data measured using an ordered or qualitative scale (any level of precision): ODA identifies the cutpoint (or category subset) of a predictor variable that yields maximum possible (weighted) prediction accuracy.[2,3] In the present D-R context this is the assignment rule that most accurately classifies the observations into their

Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

actual dose, based on the distribution of the ipsatized response variable.[1] Presently ODA tested the directional (i.e., confirmatory, *a priori*) hypothesis that a greater response (the class variable) will be elicited by a higher dose (the attribute).[1-3]

Statistical significance (p value) for ODA models is computed as a permutation probability, so no distributional assumptions are required of the data and p values are exact. A Sidak Bonferroni-type multiple comparisons methodology is used to prevent “alpha inflation” guaranteeing the desired experimentwise p value (herein, $p < 0.05$), and inhibiting over-fitting.[2,10] Ecological significance of ODA models is assessed using the effect strength for sensitivity (ESS) statistic, a normed measure of predictive accuracy that is chance- and maximum-corrected.[2,3,11] Generalizability of ODA models (which indicates how well the models classify individuals other than those utilized for developing the model—i.e., new patients, or patients in different settings[12]) is accomplished herein using “leave-one-out” (LOO), n-fold cross-validation where n is the number of observations in the dataset, and accuracy is determined as success or failure in predicting the actual class membership across all held-out observations. The results of all n predictions are cumulated to calculate LOO (validity) accuracy, which is then compared to total sample (training model) accuracy. An identical ESS value in both the training and LOO analyses suggests that the ODA model may cross-generalize without a reduction in the predictive accuracy when the model is applied to classify an independent sample.[1-3]

Ipsative standardization is accomplished using the following formula for transforming a raw score into an ipsative z-score: z = (observation’s score – Mean Score) / Standard Deviation (SD) of scores. This same formula is used to compute both a normative (z_N) z-score and an ipsative (z_I) z-score. The difference between z_N and z_I lies in how the Mean Score and SD are derived, and in the conceptual meaning of the two types of z-scores.

For normative z-scores, z_N, Mean Score and SD are calculated for the entire sample. Conceptually z_N measures the magnitude of an individual’s score relative to the population of scores for all observations. That is, z_N scores conceptualize the score of an individual in the context of its relative magnitude in the distribution of such scores in the population. A z_N forearm blood flow index score of a given magnitude conveys meaning regarding one’s location on the (raw or normatively standardized) forearm blood flow index dimension relative to the population: how large is my score—relative to all others?

In contrast, for ipsative z-scores, z_I, Mean Score and SD are based on data only from the individual’s observations. Conceptually z_I measures the magnitude of an individual’s score relative to all scores for that individual.[13-16] For example a z_I score of a given magnitude conveys meaning regarding one’s location on the dimension relative to one’s personal distribution of z_I measurements across time: how large is my score—relative to myself?

The *analytic approach* used presently involved conducting all analyses employing the same methods used in the original study[1,8], however the response data were first transformed using an ipsative standardization conducted separately for every observation.

For the conventional statistical approach, we estimated a GEE model with z_I scores treated as the response (dependent) variable.[1] For the ODA analyses, three separate models were generated—two separately by race, and one pooled. All three models used z_I scores (attribute) to predict assignment to each dose level (class variable). Models were directional (i.e., “one-sided”), with the *a priori* hypothesis that dose would increase with increasing forearm blood flow. Exact p values were estimated using 25,000 Monte Carlo experiments. Omnibus Type I error rate for the effect of multiple tests of statistical hypotheses was controlled using a Sidak multiple comparisons procedure[2] to ensure experimentwise $p < 0.05$. LOO analysis was used to assess the potential cross-generalizability of each ODA model when used to classify individuals other than those in the original study sample.[1-3]

Analytic agreement across corresponding comparisons was assessed using ODA, treating method (GEE versus ODA) as the class variable, and result (experimentwise $p \leq 0.05$ versus $p > 0.05$) as the attribute.[3]

Stata 14.1 (StataCorp., College Station, TX, USA) was used to conduct all GEE analyses, and the Sidak adjustments for multiple testing after generating the GEE and ODA models.[17] ODA analyses were performed using ODA software.[2]

Results

Examination of 95% confidence interval (CI) overlap for black and white participants reveals that the results, as reflected by raw (Figure 1) versus by ipsatively standardized (Figure 2) response data, are incongruous. CIs for raw data indicate that forearm blood-flow responses to Isoproterenol were attenuated in normotensive blacks beginning with 20 ng/min doses (Figure 1). In contrast, CIs for ipsatized data indicate that standardized forearm blood-flow responses of blacks and whites were indistinguishable through 150 ng/min doses, and CIs for ipsatized responses of blacks at 300 ng/min dose were elevated—not attenuated (Figure 2). Ipsatized responses of blacks at 400 ng/min dose regressed to response levels observed at 150 ng/min dose, while whites at 400 ng/min dose had the greatest observed response levels.

