DOI: 10.1113/EP087712

Accepted Article

Issues in the determination of "responders" and "non-responders" in physiological research

Greg Atkinson[1], Philip Williamson[2] and Alan M Batterham[1]

[1]*School of Health and Social Care Teesside University, Middlesbrough, UK.*

[2]*Faculty of Health Sciences, School of Life Sciences, University of Hull, Hull, UK*

All correspondence to:

Greg Atkinson

Health and Social Care Institute

School of Health & Social Care

Parkside West

Teesside University

MIDDLESBROUGH

Tees Valley TS1 3BA

United Kingdom

Tel: 00-44-1642 342771

Email: g.atkinson@tees.ac.uk

This is an Accepted Article that has been peer-reviewed and approved for publication in the Experimental Physiology, but has yet to undergo copy-editing and proof correction. Please cite this article as an Accepted Article; doi: 10.1113/EP087712.
This article is protected by copyright. All rights reserved.

Accepted Article

New findings

- **What is the topic for this review?**

  The dichotomisation of continuous-level physiological measurements into "responders" and "non-responders", when interventions/treatments are examined in robust parallel-group studies

- **What advances does it highlight?**

  Sample responder counts are biased by pre-to-post within-subjects variability. Sample differences in counts may be explained wholly by differences in mean response, even without individual response heterogeneity, and even if test-retest measurement error informs the choice of response threshold. A less biased and more informative approach employs the SD of individual responses to estimate the chance a new person from the population of interest will be a responder.

Abstract

As a follow-up to our 2015 review, we cover more issues on the topic of "response heterogeneity", which we define as clinically-important individual differences in the physiological responses to the same treatment or intervention that cannot be attributed to random within-subjects variability. We highlight various pitfalls with the common practice of counting the number of "responders", "non-responders" and "adverse responders" in samples that have been given certain treatments/interventions for research purposes. We focus on the classical parallel-group randomised controlled trial (RCT) and assume typical good practice in trial design.

We show that sample responder counts are biased because individuals differ in terms of pre-to-post within-subjects random variability in the study outcome(s) and not necessarily treatment response. Ironically, sample differences in responder counts may be explained

This article is protected by copyright. All rights reserved.

Accepted Article

wholly by sample differences in mean response, even if there is no response heterogeneity at all. Sample comparisons of responder counts also have relatively low statistical precision. These problems do not depend on how the response threshold has been selected, e.g. on the basis of a measurement error statistic, and are not rectified fully by the use of confidence intervals for individual responses in the sample.

The dichotomisation of individual responses in a research sample is fraught with pitfalls. Less biased approaches for estimating the proportion of responders in a population of interest are now available. Importantly, these approaches are based on the standard deviation for true individual responses, directly incorporating information from the control group.

Keywords: Response heterogeneity, Inter-individual differences, standard deviation, responders, within-subject random variability

This article is protected by copyright. All rights reserved.

Accepted Article

1. Introduction and Background

In a previous issue of *Experimental Physiology*, the paper, "*True and false individual differences in the physiological response to an intervention*" was published (Atkinson and Batterham, 2015). This earlier review was written because we observed that many claims of individual differences in the physiological responses to exercise training and other interventions were based solely on intervention group data, sometimes even if these data were collected as part of a parallel arm randomised controlled trial (RCT). In the context of personalised (precision) medicine, we showed how common plots and analyses of pre-to-post changes (measured on a continuous scale) can be misleading when undertaken only on data from the intervention/treatment group because of unavoidable within-subject random variation between the baseline and follow-up time-points. This source of variation is inevitable even with 'gold-standard' measurement tools/protocols that happen to have excellent short-term (over hours or days) repeatability. For example, the short term test-retest coefficient of variation (CV) of body mass is probably less than 0.5%, but the pre-to-post CV in body mass over a 6-12 month period (a typical intervention duration for a weight management service) will be considerably larger (Abe et al., 2019; Atkinson and Batterham, 2017). This differential amount of variability, depending on duration of retest (follow-up) is likely to be present for most physiological measurements and functional tests (Abe et al., 2019). We showed that this component of variance between baseline and follow-up may be so substantial that it can sometimes explain all the perceived individual response differences, as observed solely in the treatment/intervention group (Atkinson and Batterham, 2015).

In our previous review, we presented a "roadmap" for research (particularly RCTs) on physiological response heterogeneity, which included how to quantify individual response differences via a relatively simple comparison of the standard deviation (SD) of changes (baseline – follow-up values) between intervention and comparator arms in a parallel group study. We termed this difference between SDs of change in intervention and control groups the standard deviation for individual responses (SDir). Any substantial treatment effect

This article is protected by copyright. All rights reserved.

Accepted Article

heterogeneity that is larger than the heterogeneity in the data due to random within-subject variability over time would be revealed if the SD of changes in the treatment group is larger than that in the comparator group (Hopkins, 2015; Atkinson and Batterham, 2015; Cortes et al., 2018; Atkinson et al., 2018). When these SDs are similar, any individual response differences to the treatment itself are not large enough to be detected beyond the typical and inevitable within-subjects random variability in the study outcome. Therefore, further analyses, e.g. identification of specific response moderators, may be unwarranted and wasteful of resources. In essence, this reflects the "counterfactual" nature of the control group, which must always be appreciated in parallel group trials, whether one is researching mean or individual treatment effects (Senn, 2015; 2018).

In Panel 1, we present various questions which have been raised in conferences and communications about the SDir approach since our last publication, and we have endeavoured to answer these questions. Like the estimation of mean treatment effects, any SD of changes comparison is contingent on a well-designed, adequately powered and executed RCT. We preferred to interpret the SDir against a minimal clinically important difference (MCID), while retaining the original units of measurement. Recently, Cortes et al. (2018) compared the SD of changes using a "relative" and unit-free F-ratio (treatment SD / control SD) approach. Both approaches could be reported in any RCT.

Another practice that is common in research of this nature is counting the number (or proportion) of people in the study sample(s) who show observed responses above (responders) or below (non-responders) certain thresholds. In this, our update paper, we focus on the question of how robust the various responder identification type approaches are in physiological research. We maintain that there are also many pitfalls in these approaches, the worst scenario being that true clinically-important response heterogeneity has not been quantified and the responder counting analyses are not robust, leading to false inferences and recommendations about individuals who are deemed to be responders/non-responders in a given sample. Our primary aim, in this update review, is, therefore, to highlight these

This article is protected by copyright. All rights reserved.

Accepted Article

pitfalls about sample responder counting and make recommendations on how to avoid these pitfalls.

