# Directions for Exercise Treatment Response Heterogeneity and Individual Response Research

Authors

**Travis J. Hrubeniuk[1, 2], Jacob T. Bonafiglia[3], Danielle R. Bouchard[2, 4], Brendon J. Gurd[3], Martin Sénéchal[2, 4]**

Affiliations

1 Interdisciplinary Studies, University of New Brunswick, Fredericton, Canada
2 Cardiometabolic Exercise and Lifestyle Laboratory, University of New Brunswick, Fredericton, Canada
3 School of Kinesiology and Health Studies, Queen's University, Kingston ON, Canada
4 Faculty of Kinesiology, University of New Brunswick, Fredericton, Canada

Key words

individual response, exercise physiology, physical activity, non-responder

**accepted** 30.06.2021
**published online** 16.08.2021

Bibliography

Int J Sports Med 2022; 43: 11–22
**DOI** 10.1055/a-1548-7026
**ISSN** 0172-4622
© 2021. Thieme. All rights reserved.
Georg Thieme Verlag, Rüdigerstraße 14,
70469 Stuttgart, Germany

Correspondence

Martin Sénéchal
Kinesiology, University of New Brunswick,
90 Mackay Drive,
Fredericton
E3B 5A3,
Canada
Tel .: 506451-6889, Fax: (506) 453-3511
msenecha@unb.ca

ABSTRACT

Treatment response heterogeneity and individual responses following exercise training are topics of interest for personalized medicine. Proposed methods to determine the contribution of exercise to the magnitude of treatment response heterogeneity and categorizing participants have expanded and evolved. Setting clear research objectives and having a comprehensive understanding of the strengths and weaknesses of the available methods are vital to ensure the correct study design and analytical approach are used. Doing so will ensure contributions to the field are conducted as rigorously as possible. Nonetheless, concerns have emerged regarding the ability to truly isolate the impact of exercise training, and the nature of individual responses in relation to mean group changes. The purpose of this review is threefold. First, the strengths and limitations associated with current methods for quantifying the contribution of exercise to observed treatment response heterogeneity will be discussed. Second, current methods used to categorize participants based on their response to exercise will be outlined, as well as proposed mechanisms for factors that contribute to response variation. Finally, this review will provide an overview of some current issues at the forefront of individual response research.

## Introduction

Health organizations provide exercise recommendations aimed at reducing chronic disease and premature mortality [1–3]. While regular exercise is generally associated with health benefits, studies dating back to the 1980s report the magnitude of change in a given outcome varies widely between individuals, to the degree that some are seemingly unable to garner the expected benefits [4, 5]. The seminal HEalth, RIsk factors, exercise Training And Genetics (HERITAGE) Family Study was among the first to demonstrate heterogeneity in the observed responses following a standardized aerobic exercise training intervention [6, 7]. A recent review by Williamson and colleagues [8] critiqued these findings mainly because the HERITAGE Family Study lacked a control group and referred to short-term reliability – a limitation discussed in more detail in the next section. Although many studies following HERITAGE also lacked control groups [8], the number of reports investigating heterogeneity in observed responses has constantly grown [5, 9–17].

A by-product of investigating exercise treatment response heterogeneity has been a more clinically applicable focus on the individuals who struggle to experience the anticipated benefits, typically referred to as 'non-responders'. Accurate interpretation of an individual's ability to experience the targeted benefits – or not – from a prescribed exercise program would allow clinicians to make an informed decision and adapt the exercise program parameters or implementing a different treatment regimen. Moreover, identifying the mechanisms contributing to an individual's ability to benefit from an exercise program may provide insight to the true effectiveness of exercise and allow for personalized exercise prescriptions, moving beyond the standard recommendation of achieving current physical activity guidelines in a clinical setting [18, 19].

It is important to recognize that investigating the influence of exercise training on treatment response heterogeneity, and determining if a participant responded to an intervention (or not) are two distinct analytical approaches, designed to answer different questions. Conflating the two may lead to inappropriate interpretation of results. Common errors include interpreting improvements in the group mean as an exercise intervention increasing the 'response rate', or associating higher response rates with reduced treatment response heterogeneity [20, 21]. Moreover, as the desire to understand exercise treatment response heterogeneity and individual response has grown, the number of proposed methods for conducting such research has significantly inflated. Accordingly, there is an ongoing debate on the most appropriate, rigorous and feasible techniques [8, 19, 20, 22–28]. Various questions and concerns have emerged, including the ability to truly isolate the impact of exercise training, the nature of individual responses in relation to mean group changes, and the utility of individual response research [20, 29–32].

The purpose of this review is threefold. First, the strengths and limitations associated with current methods for quantifying the contribution of exercise to observed treatment response heterogeneity will be discussed. Second, current methods used to categorize participants based on their response to exercise will be outlined, as well as proposed mechanisms to identify factors that contribute to response variation. Finally, this review will provide an overview of some current issues at the forefront of individual response research.

## Sources of Variation Contributing to an Individual's Observed Response

Prior to discussing exercise treatment response heterogeneity or individual response categorizations, it is important to understand that individual observed changes following an exercise intervention are the product of some combination of random variation, within-subject variation, and the subject-by-training interaction. As these sources of variation have been explained in depth previously, below we provide a succinct overview [19, 22, 23, 25, 28, 33, 34]. It should be noted the terminology used in this review – as defined below – is inconsistently applied across the literature, with alternative terminology outlined by Bonafiglia et al. [23].

### Random variation

Random variation is comprised of the error introduced by the measurement instrument (technical error) and day-to-day biological variation. When multiple measures are taken over a period of time within which the true value is not expected to change, the noise introduced by random variation will result in a cluster of observed scores that are normally distributed around the true value. Using the average value of multiple measurements for analysis can reduce the influence of random variation [33]. Likewise, random variation can be estimated by collecting repeated measurements on a group of individuals and calculating the typical error of the measurement ($TE_M$) [26, 28, 33].

### Within-subject variation

Within-subject variation is inconsistent variance introduced by changes in the environment or behaviour unrelated to the intervention (e. g., short-term changes in eating patterns, seasonal changes influencing behaviour or mood) [8]. Theoretically, if the same individual were provided the same intervention at a different time, within-subject variation would be responsible for much of the variance in the observed difference in the change scores, if random variation was removed. The impact of within-subject variation on a participant change score may be dependent on the duration of the intervention, with longer duration trials expected to increase its influence [23].

### True responses to exercise training

The true response to exercise represents the 'trainability' of an individual, or the consistent, repeatable changes experienced in relation to the provided intervention. Although genetic endowment may contribute to an individual's trainability, stable characteristics or traits such as lived experiences, lifestyle habits (e.g., exercise, diet), and epigenetic modifications may also influence the ability of an individual to respond to the training stimulus [11, 25, 26]. The subject-by-training interaction refers to the degree to which true responses to training (i.e. trainability) differs across a group of participants.

