Review

Experimental Physiology

True and false interindividual differences in the physiological response to an intervention

Greg Atkinson and Alan M. Batterham

Health and Social Care Institute, School of Health and Social Care, Teesside University, Middlesbrough, UK

New Findings

- **What is the topic of this review?**
  In 'personalized medicine', various plots and analyses are purported to quantify individual differences in intervention response, identify responders/non-responders and explore response moderators or mediators.
- **What advances does it highlight?**
  We highlight the impact of within-subject random variation, which is inevitable even with 'gold-standard' measurement tools/protocols and sometimes so substantial that it explains all apparent individual response differences. True individual response differences are quantified only by comparing the SDs of changes between intervention and comparator arms. When these SDs are similar, true individual response differences are clinically unimportant and further analysis unwarranted.

Within the 'hot topic' of personalized medicine, we scrutinize common approaches for presenting and quantifying individual differences in the physiological response to an intervention. First, we explain how popular plots used to present individual differences in response are contaminated by random within-subject variation and the regression to the mean artefact. Using a simulated data set of blood pressure measurements, we show that large individual differences in physiological response can be suggested by some plots and analyses, even when the true magnitude of response is exactly the same in all individuals. Second, we present the appropriate designs and analysis approaches for quantifying the true interindividual variation in physiological response. It is imperative to include a comparator arm/condition (or derive information from a prior relevant repeatability study) to quantify true interindividual differences in response. The most important statistic is the SD of changes in the intervention arm, which should be compared with the same SD in the comparator arm or from a prior repeatability study in the same population conducted over the same duration as the particular intervention. Only if the difference between these SDs is clinically relevant is it logical to go on to explore any moderators or mediators of the intervention effect that might explain the individual response. To date, very few researchers have compared these SDs before making claims about individual differences in physiological response and their importance to personalized medicine.

(Received 30 January 2015; accepted after revision 27 March 2015; first published online 30 March 2015)

Corresponding author G. Atkinson: Health and Social Care Institute, School of Health and Social Care, Parkside West, Teesside University, Middlesbrough, Tees Valley TS1 3BA, UK. Email: g.atkinson@tees.ac.uk

DOI: 10.1113/EP085070

A hypothetical scenario: common yet compromised

A hypothetical physiologist is aware of the growing importance of personalized or precision medicine (McCarthy, 2015) and is, therefore, interested in examining how the responses of diastolic blood pressure to an exercise training programme might vary for the individual participants in her study sample. The physiologist knows that the health benefits of physical activity are typically evaluated with reference to the mean response of a sample; an approach that has been said to 'fail to recognize that there are considerable interindividual differences in responses to any exercise program' (Bouchard *et al.* 2015; p. 2). Therefore, let us see how individual differences in intervention response are typically analysed and reported.

The physiologist manages to recruit a sample of 2000 participants, some with normal and some with raised diastolic blood pressure. Her study design is a two-arm randomized controlled trial, with a 3 month exercise programme implemented between the baseline and follow-up measurements (Fig. 1). After obtaining measurements at baseline (to minimize any reactivity influences, such as resentful demoralization), she randomizes the participants into the intervention (exercise training) arm ($n = 1000$) or the comparator arm ($n = 1000$). Reporting of randomization procedures is fully transparent, and other best-practice aspects of a trial are followed according to the Consolidated Standards of Reporting Trials (Schulz *et al.* 2010). The large sample sizes in each study arm also offer good precision for the estimate of intervention effect size (Batterham & Atkinson, 2001). *A priori*, the physiologist defines the minimal clinically important decrease in diastolic blood pressure as 5 mmHg; around one-quarter of the between-subject SD for this population.

At the end of the study, the physiologist finds that mean (SD) diastolic blood pressure decreased from 78.9 (11.2) to 73.7 (11.4) mmHg in the intervention arm, compared with 79.4 (11.3) and 79.2 (11.6) mmHg for baseline and follow-up in the comparator arm. An analysis of covariance (ANCOVA) model was used to derive the confidence interval for the baseline-adjusted mean reduction in diastolic blood pressure (Vickers & Altman, 2001). This confidence interval was −6.1 to −4.8 mm Hg, indicating a likely clinically important mean effect of the exercise programme on diastolic blood pressure. Nevertheless, the physiologist notes considerable individual variation in the change in blood pressure for the participants in the intervention arm of the study. Therefore, in keeping with similar studies (Church *et al.* 2009; Caudwell *et al.* 2012; Buford *et al.* 2013), the physiologist plots the blood pressure reduction (in order of magnitude) for each participant in the intervention arm (Fig. 2). It can be seen from Fig. 2 that the blood pressure of some participants seems to reduce markedly in response to the intervention, whereas the blood pressure of other participants seems to remain unchanged or even increase at follow-up.

The physiologist is aware of recent reviews on individual 'responders' and 'non-responders' to an intervention (Mann *et al.* 2014). Therefore, she records that 495 individuals (nearly half of the whole intervention sample) are 'non-responders' in that their diastolic blood pressure did not reduce by more than 5 mmHg in response to the exercise training. Consequently and, in keeping with other recent studies in humans (Currie *et al.* 2014; Green *et al.* 2014; Loenneke *et al.* 2014) and animals (Rossi *et al.* 2013), the physiologist examines whether the magnitude of blood pressure response depends on the blood pressure status measured at baseline. She finds a negative correlation of −0.30 (95% confidence interval: −0.36 to −0.24) between baseline blood pressure and change in blood pressure (Fig. 3). She also finds that the mean (SD) diastolic blood pressure at baseline is 81.5 (11.0) mmHg for the 'responders' compared with only 76.3 (10.9) mmHg for the 'non-responders'.

The physiologist concludes that there is substantial individual variation in the blood pressure response to an exercise training programme, with just over half the sample showing a training-induced reduction in blood pressure that is more than the minimal clinically important reduction. The physiologist also concludes that the exercise training is most effective for people who have existing high blood pressure at baseline, which is a common finding in the literature for a number of health-related outcomes (Atkinson *et al.* 2001, 2010; Atkinson & Taylor, 2011; Taylor *et al.* 2010; Atkinson, 2014). Therefore, her study is perceived to have impact on the treatment of people with existing hypertension. In the context of personalized medicine, the researcher then considers various follow-up studies on her participants in keeping with the research design frameworks presented in the literature (Buford *et al.* 2013).