The effect is exacerbated by GEE analysis (Table 1). For example, all three paired-comparisons between responses to 150, 300, and 400 ng/min doses were statistically significant for whites (150 < 300 < 400), and for blacks (150 < 400 < 300). In contrast, in ODA analysis only the (150 < 300) comparison for blacks, and the (150 < 400) comparison for whites, achieved the criterion for experimentwise statistical significance (Appendix I).

Figure 2: Forearm Blood Flow Responses (*Ipsatively Standardized z-Score Units*) to Isoproterenol in Normotensive Blacks (N=9) and Whites (N=12): Values Shown are Means and 95% Confidence Intervals

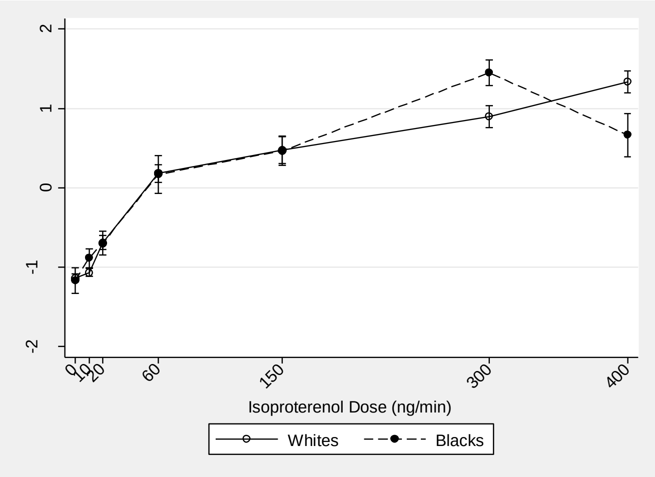


Table 2 gives the results of ODA-based analysis, including cut-points (decision thresholds) on the z_I scores that are associated with each dose of Isoproterenol (these cutpoints explicitly maximize training model ESS), and the corresponding model sensitivity (true positive rate)—the proportion of individuals correctly predicted by the ODA model to be on each specific dose.[18] For clarity, we use the 20 ng/min dose of the pooled data as an example. The ODA model predicts an individual was on a dose of 20 ng/min if their ipsatively standardized (z_I) forearm blood flow was $-0.878 < z_I \leq -0.273$ (values are the ipsatively standardized equivalent of ml/min/dl-units). The ODA model correctly classified 61.90% of individuals at this dose in training analysis (Table 2). In LOO analysis model sensitivity at this dose declined modestly to 57.14%, suggesting the model can predict with relatively strong accuracy which of newly tested subjects is on the 20 ng/min dose.

Table 1: Ipsatively Standardized Forearm Blood Flow Responses to Isoproterenol in Normotensive Blacks (N=9) and Whites (N=12)[1]. Results are From a Generalized Estimating Equation (GEE) Model in which Forearm Blood Flow was Regressed on Dose, Race, and an Interaction Term of Dose and Race

	Dose of Isoproterenol (ng/min)							
	0	10	20	60	150	300	400	*p* value
All								
Mean	-1.149	-0.988	-0.692	0.176	0.472	1.134	1.047	<0.0001
SE	0.038	0.030	0.042	0.061	0.063	0.053	0.072	
95% CI	-1.224, -1.074	-1.047, -0.930	-0.774, -0.610	0.056, 0.296	0.348, 0.596	1.030, 1.238	0.906, 1.188	
White								
Mean	-1.134	-1.065	-0.688	0.181	0.476	0.897	1.332	<0.0001
SE	0.026	0.026	0.045	0.056	0.085	0.069	0.070	
95% CI	-1.185, -1.084	-1.116, -1.014	-0.776, -0.599	0.072, 0.290	0.309, 0.642	0.762, 1.032	1.195, 1.470	
Black								
Mean	-1.168	-0.886	-0.698	0.168	0.467	1.450	0.666	<0.0001
SE	0.083	0.060	0.077	0.122	0.094	0.083	0.140	
95% CI	-1.330, -1.006	-1.004, -0.767	-0.848, -0.548	-0.071, 0.408	0.283, 0.652	1.287, 1.614	0.392, 0.939	
Diff (Black vs White)	-0.034	0.179	-0.010	-0.013	-0.009	0.553	-0.667	
SE	0.086	0.066	0.089	0.134	0.127	0.108	0.156	
95% CI	-0.204, 0.135	0.050, 0.308	-0.184, 0.164	-0.276, 0.250	-0.257, 0.240	0.341, 0.765	-0.973, -0.361	
p value	0.693	0.007	0.912	0.925	0.946	<0.0001	<0.0001	