## Methods for Quantifying Exercise Treatment Response Heterogeneity

Exercise treatment response heterogeneity exists when the true training-induced changes experienced across a sample of participants differ to a degree that can be considered meaningful [20, 23]. Quantifying exercise treatment response heterogeneity requires the subject-by-training interactions to be isolated from within-subject and random variation [25, 34, 35], necessitating a crossover trial with multiple intervention and control periods, each separated by an adequate washout [35]. The true value of the subject-by-training interaction can then be calculated using a linear mixed model approach [25]. While this study design may be possible in certain research areas [36], conducting such a trial using an exercise intervention presents several limitations: high operating costs, heavy resource requirements, significant time investments, and challenges with participant recruitment and compliance. Moreover, the carry-over effects of an initial training intervention are not

▶ **Table 1** Proposed methods for assessing variation†.

| Method | Statistical Approach | Outcome | Potential Limitations |
|---|---|---|---|
| Repeated measurements at individual timepoint | $TE_M$ ($TE_M = SD_{diff} / \sqrt{2}$) | Estimate of random variation | Does not account for within-subject variation |
| Replication of a cross-over trial | Linear mixed model Fixed Effect: Training Group Random Effect: Subject Identity | Determines individual subject-by-training interaction | Feasibility Unknown wash-out duration |
| Comparator Group: Control Group | $SD_{IR} = \sqrt{(SD_{EXP}^2 - SD_{CON}^2)}$ | Standard deviation of individual responses; Group-based estimate of the variance across participants' subject-by-training interaction | Potential differences in variation between the control and training groups |
| | $TE_\Delta$ ($TE_\Delta = SD_{diff} / \sqrt{2}$) | Estimate of within-subject variation. | May also contain random variation |
| Comparator Group: Reliability Data | $SD_{IR} = \sqrt{(SD_{EXP}^2 - [\sqrt{2} * TE_M])}$ or<br>$SD_{IR} = \sqrt{(SD_{EXP}^2 - [\sqrt{2} * CV])}$ | Standard deviation of individual responses; Group-based estimate of the variance across participants' subject-by-training interaction | Transferability of sample used to calculate the reliability data Duration of previous trial |
| Repeated measurements throughout the intervention | Linear mixed model Fixed effect: measurement number Random effect: subject ID-by-measurement number interaction Dependent variable: measured value | Individual estimates of subject-by-training interaction | Increased demand on participants and resources Potential influence of accumulating tests |

†Interpreted based on Table 3 from Hecksteden et al., 2015 and Table 2 from Ross et al., 2019. $TE_M$ = typical error of a measurement; $SD_{diff}$ = standard deviation of the difference score; $SD_{IR}$ = standard deviation of individual responses; $SD_{EXP}$ = standard deviation of the change scores from the experimental group; $SD_{CON}$ = standard deviation of the change scores from the experimental group; $TE_\Delta$ = typical error of a change score; CV = coefficient of variation.

well enough understood to accurately suggest an adequate washout period between phases [8].

Alternatives for estimating exercise treatment response heterogeneity have been proposed, each of which must complete two steps to increase accuracy. First, the observed heterogeneity among participant change scores must be shown to be attributable to exercise training *per se*, and therefore not primarily a product of random or within-subject variation. Second, the magnitude of heterogeneity throughout the sample should be contextualized in relation to the chosen outcome measure [27]. Contextualizing the magnitude of the heterogeneity helps to determine if the variance introduced by exercise is meaningful in relation to the anticipated variation associated with the measurement technique. ▶ **Table 1** outlines the various statistical methods proposed for estimating treatment response heterogeneity, the subject-by-training interaction, random variation, and within-subject variation. The following sub-sections will provide additional detail for some of the notable methods.

## Control group designs: $SD_{IR}$ method

If exercise training meaningfully contributes to treatment response heterogeneity, exercising participants will display greater variation in their pre-post difference scores than a non-exercising control group [27]. As the control group did not participate in the intervention, it is assumed variance among the difference scores is a product of random and within-subject variation. Conversely, the intervention change scores also include variance introduced by individual responses to training (i.e. subject-by-training interactions). Therefore, the magnitude of variation introduced by participating in the exercise program can be estimated by subtracting the variance of the control group change scores, from the variance of the intervention group changes [23, 27, 28]. The resulting value is referred to as the standard deviation of individual responses ($SD_{IR}$), which estimates the influence of individual responses to exercise training on overall response heterogeneity, accounting for the influence of random and within-subject variation [8, 22, 23, 27]. If the standard deviation of the changes in the intervention group are not substantially larger than the control group, theoretically the exercise training has not contributed to the observed variance [8, 22, 27].

The $SD_{IR}$ method relies on the assumption that larger variance among the change scores in the experimental group, when compared to controls, is sufficient to estimate the magnitude of treatment response heterogeneity attributable to exercise training. To improve the accuracy of the $SD_{IR}$, steps should be taken to improve estimates of within-subject and random variation. Ensuring the control group follows the same time interval as the intervention group [8], and taking multiple measurements at every timepoint (for both the control and intervention groups) can reduce these influences [33]. Simply stated, researchers should follow proper randomized control trial practices to ascertain the impacts of participating in the provided intervention [20].

A number of additional assumptions and limitations require consideration when using the $SD_{IR}$. Most notably, it is important to consider that the $SD_{IR}$ relies on the assumption that the combined effect of random and within-subject variation is equal between the intervention and control groups [23]. Even with randomized allocation to control and intervention groups, an inability to calculate within-subject variation across each group leaves the potential for its influence to differ. Accordingly, this assumption should be reported as a limitation whenever the $SD_{IR}$ is used.

Despite a necessary reliance on the assumptions associated with implementing randomized control trials in exercise science [23], the $SD_{IR}$ remains the preferred method for estimating the influence of exercise training on observed treatment response heterogeneity. This is in large part due to the $SD_{IR}$ being, to our knowledge, the only method able to estimate the magnitude of confounding sources of variation in parallel-arm randomized control trials [17, 24]. Nonetheless, it is important to recognize there are many statistical methods designed to compare the equality of variance across groups. For example, Leifer et al. [37] compared variability in observed responses between exercise and control groups using Levene's test. We find the $SD_{IR}$ to be preferred over such tests as it estimates the magnitude of treatment response heterogeneity in units of the measured outcomes rather than relying on the interpretations of a *p*-value. However, to maximize confidence in the utility of the $SD_{IR}$, potential threats to major assumptions need to be avoided, and if present, reported.

### Using reliability data

If a control group is not available, using data from a relevant reliability study (i.e., test, re-test) has been proposed as an alternative [25–27]. When using reliability data, it is suggested the $SD_{IR}$ can be estimated by subtracting the variance of test, re-test change scores from the variance in change scores following exercise training. However, 'replacing' a control group should not be done without concern, as the limitations when using reliability data to calculate the $SD_{IR}$ far outweigh the benefits. Data from a relevant reliability trial will solely allow for an estimate of random variation, meaning within-subject variation *remains unaccounted for*. Accordingly, the estimate of what would have happened to intervention participants had they not participated in the intervention is no longer valid; meaning the $SD_{IR}$ cannot distinguish variability in true changes attributable to exercise training from the changes resulting from behavioural and/or environmental factors. Additionally, differences between the intervention participants and the reliability group can significantly impact the results. As such, we recommend against using reliability data to calculate the $SD_{IR}$.

### Repeated measures design

Hecksteden et al. [25] introduced a longitudinal approach using the collection of repeated measurements throughout the duration of an intervention as an alternative to calculating the $SD_{IR}$. The concept has been subsequently demonstrated on two occasions [26, 38]. When using the repeated measurement design, alternative tools for detecting exercise treatment response heterogeneity become available. Namely, true response estimates are derived from the slope of each individual's regression line of the measured values throughout the duration of the intervention. As opposed to calculating the $SD_{IR}$ based on the variance among the change scores from the intervention and control groups or reliability trials, taking repeated measures allows for the $SD_{IR}$ to be estimated by calculating the between-subject standard deviation of individual slopes or by using a linear mixed model. Importantly, and in line with any single-group design, this method cannot account for the counterfactual, and therefore cannot discern whether the variance introduced was the product of exercise training or behavioural/environmental changes that occurred during the intervention, *per se*. Moreover, this method assumes that responses generated over time will be linear, which may not be the case for certain physiological variables (*e.g.* cardiorespiratory fitness) [38]. Frequent testing may also introduce learning effects, carry-over effects or performance biases, masking results. Therefore, using a control group to calculate the $SD_{IR}$ remains the preferred method.

### Contextualizing the magnitude of exercise treatment response heterogeneity

Once the $SD_{IR}$ has been calculated, its magnitude should be contextualized to determine if the variance introduced by exercise is meaningful. This can be done via standardization, or by comparing the $SD_{IR}$ to a predetermined threshold value. Standardization consists of dividing the $SD_{IR}$ by the standard deviation of all subjects at baseline, and comparing them to threshold values of 0.1, 0.3, 0.6, 1.0, and 2.0 (representing small, moderate, large, very large, and extremely large effects) [27, 39]. Alternatively, the $SD_{IR}$ can be viewed in relation to a predetermined threshold such as the minimal clinically important difference, or the smallest worthwhile difference [28, 33].