Random error masquerading as response differences

While the above hypothetical scenario is in keeping with real studies and the above findings are consistent with the results of those real studies, all the above inferences that the physiologist made about individual differences in blood pressure response are completely false. Consequently, the proposed follow-up studies designed to explore responders and non-responders are unwarranted. False conclusions were arrived at, even though the physiologist started her study with a robust randomized controlled trial design. There are, in reality,

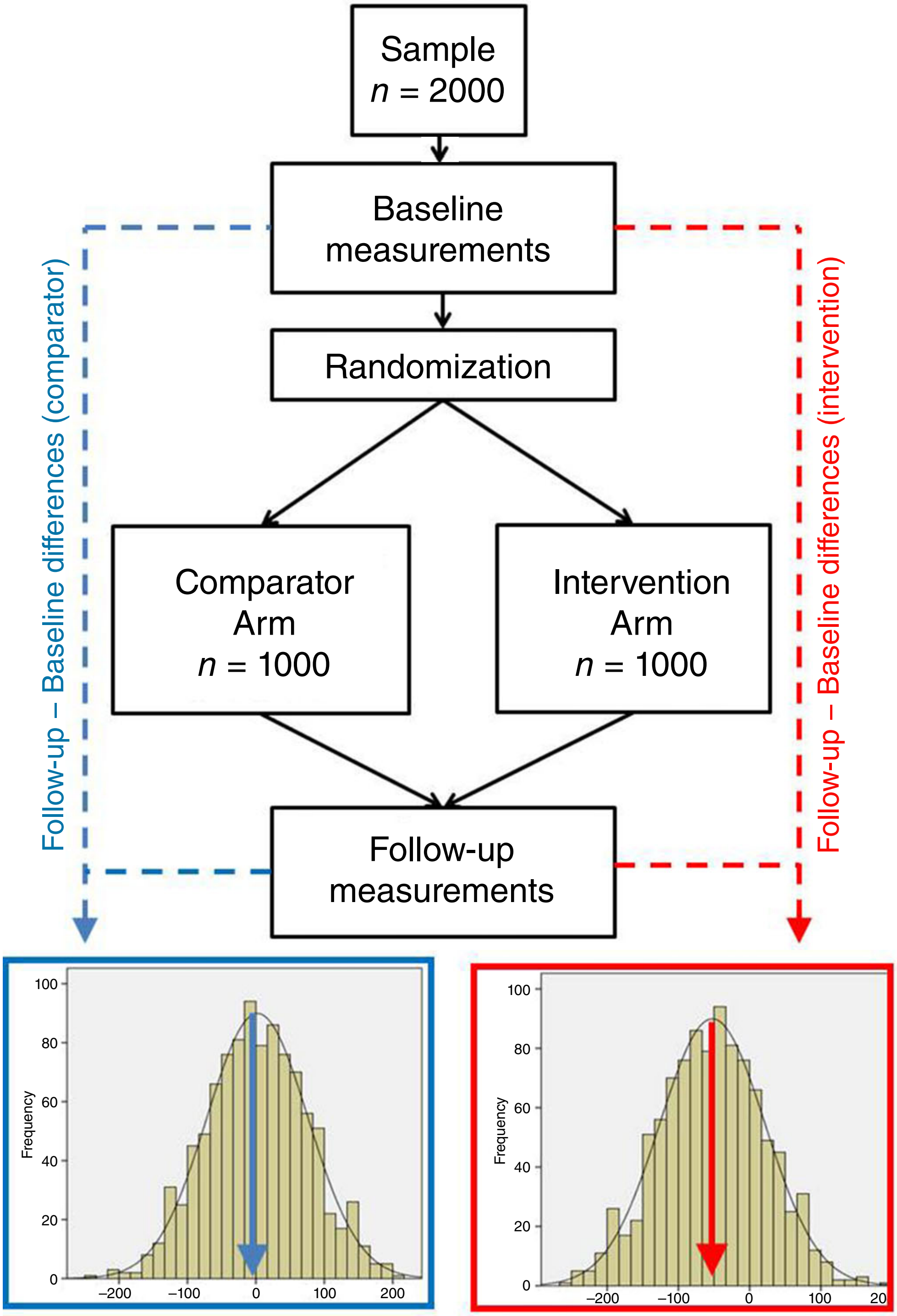


Figure 1. A typical randomized control trial involving an intervention study arm and a comparator arm
Of primary interest is the mean change in the intervention sample/condition *versus* the mean change in the comparator arm. Researchers may proceed to examine the individual responses of each participant in the intervention arm, overlooking the fact that there are also individual differences in the observed difference between baseline and follow-up in the comparator arm of the study (see histograms for the two samples).

no true individual differences in blood pressure response at all in the data that were analysed. The above findings can be explained completely by random within-subject variation from the 'true' blood pressure values at both baseline and follow-up time points (Atkinson & Nevill, 1998) and by the associated artefacts of regression to the mean and mathematical coupling (Pearson, 1896; Atkinson & Taylor, 2011; Taylor *et al.* 2010).

The root cause of the physiologist's inferential error in this example is neglecting to analyse the data from the comparator sample group when assessing the individual differences in response. In a randomized controlled trial, the mean intervention effect represents what happened on average to participants in the intervention group compared with what would have happened if, 'contrary to the fact', they had been in the control group, i.e. the mean change in the intervention group minus the mean change in the control. This basis of the randomized trial indicates that one cannot identify responders and non-responders by inspection of changes in the intervention group alone. For example, we might observe a 5 mmHg reduction in diastolic blood pressure in a participant in the intervention sample. Nevertheless, we cannot legitimately label this person as a 'responder' because we do not know what would have happened to this person if they had been in the comparator arm. What we can do in a parallel group randomized controlled trial, however, is quantify the typical interindividual variability in the intervention response. This quantification is crucial for elucidating whether clinically relevant interindividual differences in response exist at all.