Table 2: Ipsatively Standardized Forearm Blood Flow Responses to Isoproterenol in Normotensive Blacks (N=9) and Whites (N=12).[1] Results are From an Optimal Discriminant Analysis (ODA): Values Represent Optimal (Maximum-Accuracy) Thresholds (Cut-Points) on the Response Variable (Forearm Blood Flow)

	Dose of Isoproterenol (ng/min)							ESS	
	0	10	20	60	150	300	400	(%)	*P* value
All									
Cutpoints	<= -1.093	> -1.093 & <= -0.878	> -0.878 & <= -0.273	> -0.273 & <= 0.294	> 0.294 & <= 0.845	> 0.845 & <= 0.914	> 0.914		
Sens-train (%)	61.90	61.90	90.48	66.67	71.43	14.29	66.67	55.56	<0.001
Sens-LOO (%)	61.90	57.14	85.71	66.67	57.14	9.52	4.76	40.48	
White									
Cutpoints	<= -1.125	> -1.125 & <= -0.895	> -0.895 & <= -0.2648	> -0.2648 & <= 0.294	> 0.294 & <= 0.543	> 0.543 & <= 0.986	> 0.986		
Sens-train (%)	58.33	83.33	100.00	75.00	33.33	66.67	100.00	69.44	<0.001
Sens-LOO (%)	58.33	75.00	100.00	75.00	8.33	25.00	100.00	56.94	
Black									
Cutpoints	<= -1.079	> -1.079 & <= -0.739	> -0.739 & <= -0.273	> -0.273 & <= 0.173	> 0.173 & <= 1.014	> 1.014 & <= 1.680	> 1.680		
Sens-train (%)	66.67	66.67	66.67	55.56	88.89	77.78	11.11	55.56	<0.001
Sens-LOO (%)	66.67	55.56	66.67	55.56	66.67	66.67	11.11	48.15	

Notes: Sens = sensitivity; LOO = leave one out cross validation; ESS = effect strength for sensitivity

As seen, use of ipsatized data increased the normed predictive accuracy of ODA models. ODA models of raw response data yielded moderate effects for blacks and whites in training (ESS = 40.74 and 40.28) and LOO validity analysis (ESS = 29.63, 27.78). Using ipsatively standardized response data ODA models yielded relatively strong effects for both blacks and whites in training (ESS = 55.56 and 69.44), and moderate to relatively strong effects in LOO validity analysis (ESS = 48.15, 56.94). The weakest (least accurate) model obtained for ipsatively standardized data emerged in LOO analysis (ESS = 48.15), and it was 18.2% more accurate than the strongest (most accurate) model obtained for raw data that emerged in training analysis (ESS = 40.74).

Examination of training and LOO sensitivities (Table 2) reveals that the ODA model for whites has relatively strong training predictive accuracy for all doses studied except 150 ng/min, and relatively strong LOO predictive accuracy for all doses studied except 150 and 300 ng/min. Training and LOO classification of the ODA model for blacks is relatively strong for all doses studied except 400 ng/min.

The plurality of the standardized responses of men to doses between 0 and 150 ng/min (Figure 2) not only yielded ODA models having greater normed predictive accuracy (ESS) in training and LOO analysis (Table 2) as compared with corresponding models developed using raw response data.[1] The comparative uniformity of standardized (versus raw) response also reduced inter-subject variability (heterogeneity) and thus increased statistical power.[3] For blacks, the number of statistically unreliable pairwise comparisons for GEE analysis was reduced from 8 using raw data to 5 using standardized data, and the number of statistically unreliable pairwise comparisons for ODA analysis was reduced from 13 using raw data to 7 using standardized data.