### Summary

Understanding the contribution of exercise to the observed heterogeneity following an intervention can aid researchers in interpreting the effects of exercise. Currently, there are a number of limitations restricting the ability to accurately quantify exercise treatment response heterogeneity. While multiple methods have been proposed to provide estimates of the influence of exercise training on the observed variance, the $SD_{IR}$ remains the preferred metric. It is important to emphasize that a negative $SD_{IR}$, or a suggestion that treatment response heterogeneity was not only a product of exercise training; does not preclude the possibility of responders and non-responders being identified [16]. A negative $SD_{IR}$ could suggest the observed heterogeneity was largely influenced by other factors than participation in the prescribed exercise training. Alternatively, negative $SD_{IR}$ values may reflect sampling errors in estimates of observed variability due to small sample sizes, which remains a major challenge for exercise training studies aiming to investigate treatment response heterogeneity.

## Categorization of Participants as Responders or Non-responders

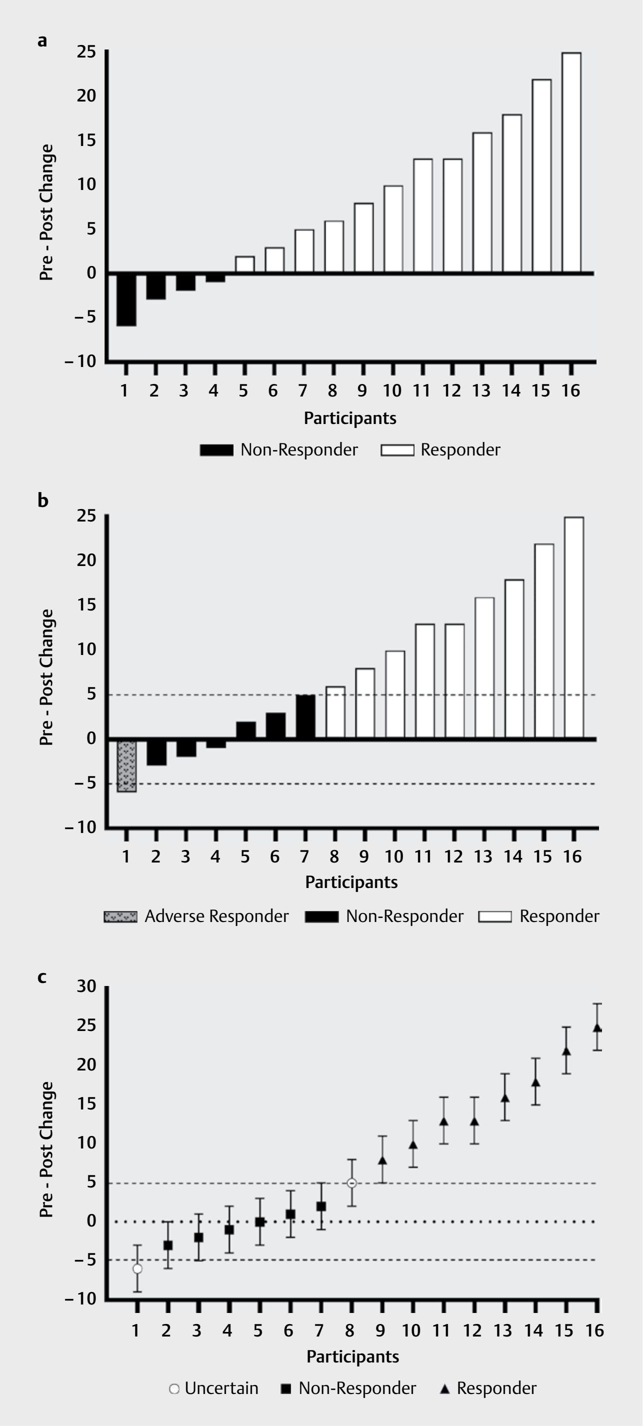
Categorizing participants as responders or non-responders speaks to the ability of each individual to improve beyond a threshold value for a selected outcome, as a result of participating in a specific exercise intervention. Such a categorization requires a consideration and comparison to a control condition. It is important to clearly define what being categorized as a 'responder' means for the current investigation, prior to conducting an analysis. Categorization may occur in a variety of ways, each requiring a response threshold to be selected, and a method to account for the variance limiting the ability to directly quantify an individual's true response [26]. The method used to categorize individuals should be carefully selected to answer the research question and be constructed in a way that accounts for the desired sources of variance. Categorizations may be made based on estimates of random variation (see below), and it can be argued these participants did experience benefit. However, without comparison to a control group the observed changes cannot be assigned to the exercise intervention, *per se*. Accordingly, we suggest participants only be categorized as responders/non-responders when a control condition is part of the design and considered when making the categorization. If a control group is not considered, participants can only – at best – be categorized as experiencing a benefit, or not experiencing a benefit.

Various types of response thresholds have been used throughout the literature, including: (1) zero change as a fixed value, (2) the upper limit of observed differences expected as a result of variation, and (3) the lower limit of clinically or practically meaningful differences. Each threshold type differs in if, and to what degree, a method of accounting for the variance masking an individual's true change score is intrinsically applied. If an estimate of the potential influence of variance is not built into the threshold, additional steps should be taken to ensure the accuracy of a response estimate prior to the response categorization. Moreover, a threshold which is too permissive to answer the research question may result in inaccurate response categorizations, and a threshold which is too conservative may overstate the non-response rate [26]. Regardless of the method used, the threshold will always remain arbitrary and debated [40]. The following section will outline the thresholds used to categorize participants, and the methods which may be applied to account for the variance limiting the accuracy of an individual response estimate.

## Zero change as a fixed value

Here the response threshold is set at zero change, defining non-response as a difference from pre- to post-intervention as zero or less (▶ **Fig. 1a**), [11, 41]. While straightforward, using zero change as a threshold does not account for the limited accuracy of a response estimate, allowing random and/or within-subject variation to bias the categorization. To improve the accuracy of the categorization, an individual confidence interval for each participant's change score can be calculated using an intervention-specific estimate of error [24, 28]. Individuals can then be categorized as successful or not based on whether the confidence interval for the true change lays across or beyond the set threshold. Using a zero-based threshold, individuals can be categorized as having a positive response, a negative response, or an uncertain response, dependent on the interval laying entirely above, below, or crossing zero, respectively.

The interpretation of said categorization will be highly dependent upon the estimate of error used to calculate the confidence interval. Error estimates based on repeat baseline tests ($TE_M$) can be used if the goal is to categorize participants based on changes beyond an estimate of random variation. Accordingly, these participants should be referred to as having 'experienced benefit' or not. Conversely, if the goal is to more confidently categorize responders based on changes beyond both random variation and within-subject variation – thereby increasing the likelihood that observed changed resulted from participation in the provided exercise intervention – a time-matched control group should be used to estimate the combined influence of random and within-subject variation by calculating the typical error of the change score ($TE_\Delta$) [8, 26, 28], and using it to estimate the influence of error. Subsequently, participants can be categorized as responders or non-responders to the intervention.



▶ **Fig. 1** Participants categorized based on **(a)** zero change as a fixed value, **(b)** estimates of variation or clinical relevance, and **(c)** individual confidence intervals.

Swinton et al. [28] provides adjusted multiples that can be used to calculate the width of the confidence interval. The width of the confidence interval is at the discretion of the user, understanding that larger widths increase the risk of making type 2 errors (incorrectly categorizing an individual as uncertain when they are likely to be responders or non-responders), whereas smaller widths increase the risk of making type 1 errors (incorrectly categorizing as a responder or non-responder).

### Thresholds based on estimates of variation

The aim when using an estimate of variation is to set the threshold at the highest possible change that may occur as a result of extraneous variation. Similar to using confidence intervals, the interpretation of response categorizations will depend upon the estimate of error used to set the response threshold (▶ **Fig. 1b**) [42–46].

