# Using Machine Learning to Model Dose-Response Relationships

Ariel Linden, DrPH[1,2], Paul R. Yarnold, PhD[3,4], Brahmajee K. Nallamothu, MD, MPH[5]

[1] President, Linden Consulting Group, LLC, Ann Arbor, Michigan, USA

[2] Research Scientist, Division of General Medicine, Medical School - University of Michigan, Ann Arbor, Michigan, USA

[3] Principal Scientist, Optimal Data Analysis, LLC

[4] Statistician, Southern Network on Adverse Reactions (SONAR), College of Pharmacy, University of South Carolina, Columbia, South Carolina, USA

[5] Professor, Division of Cardiovascular Diseases, Department of Internal Medicine, Medical School - University of Michigan, Ann Arbor, Michigan, USA

**Corresponding Author Information**:
Ariel Linden, DrPH
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI USA 48103
Phone: (971) 409-3505
Email: alinden@lindenconsulting.org

**Key Words:** machine learning, data mining, dose-response, efficacy, adherence

**Running Header**: Machine learning for estimating dose-response

**Acknowledgement:** We wish to thank Julia Adler-Milstein for her review and feedback on the manuscript. Dr. Yarnold's work was partly funded by a grant from the National Cancer Institute (1R01CA165609-01A1).



This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jep.12573

## ABSTRACT

Rationale, aims and objectives: Establishing the relationship between various doses of an exposure and a response variable is integral to many studies in health care. Linear parametric models, widely used for estimating dose-response relationships, have several limitations. This paper employs the optimal discriminant analysis (ODA) machine-learning algorithm to determine the degree to which exposure dose can be distinguished based on the distribution of the response variable. By framing the dose-response relationship as a classification problem, machine learning can provide the same functionality as conventional models, but can additionally make individual level predictions, which may be helpful in practical applications like establishing responsiveness to prescribed drug regimens.

Method: Using data from a study measuring the responses of blood flow in the forearm to the intra-arterial administration of isoproterenol (separately for 9 black and 13 white men, and pooled), we compare the results estimated from a generalized estimating equations (GEE) model with those estimated using ODA.

Results: GEE and ODA both identified many statistically significant dose-response relationships, separately by race and for pooled data. *Post-hoc* comparisons between doses indicated ODA (based on exact *P* values) was consistently more conservative than GEE (based on estimated *P* values). Compared to ODA, GEE produced twice as many instances of paradoxical confounding (findings from analysis of pooled data that are inconsistent with findings from analyses stratified by race).

Conclusions: Given its unique advantages and greater analytic flexibility, maximum-accuracy machine learning methods like ODA should be considered as the primary analytic approach in dose-response applications.

## 1. INTRODUCTION

Establishing the relationship between various doses of an exposure and a response variable (i.e. outcome) is integral to many studies in health care, whether it is for determining safety (e.g. environmental hazards, drug toxicity), efficacy (e.g. a new drug, multivalued treatments), or adherence/responsiveness (e.g. treatment plan, intervention regimen). Correspondingly, failure to define dose-response relationships may lead to unacceptable toxicity or adverse-effect rates, marginal evidence of effectiveness, and a lack of information on how to individualize dosing regimens [1].

Linear statistical models, such as analysis of variance (ANOVA), generalized estimating equations (GEE), or multilevel models are widely used for estimating dose-response relationships. As a family, these parametric models share several drawbacks when used in dose-response studies. First, they assume a linear relationship exists between the dose and the response. Given that orderly, linear relationships rarely exist in health care data, such models may over- or under-estimate the true dose-response relationship at various points across the range of doses studied. Second, only a limited number of doses are typically tested in most dose-response studies, thus requiring interpolation or extrapolation for any dose not studied. Third, conventional statistical methods are intended for estimating treatment effects at the population level, are generally inaccurate when applied to small samples, and are inappropriate when used for making point predictions concerning the response effect for individuals. Several innovative modeling approaches and software have been introduced to account for non-linear relationships

between dose and response (e.g. see [2,3]). However these approaches are implemented within a parametric framework, and thus many of the same limitations still apply.

In this paper, we introduce a novel machine-learning approach to establish dose-response relationships. This methodology employs an algorithm called optimal discriminant analysis (ODA) [4,5] to determine if, and to what degree, doses of the exposure can be distinguished based on the distribution of the response variable. By framing the dose-response relationship as a classification problem (i.e., how accurately does the response variable classify patients into specific dose levels), many of the aforementioned limitations of conventional statistical models are overcome, while obtaining several additional benefits. Specific advantages of ODA as compared to the conventional statistical approach in assessing dose-response relationships include the ability to handle a response variable measured using any metric (from categorical to continuous) and number of doses, insensitivity to skewed data or outliers, and the use of accuracy measures that can be widely applied to all classification analyses. ODA also offers the unique ability to ascertain if individuals are likely to be responding to the dose as prescribed based on optimized (maximum-accuracy) cut-points on the response variable. Moreover, ODA accepts analytic weights, thereby extending the assessment of a dose-response relationship to observational studies where weights are used for covariate adjustment [6]. Finally, ODA provides the capability to use cross-validation in assessing the generalizability of the model to other individuals outside of the original study sample, or to identify solutions that cross-generalize with maximum accuracy when applied across multiple samples.

To illustrate the ODA approach and compare it to a conventional statistical approach (we use a GEE model) for estimating dose-response relationships, we organized the paper as follows. In Section 2 we describe our methods including the data source, an introduction to ODA, and the analytic strategy employed in the study. Section 3 reports the results of each approach and a comparison between them. Finally, Section 4 discusses the specific advantages of ODA in assessing dose-response relationships compared to the conventional approach.

