Sports Medicine (2020) 50:431–434
https://doi.org/10.1007/s40279-019-01249-9

LETTER TO THE EDITOR

Comment on: "A Method to Stop Analyzing Random Error and Start Analyzing Differential Responders to Exercise"

Matthew S. Tenan[1] · **Andrew D. Vigotsky**[2,3] · **Aaron R. Caldwell**[4]

Published online: 18 December 2019

Dear Editor,

We read with great interest the Current Opinion Dankel and Loenneke [1]—a paper that introduces an analytical approach (herein, the Dankel–Loenneke (DL) method) to classifying "differential responders" in exercise science studies. We applaud the authors' encouragement of exercise scientists to include a control in addition to an experimental group (i.e., parallel groups design). However, the DL method itself has unintended, undesirable statistical properties. Long-standing critiques of differential responder analyses aside [2, 3], the focus of the current letter is on the error rates of the DL method. Here, we demonstrate how the DL method performs poorly, including error rates far above 5%.

1 Simulations

Reference [1] describes a trichotomous discretization of continuous responses to bin participants into groups (i.e., "low," "average," and "high" responders), with the purpose of using these groups for subsequent analyses. For categorizing participants into these groups, Ref. [1] asserts that, using their approach, "approximately 5% of the total sample […] will be incorrectly classified as differential responders," but they provided neither proofs nor simulations to support this assertion. We systematically evaluated this assertion via simulation and mathematical derivation. We simulated studies that closely resemble the properties of the example studies that [1] presents, with constant Gaussian random error that is orthogonal to true effect magnitude and variance. Our simulations show that the DL method is not robust to random error and does not have constant error rates as the authors describe (Fig. 1). What is more, the accuracy of the DL method is dependent upon sample size and the relationship between true effect variance and random error. Even in the best of circumstances—in which sample sizes are large and error is homogeneous, independent of effect magnitude, and equal on the individual and aggregate levels—the DL method is capable of miscategorizing at a rate greater than the claimed 5%. Our mathematical evaluation of the error rates is in agreement with these simulation results, and further, they provide a mathematical rationale as to why the DL approach fails to maintain the claimed error rates (see DL_TypeI_Error_Rate_Math.pdf).

Next, we employed a widely published model for indirect calorimetry minute ventilation (VE), which incorporates the nonlinear differential measurement error inherent in many electronic measuring devices used in exercise science/sports medicine, to assess the performance of the DL method [4–7]. The code used to create this simulation is available (see Differror_VE_LoennekeMethod.pdf) and results can be seen in Fig. 2. Even when there is no heterogeneous effect of the intervention, the method may have a statistically significant Levene's test and incorrectly categorize participants as differential responders. Because there is no true response heterogeneity, the product of the rate of misclassification and Levene's test power (Fig. 2, top and middle, respectively) can be used to obtain the total probability of misclassification (Fig. 2, bottom).

Details of our math and simulations can be found on OSF: https://osf.io/2r5ev/.

This comment refers to the article available at https://doi.org/10.1007/s40279-019-01147-0.

✉ Andrew D. Vigotsky
avigotsky@gmail.com

[1] Optimum Performance Analytics Associates, Apex, NC, USA

[2] Department of Biomedical Engineering, Northwestern University, Evanston, IL, USA

[3] Department of Statistics, Northwestern University, Evanston, IL, USA

[4] Exercise Science Research Center, University of Arkansas-Fayetteville, Fayetteville, AR, USA