2. A definition of response heterogeneity in the context of precision medicine

As in our previous review, we focus on a definition of response heterogeneity that is relevant to study variables (outcomes) measured on a continuous scale and that is most relevant to precision (personalised) medicine. We highlight the fact that the response heterogeneity we cover here is distinct from other aspects of personalised medicine, e.g. a personal risk profile, based on genes or other information, to predict future conditions or diseases. So, by "treatment response heterogeneity", we refer to clinically important individual differences in the magnitude of response to *the same* treatment/intervention. We use these latter terms interchangeably. We refer to individual differences in treatment response that are not explained by random within-subject variability over time. We also emphasise that this definition refers to the responses to *the same* treatment prescribed to a sample of individuals, rather than observing how many people in *different* samples respond to *different* interventions beyond a certain threshold response value. We discuss later, and in more detail, how this latter approach tells us little about true response heterogeneity in the context of personalised or precision medicine.

By "clinically important", we mean a magnitude of response that, ideally, is anchored to a meaningful change in the risk of morbidity and/or mortality, or an overall endpoint that matters like the chance of winning an Olympic medal in an exercise science context. There are various ways in which "target differences" or minimal clinically important differences (MCID) can be arrived at, and we encourage readers to refer to the guidelines laid down in the DELTA1 and DELTA2 publications (Cook et al., 2014; 2018). Later, we also highlight the difference between this MCID and "minimal detectable change", which is the magnitude of response that surpasses, with a certain probability, measurement error. Such a change may be "statistically significant" or be detectable amongst a background of experimental noise, but it may not necessarily be clinically important and be the same as the MCID (de Vet et al., 2006).

This article is protected by copyright. All rights reserved.

Accepted Article

3. How large should the mean treatment response be for response heterogeneity to matter?

In our previously-reported "roadmap", we showed how the SDir (derived from the standard deviations of change for both intervention and control groups) can be used to inform the magnitude of response heterogeneity that is adjusted for any within-subjects random variability (present in both treatment and control groups). The SDir is compared to a target clinically-important threshold, alongside the magnitude of mean treatment effect. One question within this approach is how likely is it that response heterogeneity is clinically important, if the mean intervention effect is already known to be close to zero?

Harrell (2018) maintained that, if response heterogeneity is present when the mean treatment effect is tiny, it follows that there must be patients or subgroups for whom the treatment worsens the outcome. Harrell (2018) thought it not to make sense to perform further studies on a treatment known to be, on average, not beneficial at all just to gain further knowledge that the treatment could actually also be detrimental to some people. In the context of exercise training, this notion would translate to a researcher wanting to find out whether a certain training intervention worsens health for some people, even though it is already known that the intervention has close to a zero average effect.

In theory, if a certain intervention leads to a clinically important benefit on average, and there is actually very little response heterogeneity, then this is preferable to the situation where response heterogeneity is present but only relatively few people will benefit from the intervention. Such interpretations of response heterogeneity, especially in the context of health economics, have been reported to be under-appreciated (Senn, 2015; 2018). Nevertheless, if the mean treatment response is small but response heterogeneity between people is indeed very large, then we believe that this finding is important to confirm robustly, as the treatment could benefit a reasonably large proportion of people in the population of interest. Therefore, we think it is interesting to know what proportion of people in a population of interest may be responders, non-responders or adverse responders. Nevertheless, there does not seem to be, at present, any agreed thresholds for these

This article is protected by copyright. All rights reserved.

Accepted Article

proportions in order to guide research and practice, e.g. whether a certain treatment is worth rolling out if the estimated proportion of responders is say 20%.

4. A hypothetical study and dataset

In keeping with our previous review, we can communicate some pitfalls and issues *via* a hypothetical RCT and a data simulation involving large (n=1000) samples. This is not a simulation that is designed to illustrate the impact of sampling error on a certain statistic. We merely aim to scrutinise various responder counting approaches in hypothetical large samples with characteristics we can stipulate *a priori*. In this respect, we generated our samples using the popular package, Microsoft Excel, rather than dedicated code-dependent software such as R, with the belief that any researcher may like to reproduce, or formulate their own, data simulation. The overall benefits of this approach are that we know exactly what the parameters of the data are, e.g. Normal distribution of pre-to-post changes, we can pre-specify means and SDs that are realistic, and we can make inferences with decent precision with a sample size of 1000 (Morris et al., 2019).

Obviously, all the usual pre-specified trial design and data analysis considerations are critical, whether it is treatment response heterogeneity that is of interest and/or the mean treatment response. Both these outcomes are reliant on good trial design. These issues are covered comprehensively in the CONSORT explanation and elaboration document (Moher et al., 2010). If there are deviations from typical good practice in trial design (Moher et al., 2010), then the least of the researchers problems is how to robustly undertake a response variance comparison, or indeed any data analysis (Panel 1). No statistical analysis approach, whether it is designed to quantify mean treatment response or response heterogeneity, can retrieve poor study design characteristics (Campbell and Machin, 1993).

In Table 1, we present the results of the data simulation for a three-group parallel arm intervention study designed to quantify the effects of two exercise interventions vs a control group on maximal oxygen uptake (VO_2max). This design and context is similar to those reported recently by Williams et al. (2019). The three study groups are; control (zero change in true VO_2max for all participants), Intervention 1 (a 3.6 ml/kg/min increase in true VO_2max

This article is protected by copyright. All rights reserved.

Accepted Article

for all participants) and Intervention 2 (a 2.0 ml/kg/min increase in true VO_2max for all participants). We highlight the fact that the true change in VO_2max is a constant value for every hypothetical participant, and only varies according to the study group they are in. In this way, we wish to set up the "null position" of no true treatment response heterogeneity. In this situation, it follows that there are no responders whatsoever in a sample when the mean response is below the response threshold. Later, we discuss the basis of how such a threshold should be selected.

Within-subjects random measurement variability is inevitable in physiological research. Therefore, a random amount of within-subjects variability was added to each of the "true values" and this "error" had an approximate mean (SD) of 0 (3) ml/kg/min in order to provide each group's observed baseline and follow-up measurements. These errors led to the SD of change in each study arm to be 4.3-4.4 ml/kg/min, which are similar to the SDs of change reported by Prud'Homme et al. (1984). These values of SDchange are expected because of the mathematical relationship between the SD of change and the within-subjects SD or "typical error", e.g. SD of change = within-subjects SD x $\sqrt{2}$ (Atkinson and Nevill, 1996). As is likely the case for real data, we assumed that the distribution of these within-subjects errors is Gaussian. Most measurement error statistics are reliant on this assumption (Atkinson and Nevill, 1996), which should, nevertheless, always be verified for any data analysis. Interestingly, non-Gaussian distribution of responses has been claimed for certain measurements of pain (BMJ, 2019), although whether this is the case in general has not yet been confirmed for studies on pain outcomes. Irrespective of this assumption, we highlight the fact that individuals can differ in how much random within-subjects variability influence the measurements made at baseline and follow-up timepoints, as is the case for real data. That is, in any study, some participants show higher amounts of random test-retest variability than other participants. This common measurement characteristic is important for explaining some of our observations and conclusions later in this review. The Excel spreadsheet for these data is available as a supplementary file.