## 2. METHODS

### *2.1 Data*

We use data from Lang et al [7] that measures the responses of blood flow in the forearm to the intra-arterial administration of isoproterenol (in seven escalating doses: 0, 10, 20, 60, 150, 300 and 400 ng/min) in 9 normotensive black men and 13 normotensive white men. This study found that forearm blood-flow responses to isoproterenol were markedly attenuated in normotensive blacks, whereas the responses were approximately linear in white subjects. It was hypothesized that the mechanisms responsible for blunted vasodilatation in response to the administration of isoproterenol may contribute to enhanced vascular reactivity and influence the pathogenesis of hypertension in blacks.

These data are ideal for illustrating many of the problems typically encountered in dose-response studies: small sample size, limited number of doses, non-linearity in the dose-response relationship, and differential effects by subgroup. The dataset was accessed as a supplement to

the book *"Statistical Modeling for Biomedical Researchers"* [8] (found at: http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/index.html#datasets).

*2.2 A Brief Introduction to Optimal Discriminant Analysis (ODA)*

ODA is a machine learning algorithm developed more than 25 years ago [9,10] that is capable of analyzing data measured on a continuous or interval-level scale, on an ordered scale having relatively few levels, or on a qualitative scale with two or more categories [4,5]. As opposed to parametric statistical models that model a dose-response relationship using a linear function that maximizes variance explained or the value of the likelihood function, the ODA algorithm identifies the cutpoint (or category subset) of the response variable that yields maximum predictive accuracy in classifying observations into their actual dose. Explicitly maximized classification accuracy may be either the effect strength for sensitivity (ESS) (described in the next Section), or overall percent accuracy in classification (PAC) depending on whether or not the investigator chooses to weight the data by prior odds [4,5]. As used herein, for an ordered or continuous response variable and multiple discrete doses (exposure variable), the ODA model has the form: if response variable score $\leq$ (threshold value$_1$) predict that the observation is from dose A; if response variable score > (threshold value$_1$) and $\leq$ (threshold value$_2$) predict that the observation is from dose B; or else if response variable score > (threshold value$_2$) predict that the observation is from dose C, and so forth.

*2.2.1 Assessing statistical significance of ODA models*

Statistical significance (*P* value) for ODA models is computed via Monte Carlo simulation and reported as a permutation probability, such that no distributional assumptions are required of the

data and *P* values are exact [5,9]. In study designs involving more than one test of statistical significance (for example, multiple pairwise comparisons between doses), a Sidak Bonferroni-type multiple comparisons methodology is employed to prevent "alpha inflation" and ensure the desired experimentwise *P* value (here, $P < 0.05$), and to inhibit over-fitting [5,11].

*2.2.2 How GEE and ODA deal with repeated observations*

The statistical model used to evaluate dose-response relationships must be able to account for autocorrelation when individuals are subjected to multiple doses of the exposure (e.g. increasing doses of a drug). Autocorrelation indicates that any variable measured over time is potentially influenced by previous observations [12]. So in the case of an individual exposed to multiple doses, an individual's current response at the current dose is likely to be correlated with their previous response to the previous dose. This is contrast to studies in which individuals are only exposed to a single dose, and thus all observations are independent of each other.

If the response score is autocorrelated in a consistent manner within and between individuals, then estimating a GEE model that accounts for the within-group correlation structure will increase the fit of the model for predicting the response when individuals are tested across multiple doses. This assumes that "detecting" (or failing to detect) autocorrelation is statistically warranted for ubiquitous small samples (e.g., with respect to statistical power, and to the veracity of assumptions upon which validity depends), and that autocorrelation-based models retain predictive accuracy when applied in independent validity and generalizability assessments.

In its simplest application, the ODA approach may be used to determine if response scores can be reliably discriminated between two different levels of a stimulus (e.g., dose),

without considering the possible existence of autocorrelation in the response score. In this case the ODA algorithm identifies the model that maximizes ESS achieved in testing the non-directional (i.e., exploratory, *a posteriori*) alternative hypothesis that response differs between the two doses. If instead it is hypothesized that there is some specific direction of influence of autocorrelation (either positive or negative), then this may be modeled in ODA by testing a directional (i.e., confirmatory, *a priori*) hypothesis: for example that a greater response will be elicited by a higher dose, or an even more precise *a priori* hypothesis specifying the exact thresholds and decision-making criteria [4,5]. Finally, if instead one is interested in identifying (vis-à-vis exploratory research) or in evaluating (vis-à-vis confirmatory research) ODA-based *dynamic models*, maximum-predictive-accuracy methods have been developed and found to be more accurate and parsimonious than alternative methods in applications involving single-case as well as multiple-observation time series, and traditionally analyzed using methods such as Markov models, turnover tables, test-retest reliability cross-classification methods, and so forth. [4,5,11]

*2.2.3 Ecological significance of ODA models*

The issue of how best to interpret the predictive accuracy achieved by a statistical model has been extensively discussed, and wide and broadening consensus asserts that evaluating model performance based solely on statistical significant findings (*P* values) is inappropriate [4,5]. Instead, it is argued that assessment should also include the ability of the model to achieve "clinically" [13,14] or otherwise context-relevant "ecologically" meaningful levels of predictive accuracy [15,16,17,18].