The truth behind the data

The data for the intervention participants used to plot Figs 2 and 3 can be found in the supplementary data S1. A few observations and the sample statistics for these data are also presented in Table 1. These data are randomly generated but are physiologically realistic (Jones *et al.* 2006). Such data simulations are useful for illustrating how statistical biases can be introduced into an analysis approach resulting in incorrect inferences (Taylor *et al.* 2010; Atkinson & Batterham, 2013). First, for all participants in any study, there is a 'true' baseline value of a physiological variable that is free of within-subjects variability (Table 1). There is, of course, between-subject variability in these true baseline values of diastolic blood pressure. The mean (79 mmHg) and SD (10 mmHg) of these true baseline values were similar to those reported by Jones *et al.* (2006) for people being investigated in a hypertension clinic. Some of these people were diagnosed with hypertension and some were deemed normotensive. Importantly, we subtracted exactly 5 mmHg from every participant's true baseline value, which provides the true values of diastolic blood pressure for every participant at follow-up. This consistent subtraction of exactly 5 mmHg from the true baseline data means that there are no individual differences at all in the true reduction in diastolic blood pressure from baseline to follow-up in the data set. Unfortunately, these hypothetical true values, like every physiological variable, are always subject to random within-subject variation when measured in the real world. These random errors have two main components; the

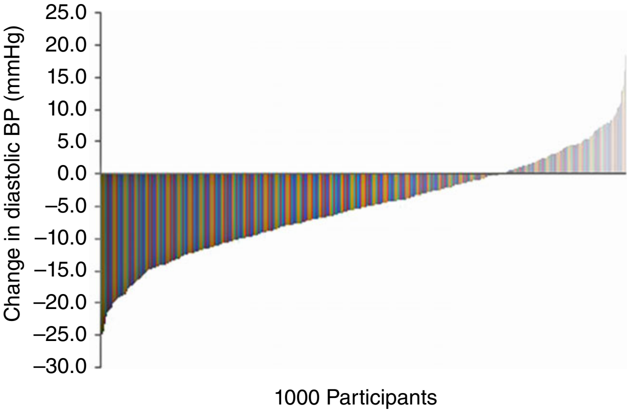


Figure 2. Individual changes in diastolic blood pressure (BP) plotted for a hypothetical sample of 1000 participants
In the data simulation, the 'true' blood pressure of all participants decreased by 5 mmHg. The individual differences in blood pressure change suggested by the figure are due entirely to within-participant variation in baseline and follow-up measurements and regression to the mean.

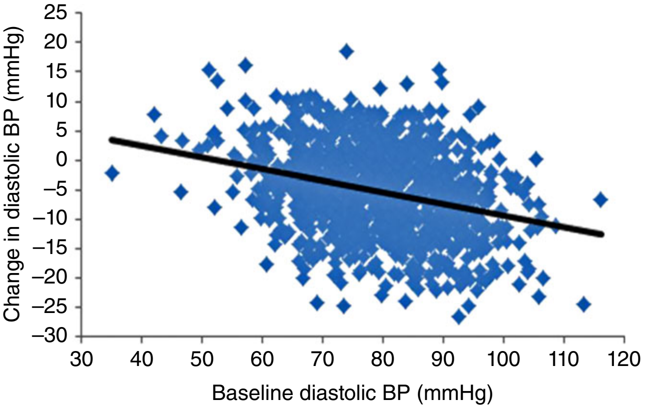


Figure 3. Individual changes in blood pressure plotted against baseline blood pressure
The negative slope so often seen in this type of plot can be due entirely to regression to the mean and mathematical coupling. Participants with a relatively high measured blood pressure at baseline will naturally regress towards the mean so that follow-up measurements are lower, and vice versa for participants with a relatively low blood pressure at baseline. This regression to the mean leads to the artefact of a negative correlation between change and initial value (or any other variable that is correlated with the initial value).

Table 1. Example observations and sample statistics for the simulated data set

Participant number	True baseline	True follow-up	Error in baseline	Error in follow-up	Observed baseline	Observed follow-up	Observed change
1	64.44	59.44	−1.50	−0.77	62.94	58.67	−4.27
2	79.83	74.83	−6.39	−4.61	73.44	70.22	−3.22
3	81.31	76.31	1.22	1.64	82.54	77.96	−4.58
4	66.56	61.56	6.38	11.33	72.94	72.89	−0.06
5	62.33	57.33	5.99	−5.82	68.33	51.51	−16.82
994	75.98	70.98	−5.09	6.38	70.89	77.36	6.48
995	81.46	76.46	−0.77	7.43	80.69	83.89	3.20
996	80.31	75.31	1.27	−8.58	81.58	66.74	−14.85
997	85.15	80.15	−1.31	11.62	83.83	91.77	7.93
998	65.18	60.18	2.42	−7.04	67.59	53.14	−14.45
999	80.76	75.76	2.88	−4.39	83.63	71.36	−12.27
1000	80.88	75.88	1.67	2.80	82.55	78.68	−3.87
Mean	78.79	73.79	0.13	−0.10	78.92	73.69	−5.23
SD	10.25	10.25	5.11	5.12	11.23	11.43	7.39

True values for baseline and follow-up are always subject to random within-participant variation ('error') over time. An apparent change from baseline to follow-up is only an estimate, measured with error, of the true change. In this simulated data set, the true change in diastolic blood pressure for every single participant is a reduction of 5 mmHg. The apparent individual variation in observed change is due entirely to within-participant variation. Regression to the mean can then exert its influence so that it appears that participants with initially high blood pressure (or other phenotypes correlated with baseline blood pressure) show the largest observed change.

technical error from the measurement tool/protocol and the random within-subject biological variation (Box 1). This latter component of within-subject random variation is typically large, even if 'gold-standard' methods, with small short-term technical errors, are employed (Atkinson & Nevill, 1998). Random within-subject variation is especially large if there is considerable time, e.g. 3–6 months, between baseline and follow-up measurements. For example, even though body mass index can be measured very accurately by a trained person, correlations between repeated measurements of body mass index made months apart can be low, i.e. within-subject random variation from biological, behavioural and environmental sources is relatively high (Bayer *et al.* 2011).