This article is protected by copyright. All rights reserved.

Accepted Article

Mean intervention effects in RCTs are most-appropriately quantified with a general linear model, including study group as a fixed factor and baseline VO_2max as a covariate. This ANCOVA-model approach has been shown to be superior to a group x time interaction based model (Vickers, 2005), which is unfortunately often selected by physiologists and exercise scientists The estimated mean (95% confidence interval) change in VO_2max (vs control) for interventions 1 and 2 are 3.6 (3.2 to 4.0) ml/kg/min and 2.2 (1.8 to 2.6) ml/kg/min, respectively. These are conditional mean changes, whereby group differences at baseline have been adjusted for in the model. The same baseline-adjusted modelling approach can be used to also derive the SDir (see later).

5. Counting responders and non-responders in the sample using a defined response threshold

The fundamental problem with sample responder counting in a parallel arm RCT is the "counterfactual", whereby it is impossible to determine who is a responder in a treatment arm, because it is unknown what would have happened to that individual if, contrary to the fact, they had been in the control group (Senn, 2015). Consequently, there are four issues of practical validity to consider when counting the number of changes in a sample that surpass or fall short of a certain response threshold and comparing these counts between different study groups who received different interventions,

(i) The relevance of this approach to response heterogeneity in the context of precision medicine,

(ii) The sensitivity of the responder counts to group differences in mean response,

(iii) The incorporation of probability inference for the precision of identifying responders or non-responders, and what is done with this information.

(iv) How the response threshold has been selected.

This article is protected by copyright. All rights reserved.

Accepted Article

5.1. What are responder counts in samples actually telling us?

We will consider issues (i) and (ii) together, since we maintain that the sensitivity of the responder counting approach to the group mean change renders the approach irrelevant to response heterogeneity in the context of precision medicine. Note, in Table 1, that the SDs of observed changes are very similar (4.3-4.4 ml/kg/min) for all three groups, in agreement with the fact that no individual differences in change were simulated in all three groups. According to our previously-reported equation to estimate the SD for true individual response heterogeneity, this SD (95%CI) is 0.83 (-1.29 to 1.75) ml/kg/min and 0.41 (-1.47 to 1.58) ml/kg/min for Intervention 1 and 2 participants, respectively (observed response heterogeneity = random within-subject heterogeneity in the control sample). As mentioned above, these SDir estimates and confidence intervals can also be obtained using a modelling approach, adjusting for any differences at baseline (Atkinson and Batterham, 2015). In the case of our large sample random data simulation, the SDir estimates are similar between equation and modelling approach. Importantly, the SDir values are small and, therefore, not indicative of any clinically important response heterogeneity. This is, of course, exactly what was simulated. These SDir values are not exactly zero because of sampling error (even for our relatively large sample sizes of 1000 in each group) and small random variability in the random number generator in Excel. Note also that, because of the sampling error we discussed in section 1, the lower confidence limit for both SDs is negative in sign. It can be seen that even with relatively large sample sizes of 1000 cases in each study arm, sampling error is still large enough for the 95% confidence interval of a very small SDir to overlap zero.

Strikingly, the responder counts indicate that there are a number of responders and adverse responders in each sample, even in the control group. Nevertheless, we already know that the treatment response *per se* of every case is a constant value in each group and smaller than the response threshold we selected of 5 ml/kg/min. For example, the “error-free” increase for all cases in the Intervention 1 sample is 3.6 ml/kg/min. Nevertheless, the responder counts are telling us that 363 (36%) of this sample responded ≥ 5 ml/kg/min. This

This article is protected by copyright. All rights reserved.

Accepted Article

discrepancy between observed counts and the true counts is due to the inevitable random within-subjects variability between the baseline and follow-up time points in the RCT. The distribution of this random variability tends to be Normal. Therefore, there are always some people who show larger amounts of this variability than other people. In our data simulation, the apparent "responders" are actually the cases who happen to show a relatively large amount of random variability between baseline and follow-up, and this variability happens to be in a positive direction, thus rendering the baseline to follow-up change large and positive. The apparent "adverse responders" are cases for which random variability happens to be large in the other direction, rendering a substantial apparent deterioration in VO_2max. Counting the number of responders in a sample is compromised by within-subjects variation between time points (and individual differences in this within-subjects variation) and can be misleading. One may think that comparing responder counts between intervention and control groups would rectify this problem. Nevertheless, this is not the case because there can be problems also with such a comparison, and these are covered in the next section.

5.2. Comparing responder counts between samples.

Note in Table 1 and Figure 1 how the group differences in observed mean response lead to group differences in the observed numbers of responders, adverse responders, and trivial responders (according to a response threshold of 5 ml/kg/min and an adverse response threshold of -5 ml/kg/min). Therefore, although researchers have made inferences relating to response heterogeneity or "trainability" on the basis of such "responder counts" (Ross et al., 2015; Williams et al., 2019; Hammond et al., 2019; Bonafiglia et al., 2019), it is in fact the group differences in mean treatment response that explain the differences in responder counts between our groups (Figure 1), besides the fact that the responder counts cannot be correct in the first place. Therefore, such group comparisons of responder counts do not provide much information about response heterogeneity, as defined in the context of precision medicine (section 2).

This article is protected by copyright. All rights reserved.

Accepted Article

We have demonstrated that responder counts can differ between groups even if there is no treatment response heterogeneity present at all within each of the groups. The only factor we have manipulated in our simulation is the sample mean. It is incongruous for a proposed approach to quantifying individual differences in response to merely reflect differences in mean response. Interestingly, this distinction is an important aspect of the work of Geoffrey Rose, especially in how population mean characteristics underpin individual characteristics in public health (Rose, 2001). Ironically, these responder count comparisons could be telling us more about "average medicine" than personalised medicine. Another secondary problem with comparing responder counts between samples is the relatively low statistical precision or "power" of the comparison. This issue has been covered extensively by Snapinn and Jiang (2007).

5.3. Does the use of a measurement error statistic to inform the response threshold help?

It is clear that the approach of responder counting is compromised by within-subjects random variability and group differences in mean response. One question is whether this is so, irrespective of how a certain response threshold is formulated, i.e., whether it was formulated on the basis of a well-defined minimal clinically important response or in comparison to a measurement error statistic like the technical error of measurement, or a combination of both the MCID and a measurement error threshold. We selected a response threshold of 5 ml/kg/min merely for illustrative purposes knowing that all "error-free" responses in both treatment arms are below this MCID. A response threshold should be selected on the basis of clinical importance rather than measurement variability (Cook et al., 2014; 2018). Nevertheless, suppose we select 2 x the "typical error" as our threshold, as several authors have done (Ross et al., 2015). Using the information from our control group, we can calculate typical baseline to follow-up variability by dividing the SD of changes by the square root of 2, giving a typical error of about 3 ml/kg/min. Two times this value gives 6 ml/kg/min. The proportion of "responders" in each group whose change in VO_2max exceeds this threshold is 7%, 29% and 19% for control, intervention 1 and intervention 2,

This article is protected by copyright. All rights reserved.