The first step when using an estimate of variation to set a response threshold is to estimate the influence of variation on the observed changes. Therefore, a decision needs to be made regarding what sources of variation will be accounted for by the threshold value. Again, the $TE_M$ can be used to estimate the influence of random variation on a single measure. However, a change score is composed of two independent measurements, each subject to random variation. To account for this, the estimate should be multiplied by a factor greater than one prior to setting the response threshold, to increase the level of confidence in subsequent categorizations [26, 28]. Again. it must be emphasized that individuals categorized using this method only experienced a change beyond an estimate of random variation. This change cannot be attributed to participation in exercise (or any other source) *per se*; all that is known is a change greater than what was expected to occur due to random error has occurred, and categorized as having experienced benefit, or not. Alternatively, a time-matched control group can be used to estimate the combined influence of random and within-subject variation by calculating the typical error of the change score ($TE_\Delta$) [8, 26, 28]. By replicating the measurement schedule of the intervention, the $TE_\Delta$ estimates the variance that may be introduced due to within-subject variation in the absence of the prescribed exercise, while also capturing the influence of random variation [8, 28]. Utilizing a well-structured randomized control trial to set a threshold using the $TE_\Delta$ means those individuals categorized as responders were likely to have experienced a beneficial change as a result of the provided intervention.

Once the influence of extraneous variation is estimated, the threshold can be set. Although the $TE_\Delta$ theoretically estimates the majority of variance introduced throughout the duration of an intervention, some suggest this estimate should also be multiplied by a factor greater than one to increase the level of confidence in response categorizations [26, 28]. However, multiplying the $TE_\Delta$ may inappropriately increase the threshold to a value which inaccurately categorizes individuals as non-responders. Despite the estimate of variance being empirically defined, the factor used to set the limits for response introduces a potential limitation. As the chosen factor can only be legitimized by convention and/or recommendation across the scientific community, disagreement will lead to a variety of factors being implemented and inconsistent thresholds throughout the literature. As a result, comparing results or responder prevalence across trials can pose a great challenge.

### Lower limit of practical relevance

A practically relevant response threshold exists at the border between a trivial and a meaningful difference. The preferred method for setting a threshold at the lower limit of practical relevance is to use a well-established minimal clinically important difference. While using the smallest worthwhile difference (0.2 multiplied by the standard deviation of baseline values) is an acceptable alternative, the calculated value will be sample-specific and restrict generalizability. An example of categorizations made based on the lower limit of practical relevance can also be seen in ▶ **Fig. 1b**, [16, 38].

Similar to using zero change, a threshold based on practical relevance has no inherent ability to account for the limited accuracy of a single response estimate. As such, an individual confidence interval for each participant's observed change score should be calculated with the aforementioned considerations and interpretations in mind [28]. Assuming the utility of a control group to construct the confidence intervals, individuals can then be categorized as having a clinically meaningful positive response (*responder*; confidence interval lays completely beyond the threshold in the positive direction), a meaningfully negative response (*adverse responder*; confidence interval lays completely beyond the threshold in the negative direction), a non-response (*non-responder*; confidence interval lays completely below the positive threshold and above the negative threshold), or uncertain (confidence interval partially overlaps the threshold) (▶ **Fig. 1c**), [24, 28].

### Repeated measurements to calculate individualized confidence intervals

As an alternative to using group-based estimates of error to calculate individual confidence intervals, Hecksteden et al. [26] outlines how to produce individualized response estimates based on repeated measurements taken throughout a longitudinal intervention. This method eliminates the need to assume equal variance throughout the sample, permitting a more personalized analysis. Response estimates are calculated as the slope of the individual's regression line of observed values vs. time, with the scatter of the observed values around the individual regression line ($TE_{SLOPE}$) providing a method to calculate the uncertainty. These values can then be used to calculate individualized confidence intervals.

### Summary

Undeniably, there are numerous options available for setting response thresholds and accounting for degrees of variance. The decision of which threshold to apply or which sources of variance to take into account will significantly impact both the categorization of numerous participants and the interpretation of the results. Importantly, if the goal is to attribute a response categorization directly to the provided exercise intervention, a high-quality randomized control trial design must be used. As such, decisions regarding how response categorizations will be interpreted must be made prior to study initiation, as post-hoc decision making can lead to interpretation errors.

## Methods for Investigating Factors Influencing Response Variation

An important component to investigating exercise treatment response heterogeneity and individual response is to determine factors contributing to the differences in response. Several reviews have discussed factors which may influence response heterogeneity following exercise [47–49]. Here, we will outline the methods commonly used to identify the moderators and mediators of exercise treatment response heterogeneity and individual responses.

## Moderators of the $SD_{IR}$

As described in a series of articles from Atkinson and Batterham [22], Hecksteden et al. [25], and Hopkins [27], any variable influencing treatment response heterogeneity would logically impact the magnitude of the $SD_{IR}$. As such, Atkinson and Batterham [17] proposed a method to determine the $SD_{IR}$ and identify its moderators using a modelling approach and adjusting for identified covariates at baseline. Using a linear mixed model the study arm (intervention or control) is entered as the fixed effect, an additional binary 'dummy' variable is entered as a random effect (explained as allowing for extra variance in the change scores in one group versus the other), and the baseline value of the outcome entered as a covariate. The $SD_{IR}$ may then be derived from the parameter estimate. Subsequently, potential moderators can then be tested by including them in the model and interpreting the resultant changes in the $SD_{IR}$. The authors recommend consulting with a statistician to ensure models are properly applied.

The Atkinson and Batterham [22] method was subsequently implemented by Hammond et al. [50] who, despite large effect sizes, were unable to detect any statistically significant predictors of exercise induced treatment response heterogeneity. However, the authors provided meaningful commentary on numerous limitations associated with this method. Most prominently, the utility of a linear mixed model – while in line with good analytical practice – will likely require large sample sizes to find statistically significant results. As an example, the authors referred to their inadequate power with 181 participants, calculating that 504 participants were required to achieve statistical significance. Although samples of that size may pose limitations to individual research groups, it is important to note a requirement of large sample sizes is not unique to this modelling approach; small sample sizes are a challenge for all methods estimating $SD_{IR}$ values and/or moderators treatment response heterogeneity. High-quality analytical approaches are necessary, to accurately elucidate moderators of exercise treatment response, meaning successful collaborative efforts, or innovative, practical alternatives should be pursued.

## Categorization of 'high' and 'low' responders

Numerous authors have instead opted to divide their sample into groups of 'high' and 'low' responders before identifying moderators and/or mediators contributing to the observed differences in the magnitude of change [48]. The two key steps to this method are deciding how to divide the sample, and choosing a statistical method to identify the factors contributing to the differences in the observed changes.

A common method used to separate participants into categories of high and low responders is to break the sample into quartiles or quantiles, based on the magnitude of each individual's observed change [51–54]. Selecting a fixed proportion of participants with the highest (or lowest) differences as the high (and low) responders provides balanced groups with easily identifiable differences in response. Alternatively, K-means clustering has been used [55–57]. While these methods are common, it is important to consider that breaking participants into predetermined groups will not adequately account for the limited accuracy of a response estimate and is only meaningful to the current sample. As a result, it remains possible that participants in the 'high responder' group may have not truly responded to the intervention, or that participants in the "low responder" group did have a meaningful response to the intervention.

Once divided, various techniques have been used to identify key characteristics contributing to the differences in change scores, including regression analyses [53, 54] and ANOVA [52, 55–57]. Alternatively, one group has utilized Principal Component Analysis [51]. It is important to reiterate that these approaches do not consider the influence of random or within subject variability when categorizing individuals, nor do they quantify the magnitude of observed treatment response heterogeneity, or attempt to clarify the contribution of exercise, per se. Therefore, while these approaches have been used previously and can indicate potential moderators or mediators for the response categorizations, the observed changes may be influenced by variation, or only be truly representative of the current intervention.