There were five instances of paradoxical confounding in GEE pairwise comparisons of ipsatized response rate: there was no effect for the pooled sample for the 300-400 ng/min dose comparison, but there were significant effects for both cohorts (positive for whites, negative for blacks); and a significant effect emerged for the pooled and white samples, but not for the black sample, for the 10-20, 60-150, 60-400, and 150-400 comparisons (Appendix I). There also were five instances of paradoxical confounding in ODA pairwise comparisons of ipsatized response rate: a significant effect emerged for pooled and white samples, but not for the black sample, for 0-20, 10-20, 60-400, and 150-400 comparisons; and a significant effect emerged for pooled and black samples, but not for the white sample, for the 150-300 comparison (Appendix I).

Discussion

The choice of utilizing raw or ipsatively standardized values depends on the focus and objective of the application (i.e., research or clinical practice). In the D-R research discussed herein, while analysis of raw data showed strong differences between blacks and whites beginning at a low dose[1], ipsative standardization showed an indistinguishable response between blacks and whites at low to moderate doses, and a modest response difference at higher doses.

Ipsative standardization is clearly the appropriate perspective in this D-R study because there was evidence of strong agreement between methods for blacks (Appendix II), whereas there was no evidence of reliable agreement concerning statistical conclusions obtained by GEE versus by ODA analyses for raw response measures.[1] Failure to achieve statistically reliable inter-method agreement for whites was attributable to the meager statistical power afforded by nearly perfectly homogeneous statistical conclusions (i.e., $p < 0.05$) reached by GEE analysis (Appendix II).

Considered from the perspective of maximizing predictive accuracy of point predictions made using the present models—

Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

involving either raw or ipsatively standardized response scores, the predictive accuracy of the ODA model for the combined sample never exceeds and rarely equals the predictive accuracy attained by the models developed separately for blacks and whites. Therefore the race-specific models developed presently are appropriate, versus the model for the pooled data, as a means of maximizing model predictive accuracy and circumventing paradoxical confounding in future research conducted in this area.

In summary, this study adds to a growing and converging literature that consistently demonstrates that ipsative standardization of the response variable studied in single-case or multiple-observation "repeated measures" designs yields generalizable models that generate the most accurate predictions analytically possible for the data.

References

[1]Linden, A., Yarnold, P.R. & Nallamothu, B. (*In Press*) Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*.

[2]Yarnold, P.R., & Soltysik, R.C. (2005) *Optimal data analysis: A Guidebook with Software for Windows* Washington, DC: APA Books.

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

[4]Linden, A., & Yarnold, P. R. (2016) Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12538

[5]Linden, A., & Adams, J. (2010a). Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice*, *16*, 175-179.

[6]Linden, A., & Adams, J. (2010b). Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice*, *16*, 180-185.

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

[8]Lang, C.C., Stein, C.M., Brown, R.M., Deegan, R., Nelson, R., He, H.B., Wood, M. & Wood, A.J. (1995) Attenuation of Isoproterenol-mediated vasodilatation in blacks. *New England Journal of Medicine*, 333, 155-160.

[9]Dupont, W. D. (2009) *Statistical Modeling for Biomedical Researchers.* Cambridge, U.K.: Cambridge University Press.

[10]Linden, A., & Yarnold, P.R. (2016) Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12544

[11]Linden, A., & Yarnold, P.R. (2016) Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12515

[12]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.

[13]Yarnold, P.R. (1988). Classical test theory methods for repeated-measures N=1 research designs. *Educational and Psychological Measurement*, *48*, 913-919.

[14]Yarnold, P.R. (1992). *Statistical analysis for single-case designs*. In: F.B. Bryant, L. Heath, E. Posavac, J. Edwards, E. Henderson, Y. SuarezBalcazar, S. Tindale (Eds.), *Social Psychological Applications to Social Issues, Volume 2: Methodological Issues in Applied*

Social Research. New York, NY: Plenum, pp. 177-197.

[15]Mueser, K.T., Yarnold, P.R., & Foy, D.W. (1991). Statistical analysis for single-case designs: Evaluating outcomes of imaginal exposure treatment of chronic PTSD. *Behavior Modification*, *15*, 134-155.

[16]Yarnold, P.R., Feinglass, J., Martin, G.J., & McCarthy, W.J. (1999). Comparing three pre-processing strategies for longitudinal data for individual patients: An example in functional outcomes research. *Evaluation and the Health Professions*, *22*, 254-277.

[17]Newson, R.B. (2010) Frequentist q-values for multiple-test procedures. *The Stata Journal*, 10, 568-584.

[18]Linden, A. (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12, 132-139.