Accepted Article

respectively. Again, there is a suggestion of "responders" within each sample, when in fact there are no cases which exceed a treatment response *per se* of 6 ml/kg/min in all three samples. Moreover, the differences in responder counts are due almost solely to the differences in mean response between the groups. The responder count differences are not an indication that response heterogeneity differs between the groups, even though the response threshold was selected on the basis of a random within-subjects test-retest or baseline to follow-up statistic.

Whatever the threshold value is, and however it is selected, this would not alter the fact that it is the group differences in mean response, and not response heterogeneity, which are explaining the group differences in "responder counts" in our simulation. This relationship between the difference between two group means and the difference in area under the Normal curve of changes is well known, has mathematical underpinnings, and can be shown by inputting values in to this useful on-line calculator, http://onlinestatbook.com/2/calculators/normal_dist.html. For example, if a mean of -0.11 ml/kg/min and an SD of change of 4.3 are entered into the calculator and the area under the Normal curve (AUC) above the threshold of 5 ml/kg/min is calculated, this AUC is 12%, which agrees reasonably well with the 13% in our simulated control group (Figure 1). If the mean is altered to the 3.6 ml/min/kg and SD = 4.4 ml/kg/min observed for the Intervention 1 group, then the AUC above 5 ml/kg/min becomes 38%, which, again, agrees well with our simulation results (Table 1, Figure 1).

We maintain that the dependency of responder count comparisons on group differences in mean change is not fully appreciated by researchers, even though it is clearly illogical for inferences on individual response heterogeneity (in the context of precision medicine) to be made entirely on the basis of the magnitude of group mean response. For example, Ross et al. (2015) studied what they claimed was the "individual cardiorespiratory fitness response" to different types of exercise interventions (undertaken by different groups). The number of "responders" was found to increase as the group mean response increased. This approach to responder counting has also been adopted by researchers who defined their study topic

This article is protected by copyright. All rights reserved.

Accepted Article

as “trainability” (Williams et al., 2019). Most definitions of this term encompass the notion that individuals differ in their response to the same or similar interventions. Therefore, it is unclear how this approach to responder counting relates to this primary question of interest. We repeat; observing more individual changes in a sample that surpass a certain response threshold when the mean change of that sample is higher has very little to do with individual response heterogeneity in the context of precision medicine.

5.4. Quantifying the probability of being a responder/non-responder in the sample of interest

Approaches have been developed for quantifying the probability that a particular person in the sample of interest is a responder or non-responder (or a “trivial responder”). The context of our review is research and, particularly, an RCT in which parallel samples of participants are measured on a particular study outcome at baseline and at a later follow-up time-point. This context is not the same as clinical decision making on individual patients nor when monitoring individual or team athletes. Therefore, if responders and non-responders can be identified in a particular study, it is important for the researcher to communicate exactly what will be done with this information, especially with governance and ethics in mind (Harriss et al., 2017). According, to the UK Health Research Association (Health Research Authority, 2018), any information communicated to participants about their research results should be in line with the arrangements agreed by the original ethics committee that approved the study. This means that full details about how the researcher feeds back information and advice to a non-responder or adverse responder should be transparent in the ethics approval process and be present on any participant information sheet.

Bonafiglia et al. (2019) cited the paper by Swinton et al. (2018) in order to attach a probability interval around each participant’s response in the intervention sample itself. For example, assuming a large sample, Swinton et al, (2018) reported that an interval of 95% width is calculated by response ± SDchange for control x 1.96. This 1.96 multiplier can be replaced by values from the t distribution for smaller sample sizes. In Table 2, we show the results of applying the similar approach reported by Bonafiglia et al. (2019) to our simulated

This article is protected by copyright. All rights reserved.

Accepted Article

data. In keeping with their approach, we set the response threshold to 1 MET (3.5 ml/kg/min). Each individual response interval was calculated according to the equation presented by Bonafiglia et al. (2019), whereby Individual 95% CI = Response estimate ± (1.96 × TE), The typical error term (TE) is SDchange in the control group divided by √2 = 4.32/√2 = 3.06. Note that TE itself was used by Bonafglia et al. (2019) rather than the SDchange advised by Swinton et al. (2018). Nevertheless, in each study group, we counted the number of responders, "uncertain responders", and adverse responders on the basis of each individual's whole confidence interval being higher, overlapping or lower than the response threshold.

We can compare the responder counts presented in Table 2 with what we would be expecting already knowing the exact nature of our simulated data. For example, we already know that the mean treatment effect for intervention 1 is 3.6 ml/kg/min and there that is no individual heterogeneity in treatment response in this study group. When the response threshold is selected to be 3.5 ml/kg/min, we would, therefore expect about 50% of the intervention 1 participants to be above this threshold and 50% of the sample responses to be below this threshold. This is because the threshold is close to the mean treatment effect. Nevertheless, we do not observe these expected counts because in Table 2, only about 8% of the participants have a response that is above 3.5%, according to the approach reported by Bonafiglia et al. (2019). This approach is clearly erroneous for the robust identification of responders and non-responders.

Unfortunately, the approach reported by Bonafiglia et al. (2019) is also sensitive to group differences in mean response, which compromises its usefulness for indicating response heterogeneity or group differences in response heterogeneity. The fact is that only the mean treatment effect differs between groups in our simulation. Again, when a mean treatment effect is different between samples, then naturally so is the number of people in each sample whose response is higher or lower than a certain threshold value, and this is also the case here when individual confidence bands are estimated for each individual response.

This article is protected by copyright. All rights reserved.

Accepted Article

5.5. Sample response “dichotomania”

There are some other important factors to consider when a researcher is interested in using response thresholds on ratio or interval data, to categorise sample participants as responders or non-responders. Most importantly, the act of converting measurements on a continuous ratio or interval scale into a binary (response/no response) variable has received much criticism amongst statisticians, some of whom have labelled this procedure as “dichotomania” (Senn, 2005). Besides the issue of poorer statistical power for dichotomisation vs analysis of the original continuous data, dichotomisation leads to problems in adjusting for baseline differences between study groups. Senn (2005) also showed how some responder threshold definitions lead to illogical and inconsistent labelling of a “responder”, especially if these definitions are based on multiple outcomes, e.g. both systolic and diastolic blood pressures, and are dependent on the initial status of the outcome, e.g. being in a higher hypertensive category than a lower category and/or using a percentage change as the response threshold.