Box 1: There is no escape from random within-subject variation

In every study, physiologists make observations on human volunteers or animals. Most of these observations are interval or ratio in nature, meaning that they are measured on a certain continuum or scale. Physiological measurements on humans are never perfectly reproducible over time, especially over relatively long time periods measured before and after a longer-term intervention, such as exercise training or a particular diet. Even if a measurement tool or protocol is associated with no error whatsoever on a short-term test–retest basis and even if the participant or animal is well accustomed to the measurement protocol, there will always be test–retest variability that is attributable to random biological and behavioural fluctuation (Atkinson & Nevill, 1998). It is this random within-subject variability that makes it appear that individuals have responded differently to an intervention, when in fact there might be clinically unimportant interindividual differences in the true response to an intervention.

The next two spreadsheet columns in the data file and Table 1 represent the random error component between the true and observed baseline measurement and between the true and observed follow-up measurement. Therefore, what really varies between individuals is diastolic blood pressure at baseline and the degree of random within-subject variation in both baseline and follow-up measurements, rather than any individual difference in the blood pressure response to the intervention. The random variation is simply added to the true values to obtain the measured values of baseline blood pressure, follow-up blood pressure and the change in blood pressure, and it is these data that are used to plot Figs 2 and 3. The within-subject variation that has been simulated is relatively small in magnitude. For example, the correlation between baseline and follow-up measurements is a 'high' value of 0.79, and the within-subject SD (standard error of measurement or 'typical error') between baseline and follow-up measurements is ∼5 mmHg. If the variation between repeated observations over time during a study is large, which is the case for many physiological variables (Atkinson & Nevill, 1998; Bayer *et al.* 2011), then the apparent, yet false, individual difference in response would also be large.

Therefore, what lies beneath the apparent individual differences in blood pressure response is random within-subjects variation, as illustrated in the control arm histogram in Fig. 1. But the complete explanation also involves the regression to the mean and mathematical coupling artefacts. The influence of these artefacts has been discussed in the context of blood pressure measurements by Taylor *et al.* (2010). If one study arm from a trial (e.g. the intervention sample) is selected and the difference between baseline and follow-up measurements is analysed, then regression to the mean can exert an influence (Taylor *et al.* 2010). The participants in the intervention arm, who had a relatively low or high observed blood pressure at baseline, will regress towards the mean value naturally because of within-subject variation. While it appears that a systematic difference in blood pressure response is there for certain individuals, this difference is an artefact of measurement error being present and the people initially at the extreme tails of the data distribution regressing to the mean at follow-up (Taylor *et al.* 2010). In our hypothetical example, the researcher has erred in thinking that there are individual differences in intervention response in the first place and erred in thinking that there are moderators of this individual difference in response; a wholly unsatisfactory situation in itself. This situation could be worse if the researcher undertakes follow-up studies on the responder and/or non-responder participants (as suggested by Buford *et al.* 2013) because this would be on the basis of observed yet false interindividual differences in response.

Looking before leaping to explore individual responses

Although the above quote from Bouchard *et al.* (2015) essentially assumes that considerable interindividual differences in response exist, this assumption may or may not be true in any particular study. It is imperative that the true individual differences in response, free from the influence of random within-subject variability, are quantified first. Then this true individual difference in response can be appraised for its clinical importance. First, the size of the SD for individual responses and its uncertainty (confidence interval) may be interpreted in relation to the baseline between-subject SD. The typical distribution-based thresholds for interpreting standardized mean changes (0.2, 0.6, 1.2, 2.0 and 4.0 SDs for small, moderate, large, very large and extremely large effects, respectively; Hopkins *et al.* 2009) need to be halved (0.1, 0.3, 0.6, 1.0 and 2.0) for interpreting the magnitude of standardized SDs (Smith & Hopkins, 2011), including those representing individual responses (Hopkins *et al.* 2015). Second, we also advocate assessing the size of the SD for individual responses relative to the mean intervention effect. The SD for individual responses represents the typical variability around the mean intervention effect for a participant in the intervention group (*versus* control). Therefore, the typical overall effect of an intervention on an individual may be summarized as ranging from the mean effect minus SD for individual responses to the mean effect plus SD for individual responses. For example, if the mean effect of an intervention (*versus* control) was a reduction in diastolic blood pressure of 5 mmHg, with an SD for individual responses of 10 mmHg, then the effect of the intervention on a randomly drawn individual from the intervention group in this sample would range typically (±1 SD) from a reduction of 15 mmHg to an increase of 5 mmHg. Assuming a minimal clinically important difference of 5 mmHg, the typical effect of this intervention for the individuals ranges from moderate-to-large benefit to borderline harm. Such a finding would indicate substantial individual responses to treatment, with some individuals benefiting, some getting worse and others for whom the treatment was ineffective. Importantly, only if the true individual response is clinically important is it logical to go on to explore individual moderators or mediators of the treatment effect that might explain the observed individual response.

Researchers have rarely followed the above logical framework when investigating the individual response to interventions, even though the importance of random within-subject variability is clearly demonstrated in several past studies. For example, Church *et al.* (2009) presented plots similar to that of our Fig. 2. These plots are reproduced with permission in Fig. 4. It can be seen that there are apparent individual differences in the change in body mass over a 6 month follow-up period in all three exercise intervention samples. But there is also clear individual variation in the change of body mass in the non-exercising control sample (top left plot). This individual difference is present, even though body mass is a relatively simply variable to measure precisely. As presented in Box 1 and discussed above, the influence of random within-subject variability is inevitable, especially over follow-up periods as long as 3–6 months, and irrespective of the magnitude of random technical errors associated with the measurement method itself.