Ecological significance of ODA models in particular, and of all classification methodologies in general, is assessed using the effect strength for sensitivity (ESS) statistic. ESS is a normed measure of predictive accuracy that is both chance-corrected (0 = the level of predictive accuracy expected by chance), and maximum-corrected (100 = perfect prediction) [4,5,6,19]. The cut-points (or category subsets) identified by ODA explicitly maximize the ESS obtained by the ODA model for the total ("training") sample. Using ESS, investigators may directly compare the predictive accuracy of different models (corrected for chance) -- developed using the same and/or different samples, regardless of structural features such as sample size, skew, or "outliers" [5]. Established via simulation research, ESS $\leq$ 25% conventionally indicates a relatively weak effect; values >25% to $\leq$ 50% indicate a moderate effect; values >50% to $\leq$ 75% indicate a relatively strong effect; values >75% to $\leq$ 90% or less indicate a strong effect; and ESS values > 90% indicate a very strong effect [4].

*2.2.4 Assessing generalizability of ODA models*

Cross-validation in the dose-response context denotes the generalizability of the model when it is used to classify a sample of individuals other than those utilized for developing the model -- for example new patients prescribed the medication [20]. Commonly-used algorithms for estimating model generalizability include bootstrapping, *k*-fold cross-validation, and leave-one-out jackknife (LOO) cross-validation [4,19,21,22]. In this paper, we use ODA to implement the LOO approach -- *n*-fold cross-validation, where *n* is the number of observations in the dataset. Each observation is in turn held out, predicted class membership (i.e., dose) is obtained for each held-out observation, and accuracy is determined as success or failure in predicting the actual class

membership across 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. In contrast, an ESS value that is lower in the LOO analysis than in the training analysis suggests that application of the ODA model may yield lower normed predictive accuracy when used to classify independent samples.

*2.3 Analytic approach*

All analyses were conducted using the data analyzed in the original study [7]. For the conventional statistical approach, we estimated a GEE model with forearm blood flow (ml/min/dl) treated as the response (dependent) variable. As in the original analysis, the covariates included Isoproterenol dose treated as a categorical ordinal variable with values of 0, 10, 20, 60, 150, 300 and 400 (ng/min); race (black/white) treated as a binary categorical variable; and an interaction term between dose and race to allow for comparisons of the response between race and dose. The GEE model was estimated using a Gaussian family and identity link with an exchangeable within-group correlation structure to account for autocorrelation within subjects across multiple doses, and robust standard errors [23,24]. All pairwise contrasts (between doses) were adjusted for multiple comparisons using Sidak's method [25].

For the ODA analyses, three separate models were generated -- two separately by race, and one pooled. All three models used forearm blood flow (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. We controlled the omnibus Type I error rate for the effect of multiple tests of statistical hypotheses by performing a Sidak multiple comparisons procedure to ensure an experimentwise $P < 0.05$ [4], and LOO analysis was conducted to assess the potential cross-generalizability of each ODA model when used to classify individuals other than those in the original study sample.

The presence of paradoxical confounding (also called Simpson's Paradox) was assessed separately for GEE and ODA analyses. Paradoxical confounding exists when findings of analyses conducted for pooled data (e.g., $P \leq 0.05$ or $P > 0.05$) are inconsistent with the findings of analyses conducted separately for any of the constituent groups [5,26].

Finally, we evaluated the analytic agreement across corresponding comparisons using ODA, with method (GEE versus ODA) treated as the class variable, and finding (experimentwise $P \leq 0.05$ versus $P > 0.05$) treated as the attribute [5]. That is, we compared corresponding ODA and GEE statistical conclusion findings pooled across all 21 multiple comparisons for the blacks-only data, for the whites-only data, and for the combined (blacks and whites) data. If GEE and ODA methods agree on statistical conclusions regarding inter-dose response differences, this indicates cross-analytic-generalizability of the findings. However, if GEE and ODA methods disagree on the statistical conclusions regarding inter-dose response differences, this may indicate that: (1) even though significantly different mean response scores exist between two doses (GEE), dose is not an accurate predictor of individual response (ODA); (2) that *P* values generated by GEE (which must satisfy distributional assumptions in order to be valid) are

misleading (ODA uses permutation *P* values that require no distributional assumptions and are always valid); and/or (3) there is inadequate statistical power (e.g., due to small sample, class sample size imbalance, and/or sparse data) to warrant analytic comparison [5].

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 [27]. ODA analyses were performed using *ODA Software* [4].

## 3. RESULTS

The Figure illustrates means and 95% confidence intervals of the response variable (forearm blood flow) for each dose, separately for white and black subjects. Table 1 presents the numeric results of the GEE-based analysis. As reported by Lang et al [7], forearm blood flow increases with escalating dose in the white subjects, while the dose-response relationship appears blunted in the black subjects. Examination of confidence interval overlap indicates that the difference between the black and white cohorts in forearm blood flow becomes statistically significant commencing at a dose of 20 ng/min. Interestingly, the pooled data show a similarly strong dose-response relationship to that of the white cohort (Sidak adjusted *P* values for the GEE pairwise comparisons between doses are presented in Appendix Tables 1-3). In the white cohort, all pairwise dose-response comparisons are statistically significant at the experimentwise $P < 0.05$ level except for the 0 to 10 ng/min comparison (Appendix Table 2). Conversely, in the black cohort, only 13 of the 21 pairwise comparisons are statistically significant at the experimentwise criterion (Appendix Table 3). When the data are pooled, all pairwise comparisons are statistically

significant except for the 300 to 400 ng/min comparison (Appendix Table 1). This clearly indicates the presence of paradoxical confounding [26]. In total, there were eight instances of confounding in the GEE analyses: a significant effect for the pooled sample for the 0-10 dose comparison, but no effect for either cohort; a significant effect for the pooled and white samples, but not for the black sample, for the 10-20, 10-400, 20-400, 60-150, 60-400, and 150-400 comparisons; and a significant effect for the white sample, but not for the pooled or black samples, for the 300-400 comparison.