## Summary

A primary purpose of individual analyses is to identify factors contributing to an individual's response categorization. However, in line with many aspects of this research field, conducting such analyses requires a great deal of assumptions. Likewise, researchers who choose to investigate these outcomes should be prepared to experience a number of challenges. While the prospect of identifying moderators and mediators of exercise response is intriguing, identifying rigorous study design and analytical methods for doing so represents one of the areas for future development within treatment response heterogeneity and individual response research.

# Current Questions Facing Exercise Treatment Response Heterogeneity and Individual Response Research

## Can we move beyond a categorical approach when describing individual responses?

There are advantages to identifying individuals who significantly benefit, and seemingly struggle to benefit, from an exercise trial and categorizing them accordingly. Doing so can provide clear cut points from which subsequent analysis can occur; such as investigating the underlying mechanisms contributing to these differences, or potentially working towards a future with personalized exercise prescriptions. However, categorizing individuals as responders or non-responders following a single trial comes with notable limitations. First, any categorization only holds true in the context of the provided exercise intervention, the selected response threshold, and the outcome of interest. Adapting the exercise protocol, choosing an alternative response threshold, or focusing on a different outcome can result in a different categorization [10, 16, 26, 44, 45, 58]. As such, the generalizability of findings may be severely limited. Second, categorizing an individual fails to consider the continuous nature that probabilities of response may provide [24, 28]. As a result, individuals with a high likelihood of response may be classified as non-responders simply due to their inability to improve beyond a subjectively chosen threshold, and an individual who only slightly surpassed the response threshold is considered equal to an individual who surpassed the threshold fivefold.

Swinton et al. [28] proposes a method to address the limitations imposed by the categorical approach and move towards likelihood-based classification, which were used by Bonafiglia et al. [24]. It should be noted that these methods make use of the Magnitude-Based Decision Making technique [59], which has received much criticism [60–67]. However, these critiques are aimed at the group-level utility of these procedures, with no current debate regarding the application of these methods on individual analyses. Nonetheless, research attempting to transition towards an accepted, reliable mechanism for likelihood-based decision making is warranted.

## Are we truly able to identify exercise-training related changes on an individual level?

One of the most important assumptions associated with individual analyses is that the true response to exercise training is an identifiable, consistent, reproducible trait. The origins of this assumption stem from studies using selectively bred rats, monozygotic twins, and nuclear families, highlighting a genetic component to the changes experienced following exercise training, particularly in reference to measures of cardiorespiratory fitness [7, 9, 12, 13, 68, 69]. While these studies suggest some degree of reproducibility in the true response to exercise, evidence of this reproducibility remains limited. Lindholm et al. [32] detected poor correlations for individual changes in exercise performance following two identical training sessions separated by a washout period, Islam et al. [30] reported poor reproducibility among acute changes in gene expression following repeated application of exercise, Islam et al. [31] outlined non-significant correlations among skeletal muscle adaptations in individuals who repeatedly completed an identical training regimen, and Del Giudice et al. [29] found changes in $VO_2$max and time to fatigue were not reproducible and led to some participants whose response categorization changed following identical four-week high-intensity training regimens, separated by a three-month washout period. There are several potential explanations for these findings, each with different repercussions for individual response research.

First, it is possible the implemented washout periods were unable to account for potential carry-over effects from the initial training phase, or that participant behaviour was different prior to each of the training interventions. Although some authors report similar performance metrics at the initiation of each training phase [29, 31], it remains possible that undetected physiological alterations influenced the observed adaptation in subsequent training periods. Improved understanding of the carry-over effects following exercise training (including both a more holistic physiological outlook across various outcomes and a better understanding of the duration of these effects) would allow for more accurate washout lengths. Regardless, participants experiencing differences in response following subsequent, identical training periods highlights that a response categorization is highly specific to the provided exercise intervention.

Second, it may hold true that the true response to exercise is a constant, repeatable trait, and the current methods for separating random and/or within-subject variation from the true response are not sensitive enough to fully account for their influence. As such, our ability to isolate the subject-by-training interaction may be inadequate. Regardless of the threshold used, current methods for categorizing an individual as a responder or non-responder neglect to account for what would have happened if that *individual* had not participated in exercise, making it impossible to definitively know if an individual responded to the exercise *per se* [20, 34]. Proposed mechanisms for setting response thresholds or calculating confidence intervals provide *estimates* for extraneous variation, but we currently cannot be certain that these influences are entirely accounted for. This explanation would suggest a need for more accurate estimates of the true response to exercise training, or improved methods for accounting for random and/or within-subject variation. Doing so would provide a more accurate indication of the influence of exercise training on individuals. A better understanding of the carry-over effects of an exercise trial and collaboration across research labs may improve the feasibility of conducting high quality cross-over trials to address this concern. Until that time, it may be wise to reconsider the 'exercise responder' terminology. We recommend shifting away from categorizing participants as responders or non-responders to exercise, and re-phrasing these determinations to who responded beyond an estimate of random or within-subject variation.

Lastly, these findings may suggest there is true intra-individual variation in response to exercise training, meaning the true response to exercise training is not a stable, reproducible trait. This would pose a great challenge for the future of personalized exercise prescription, and emphasizes the importance of not categorizing individuals following a single trial. Instead, it may be more worthwhile for research to focus on identifying the factors contributing to higher and lower change scores among participants. Moreover, as opposed to individualizing exercise prescription, practitioners may instead focus their attention on those who do not benefit from an initial training intervention and adapt exercise to garner improvements in the future [58].

It is important to discern the genuine nature of the true response to exercise training, and our ability to accurately identify it. Notably, much of the research questioning the identifiability of the subject-by-training interaction has focused on outcomes related to fitness or muscle health in relatively young, healthy men. Future investigations should aim to include various outcomes and populations to confirm these findings and move the field forward.

## Are response rates reflective of the individual, or truly a group statistic?

The vast majority of research reporting response rates do so by counting the total number of responders throughout the sample [41–43, 45, 46, 70]. These numbers are often used to compare interventions and determine the preferred method for reducing the quantity of non-responders. Results often show higher volumes of exercise leading to reduced non-response rates, or suggest increases in training volume reduce treatment response heterogeneity and eradicate non-response [43, 45, 46]. Atkinson et al. [16] challenge this assertion and argue responder counts are highly sensitive to changes in the group mean; concluding that these metrics are truly representative of changes throughout the group – rather than individualized analyses – and should be treated as such. Subsequent analyses conducted by Bonafiglia et al. [21] supported this argument highlighting how 'response rates' are reflective of differences in mean group changes, not in treatment response heterogeneity or true individual responses. Recent studies have hypothesized that, compared with exercising at relative intensities (*e.g.* a percentage of $VO_{2max}$), exercising above physiological thresholds reduces heterogeneity in metabolic stress and thus decreases interindividual variability in observed responses to exercise training [71–73]. Although these studies reported that threshold-based prescriptions increased response rates compared with relative-intensity prescription, they did not statistically compare the variability in observed response between groups. It is therefore unclear whether larger response rates following threshold-based prescription are explained by larger mean changes in the absence or presence of reduced interindividual variability [21]. Future work should adopt a statistical test (*e.g.* Levene's test) to compare variability in observed responses to exercise training prescribed at a relative and threshold-based intensity.

Atkinson et al. [20] suggest avoiding response counts when comparing interventions, instead recommending an approach described by Swinton et al. [28] to estimate the proportion of responders in the population of interest. Here, the $SD_{IR}$ is used as a parameter for the distribution of true responses around the mean treatment effect. The proportion of individuals predicted to be above or below the selected response threshold is then estimated using the characteristics of a normal distribution. Simulations run by Atkinson et al. [20] outline the superior accuracy of this method to reflect the proportion of responders following an intervention. The authors recommend future researchers use the Swinton method to estimate the proportion of response and infer to a population of interest to reduce the influence of bias associated with response categorizations.