Lastly, if one is interested in designing studies to inform precision medicine in general, one needs to question the efficiency and utility of identifying responders and/or non-responders merely in the study sample, even if this identification process was robust. For example, if 4 people (10%) from a sample of 40 people who received a certain exercise intervention were found to be “non-responders”, is the researcher obliged to undertake further studies on these 4 people to see what does “work” for them? Such an approach could be very costly relative to the scope of the research impact. In this respect, we believe that researchers seem to be confusing empirical trials of effectiveness in a research context with exercise performance support work, e.g. sports science support or coaching. Again, the most relevant question in a research context is not necessarily which individuals in a sample itself are responders/non-responders, but what are the chances a new person from the population of interest is a responder or a non-responder, that is, statistical inference, and not necessarily participant identification in the particular study sample itself.

This article is protected by copyright. All rights reserved.

Accepted Article

6. Estimating the proportion of responders in a population of interest.

From the arguments presented above, we maintain that identifying people from observed values in a single or multiple intervention sample as “responders” or “non-responders” is fraught with pitfalls, Although there are approaches for observing individual change for a person in an intervention/treatment sample and estimating confidence intervals for their “true” individual change, we have shown that the observed change itself can be contaminated by within-person random variability between baseline and follow-up measurements. Also, we do not think that robust conclusions can be derived by comparing responder or non-responder counts between different samples because such comparisons lack statistical power and may merely be proxies for sample differences in mean response (Senn, 2005; 2015; 2018; Snappin and Jang, 2007). We, therefore, favour approaches that do not involve the identification of responders or non-responders in the particular sample(s) of interest, but estimate the proportion of responders or non-responders in the population of interest. An analogous estimation would be the chance that any new person from the population of interest would be a responder or not.

We maintain that an estimation of how many people in a population of interest who may benefit or not from an intervention can be useful. Approaches for this notion have been forwarded recently by Swinton et al. (2018) and Hopkins (2018). Importantly, these approaches involve the SDir, directly accounting for the random within-subjects variability that is present. Essentially, this SDir is considered a parameter for the distribution of true responses in the population of interest alongside the mean treatment effect (Figure 2). Then the proportion of people predicted to be above or below a certain response threshold is estimated using the characteristics of the Normal distribution. Again, there are online calculators for this step like the one we mentioned in section 5.3, as well as dedicated spreadsheets (Swinton et al., 2018). Only with this approach, does one get close to what was actually defined in our simulated datasets (Table 3). For example, the mean treatment response for intervention 1 is 3.6 ml/kg/min. Let us assume an MCID of 3.5 ml/kg/min (Bonafiglia et al., 2019). Because the mean intervention response is very similar to the

This article is protected by copyright. All rights reserved.

Accepted Article

selected MCID of 3.5 ml/kg/min, and because the SD for true individual response differences is small (0.83 ml/kg/min), it is not surprising that just over half (55%) of all people in the population are estimated to be responders and 45% are trivial responders, with no adverse responders present. For intervention 2, the SD for true individual responses is 0.41 ml/kg/min and the mean treatment effect is 2.2 ml/kg/min. Therefore the number of responders in the population of interest above an MCID of 3.5 ml/kg/min is estimated to be zero for intervention 2, with zero people being lower than the adverse response threshold of -3.5 ml/min. Therefore, everyone's (100%) response in the population of interest is expected to be trivial for intervention 2. This is of course what we simulated; a mean intervention response of approximately 2 ml/kg/min and no individual differences in response.

Hopkins (2018) suggested a similar approach to that of Swinton et al. (2018). Confidence intervals for these proportions are best derived using bootstrapping (Swinton et al., 2018), preferably the bias-corrected and accelerated bootstrap. However, bootstrapping the whole analytical process involving such a relatively complex linear mixed model incorporating baseline values of the outcome and perhaps other covariates can be computer-intensive for this standard deviation estimation problem. Analytic formulae are available (Mathur and VanderWeele, 2019), but are not robust when the proportion is <0.15 or >0.85. Note that these formulae were derived for application to meta-analyses, but are directly transferable to deriving confidence intervals for proportions of individual responders, rather than proportions of individual studies. We maintain that only these approaches, which use the SDir, give estimated population proportions that are relatively unbiased.

Summary

We have followed up our earlier review on this research topic by highlighting some additional pitfalls in the analysis of individual physiological responses to an intervention or treatment. We have focussed particularly on the act of counting the participants in a study group whose individual response is above or below a certain response threshold deemed to be important. Before doing this, researchers need to ask themselves the following sets of questions;

This article is protected by copyright. All rights reserved.

Accepted Article

1. What is the goal in identifying research participants as responders, non-responders, or adverse responders? Are these people to be followed up with further study? Who will fund such studies? How should this information be fed back to participants and what will the participants' likely reactions be?
2. What is the response threshold that is clinically or practically important? Given that such a threshold might not coincide with the minimal detectable change (as indicated by a measurement error statistics), how should this threshold be rationalised? Can a response threshold be formulated in relation to a robust anchor of morbidity and/or mortality or can it be rationalised on the basis of the fraction of a between-subjects standard deviation?

Once a researcher is comfortable that these questions have been answered, we recommend that the approaches of Swinton et al. (2018) and Hopkins et al. (2018) are followed because these approaches use the SDir and infer to a population of interest rather than the study sample participants. Importantly, the approaches by Swinton et al. (2018) and Hopkins (2018) were the only ones that fully reflected the underlying characteristics of our data simulations. The approaches based on counting responders in each sample of interest are biased relative to the "truth" of our simulation. This bias is not resolved by selecting a response threshold based on measurement error or by calculating confidence intervals for individual response values.

This article is protected by copyright. All rights reserved.

Accepted Article

References

Abe T, Dankel SJ, Buckner SL, Jessee MB, Mattocks KT et al. (2018). Short term (24 hours) and long term (1 year) assessments of reliability in older adults: can one replace the other? *J Aging Res Clin Practice* 18, 82-84

Atkinson G, Batterham AM (2015) True and false interindividual differences in the physiological response to an intervention. *Exp Physiol* 100, 577-88.

Atkinson, G. and AM. Batterham (2017) *The Impact of Random Individual Differences in Weight Change on the Measurable Objectives of Lifestyle Weight Management Services. Sports Med* 47, 1683-1688.

Atkinson G, Batterham AM, and Williamson P (2018). *Comments on "Predictors of Change in Physical Function in Older Adults in Response to Long-Term, Structured Physical Activity: The LIFE Study". Arch Phys Med Rehab* **99**, 408.