Table 2 presents the results of the ODA-based analysis, including the cut-points (decision thresholds) on the forearm blood flow variable 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 that are correctly predicted by the ODA model to be on each specific dose [28]. To facilitate interpretation of these values, we use the 20 ng/min dose of the pooled data as an example. The ODA model predicts that an individual was on a dose of 20 ng/min if their forearm blood flow was > 2.945 (ml/min/dl) and ≤ 6.855 (ml/min/dl). The ODA model correctly classified 71.43% of individuals at this dose in the training analysis (Table 2). In the LOO analysis, the sensitivity at this dose remained unchanged, suggesting that the model can predict with relatively strong accuracy, which among newly tested subjects is on the 20 ng/min dose.

In the white cohort (Table 2), the sensitivity for the overall sample ranges between 33.33% at doses of 150 and 300, and 66.67% at a dose of 20 ng/min. This corresponds to an ESS

value of 40.28%, indicating a moderate effect [4]. LOO analysis resulted in a decline in ESS to 27.78%, suggesting moderate generalizability to subjects outside of the training cohort.

In the black cohort (Table 2), the sensitivity for the overall sample ranges from 11.11% at the 400 level dose, to 100% at the 300 level dose. This corresponds to an ESS value of 40.74%, indicating a moderate effect. Similar to the white cohort, LOO analysis in the black cohort also resulted in a decline in ESS (to 29.63%) suggesting moderate generalizability.

When the data are pooled (Table 2), a different pattern in the dose-response relationship emerges. Here, sensitivity ranges from a low of 4.76% at a dose of 60 ng/min, to a high of 71.43% at a dose of 20 ng/min. This corresponds to a moderate ESS of 31.75% -- that is substantially lower than the ESS that was obtained when either cohort was evaluated separately (paradoxical confounding, by definition). As in the analysis for the separate cohorts of white and black observations, LOO analysis ESS declined (26.98%) suggesting moderate generalizability.

All Sidak adjusted *P* values for the ODA pairwise comparisons between doses are presented in Appendix Tables 4-6. In the white cohort, 13 of the 21 pairwise comparisons are statistically significant (Appendix Table 5), whereas in the black cohort, only 8 of the 21 pairwise comparisons are statistically significant (Appendix Table 6). When the data are pooled, 12 of the 21 pairwise comparisons are statistically significant (Appendix Table 4). In total, there were four instances of paradoxical confounding in the ODA analyses. There was a significant effect for the pooled sample for the 20-150 dose comparison, but no effect for either cohort; a significant effect for the white sample, but not for the pooled or black samples, for the 10-20

comparison; and a significant effect for the white and the pooled samples, but not for the black sample, for the 0-20 and 20-400 comparisons.

The final analyses evaluate inter-method agreement across the multiple inferential paired-comparisons tests performed (Table 3). Isomorphic results were obtained for statistical conclusion findings obtained in analysis of data for the pooled sample, and in analysis of data for whites-only. For these samples only one of the eight paradoxical findings for the GEE model (the 20-400 paired-comparison) has the same corresponding paradoxical finding for the ODA model. Comparing GEE (Appendix Tables 1 and 4) and ODA (Appendix Tables 2 and 5), statistical conclusions across the 21 paired-comparisons yields ESS = 60.0, $P \leq 0.43$ (Table 3). Thus there is no evidence of statistically significant agreement (i.e., inter-method reliability [5]) of the statistical conclusions reached by GEE and ODA when aggregated over 21 paired-comparisons for pooled or whites-only data. However, comparing the 21 GEE and ODA paired-comparisons for blacks-only (Appendix Tables 3 and 6) yields ESS = 41.4, $P \leq 0.09$ (Table 3). Thus there is evidence of marginally reliable agreement of statistical conclusions reached by GEE and ODA methods, aggregated over 21 paired-comparisons, for blacks-only data.

These inter-method agreement analyses had low statistical power due to the small number of times that both methods agreed that a multiple comparison wasn't statistically significant (Table 3) [5]. A consistent point of disagreement involved multiple comparisons for which GEE reported $P \leq 0.05$ (accept the alternative hypothesis), and ODA reported $P > 0.05$ (reject the alternative hypothesis). This suggests that, in the present data, ODA (providing exact $P$ values with unquestioned validity [4,5]) produced more conservative estimates of statistical significance

than GEE -- for which the validity of *P* depends upon underlying distributional assumptions being satisfied [8].

## 4. DISCUSSION

Classification is the most popular data mining application in healthcare research, and it has been used to improve diagnostic accuracy, identify high-risk patients, and to extract concepts in unstructured data, for example [29]. The present paper focuses on the use of classification for identifying dose-response relationships. Using a published dataset possessing characteristics exemplifying many dose-response studies (small sample size, limited number of doses, non-linearity in the dose-response relationship, differential effects by subgroup), the present article demonstrates that ODA can perform the usual functions of conventional linear models (i.e., testing for statistical significance of the overall model, *post-hoc* pairwise comparisons between doses), while offering several key advantages over those models.

First, the ODA algorithm, with its associated measure of classification performance (ESS) and non-parametric permutation tests, can be universally applied to any response variable type and number of doses, and is not affected by skewed (non-linear) data, or outliers -- a concern that may arise in the context of meeting assumptions underlying the validity of the estimated *P* value obtained using the conventional approach. In the current example, ODA was consistently more conservative than GEE when comparing between models across the multiple inferential tests performed.