Estimating the proportion of response can help remove the influence of mean changes when comparing interventions, and may provide a more accurate representation of how many individuals will benefit from a provided intervention. However, this method poses practical application limitations. Specifically, the ability to consider the implications of an intervention at the individual level have been removed. As such, individual non-responders cannot be identified, meaning subsequent decision making or exercise prescription adaptations cannot be completed. Therefore, prior to selecting the analytical method, the purpose of the analysis and utility of the outcomes must be considered, as these two streams of research (individual analysis vs. group based analysis) propose different theoretical approaches.

## Can multiple outcome measures be simultaneously considered?

Human physiology is complex. As such, categorizing an individual as a 'universal' responder or non-responder based solely on the changes experienced in a single outcome measure following an exercise intervention does not adequately reflect the complexity of the physiological response. Current response categorizations are specific to the selected outcome measure and the provided intervention. However studies have shown intraindividual variance and inconsistency across response categorizations when numerous outcomes are considered [10, 16, 74]. Moreover, the current method of categorizing individuals using an 'outcome by outcome' basis does not allow for conceptual outcomes, in which several outcomes are considered simultaneously (such as physical function, frailty, or metabolic syndrome), to be utilized. Therefore, methods which allow for the consideration of numerous outcomes simultaneously should be investigated. This would allow for more global categorizations of responders and non-responders.

## What degree of confidence should we use when categorizing individuals?

The confidence with which response categorizations are made is highly varied, including categorizations made based solely on estimates of variation [40, 46, 71, 73], using 50 % confidence intervals [24, 40], and 90 % confidence intervals [24, 75]. Generalizing the confidence of response categorizations will likely be difficult, as it will be highly influenced by the desired threshold value and the variance the researcher or clinician wants to account for, the outcome measure being focused on, and the acceptable degree of risk associated with an error in the categorization. Additional commentary on the acceptable degree of confidence is necessary, but will always be contended. Currently, we proposed future research provide a complete justification for the selected level of confidence, and outline the associated strengths and limitations.

## What are the economic costs of 'precision medicine'?

Throughout this review a number of recommendations and limitations associated with assessing variance and analyzing on an individual level have been noted, including larger sample sizes, taking multiple measures at each time point, and conducting repeated crossover trials with extensive washout periods. Addressing these recommendations and limitations will require significant resources and financial support. While we believe conducting these trials is worthwhile, and results can be directly transferred into practice, the economic costs are not trivial. To the best of our knowledge, there is minimal evidence for, nor has a cost-benefit analysis been conducted, suggesting implementing personalized exercise prescriptions will provide superior, financially responsible outcomes compared to the current 'mean improvement' based exercise prescription model.

Understandably, cost-benefit analyses and longitudinal trials comparing the effects of precision exercise prescription to mean improvement based prescription cannot be completed until previous questions are answered and optimal procedures are accepted. However, resource requirements and financial needs should be taken into consideration while attempting to develop and implement these models.

## What should be done once a response categorization is made?

A major question often left unanswered is what should be done with individuals following a response categorization. The answer will likely depend on the setting within which the categorization is made (i.e. clinical or research). A potential course of action for a clinical setting will be described in the subsequent section. From a research perspective, a decision on how to progress following response categorizations should be made prior to the initial analysis. As individual responses are highly specific to the outcome of interest, selected response threshold, and provided intervention [26], and reporting 'response rates' based on responder counts has been shown to more accurately reflect mean group changes [20, 21], it is hard to justify simply reporting individual responder and non-responder categorizations as generalizable, novel, or helping advance the field. Therefore, we propose future research should look beyond categorizing Approaches could include investigating any one of the questions provided throughout this section, investigating the proportion of responders produced following an exercise intervention and comparing it to alternate interventions using the Swinton method [20, 28], with the goal of improving the likelihood of participants experiencing beneficial changes, attempting to identify the root causes of heterogeneity across the sample, or exploring if subsequent/alternative exercise interventions may 'rescue' individuals categorized as non-responders and generate beneficial adaptations within the targeted outcome [58, 75]. These are topics worthy of additional attention, allow for a more thorough analysis beyond the response categorization, and may help move us closer to precision exercise prescriptions.

## Translating Treatment Response Heterogeneity and Individual Response Research to Practice

It is worth outlining how research investigating treatment response heterogeneity and individual responses to exercise training could improve the application of exercise. Currently, exercise is prescribed based on broad guidelines, designed to provide a mean improvement to various aspects of overall health. The goal of the methods described throughout this review is to provide clinicians a method for implementing a targeted, personalized approach when prescribing exercise, while recognizing the complexity of human physiology implies benefits will extend beyond a single outcome.

The utility of treatment response heterogeneity and individual response research differ; however, they may be applied in unison when prescribing exercise. Understanding treatment response heterogeneity and subsequently estimating the proportion of response (as directed by Swinton et al. [28] and Atkinson et al. [20]) will allow clinicians to prescribe exercise based on the likelihood of an individual to experience a change greater than a selected threshold [20]. This means the clinician may be able to target a specific outcome of importance for the patient (e.g., cardiorespiratory fitness, glycated hemoglobin concentration, or systolic blood pressure), a desired degree of improvement, and prescribed exercise intensity, modality, and mode based on which combination would provide the greatest likelihood of experiencing the targeted improvement. Subsequently, individual response research may allow for clinicians to more accurately interpret the individual's ability to experience the targeted benefits – or not – following the prescribed exercise, meaning the clinician can react appropriately at follow-up. If the individual was determined to be a non-responder following the provided intervention, the prescription could be adapted, or an alternative prescription could be provided based on the estimated likelihood of success. It is our view that these methods provide an alternative, and potentially more beneficial, approach to the current model of prescribing exercise.

As research progresses and the factors contributing to an individual response categorization are better understood, the accuracy of an initial prescription or subsequent adaptations may improve. Moreover, the aforementioned questions facing treatment response heterogeneity and individual response research must be addressed for such a future to become reality. Nonetheless, we believe it is a worthwhile pursuit to advance the utility of exercise.

## Conclusion

Research investigating exercise treatment response heterogeneity and individual responses will surely continue to proliferate as interest in personalized medicine grows. While many questions remain unanswered, they represent areas for future research and growth required to advance the field and ensure rigour. Given the current limitations, researchers analyzing treatment response heterogeneity and categorizing participants must consider the context of the research question, how categorizations may be used, and make methodological decisions prior to conducting a trial. Subsequently, results should be interpreted within the selected method's capabilities, and applicable limitations clearly outlined.

## Funding

Salary Support for this project was provided by the New Brunswick Health Research Foundation and the Maritime SPOR Support Unit (T.J.H). J.T. Bonafiglia was supported by a NSERC Vanier CGS Scholarship.

### Conflict of interest

The authors have no conflict of interest to declare. The manuscript was built based on the International Journal of Sports Medicine ethical standards [76].

### References

[1] Piercy KL, Troiano RP, Ballard RM et al. The physical activity guidelines for americans. JAMA 2018; 320: 2020–2028. doi:10.1001/jama.2018.14854

[2] Tremblay MS, Carson V, Chaput J-P. Introduction to the canadian 24-hour movement guidelines for children and youth: An integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab 2016; 41: iii–iv. doi:10.1139/apnm-2016-0203

[3] Bull FC, Al-Ansari SS, Biddle S et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020; 54: 1451–1462. doi:10.1136/bjsports-2020-102955

[4] Lortie G, Simoneau JA, Hamel P et al. Responses of maximal aerobic power and capacity to aerobic training. Int J Sports Med 1984; 05: 232–236. doi:10.1055/s-2008-1025911

[5] Prud'homme D, Bouchard C, Leblanc C et al. Sensitivity of maximal aerobic power to training is genotype-dependent. Med Sci Sports Exerc 1984; 16: 489–493

[6] Bouchard C, Leon AS, Rao DC et al. The HERITAGE family study. Aims, design, and measurement protocol. Med Sci Sports Exerc 1995; 27: 721–729

[7] Bouchard C, An P, Rice T et al. Familial aggregation of Vo2 max response to exercise training: Results from the HERITAGE Family Study. J Appl Physiol (1985) 1999; 87: 1003–1008