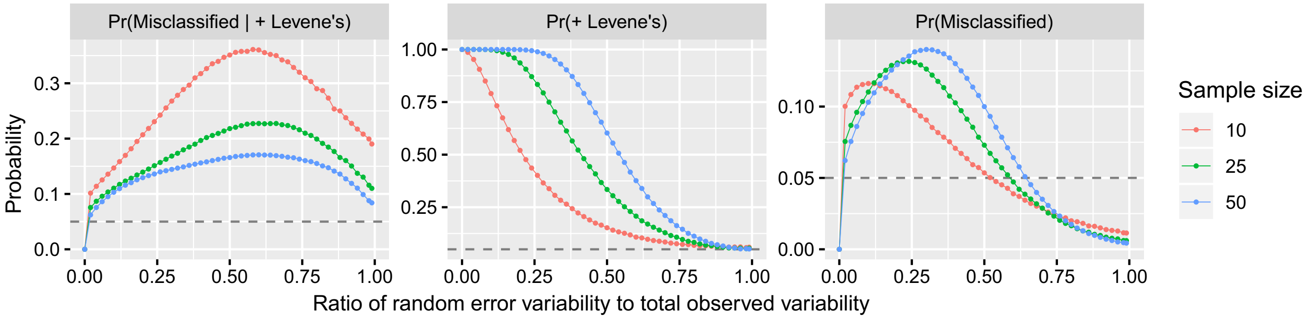

Fig. 1 Probability of classifying an "average responder" as a "differential responder" using the Dankel–Loenneke (DL) method under constant Gaussian error. 1,000,000 simulations were run for groups with $n = \{10, 25, 50\}$ for a range of variance ratios ($\sigma_\epsilon^2/(\sigma_\epsilon^2 + \tau^2)$, where σ_ϵ^2 is the variance of random error, τ^2 is the variance of the treatment effect, and their sum is the observed variance in the experimental group). A variance ratio of 0 indicates no random error (pure treatment heterogeneity), while a variance ratio of 1 indicates pure random error (no treatment heterogeneity). Each participant's true score (not including random error) and observed score (including random error) were compared to the thresholds for classification as determined by the DL method. A misclassification was noted for any "average responder" whose observed score fell into "low" or "high" responder categories—this difference is strictly due to random error, including constant Gaussian biological variability and measurement error. Importantly, unlike traditional false positive rates, and in accordance with DL, the total sample size was used as the denominator rather than the total number of average responders; this leads to a *lower* error rate than a traditional false positive rate. Left panel: the probability of misclassifying an average responder as a differential responder, given a positive Levene's test. When Levene's test is positive, the DL fails to maintain a 5% misclassification rate for responder classification. Because Levene's test is serving as a filter, smaller sample sizes perform more poorly because they are noisier. Center panel: the probability of a positive Levene's test. As the ratio approaches 0, the variance of the treatment effect dominates the variance of random error, increasing the probability of a positive Levene's test. (Right panel) Total probability of being misclassified. This is the product of the left and center panels and thus takes into account Levene's test. Even when using Levene's test as a filter, the misclassification rate is unstable and is a function of sample size—greater error rates with more data—and the magnitudes of the treatment and error variances. Dashed grey lines indicate $P = 0.05$

2 Discussion

We have presented evidence that the DL method is prone to error rates well beyond the claimed 5% and is exacerbated when measurement error is not constant. In addition to our statistical concerns about the DL method, we wish to note that more general concerns about differential responder analyses are discussed extensively in the applied statistics literature [2, 8, 9], and more generally, their usefulness and philosophical grounding have been called into question [3].

Our simulations clearly demonstrate that this method fails in its goal to categorize response magnitude, and in doing so, has unacceptably high error rates. The interested reader is strongly encouraged to explore the established statistics literature when designing studies where a "responder analysis" is desired [2, 8, 10]. In such cases, researchers should focus on the subject-by-treatment interaction, and as Ref. [1] suggest, this may not always be possible to calculate without a crossover replicate design [9]. If subsequent analyses are of interest, we suggest that continuous errors-in-variables models are more efficient and have been properly vetted [11]. Finally, as a general practice, we, like others [2, 3], advocate for researchers to avoid "classifying" participants as responders or non-responders, and instead, identify theoretical justifications for heterogeneous response magnitudes.

Unless Dankel and Loenneke provide clear and unambiguous mathematical proofs and reproducible data simulations substantiating their claimed error rates, the incorrectly claimed error rates constitute an "honest error" where we have provided "clear evidence that the findings are unreliable", as per the Committee on Publication Ethics (COPE) guidelines [12]. However, based on our proofs and simulations, this seems impossible. Our field should no longer accept statistically sounding rationale for "novel statistical methods" when mathematical proofs are the gold standard in statistics journals.