Second, ODA models explicitly, by mathematical formulation, maximize the accuracy of point predictions made at the individual level (this may be compared to the context of the population level that is assumed by conventional parametric statistical methods) for the model that maximizes (weighted) ESS [4]. This feature of ODA has tremendous value for determining whether patients are adherent, or responding, to their prescribed drug regimens (see for example [30,31]). Physicians monitoring the level of a patient's response variable (i.e., blood or urine levels of a metabolite, etc.) can cross reference those values to the cutpoint range generated by ODA to determine the likelihood that the patient is responding to the prescribed dose. The sensitivity of the ODA model observed for each dose provides the physician with the level of predictive accuracy to expect at any particular dose. High model sensitivity would provide the physician with more confidence that the patient is actually on the prescribed dose, than if the model sensitivity was low. Likewise, the sensitivity estimates derived from the LOO analysis allows the physician to further consider the likelihood of adherence to that dose for new patients, existing patients newly prescribed the drug, or patients with somewhat different characteristics than those from the original study population. In the current data, sensitivities for the various doses were of moderate strength (for training and LOO analyses), suggesting that the models provide moderately accurate predictions of dose "assignment" based on the distribution of the response variable.

Third, ODA accepts analytic weights, thereby extending this individual-level assessment of a dose-response relationship to observational studies where weights are used for covariate adjustment [32,33,34,35,36,37]. This feature is particularly valuable, for example, in after-

market drug studies, exposure to environmental hazards, or multivalued interventions in which self-selection is likely to bias the outcome.

ODA clearly offers important advantages and presents greater analytic flexibility than conventional parametric approaches for estimating dose-response relationships. Nevertheless, ODA and conventional approaches are subject to two limiting factors relating to data. First, there is the common limitation of a small sample. Small and highly imbalanced sample sizes offer little statistical power for testing *a priori* hypotheses, and can limit the predictive accuracy of the model -- particularly when applied in cross-generalizability analysis. Second, the constitution of the training sample, and of any independent validation samples, warrants consideration. Combining data from observations representing two or more strata (e.g., 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. Such paradoxical confounding occurs in studies involving multiple subjects and in studies involving single-case longitudinal (time-series) designs. Sometimes two or more groups simply cannot be combined without inducing confounding, due to structural differences among the groups [5].

In summary, this paper introduced a novel approach for modelling dose-response relationships that uses a machine learning algorithm to determine the degree to which doses of the exposure can be distinguished based on the distribution of the response variable. The methodology offers important advantages and presents greater analytic flexibility than conventional parametric approaches, and thus should be considered as an alternative -- if not the preferred approach -- in dose-response applications.

## REFERENCES


1. Peck, C.C., Barr, W.H., Benet, L.Z., Collins, J., Desjardins, R.E., Furst, D.E., Harter, J.G., Levy, G., Ludden, T., Rodman, J.H. & Sanathanan, L. (1992) Opportunities for integration of pharmacokinetics, pharmacodynamics, and toxicokinetics in rational drug development. *Journal of Pharmaceutical Sciences*, 81, 605-610.
2. Royston, P. (2014) A smooth covariate rank transformation for use in regression models with a sigmoid dose-response function. *Stata Journal*, 14, 329-341.
3. Di Veroli, G.Y., Fornari, C., Goldlust, I., Mills, G., Koh, S.B., Bramhall, J.L., Richards, F.M. and Jodrell, D.I. (2015) An automated fitting procedure and software for dose-response curves with multiphasic features. *Scientific reports*, 5, 1-11.
4. Yarnold, P.R., & Soltysik, R.C. (2005) *Optimal data analysis: A Guidebook with Software for Windows* Washington, DC: APA Books.
5. Yarnold, P.R., & Soltysik, R.C. (2016) *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286
6. Linden, A., Yarnold, P. R. (*In Print_B*) Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*
7. Lang, C.C., Stein, C.M., Brown, R.M., Deegan, R., Nelson, R., He, H.B., Wood, M. and Wood, A.J. (1995) Attenuation of Isoproterenol-mediated vasodilatation in blacks. *New England Journal of Medicine*, 333, 155-160.
8. Dupont, W. D. (2009) *Statistical Modeling for Biomedical Researchers.* Cambridge, U.K.: Cambridge University Press.

9. Yarnold, P.R., & Soltysik, R.C. (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences, 22*, 739-752.
10. Carmony, L., Yarnold, P.R., & Naeymi-Rad, F. (1997). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research, 74*, 223-238.
11. Linden, A., Yarnold, P.R. (*In Print_C*) Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*
12. Linden, A., Adams, J., & Roberts, N. (2003) Evaluating disease management program effectiveness: An introduction to time series analysis. *Disease Management,* 6, 243-255.
13. Feinstein, A.R (1988) Statistical significance versus clinical importance. *Quality of Life and Cardiovascular Care, 4*, 99-102.
14. Kraemer, H.C. (1992). *Evaluating medical tests*. Newbury Park, CA: Sage.
15. Baus, J.W., & Gose, E.E. (1995). Leukocyte pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-2*, 513-526.
16. Eisenbeis, R.A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, *32*, 875-900.
17. Nishikawa, K., Kubota, Y., & Ooi, T. (1983). Classification of proteins into groups based on amino acid composition and other characters, II: Grouping into four types. *Journal of Biochemistry*, *94*, 997-1007.
18. Yarnold, P.R. (1992). Statistical analysis for single-case designs. In F.B. Bryant, L. Heath, E. Posavac, J. Edwards, S. Tindale, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Social*

*psychological applications to social issues: Vol. 2. Methodological issues in applied social research* (pp. 177-197). New York, Plenum.