In view of the logical framework described above, we will concentrate on how true individual differences in response can be quantified robustly. Approaches are available to explore moderators of individual response in a single intervention arm study while adjusting for confounders such as the regression to the mean demonstrated in Fig. 3 (Taylor *et al.* 2010). Nevertheless, in keeping with the above logical framework, such analyses should be undertaken only if it is already known that the true individual difference in response is clinically important in the first place. Given that a comparator arm/prior relevant repeatability study is necessary to derive this knowledge, analysing data solely from the intervention arm can be misleading. Unfortunately, such an approach is common, even in the most recently published studies (Currie *et al.* 2014; Green *et al.* 2014; Loenneke *et al.* 2014).

Quantifying the true magnitude of individual response

The SD describes the 'typical' individual variation between participants measured at the same point in time or 'within participants', i.e. repeated measurements made over time. For example, the SD of the changes in blood pressure observed in the intervention arm of our hypothetical study is 7.4 mmHg. This SD indicates that the blood pressure of ∼68% of the intervention sample changed between 5.3 ± 7.4 mmHg, the mean reduction in blood pressure being 5.3 mmHg in the intervention arm. The blood pressure of the vast majority (95%) of the participants changed between 5.3 ± 1.96 SDs. Therefore, the blood pressure of almost all the participants changed by a value within a range bounded by a reduction of 19.8 mmHg to an increase of 9.2 mmHg. A 95% confidence interval can also be calculated for this SD (Zar, 1999) which, for the sample size of 1000 participants in the intervention arm, is a reassuringly precise 7.1–7.7 mmHg.

Nevertheless, the SD of changes in the intervention arm is not an estimate of the SD in true response *per se*, because it includes the components of within-subject variation and measurement error, which are also present in the control arm of the study (Fig. 1). Given that the true magnitude of response in our data simulation is 5 mmHg for all individuals in our data set, the SD for the true interindividual variation in response should in fact be zero. What is required is parallel information from the comparator arm of the randomized controlled trial. If the study does not have a comparator arm (unusual as this study design would be) then the SD

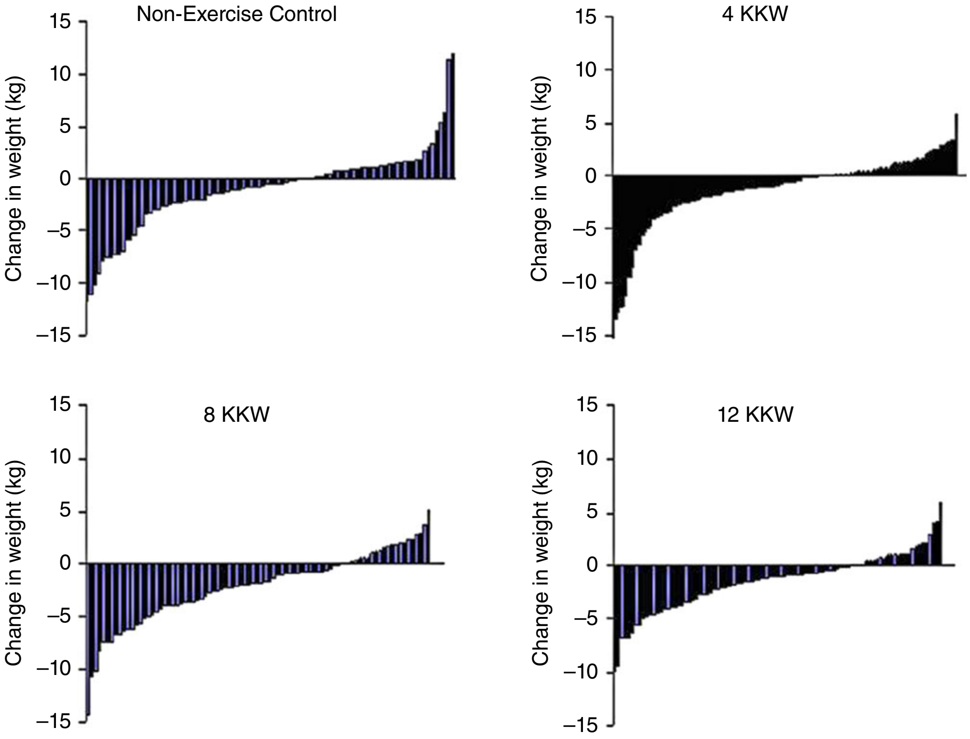


Figure 4. The distributions of body mass change for the four groups studied by Church *et al.* (2009)
Three of the groups undertook physical activity interventions of varying amounts for 6 months. There are clearly individual differences in the observed change in body mass. Nevertheless, very similar amounts of individual variation in change can be seen in the non-exercise comparator sample (top left). The SD of changes can be estimated to be 3.3 kg in every study arm, including the control participants. True individual differences in the response to exercise training are unlikely to be present because the SD of the random within-subject variability (derived from the control sample) is as large as the SD of changes in the intervention samples. KKW - kcal/kg/week

of test–retest change derived from a prior investigation into the reliability of the study outcome may be helpful. Such a reliability study should involve participants drawn from the same population and match the same time scale between baseline and follow-up measurements as the particular intervention study of interest (Atkinson & Nevill, 1998). We note that the reliability of many outcomes selected in health intervention studies can be poor, even with gold-standard measurement devices because random within-participant variation is often larger than device measurement errors (Atkinson & Nevill, 1998). If test–retest reliability is poor, then the SD of the change scores in any intervention arm would need to be relatively large in order to indicate clinically relevant interindividual differences in response *per se*.