[8] Williamson PJ, Atkinson G, Batterham AM.. Inter-individual responses of maximal oxygen uptake to exercise training: A critical review. Sports Med 2017; 47: 1501–1513. doi:10.1007/s40279-017-0680-8

[9] Avila JJ, Kim SK, Massett MP. Differences in exercise capacity and responses to training in 24 inbred mouse strains. Front Physiol 2017; 8: 974. doi:10.3389/fphys.2017.00974

[10] Bouchard C, Blair SN, Church TS et al. Adverse metabolic response to regular exercise: Is it a rare or common occurrence? PLoS One 2012; 7: e37887. doi:10.1371/journal.pone.0037887

[11] Bouchard C, Rankinen T. Individual differences in response to regular physical activity. Med Sci Sports Exerc 2001; 33 6 Suppl: S446–S451

[12] Despres JP, Bouchard C, Savard R et al. Adaptive changes to training in adipose tissue lipolysis are genotype dependent. Int J Obes 1984; 8: 87–95

[13] Hamel P, Simoneau J-A, Lortie G et al. Heredity and muscle adaptation to endurance training. Med Sci Sports Exerc 1986; 18: 690–696

[14] Koch LG, Britton SL. Theoretical and biological evaluation of the link between low exercise capacity and disease risk. Cold Spring Harb Perspect Med 2018; 8: a029868

[15] Simoneau JA, Lortie G, Boulay MR et al. Inheritance of human skeletal muscle and anaerobic capacity adaptation to high-intensity intermittent training. Int J Sports Med 1986; 7: 167–171

[16] Walsh JJ, Bonafiglia JT, Goldfield GS et al. Interindividual variability and individual responses to exercise training in adolescents with obesity. Appl Physiol Nutr Metab 2019; 45: 45–54. doi:10.1139/apnm-2019-0088

[17] Hrubeniuk MTJ, Hay MJL, MacIntosh MAC et al. Interindividual variation in cardiometabolic health outcomes following 6-months of endurance training in youth at risk of Type 2 Diabetes Mellitus. Appl Physiol Nutr Metab 2021; 46: 727–734

[18] Buford TW, Roberts MD, Church TS. Toward exercise as personalized medicine. Sports Med 2013; 43: 157–165. doi:10.1007/s40279-013-0018-0

[19] Ross R, Goodpaster BH, Koch LG et al. Precision exercise medicine: understanding exercise response variability. Br J Sports Med 2019; 53: 1141–1153. doi:10.1136/bjsports-2018-100328

[20] Atkinson G, Williamson P, Batterham AM. Issues in the determination of ‘responders’ and ‘non-responders’ in physiological research. Exp Physiol 2019; 104: 1215–1225. doi:10.1113/EP087712

[21] Bonafiglia JT, Preobrazenski N, Islam H et al. Exploring differences in cardiorespiratory fitness response rates across varying doses of exercise training: A retrospective analysis of eight randomized controlled trials. Sports Med 2021; 51: 1785–1797. doi:10.1007/s40279-021-01442-9

[22] Atkinson G, Batterham A. True and false interindividual differences in the physiological response to an intervention. Exp Physiol 2015; 100: 577–588. doi:10.1113/EP085070

[23] Bonafiglia JT, Brennan AM, Ross R et al. An appraisal of the SDIR as an estimate of true individual differences in training responsiveness in parallel-arm exercise randomized controlled trials. Physiol Rep 2019; 7: e14163. doi:10.14814/phy2.14163

[24] Bonafiglia JT, Nelms MW, Preobrazenski N et al. Moving beyond threshold-based dichotomous classification to improve the accuracy in classifying non-responders. Physiol Rep 2018; 6: e13928. doi:10.14814/phy2.13928

[25] Hecksteden A, Kraushaar J, Scharhag-Rosenberger F et al. Individual response to exercise training - a statistical perspective. J Appl Physiol (1985) 2015; 118: 1450–1459. doi:10.1152/japplphysiol.00714.2014

[26] Hecksteden A, Pitsch W, Rosenberger F et al. Repeated testing for the assessment of individual response to exercise training. J Appl Physiol (1985) 2018; 124: 1567–1579. doi:10.1152/japplphysiol.00896.2017

[27] Hopkins WG.. Individual responses made easy. J Appl Physiol (1985) 2015; 118: 1444–1446. doi:10.1152/japplphysiol.00098.2015

[28] Swinton PA, Hemingway BS, Saunders B et al. A Statistical framework to interpret individual response to intervention: paving the way for personalized nutrition and exercise prescription. Front Nutr 2018; 5: 41. doi:10.3389/fnut.2018.00041

[29] Del Giudice M, Bonafiglia JT, Islam H et al. Investigating the reproducibility of maximal oxygen uptake responses to high-intensity interval training. J Sci Med Sport 2020; 23: 94–99. doi:10.1016/j.jsams.2019.09.007

[30] Islam H, Edgett BA, Bonafiglia JT et al. Repeatability of exercise-induced changes in mRNA expression and technical considerations for qPCR analysis in human skeletal muscle. Exp Physiol 2019; 104: 407–420. doi:10.1113/EP087401

[31] Islam H, Bonafiglia JT, Giudice MD et al. Repeatability of training-induced skeletal muscle adaptations in active young males. J Sci Med Sport 2021; 24: 494–498. doi:10.1016/j.jsams.2020.10.016

[32] Lindholm ME, Giacomello S, Solnestam BW et al. The impact of endurance training on human skeletal muscle memory, global isoform expression and novel transcripts. PLoS Genet 2016; 12: e1006294. doi:10.1371/journal.pgen.1006294

[33] Hopkins WG. Measures of reliability in sports medicine and science. Sports Med 2000; 30: 1–15. doi:10.2165/00007256-200030010-00001

[34] Senn S.. Mastering variation: variance components and personalised medicine. Stat Med 2016; 35: 966–977. doi:10.1002/sim.6739

[35] Senn S, Rolfe K, Julious SA. Investigating variability in patient response to treatment – a case study from a replicate cross-over study. Stat Methods Med Res 2011; 20: 657–666. doi:10.1177/0962280210379174

[36] Goltz FR, Thackray AE, Atkinson G et al. True interindividual variability exists in postprandial appetite responses in healthy men but is not moderated by the FTO genotype. J Nutr 2019; 149: 1159–1169. doi:10.1093/jn/nxz062

[37] Leifer ES, Mikus CR, Karavirta L et al. Adverse cardiovascular response to aerobic exercise training: Is this a concern? Med Sci Sports Exerc 2016; 48: 20–25. doi:10.1249/MSS.0000000000000752

[38] Bonafiglia JT, Ross R, Gurd BJ. The application of repeated testing and monoexponential regressions to classify individual cardiorespiratory fitness responses to exercise training. Eur J Appl Physiol 2019; 119: 889–900. doi:10.1007/s00421-019-04078-w

[39] Smith TB, Hopkins WG. Variability and predictability of finals times of elite rowers. Med Sci Sports Exerc 2011; 43: 2155. doi:10.1249/MSS.0b013e31821d3f8e

[40] Schulhauser KT, Bonafiglia JT, McKie GL et al. Individual patterns of response to traditional and modified sprint interval training. J Sports Sci 2021; 39: 1077–1087. doi:10.1080/02640414.2020.1857507

[41] Sisson SB, Katzmarzyk PT, Earnest CP et al. Volume of exercise and fitness non-response in sedentary, post-menopausal women. Med Sci Sports Exerc 2009; 41: 539–545. doi:10.1249/MSS.0b013e3181896c4e

[42] Gurd BJ, Giles MD, Bonafiglia JT et al. Incidence of nonresponse and individual patterns of response following sprint interval training. Appl Physiol Nutr Metab 2015; 41: 229–234. doi:10.1139/apnm-2015-0449

[43] Astorino TA, Schubert MM. Individual responses to completion of short-term and chronic interval training: A retrospective study. PLoS One 2014; 9: e97638. doi:10.1371/journal.pone.0097638