19. Linden, A., & Yarnold, P.R. (*In Print_A*) Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*
20. 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.
21. Witten, I.H., Frank, E., & Hall, M.A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*. San Francisco: Morgan Kaufmann.
22. 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.
23. Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Vol. 1 of Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* 221–233. Berkeley: University of California Press.
24. White, H.L., Jr. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica,* 48, 817–838.
25. Sidak, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association,* 62, 626–633.
26. Yarnold, P.R. (1996) Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement, 56,* 430-442.

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

28. 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.

29. Iavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R., and Geissbuhler, A. (2009) Clinical data mining: a review. *Yearbook of Medical Informatics*, 121-133.

30. Couto, J., Webster, L., Romney, M., Leider, H., Linden, A. (2009) Using an algorithm applied to urine drug screening to assess adherence to an OxyContin regimen. *Journal of Opioid Management*, 5, 359-364.

31. Couto, J., Webster, L., Romney, M., Leider, H., Linden, A. (2011) Use of an algorithm applied to urine drug screening to assess adherence to a hydrocodone regimen. *Journal of Clinical Pharmacology & Therapeutics*, 36, 200-207.

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

33. Linden, A., & Adams, J.L. (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.

34. Linden, A., & Adams, J.L. (2011) Applying a propensity-score based weighting model to interrupted time series data: improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice,* 17, 1231-1238.

35. Linden, A., & Adams, J.L. (2012) Combining the regression-discontinuity design and propensity-score based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice,* 18, 317-325.

36. 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.

37. Linden, A., Uysal, S.D., Ryan, A., & Adams, J.L. (2016) Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine,* 35, 534-552.

Table 1. Forearm blood flow responses to Isoproterenol in normotensive blacks (N=9) and whites (N=12), from [7]. Results are from a generalized estimating equation (GEE) model in which forearm blood flow was regressed on dose, race, and an interaction term of the two.

| | Dose of Isoproterenol (ng/min) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 60 | 150 | 300 | 400 | *P* value |
| All | | | | | | | | |
| Mean | 2.468 | 3.057 | 5.067 | 10.610 | 12.529 | 15.424 | 16.948 | <0.0001 |
| SE | 0.253 | 0.386 | 0.588 | 1.638 | 1.911 | 1.923 | 2.404 | |
| 95% CI | 1.971, 2.964 | 2.300, 3.814 | 3.914, 6.219 | 7.399, 13.821 | 8.784, 16.274 | 11.656, 19.193 | 12.237, 21.659 | |
| White | | | | | | | | |
| Mean | 2.683 | 3.417 | 6.458 | 14.592 | 17.250 | 20.200 | 23.833 | <0.0001 |
| SE | 0.411 | 0.632 | 0.775 | 2.170 | 2.518 | 2.390 | 2.501 | |
| 95% CI | 1.877, 3.488 | 2.178, 4.655 | 4.940, 7.976 | 10.338, 18.845 | 12.315, 22.185 | 15.515, 24.885 | 18.932, 28.734 | |
| Black | | | | | | | | |
| Mean | 2.181 | 2.578 | 3.211 | 5.302 | 6.234 | 9.057 | 7.768 | 0.001 |
| SE | 0.179 | 0.238 | 0.335 | 0.698 | 0.720 | 1.301 | 1.769 | |
| 95% LCI | 1.829, 2.533 | 2.112, 3.043 | 2.555, 3.867 | 3.934, 6.670 | 4.822, 7.647 | 6.506, 11.607 | 4.300, 11.236 | |
| Diff (Black vs White) | -0.501 | -0.839 | -3.247 | -9.289 | -11.016 | -11.143 | -16.066 | |
| SE | 0.448 | 0.675 | 0.844 | 2.280 | 2.619 | 2.722 | 3.063 | |
| 95% CI | -1.380, 0.378 | -2.162, 0.484 | -4.901, -1.593 | -13.758, -4.821 | -16.148, -5.883 | -16.477, -5.809 | -22.069, -10.062 | |
| *P* value | 0.264 | 0.214 | 0.0001 | <0.0001 | <0.0001 | <0.0001 | <0.0001 | |

Table 2. Forearm blood flow responses to Isoproterenol in normotensive blacks (N=9) and whites (N=12), from [7]. Results are from an Optimal Discriminant Analysis (ODA). Values represent 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 | <= 2.53 | > 2.53 & <=2.945 | > 2.945 & <=6.855 | > 6.855 & <=7.355 | > 7.355 & <=9.85 | > 9.85 & <=21.0 | > 21.0 | | |
| Sens-train (%) | 66.67 | 28.57 | 71.43 | 4.76 | 23.81 | 57.14 | 38.10 | 31.75 | <0.001 |
| Sens-LOO (%) | 66.67 | 28.57 | 71.43 | 0.00 | 14.29 | 47.62 | 33.33 | 26.98 | |
| White | | | | | | | | | |
| Cutpoints | <= 2.15 | > 2.15 & <=3.95 | > 3.95 & <=7.3 | > 7.3 & <=12.05 | > 12.05 & <=17.45 | > 17.45 & <=21.2 | > 21.2 | | |
| Sens-train (%) | 50.00 | 58.33 | 66.67 | 41.67 | 33.33 | 33.33 | 58.33 | 40.28 | <0.001 |
| Sens-LOO (%) | 50.00 | 41.67 | 58.33 | 25.00 | 25.00 | 25.00 | 41.67 | 27.78 | |
| Black | | | | | | | | | |
| Cutpoints | <= 2.53 | > 2.53 & <=2.945 | > 2.945 & <=3.68 | > 3.68 & <=4.085 | > 4.085 & <=4.375 | > 4.375 & <=18.95 | > 18.95 | | |
| Sens-train (%) | 77.78 | 44.44 | 44.44 | 33.33 | 33.33 | 100.00 | 11.11 | 40.74 | <0.001 |
| Sens-LOO (%) | 66.67 | 33.33 | 44.44 | 33.33 | 0.00 | 88.89 | 11.11 | 29.63 | |