For quantifying the typical interindividual difference in response, the appropriate approach involves calculating the difference in SDs of the changes between intervention and control arms (Hopkins *et al.* 2009). This SD represents the typical true interindividual variation in response, adjusted for the influence of random biological variation and measurement error (removal of 'noise'), and can be calculated from:

$$\mathrm{SD_R} = \sqrt{\mathrm{SD_I^2} - \mathrm{SD_C^2}}$$

where $\mathrm{SD_R}$ is the SD of the true individual response to the intervention and $\mathrm{SD_I}$ and $\mathrm{SD_C}$ are the SDs of the pre-to-post change scores for the intervention and comparator arms of the study, respectively. For example, we already know that the SD of the changes is 7.4 mmHg in the intervention sample. Nevertheless, the SD of changes is also 7.4 mmHg in the control arm of our simulated data set. Therefore, the SD of the true individual responses to the intervention is zero, which is exactly what was simulated in the data set. Despite what appears to be a striking amount of individual variation in response when the data from the intervention arm are considered in isolation (a common tactic amongst researchers), the true individual variation in response is zero in our data set.

The SDs of the change scores in each sample can be derived by running Student's unpaired *t* test, with group (intervention or comparator) as the independent variable and the change scores as the outcome. This approach is commonly used anyway to compare the mean difference in change between intervention and comparator arms. Nevertheless, an approach involving ANCOVA, adjusting for any chance imbalance between trial arms at baseline, is known to be superior to Student's unpaired *t* test on change scores (Senn, 1994; Vickers & Altman, 2001). Therefore, a method is required to derive the SD for individual responses from the adjusted change scores in each group. This method involves a linear mixed model, with study arm (intervention or comparator) entered as a fixed effect and a binary-coded additional 'dummy' covariate entered as a random effect (slope), with the baseline value of the outcome also entered as a covariate. The dummy variable allows for extra variance in the change scores in one group *versus* the other. The SD of the true individual response, and its confidence interval, is then derived from the parameter estimate and tests of covariance parameters for the random effect from this model. The advantage of this mixed model approach is that it allows for additional covariates, e.g. sex or age, to be considered. If the individual difference in true response is potentially of clinical importance, then putative moderators or mediators of the intervention effect can be examined by including them in the model. For example, assessing the extent to which the effect of the intervention depended on the baseline value of the outcome (a potential moderator) requires the inclusion of a baseline × group interaction term in the model. This method contrasts with the naive correlation of baseline and change scores in the intervention group alone described in the scenario presented in the first section ('*A hypothetical scenario: common yet compromised*'). If any moderator or mediator variables partly account for the observed heterogeneity in response to the intervention, then the SD for individual responses will be attenuated substantially. Crucially, the approach to analysing individual differences in response outlined in this section includes the data from the comparator sample. This approach has also been described recently by Hopkins (2015), who commented on another relevant paper by Hecksteden *et al.* (2015). A statistician is best consulted to apply the above modelling methods.

A re-analysis of past data

It can be seen in Fig. 4 that individual differences in the measured change in body mass are present in all four study arms, even for the control sample studied by Church *et al.* (2009). In a figure, these authors also reported the mean and 95% confidence intervals for the observed change in body mass for each of the four study samples. These data can be extracted using the Digitizelt software (Köln, Germany) in order to derive the SD of changes in each sample, which can then be compared. Any figure from a published paper can be uploaded into the Digitizelt software, which allows enlargement of the figure for subsequent precise extraction of the minima and maxima on the *x*- and *y*-axes. Following this axis calibration, the raw values of each data point can be marked and extracted in digital formal for subsequent analysis. Once the raw values of the 95% confidence limits were extracted, they were converted to a standard error using published equations based on the *t* distribution (Higgins & Green, 2006). Standard errors can then be converted into SDs by multiplying by the square root of sample size. We extracted these data and estimated that the SD of body mass change was 3.3 kg in all four samples, including the control sample. These similar SDs are in agreement with the similar degree of individual differences in changes that can be seen clearly in Fig. 4. The 95% confidence interval for the SD for true individual responses is approximately −2 to +2 kg for all three interventions. The between-subject SD for all participants at baseline in the study by Church *et al.* (2009) was 11.9 kg. Hence, the true population individual responses could range from being slightly higher in the control sample to being slightly higher in the intervention samples (approximately $2/11.9 = 0.17$ SDs). Therefore, we conclude that the observed individual difference in the response to all three of the exercise interventions studied by Church *et al.* (2009) is practically zero, with true intervention responses that could be negative, trivial or, at most, small (given the magnitude of the confidence interval for the SD). This conclusion is arrived at because the SD of change in the exercise intervention samples is inseparable from the within-subject variation *per se* that has been quantified with the SD of change in the control sample data. Any follow-up analysis to explore potential moderators of the intervention effect to account for the individual differences in response is, therefore, unwarranted. Any follow-up studies on the same participants are also unnecessary and potentially unethical if there are no true individual differences in response to explain in the first place.

Discussion

If the SD of the changes in the intervention group is not substantially larger than that in the control arm, then it can be said that there is negligible interindividual variability in the response to the intervention. Most of the apparent variability in the pre-to-post change would in fact be attributable to random within-subject variation and measurement error ('noise'). The magnitude of the SD of true individual responses should be appraised in terms of clinical importance and not statistical significance. For example, an SD of, say, 2 mmHg, might still be clinically unimportant interindividual variation if the sample mean response happens to be a reduction of 10 mmHg. Conversely, if the SD is 7 mmHg, then the true response for some individuals is clearly so different from the mean response of 10 mmHg as to be clinically important and worth following up with analyses attempting to explain the individual responses. As in many situations, statistics are most useful for informing decisions about clinical/practical significance (Atkinson *et al.* 2008; Hopkins *et al.* 2009).

In some circumstances, the variance in individual responses, and/or its confidence limits, can be negative. This finding would imply greater variability in response in the control group *versus* the intervention group. This phenomenon may be due to the inherent imprecision in the estimation of individual responses with inadequate sample sizes and a large amount of noise and/or caused by the intervention 'homogenizing' values for the outcome variable, thus reducing the SD of the changes relative to the control group. In our experience, the latter is common, for example, in intervention studies involving blood test outcomes (e.g. lipids and inflammatory markers). Nevertheless, it is unlikely that this difference in SDs would be relevant clinically. In this situation, the conclusion would again be reached that there are no substantial individual differences in the true intervention response. Clearly, precise estimation of the SD for individual responses requires relatively large sample sizes and/or multiple measures of the outcome before and after the intervention to overcome large within-subject variability in the changes.