[44] Bonafiglia JT, Rotundo MP, Whittall JP et al. Inter-individual variability in the adaptive responses to endurance and sprint interval training: A randomized crossover study. PLoS One 2016; 11: e0167790. doi:10.1371/journal.pone.0167790

[45] Montero D, Lundby C. Refuting the myth of non-response to exercise training: 'non-responders' do respond to higher dose of training. J Physiol 2017; 595: 3377–3387. doi:10.1113/JP273480

[46] Lannoy L, de Clarke J, Stotz PJ et al. Effects of intensity and amount of exercise on measures of insulin and glucose: Analysis of inter-individual variability. PLoS One 2017; 12: e0177095. doi:10.1371/journal.pone.0177095

[47] Mann TN, Lamberts RP, Lambert MI. High responders and low responders: factors associated with individual variation in response to standardized training. Sports Med 2014; 44: 1113–1124. doi:10.1007/s40279-014-0197-3

[48] Roberts MD, Haun CT, Mobley CB et al. Physiological differences between low versus high skeletal muscle hypertrophic responders to resistance exercise training: current perspectives and future research directions. Front Physiol 2018; 9: 834. doi:10.3389/fphys.2018.00834

[49] Vellers HL, Kleeberger SR, Lightfoot JT. Inter-individual variation in adaptations to endurance and resistance exercise training: genetic approaches towards understanding a complex phenotype. Mamm Genome 2018; 29: 48–62. doi:10.1007/s00335-017-9732-5

[50] Hammond BP, Stotz PJ, Brennan AM et al. Individual variability in waist circumference and body weight in response to exercise. Med Sci Sports Exerc 2019; 51: 315–322. doi:10.1249/MSS.0000000000001784

[51] Morton RW, Sato K, Gallaugher MPB et al. Muscle androgen receptor content but not systemic hormones is associated with resistance training-induced skeletal muscle hypertrophy in healthy, young men. Front Physiol 2018; 9: 1373. doi:10.3389/fphys.2018.01373

[52] Raleigh JP, Giles MD, Islam H et al. Contribution of central and peripheral adaptations to changes in maximal oxygen uptake following 4 weeks of sprint interval training. Appl Physiol Nutr Metab 2018; 43: 1059–1068. doi:10.1139/apnm-2017-0864

[53] Sénéchal M, Swift DL, Johannsen NM et al. Changes in body fat distribution and fitness are associated with changes in hemoglobin A1c after 9 months of exercise training. Diabetes Care 2013; 36: 2843–2849. doi:10.2337/dc12-2428

[54] Sénéchal M, Rempel M, Duhamel TA et al. Fitness is a determinant of the metabolic response to endurance training in adolescents at risk of type 2 diabetes mellitus. Obesity 2015; 23: 823–832. doi:10.1002/oby.21032

[55] Bamman MM, Petrella JK, Kim J et al. Cluster analysis tests the importance of myogenic gene expression during myofiber hypertrophy in humans. J Appl Physiol (1985) 2007; 102: 2232–2239. doi:10.1152/japplphysiol.00024.2007

[56] Petrella JK, Kim J-S, Mayhew DL et al. Potent myofiber hypertrophy during resistance training in humans is associated with satellite cell-mediated myonuclear addition: a cluster analysis. J Appl Physiol (1985) 2008; 104: 1736–1742. doi:10.1152/japplphysiol.01215.2007

[57] Stec MJ, Kelly NA, Many GM et al. Ribosome biogenesis may augment resistance training-induced myofiber hypertrophy and is required for myotube growth in vitro. Am J Physiol Endocrinol Metab 2016; 310: E652–E661. doi:10.1152/ajpendo.00486.2015

[58] Marsh CE, Thomas HJ, Naylor LH et al. Fitness and strength responses to distinct exercise modes in twins: Studies of Twin Responses to Understand Exercise as a THerapy (STRUETH) study. J Physiol 2020; 598: 3845–3858; n/a. doi:10.1113/JP280048

[59] Hopkins W. Precision of the estimate of a subject's true value (Excel spreadsheet). In: a new view of statistics. 2000. Im Internet: www.sportsci.org/resource/stats/xprecisionsubject

[60] Aisbett J, Lakens D, Sainani K. Magnitude based inference in relation to one-sided hypotheses testing procedures. SportRχiv 2020. doi:10.31236/osf.io/pn9s3

[61] Barker RJ, Schofield MR. Inference about magnitudes of effects. Int J Sports Physiol Perform 2008; 3: 547–557. doi:10.1123/ijspp.3.4.547

[62] Borg DN, Minett GM, Stewart IB et al. Bayesian methods might solve the problems with magnitude-based inference. Med Sci Sports Exerc 2018; 50: 2609–2610. doi:10.1249/MSS.0000000000001736

[63] Curran-Everett D.. Magnitude-based inference: Good idea but flawed approach. Med Sci Sports Exerc 2018; 50: 2164–2165. doi:10.1249/MSS.0000000000001646

[64] Lohse K, Sainani K, Taylor J et al. Systematic review of the use of „Magnitude-Based Inference" in sports science and medicine. SportRχiv 2020. doi:10.31236/osf.io/wugcr

[65] Sainani KL. The problem with "magnitude-based inference". Med Sci Sports Exerc 2018; 50: 2166–2176. doi:10.1249/MSS.0000000000001645

[66] Sainani KL, Lohse KR, Jones PR et al. Magnitude-based inference is not bayesian and is not a valid method of inference. Scand J Med Sci Sports 2019; 29: 1428–1436. doi:10.1111/sms.13491

[67] Welsh AH, Knight EJ. "Magnitude-based Inference": A statistical review. Med Sci Sports Exerc 2015; 47: 874–884. doi:10.1249/MSS.0000000000000451

[68] Massett MP, Fan R, Berk BC. Quantitative trait loci for exercise training responses in FVB/NJ and C57BL/6J mice. Physiol Genomics 2009; 40: 15–22. doi:10.1152/physiolgenomics.00116.2009

[69] Sarzynski MA, Ghosh S, Bouchard C. Genomic and transcriptomic predictors of response levels to endurance exercise training. J Physiol 2017; 595: 2931–2939. doi:10.1113/JP272559

[70] Ross R, de Lannoy L, Stotz PJ. Separate effects of intensity and amount of exercise on interindividual cardiorespiratory fitness response. Mayo Clin Proc 2015; 90: 1506–1514. doi:10.1016/j.mayocp.2015.07.024

[71] Seward S, Ramos J, Drummond C et al. Inter-individual variability in metabolic syndrome severity score and VO2$_{max}$ changes following personalized, community-based exercise programming. Int J Environ Res Public Health 2019; 16: 4855. doi:10.3390/ijerph16234855

[72] Weatherwax RM, Harris NK, Kilding AE et al. Incidence of VO2$_{max}$ responders to personalized versus standardized exercise prescription. Med Sci Sports Exerc 2019; 51: 681–691. doi:10.1249/MSS.0000000000001842

[73] Wolpern AE, Burgos DJ, Janot JM et al. Is a threshold-based model a superior method to the relative percent concept for establishing individual exercise intensity? a randomized controlled trial. BMC Sports Sci Med Rehabil 2015; 7: 16. doi:10.1186/s13102-015-0011-z

[74] Phillips BE, Kelly BM, Lilja M et al. A practical and time-efficient high-intensity interval training program modifies cardio-metabolic risk factors in adults with risk factors for type ii diabetes. Front Endocrinol (Lausanne) 2017; 8: 229. doi:10.3389/fendo.2017.00229

[75] Hrubeniuk TJ, Bouchard DR, Gurd BJ et al. Can non-responders be "rescued" by increasing exercise intensity? A quasi-experimental trial of individual responses among humans living with pre-diabetes or type 2 diabetes mellitus in Canada. BMJ Open 2021; 11: e044478. doi:10.1136/bmjopen-2020-044478

[76] Harriss DJ, MacSween A, Atkinson G. Ethical standards in sport and exercise science research: 2020 update. Int J Sports Med 2019; 40: 813–817. doi:10.1055/a-1015-3123