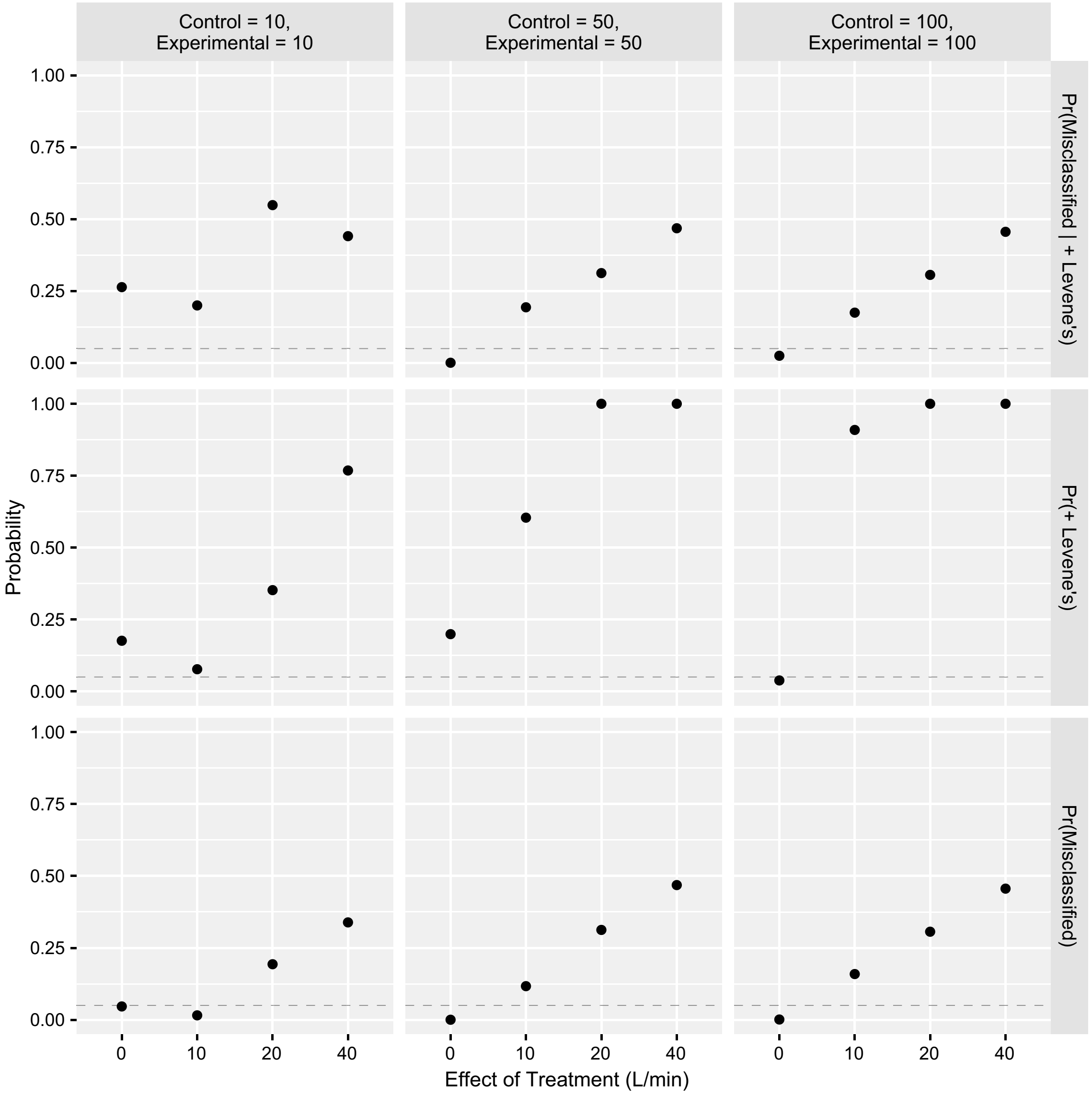

Fig. 2 Error probabilities associated with nonlinear error structure from simulations using minute ventilation. (First row) Within each simulation condition, the proportion of participants that were categorized as differential responders when a statistically significant Levene's test was calculated. (Second row) The probability of a positive Levene's test. (Third row) The total probability of misclassification, taking into account both the error of Levene's test and the error rate when Levene's test is positive. Note that this is the product of the first and second rows. These simulations are based on the Crouter and Tenan model for nonlinear differential measurement error of day-to-day variability in VE [5, 7]; this model is used to draw a participant's measured VE on trial 1 and trial 2 of a simulated study with a control arm and an experimental arm. The benefit of this model is that it enables us to easily simulate what VE a participant may have when no change occurs or when some magnitude of change occurs as a result of the intervention (i.e., there are no true "differential responders;" all participants have the exact same factual response to the intervention with simply the noise added for the nonlinear differential measurement error across days). Various sample sizes, equal and unequal, and intervention response magnitudes were simulated 1000 times with initial "true VE" measures randomly sampled between 50 and 70 L/min to obtain the above results. There is no stable pattern for the inaccuracies in their method with the slight exception that, in the case of VE, an increase in the effect of the intervention increases the probability of falsely identifying these "differential responders". Dashed grey lines indicate $P = 0.05$

Acknowledgements We would like to acknowledge Kristin Sainani and Blake McShane for their thoughtful comments.

Compliance with Ethical Standards

Funding No financial support was received for the conduct of this study or preparation of this manuscript.

Conflict of interest Matthew Tenan, Andrew Vigotsky, and Aaron Caldwell declare that they have no conflict of interest.

References