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

Table 3. Statistical conclusion agreement between GEE and ODA analyses performed for 21 paired-comparisons

| | ODA, $P<0.05$ | ODA, $P>0.05$ | ESS | $P$ Value |
|---|---|---|---|---|
| *Whites, Pooled* | | | 60.0 | 0.43 |
| GEE, $P<0.05$ | 12 | 8 | | |
| GEE, $P>0.05$ | 0 | 1 | | |
| *Blacks* | | | 41.4 | 0.09 |
| GEE, $P<0.05$ | 7 | 6 | | |
| GEE, $P>0.05$ | 1 | 7 | | |

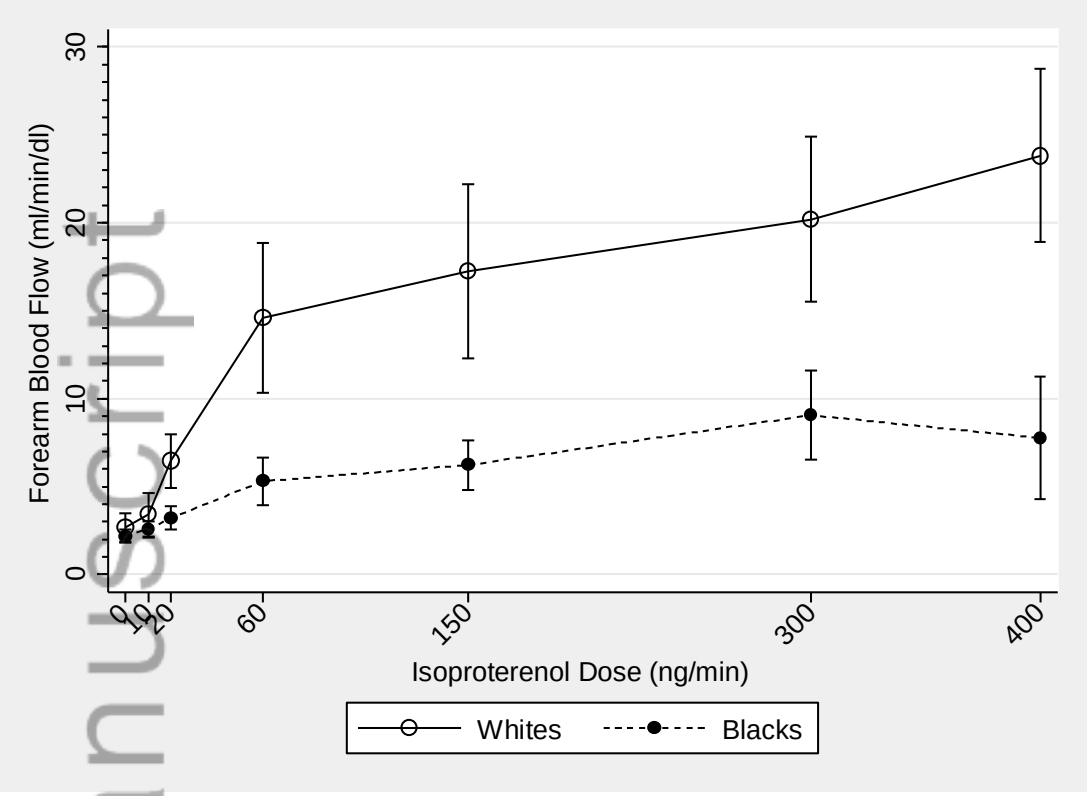



**Figure.** Forearm blood flow responses to Isoproterenol in normotensive blacks (N=9) and whites (N=12), from [7]. Values shown are means and 95% confidence intervals.

# APPENDIX

Table 1: Sidak adjusted *P* values for all pairwise comparisons following GEE for pooled data

| | Dose of Isoproterenol (ng/min) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 60 | 150 | 300 |
| 0 | | | | | | |
| 10 | 0.047 | | | | | |
| 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 | 0.375 |

Table 2: Sidak adjusted *P* values for all pairwise comparisons following GEE for Whites only

| | Dose of Isoproterenol (ng/min) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 60 | 150 | 300 |
| 0 | | | | | | |
| 10 | 0.278 | | | | | |
| 20 | < 0.001 | < 0.001 | | | | |
| 60 | < 0.001 | < 0.001 | < 0.001 | | | |
| 150 | < 0.001 | < 0.001 | < 0.001 | 0.007 | | |
| 300 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | 0.004 | |
| 400 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | 0.006 |

Table 3: Sidak adjusted *P* values for all pairwise comparisons following GEE for Blacks only

| | Dose of Isoproterenol (ng/min) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 60 | 150 | 300 |
| 0 | | | | | | |
| 10 | 0.640 | | | | | |
| 20 | 0.014 | 0.336 | | | | |
| 60 | < 0.001 | 0.001 | 0.027 | | | |
| 150 | < 0.001 | < 0.001 | < 0.001 | 0.161 | | |
| 300 | < 0.001 | < 0.001 | < 0.001 | 0.01 | 0.022 | |
| 400 | 0.028 | 0.092 | 0.178 | 0.966 | 0.999 | 0.896 |

Table 4: Sidak adjusted *P* values for all pairwise comparisons following ODA for pooled data