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported. We wish to thank Julia Adler-Milstein for reviewing the manuscript and for providing many helpful comments.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646

Appendix I

Table 1: Sidak Adjusted *p* Values for All Pairwise Comparisons Following GEE for Pooled Ipsatized Data

	Dose of Isoproterenol (ng/min)					
	0	10	20	60	150	300
0						
10	0.068					
20	< 0.001	< 0.001				
60	< 0.001	< 0.001	< 0.001			
150	< 0.001	< 0.001	< 0.001	0.001		
300	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
400	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	1.000

Table 2: Sidak Adjusted *p* Values for All Pairwise Comparisons Following GEE for Whites only, Ipsatized Data

	Dose of Isoproterenol (ng/min)					
	0	10	20	60	150	300
0						
10	0.121					
20	< 0.001	< 0.001				
60	< 0.001	< 0.001	< 0.001			
150	< 0.001	< 0.001	< 0.001	0.002		
300	< 0.001	< 0.001	< 0.001	< 0.001	0.008	
400	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.005

Table 3: Sidak Adjusted *p* Values for All Pairwise Comparisons Following GEE for Blacks only, Ipsatized Data

	Dose of Isoproterenol (ng/min)					
	0	10	20	60	150	300
0						
10	0.372					
20	0.01	0.505				
60	< 0.001	< 0.001	< 0.001			
150	< 0.001	< 0.001	< 0.001	0.449		
300	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
400	< 0.001	< 0.001	< 0.001	0.551	0.999	0.004

Table 4: Sidak Adjusted *p* Values for All Pairwise Comparisons Following ODA for Pooled Ipsatized Data

	Dose of Isoproterenol (ng/min)					
	0	10	20	60	150	300
0						
10	0.340					
20	< 0.001	< 0.001				
60	< 0.001	< 0.001	< 0.001			
150	< 0.001	< 0.001	< 0.001	0.162		
300	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
400	< 0.001	< 0.001	< 0.001	< 0.001	0.005	1.000

Table 5: Sidak Adjusted *p* Values for All Pairwise Comparisons Following ODA for Whites only, Ipsatized Data

	Dose of Isoproterenol (ng/min)					
	0	10	20	60	150	300
0						
10	0.946					
20	< 0.001	< 0.001				
60	< 0.001	< 0.001	< 0.001			
150	< 0.001	< 0.001	< 0.001	0.667		
300	< 0.001	< 0.001	< 0.001	< 0.001	0.275	
400	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.089

Table 6: Sidak Adjusted *p* Values for All Pairwise Comparisons Following ODA for Blacks only, Ipsatized Data

	Dose of Isoproterenol (ng/min)					
	0	10	20	60	150	300
0						
10	0.754					
20	0.301	0.982				
60	0.005	< 0.001	0.005			
150	< 0.001	< 0.001	< 0.001	0.754		
300	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	
400	0.005	< 0.001	< 0.001	0.298	1.000	1.000

Appendix II

This Appendix details the results of the analyses evaluating inter-method agreement across the multiple inferential paired-comparisons tests that were performed.

Pooled Sample: Comparing the 21 GEE and ODA paired-comparisons (Appendix I Tables 1 and 4) generates the following cross-classification table (ESS=94.7, *p*< 0.0143):

	ODA, *p*<0.05	ODA, *p*>0.05
GEE, *p*<0.05	18	1
GEE, *p*>0.05	0	2

Thus, there is evidence in support of the hypothesis that paired-comparison findings of GEE and ODA models developed using the ipsatized response data strongly agree, when data are aggregated across all 21 paired-comparisons for pooled data.

Whites: Comparing the 21 GEE and ODA paired-comparisons for whites (Appendix I Tables 2 and 5) generates the following cross-classification table (ESS=85.0, *p*<0.43):

	ODA, *p*<0.05	ODA, *p*>0.19
GEE, *p*<0.05	17	3
GEE, *p*>0.05	0	1

Thus, there is no statistically reliable association of results when experimentwise *p* values for GEE and ODA models are aggregated across all 21 of the paired-comparisons for whites.

Blacks: Comparing the 21 GEE and ODA paired-comparisons for blacks (Appendix I Tables 3 and 6) generates the following cross-classification table (ESS=87.5, *p*<0.0011):

	ODA, *p*<0.05	ODA, *p*>0.05
GEE, *p*<0.05	14	2
GEE, *p*>0.05	0	5

Thus, there is evidence in support of the hypothesis that paired-comparison findings of GEE and ODA models developed using the ipsatized response data strongly agree, when data are aggregated across all 21 paired-comparisons for blacks.