Atkinson G, Nevill AM (1996). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 26, 217-238

Atkinson G, Williamson P, and Batterham AM (2018) *Exercise training response heterogeneity: Statistical insights. Diabetologia* **61**, 496-497.

Bell ML, Fiero M, Horton NJ, Hsu CH. (2014). Handling missing data in RCTs; A review of the top medical journals. BMC Med Res Meth 14, doi: 10.1186/1471-2288-14-118.

Bonafiglia T, Ross R, Gurd BJ. (2019). The application of repeated testing and monoexponential regressions to classify individual cardiorespiratory fitness responses to exercise training. *Eur J Appl Physiol* https://doi.org/10.1007/s00421-019-04078-w

Campbell MJ and Machin D. (1993). Medical Statistics: A Common-sense Approach, 2nd edition. Chichester: Wiley.

This article is protected by copyright. All rights reserved.

Accepted Article

Cook JA, Hislop J, Adewuyi TE, *et al*. (2014). Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Assess* 18, 1-175. doi:10.3310/hta18280 pmid:24806703

Cook JA, Julious SA, Sones W, Hampson LV, Hewitt C, Berlin JA *et al.* (2018). DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial *BMJ* 363, k3750

Cortés J, González JA, Medina MN et al. (2018). Does evidence support the high expectations placed in precision medicine? A bibliographic review. *F1000Research* 7, 30

De Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. (2006). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 4, 54.

Hammond BP, Stotz PJ, Brennan AM, Lamarche B, Day AG, and Ross, RA. (2019). Individual Variability in Waist Circumference and Body Weight in Response to Exercise. *Med Sci Sports Exerc* 51, 315–322.

Harrell, F. (2018). Viewpoints on Heterogeneity of Treatment Effect and Precision Medicine. http://www.fharrell.com/post/hteview/

Harriss DJ, Macsween A and Atkinson G. (2017). Standards for Ethics in Sport and Exercise Science Research: 2018 Update. *Int J Sports Med* 38, 1126-1131.

Health Research Authority (2018). Ending Your Project. https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/ending-your-project/

Hopkins WG (2018) Design and analysis for studies of individual responses. *Sportscience* 22, 39-51. (sportsci.org/2018/studyir.htm)

This article is protected by copyright. All rights reserved.

Accepted Article

Hopkins WG (2015). Individual responses made easy. *J Appl Physiol*, DOI: 10.1152/japplphysiol.00098.2015. https://www.physiology.org/doi/full/10.1152/japplphysiol.00098.2015

Mathur MB, VanderWeele TJ. (2019). New metrics for meta-analyses of heterogeneous effects. Statistics in Medicine. 38, 1336-42.

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG (2010). CONONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 340, c869 doi: 10.1136/bmj.c869

Morris TP, White IR, Crowther MJ (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. 1–29, https://doi.org/10.1002/sim.8086

Newman DA and Sin H-P (2009). How Do Missing Data Bias Estimates of Within-Group Agreement? *Org Res Meth* 12, 113-147

Prud'Homme D, Bouchard C, LeBlanc C, Landry F, Fontaine E (1984) Sensitivity of maximal aerobic power to training is genotype-dependent. *Med Sci Sports Exerc* 16, 89-93.

Rose G. (2001). Sick individuals and sick populations. *Int J Epidemiol* 30: 427–432

Ross R, de Lannoy L, Stotz PJ. (2015). Separate effects of intensity and amount of exercise on interindividual cardiorespiratory fitness response. Mayo Clin Proc. 90, 1506–14.

Senn SJ. (2005). Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. In Proceedings of the International Statistical Institute, 55th Session, Sydney.

Senn S. (2018). Statistical pitfalls of personalized medicine. *Nature* 563 (7733), 619

This article is protected by copyright. All rights reserved.

Accepted Article

Senn S. (2015). Mastering variation: variance components and personalised medicine. *Stat Med.* 35, 966-77.

Snappin MS and Jang Q (2007). Responder analyses and the assessment of a clinically relevant treatment effect. *Trials* 8:31, https://doi.org/10.1186/1745-6215-8-31

Swinton PA, Hemingway BS, Saunders B, Gualano B and Dolan E (2018). A Statistical Framework to Interpret Individual Response to Intervention: Paving the Way for Personalized Nutrition and Exercise Prescription. *Front. Nutr*. 5:41. doi: 10.3389/fnut.2018.00041

Vickers A. (2005). Analysis of Variance Is Easily Misapplied in the Analysis of Randomized Trials: A Critique and Discussion of Alternative Statistical Approaches. *Psychosom Med* 67, 652–655

White IR, Horton NJ, Carpenter J, Pocock SJ. (2011). Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ 342, 910-2. doi: 10.1136/bmj.d40.

White IR, Thompson SG. (2005) Adjusting for partially missing baseline measurements in randomized trials. Statistics in Medicine. 2005;24, 993-1007. doi: 10.1002/sim.1981.

Williams VJ, Gurd B, Bonafiglia JT et al. (2019). A multi-center comparison of VO2peak trainability between interval training and moderate intensity continuous training. *Front. Physiol*. 10:19. doi: 10.3389/fphys.2019.00019