1. Dankel SJ, Loenneke JP. A method to stop analyzing random error and start analyzing differential responders to exercise. Sports Med. 2019. https://doi.org/10.1007/s40279-019-01147-0.
2. Snapinn Steven M, Jiang Qi. Responder analyses and the assessment of a clinically relevant treatment effect. Trials. 2007;8(1):31.
3. Senn Stephen. Individual response to treatment: is it a valid assumption? BMJ. 2004;329(7472):966–8.
4. Crouter Scott E, Antczak Amanda, Hudak Jonathan R, DellaValle Diane M, Haas Jere D. Accuracy and reliability of the parvomedics trueone 2400 and medgraphics vo2000 metabolic systems. Eur J Appl Physiol. 2006;98(2):139–51.
5. Tenan Matthew S. A statistical method and tool to account for indirect calorimetry differential measurement error in a single-subject analysis. Front Physiol. 2016;7:172.
6. Barnes K, Kilding A, Blagrove R, Howatson G, Boone Jan, Bourgois Jan, Fletcher J, Macintosh B, Gonzalez-Mohino F. Commentaries on viewpoint: use aerobic energy expenditure instead of oxygen uptake to quantify exercise intensity and predict endurance performance. J Appl Physiol. 2018;125(2):676–82.
7. Tenan Matthew S, Bohannon Addison W, Macfarlane Duncan J, Crouter Scott E. Determining day-to-day human variation in indirect calorimetry using Bayesian decision theory. Exp Physiol. 2018;103(12):1579–85.
8. Uryniak Tom, Chan Ivan SF, Fedorov Valerii V, Jiang Qi, Oppenheimer Leonard, Snapinn Steven M, Teng Chi-Hse, Zhang John. Responder analyses—a PhRMA position paper. Stat Biopharm Res. 2011;3(3):476–87.
9. Hecksteden Anne, Kraushaar Jochen, Scharhag-Rosenberger Friederike, Theisen Daniel, Senn Stephen, Meyer Tim. Individual response to exercise training—a statistical perspective. J Appl Physiol. 2015;118(12):1450–9.
10. Senn Stephen, Rolfe Katie, Julious Steven A. Investigating variability in patient response to treatment—a case study from a replicate cross-over study. Stat Methods Med Res. 2011;20(6):657–66.
11. Fuller Wayne A. Measurement error models, vol. 305. New York: Wiley; 2009.
12. Wager E, Barbour V, Yentis S, Kleinert S, on behalf of COPE Council. Guidelines for retracting articles. version 1. 2009. https://doi.org/10.24318/cope.2019.1.4.