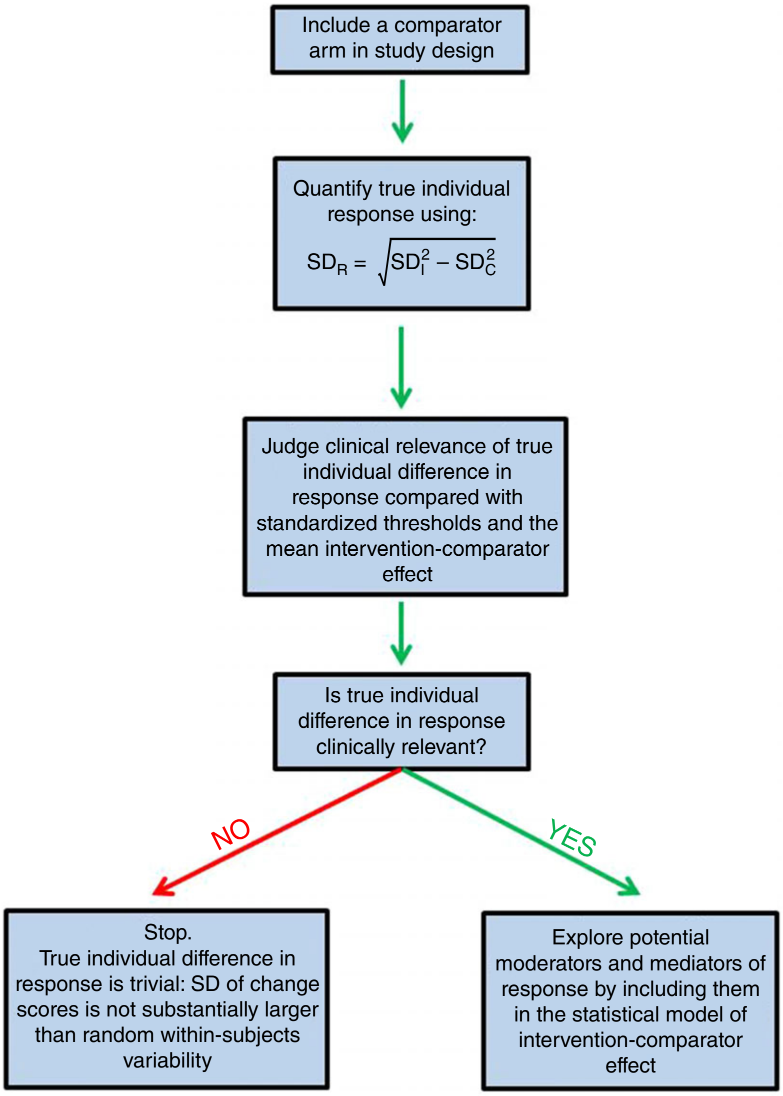
In addition to deriving the SD for individual responses, it is possible to quantify the probability of any individual participant being a responder (change greater than the minimal clinically important difference). The uncertainty in each person's change score may be expressed as a confidence interval using the SD of the change scores in the control group multiplied by the appropriate value from the *t* distribution. The probability that any given individual in the intervention group is a true responder (free from measurement error) can then be derived using reference Bayesian methods. A caveat here is that for many outcomes the uncertainty in an individual's response is greater than the minimal clinically important difference. Hopkins (2015) elaborates on this issue.

The inclusion of data from the comparator arm of the study is of paramount importance. An intervention study design that does not have a comparator arm would be assigned a very low quality score or, more likely, be excluded from any subsequent systematic review/meta-analysis. Consequently, such single-arm studies are now rare, but there are still systematic reviews of these studies in which attempts are made to quantify individual response and explore response moderators (Currie *et al.* 2014; Green *et al.* 2014; Loenneke *et al.* 2014). The analysis of individual responses solely in the intervention arm of a study even though the trial has a comparator arm is wasteful and potentially misleading. Unfortunately, there have been recently published research design frameworks for investigating individual differences in response that are flawed because information from study design comparator arms is not included. For example, Buford *et al.* (2013) presented a conceptual study design framework for evaluating heterogeneity in responsiveness to exercise and identifying potential alternative interventions for individuals perceived to be of 'low sensitivity'. No comparator samples were present in this suggested framework. Sequential design steps and subgroup analyses were suggested by Buford *et al.* (2013)

solely from the magnitude of observed change in the intervention samples. This oversight raises the possibility that participants from one study are being selected for further in-depth study merely on the basis of their personal amount of within-subject variability rather than their true intervention response magnitude. Such further study on selected participants may be wasteful and unethical if it is not reasonably certain at the outset whether the individual difference in true intervention response is genuinely of clinical importance.

In this contribution, we have focused on the parallel-group randomized controlled trial design. We have not discussed the robust quantification of individual differences in response to interventions in crossover studies, though essentially the same principles apply. We are unconcerned here with crossover designs, because our focus is on chronic, relatively permanent, adaptations to health or physiology, rather than on acute effects of short-term interventions that wash out easily and fully. For the latter scenario, Senn *et al.* (2011) and Hecksteden *et al.* (2015) present one approach to the problem of quantifying individual response in crossover studies.

In conclusion, personalized medicine is a potentially important topic for researching health interventions.


Include a comparator arm in study design

Quantify true individual response using:

$$SD_R = \sqrt{SD_I^2 - SD_C^2}$$

Judge clinical relevance of true individual difference in response compared with standardized thresholds and the mean intervention-comparator effect

Is true individual difference in response clinically relevant?

Stop.
True individual difference in response is trivial: SD of change scores is not substantially larger than random within-subjects variability

Explore potential moderators and mediators of response by including them in the statistical model of intervention-comparator effect

Figure 5. Conceptual framework for quantifying true individual differences in the response to an intervention
This framework also includes the appraisal process for judging whether individual differences in response are clinically/practically relevant, as well as the exploration of putative moderators and mediators of individual response. A comparator arm in the study design is the bedrock of this framework.