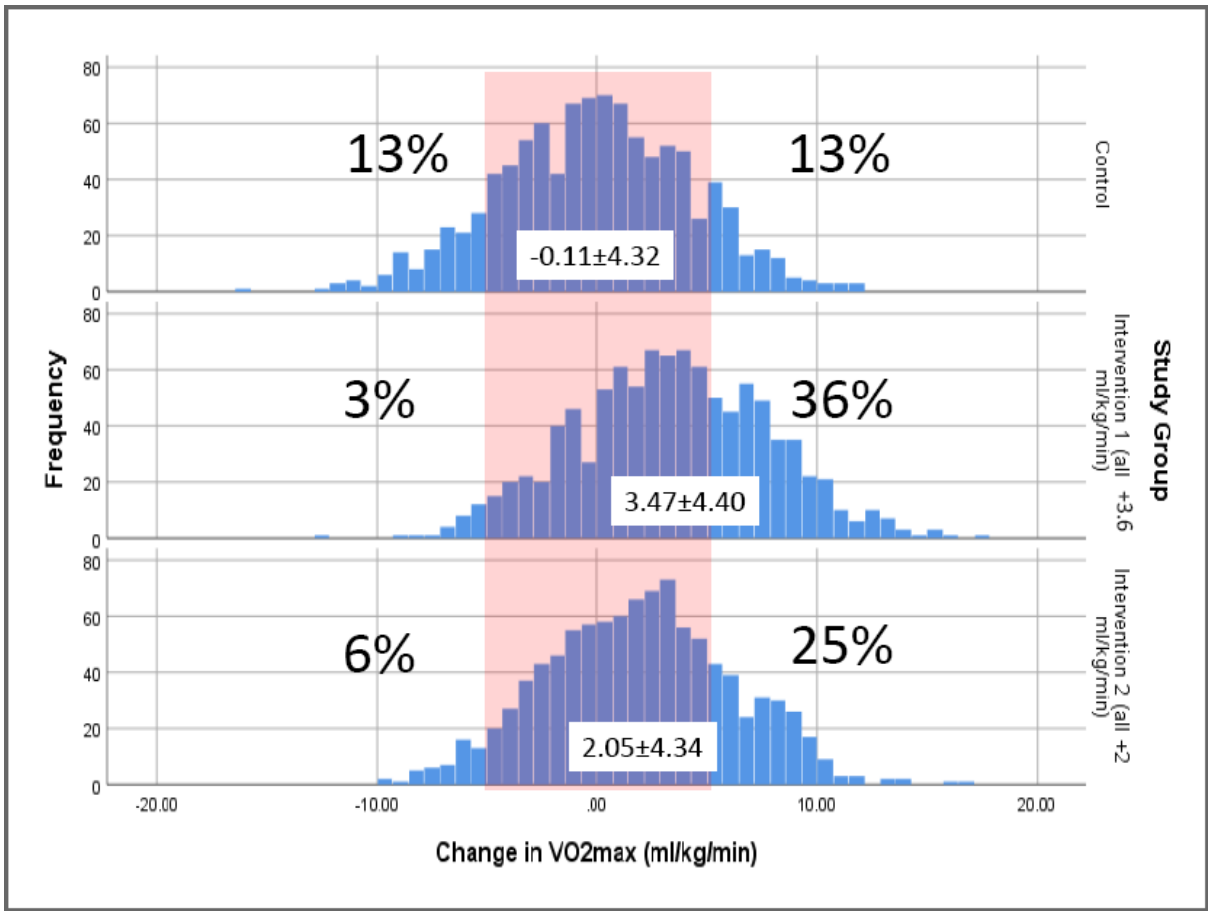
Williamson PJ, Atkinson G, Batterham AM (2017). *Inter-Individual Responses of Maximal Oxygen Uptake to Exercise Training: A Critical Review. Sports Med* 47, 1501-1513.

Zar JH. (1999). *Biostatistical Analysis*. Prentice Hall, New Jersey.

This article is protected by copyright. All rights reserved.

Accepted Article

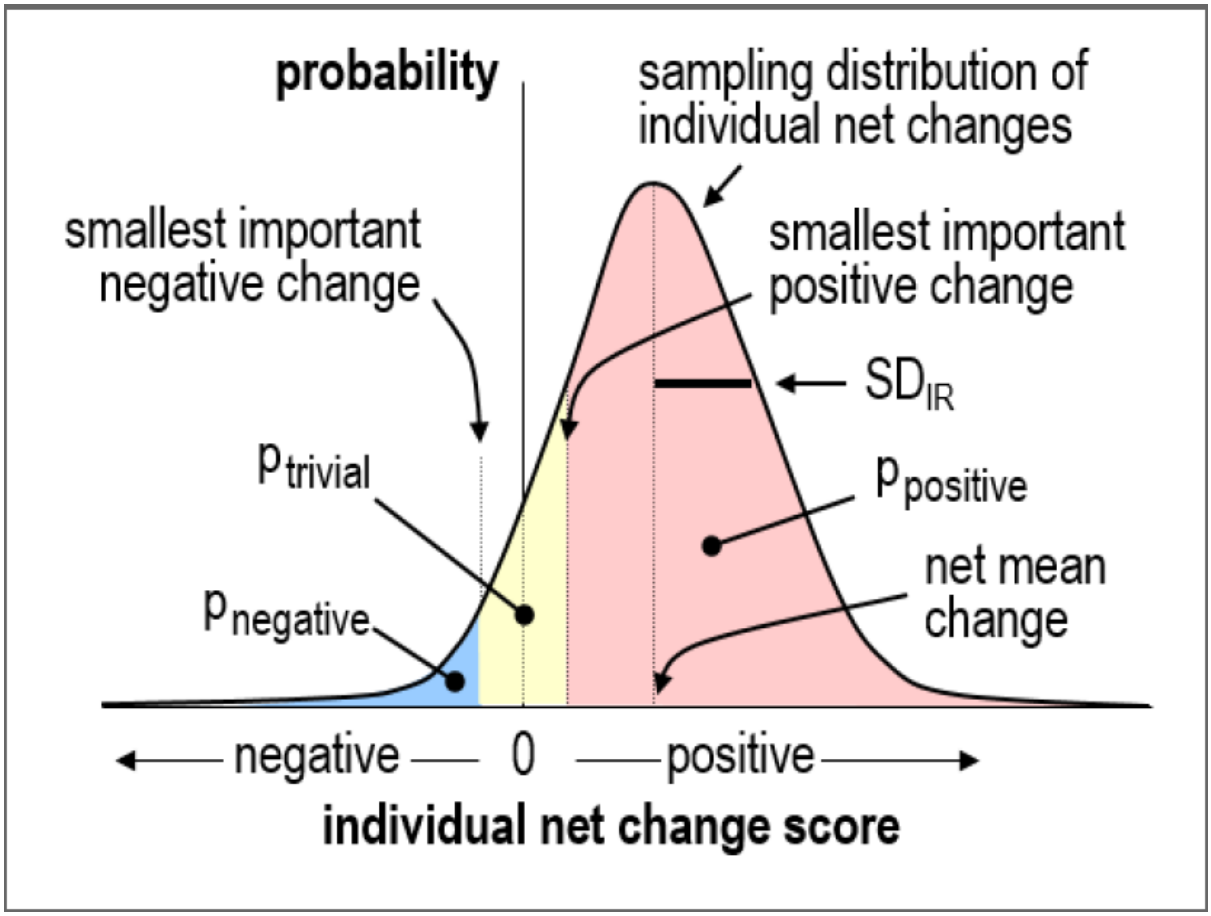
Figure 1. The distributions of individual changes for each of the three simulated study samples, together with the proportion of "responders", "trivial responders" and "adverse responders" in each sample. The MCID was an improvement in VO2 peak of 5 ml/kg/min and the MCID for adverse response was -5 ml/kg/min.



This article is protected by copyright. All rights reserved.25

Accepted Article

Figure 2. The hypothetical distribution of responses (free from random within-subjects variability) for a population of interest (Hopkins 2018). SDir = the SD for individual response heterogeneity. The net mean change is the baseline-adjusted and control group adjusted mean treatment effect for the sample. P = proportion of the population of interest.