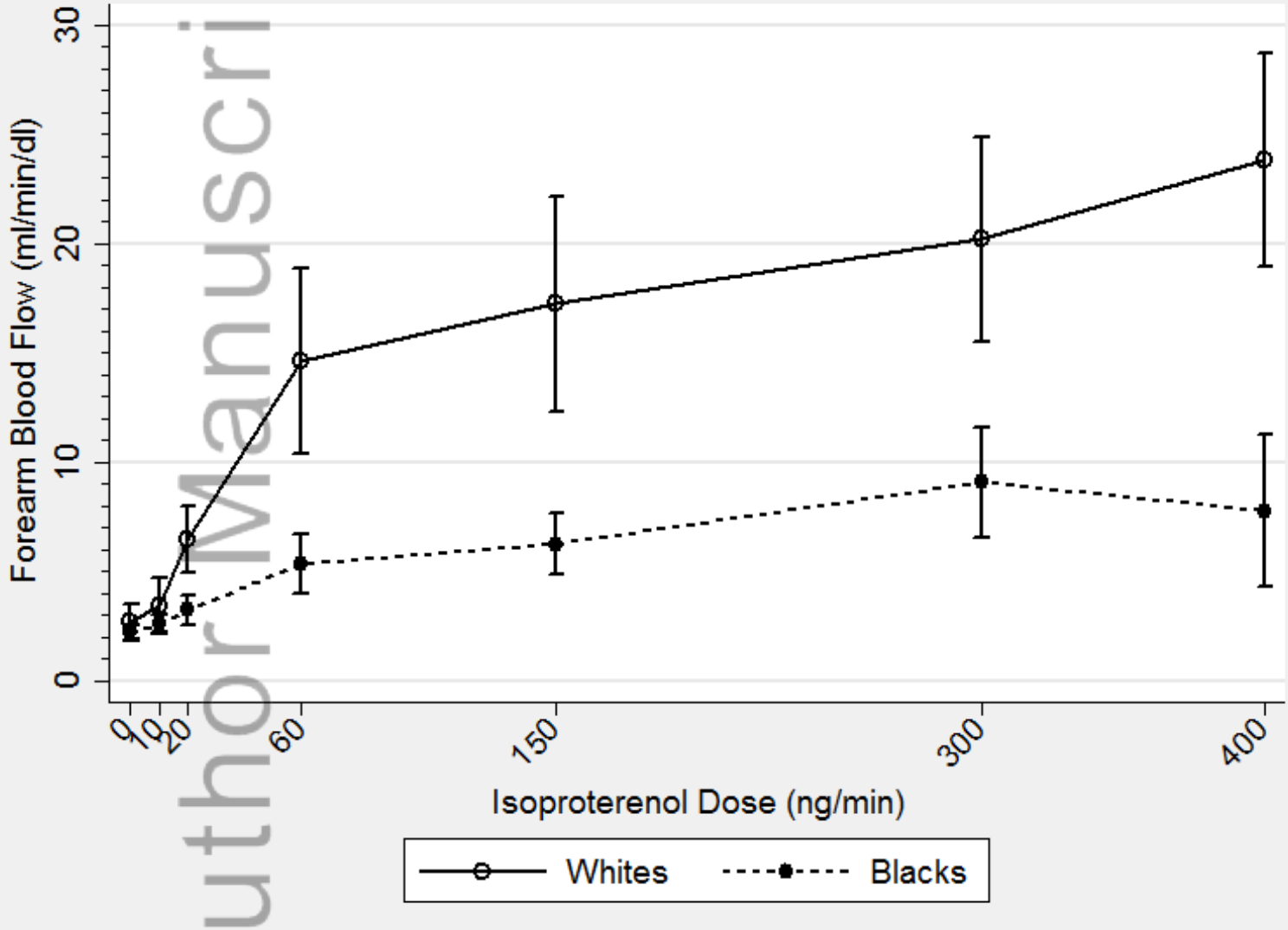
| | Dose of Isoproterenol (ng/min) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 60 | 150 | 300 |
| 0 | | | | | | |
| 10 | 0.969 | | | | | |
| 20 | < 0.001 | 0.057 | | | | |
| 60 | < 0.001 | < 0.001 | 0.167 | | | |
| 150 | < 0.001 | < 0.001 | 0.006 | 1.000 | | |
| 300 | < 0.001 | < 0.001 | < 0.001 | 0.641 | 0.883 | |
| 400 | < 0.001 | < 0.001 | < 0.001 | 0.886 | 0.980 | 1.000 |

Table 5: Sidak adjusted *P* values for all pairwise comparisons following ODA for Whites only

| | Dose of Isoproterenol (ng/min) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 60 | 150 | 300 |
| 0 | | | | | | |
| 10 | 1.000 | | | | | |
| 20 | 0.017 | 0.018 | | | | |
| 60 | < 0.001 | < 0.001 | 0.076 | | | |
| 150 | < 0.001 | < 0.001 | 0.072 | 1.000 | | |
| 300 | < 0.001 | < 0.001 | < 0.001 | 0.942 | 1.000 | |
| 400 | < 0.001 | < 0.001 | < 0.001 | 0.662 | 0.943 | 1.000 |

Table 6: Sidak adjusted *P* values for all pairwise comparisons following ODA for Blacks only

| | Dose of Isoproterenol (ng/min) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 60 | 150 | 300 |
| 0 | | | | | | |
| 10 | 0.964 | | | | | |
| 20 | 0.290 | 0.965 | | | | |
| 60 | 0.006 | 0.052 | 0.308 | | | |
| 150 | < 0.001 | 0.012 | 0.067 | 1.000 | | |
| 300 | < 0.001 | < 0.001 | 0.009 | 0.740 | 0.983 | |
| 400 | 0.006 | < 0.001 | 0.067 | 1.000 | 1.000 | 1.000 |




JEP_12573_F1.tif