This article is protected by copyright. All rights reserved.26

Accepted Article

Panel 1. Questions and Answers about between-group response variance comparisons and the SDir

1. How should a negative SDir be interpreted?

A negative sample SDir could result from several isolated factors or a combination of a number of factors in a similar way to a sample mean treatment effect that is negative (larger change in control group than in intervention group). First the population SDir may be negative. The confidence interval for the SDir will be informative in this respect, especially if the upper confidence limit itself is also negative. Nevertheless, the most likely explanation for a negative SDir is sampling error combined with a population SDir that is small. Again, the confidence interval for the SDIr will be informative here. Just as these underlying factors do not necessarily compromise the calculation of a mean treatment effect, these explanations also do not compromise the calculation of the SDir.

2. What if there are systematic changes in the mean for the control group?

A population change in the control group due to, for example, learning effects does not affect the SD of change in the control group. This fact is inherent in the mathematical underpinnings of the SD. Nevertheless, such a systematic mean change in the control group is important for a robust estimate of the mean treatment response (intervention mean change – control mean change). Recently, Hammond et al. (2019) speculated that the SD of change comparison approach is "questionable" when "*the control group is contaminated with other sources of variability, beyond that of which is random*". This statement denotes a lack of appreciation of the difference between systematic and random sources of variance. For example, if we added 1 ml/kg/min to all the true follow-up values in our control group, the SD of change would be unaffected, and this will always be so because of the term; (sample mean – observed value) within the equation for the standard deviation. This is why the

This article is protected by copyright. All rights reserved.

Accepted Article

random variability we refer to is often also termed “residual” variance, meaning the random variation that is leftover once any systematic sources of change have been partitioned. This appreciation of how a systematic constant does not affect the variance is “Rule 2” for the variance, which is covered at: http://www.kaspercpa.com/statisticalreview.htm

3. How would poor trial design affect the SDir?

Just as any analysis of mean treatment response does not rectify a poor design, then any analysis of response heterogeneity cannot retrieve a poorly designed RCT. Researchers should endeavour to design, analyse and report their RCTs in accordance with best practice guidelines like CONSORT (Moher et al., 2010).

4. How would loss of participants to follow-up affect the SDir?

Loss of data at follow-up is a common problem in randomised controlled trials, but there are principled approaches for dealing with this problem (Bell et al., 2014; White et al., 2005; 2011). Any partially missing data (for outcome or covariates), or participant withdrawal between baseline and follow-up, should be considered carefully. Such loss of data does not necessarily compromise a trial – again, irrespective of whether the mean response or response heterogeneity is of primary interest (Panel 1). The extent of any resulting problems would depend, in part, on the missing data mechanism. There is no reason to believe that data assumed to be missing at random would bias the estimate of a standard deviation of change in the exploration of response heterogeneity (Newman and Sin, 2009). The variance of change is the statistic used in the calculation of the SDir and the variance is not biased by sample size (Zar, 1999). It is good practice to pre-specify a principled approach to addressing missing data in the statistical analysis plan for the trial (see e.g., Belle et al., 2014; White et al., 2005; 2011).

This article is protected by copyright. All rights reserved.

Accepted Article

Table 1. Various mean (SD) measurements for the three parallel groups in the hypothetical study. In each group, there are no individual differences in response, merely differences between groups in the constant change value for each participant, plus random amounts of within-subjects variability between baseline and follow-up. Control = 0 change, Intervention 1 = 3.6 ml/kg/min change, Intervention 2 = 2.0 ml/kg/min change. Random within-subject variability was added to each true value of each participant in each group so that the mean (SD) random variability added was approximately 0 (3) ml/kg/min. These errors were Normally distributed. The correlation coefficient between baseline and follow-up values was 0.9 for each group. The response threshold was 5 ml/kg/min. This observed change is not from the ANCOVA model, i.e. not baseline and control group adjusted.

Measurement (ml/kg/min)	**Control Group (n=1000)**	**Intervention 1 (n=1000)**	**Intervention 2 (n=1000)**
True baseline mean (SD)	34.8 (7.8)	35.0 (8.2)	35.1 (8.3)
True follow-up mean (SD)	34.8 (7.8)	38.6 (8.2)	37.1 (8.3)
True change for all participants	0	3.6	2.0
Observed baseline mean (SD)	34.8 (8.2)	35.0 (8.8)	35.0 (8.9)
Observed follow-up mean (SD)	34.7 (8.3)	38.5 (8.8)	37.0 (8.9)
Observed change (SD)*	-0.1 (4.3)	3.5 (4.4)	2.0 (4.3)
Sample responder counts			
No. of responders	127	363	249
No. of "adverse" responders	126	27	57
No. of trivial responders	747	610	694

This article is protected by copyright. All rights reserved.

Accepted Article

Table 2. Number of responders, "uncertain" responders, and adverse responders in each of the three study groups according to the individual confidence interval approach reported by Bonafiglia et al. (2019). According their approach, the response threshold was set at 1 MET (3.5 ml/kg/min). Each individual response interval was calculated according to Individual 95% CI = Response estimate ± (1.96 × TE), where TE is SDchange in the control group divided by $\sqrt{2}$ = 4.32/$\sqrt{2}$ = 3.06.

	Control Group	**Intervention 1**	**Intervention 2**
Responders	11 (1.1%)	83 (8.3%)	38 (3.8%)
Uncertain	974 (97.4%)	916 (91.6%)	957 (95.7%)
Non-responders	15 (1.5%)	1 (0.1%)	5 (0.5%)

This article is protected by copyright. All rights reserved.

Accepted Article

Table 3. Use of the SDir to estimate the proportion of predicted responders, predicted trivial responders, and predicted adverse responders in each of the three populations of interest according to the approach reported by Swinton et al. (2018). The proportion of responders was estimated as the proportion of a Normal curve above the thresholds of 3.5 and 5.0 ml/kg/min when the Normal curve has parameters of mean treatment effect (from baseline and control group adjusted ANCOVA model) and SD = the "true" SD for response heterogeneity.

Response threshold	**Label**	**Intervention 1** Mean change = 3.6 SDIR = 0.83	**Intervention 2** Mean change = 2.2 SDIR = 0.41
3.5 ml/min/kg	**Responders**	55%	0%
	Trivial	45%	100%
	Non-responders	0%	0%
5.0 ml/min/kg	**Responders**	5%	0%
	Trivial	95%	100%
	Non-responders	0%	0%

This article is protected by copyright. All rights reserved.

Accepted Article

Author contributions

This work took place at Teesside University. All authors (GA, AMB, PW) contributed to the following aspects of the study:

1. Conception or design of the work

2. Acquisition, analysis, or interpretation of data for the work

3. Drafting of the work or revising it critically for important intellectual content

4. Approved the final version of the manuscript

5. Agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

6. All persons designated as authors qualify for authorship, and all those who qualify for authorship are listed

This article is protected by copyright. All rights reserved.