Nevertheless, it is crucial to know whether the true individual difference in response is clinically important before any attempts are made to identify 'non-responders' and explore reasons for their non-response. The first step in this logical framework (Fig. 5) must involve parallel information from a suitable comparator sample. The further identification and exploration stages must also involve the comparator sample. This framework is based on the contemporary wisdom of looking at the data in both study arms before leaping to conclusions that individual differences in response are important enough for further analysis. Without looking first, there is a danger that personalized medicine is based on a leap of faith rather than reliable evidence derived from sound study designs and appropriate statistical analysis.

References

Atkinson G (2014). Individual differences in the exercise-mediated blood pressure response: regression to the mean in disguise? *Clin Physiol Funct Imaging*, DOI: 10.1111/cpf.12211

Atkinson G & Batterham AM (2013). The percentage flow-mediated dilation index: a large-sample investigation of its appropriateness, potential for bias and causal nexus in vascular medicine. *Vasc Med* **18**, 354–365.

Atkinson G, Batterham A & Drust B (2008). Is it time for sports performance researchers to adopt a clinical-type research framework? *Int J Sports Med* **29**, 703–705.

Atkinson G & Nevill AM (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* **26**, 217–238.

Atkinson G & Taylor C (2011). Normalization effect of sports training on blood pressure in hypertensive individuals: regression to the mean? *J Sports Sci* **29**, 643–644.

Atkinson G, Taylor CE & Jones H (2010). Inter-individual variability in the improvement of physiological risk factors for disease: gene polymorphisms or simply regression to the mean? *J Physiol* **588**, 1023–1024.

Atkinson G, Waterhouse J, Reilly T & Edwards B (2001). How to show that unicorn milk is a chronobiotic: the regression-to-the-mean statistical artefact. *Chronobiol Int* **18**, 1041–1053.

Batterham AM & Atkinson G (2005). How big does my sample need to be? A primer on the murky world of sample size estimation. *Phys Ther Sport* **6**, 153–163.

Bayer O, Kruger H, vonKries R, Toschke AM (2011). Factors associated with tracking of BMI: a meta-regression analysis on BMI tracking. *Obesity* **19**, 1069–1076.

Bouchard CA, Antunes-Correa LMM, Ashley EA, Franklin N, Hwang PM, Mattsson CM, Negrao CE, Phillips SA, Sarzynski MA, Wang P-Y & Wheeler MT (2015). Personalized preventive medicine: genetics and the response to regular exercise in preventive interventions. *Prog Cardiovasc Dis* **57**, 337–346.

Buford TW, Roberts MD & Church TS (2013). Toward exercise as personalized medicine. *Sports Med* **43**, 157–165.

Caudwell P, Gibbons C, Hopkins M, King N, Finlayson G & Blundell J (2012). No sex difference in body fat in response to supervised and measured exercise. *Med Sci Sports Exerc* **45**, 351–358.

Church TS, Martin CK, Thompson AM, Earnest CP, Mikus CR & Blair SN (2009) Changes in weight, waist circumference and compensatory responses with different doses of exercise among sedentary, overweight postmenopausal women. *PLoS ONE* **4**, e4515.

Currie KD, McKelvie RS & MacDonald RJ (2014). Brachial artery endothelial responses during early recovery from an exercise bout in patients with coronary artery disease. *BioMed Res Int* **2014**, 591918.

Green DJ, Eijsvogels T, Bouts YM, Maiorana AJ, Naylor LH, Scholten RR, Spaanderman MEA, Pugh CJ, Sprung VS, Schreuder T, Jones H, Cable T, Hopman MTE & Thijssen DHJ (2014). Exercise training and artery function in humans: nonresponse and its relationship to cardiovascular risk factors. *J Appl Physiol* **117**, 345–352.

Hecksteden A, Kraushaar J, Scharhag-Rosenberger F, Theisen D, Senn S & Meyer T (2015). Individual response to exercise training - a statistical perspective. *J Appl Physiol*, DOI: 10.1152/japplphysiol.00714.2014

Higgins JPT & Green S, eds. (2006). Cochrane Handbook for Systematic Reviews of Interventions. In: The Cochrane Library, Issue 4, John Wiley & Sons, Ltd, Chichester, UK.

Hopkins WG (2015). Individual responses made easy. *J Appl Physiol*, DOI: 10.1152/japplphysiol.00098.2015

Hopkins WG, Marshall SW, Batterham AM & Hanin J (2009). Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc* **41**, 3–13.

Jones H, Atkinson G, Leary A, George K, Murphy M & Waterhouse J (2006). Reactivity of ambulatory blood pressure to physical activity varies with time of day. *Hypertension* **47**, 778–784.

Loenneke JP, Fahs CA, Abe T, Rossow LM, Ozaki H, Pujol TJ & Bemben MG (2014) Hypertension risk: exercise is medicine∗ for most but not all. *Clin Physiol Funct Imaging* **34**, 77–81.

McCarthy M (2015). Obama seeks $213m to fund "precision medicine". *BMJ* **350**, h587.

Mann TN, Lamberts RP & Lamberts MI (2014). High responders and low responders: factors associated with individual variation in response to standardized training. *Sports Med* **44**, 1113–1124.

Pearson K (1896). On a form of spurious correlation which may arise when indiced are used in the measurement of organs. *Proc Roy Soc London Lond* **60**, 489–498.

Rosen L, Manor O, Engelhard D & Zucker D (2006) . In defence of the randomized controlled trial for health promotion research. *Am J Public Health* **96**, 1181–1186.

Rossi NF, Chen H & Maliszewska-Scislo M (2013). Paraventricular nucleus control of blood pressure in two-kidney, one-clip rats: effects of exercise training and resting blood pressure. *Am J Physiol Regul Integr Comp Physiol* **305**, R1390–R1400,

Schulz KF, Altman DG & Moher D; CONSORT Group (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, c332.

Senn S (1994). Testing for baseline differences in clinical trials. *Stat Med* **13**, 1715–1726.

Senn S, Rolfe K & Julious SA (2011). Investigating variability in patient response to treatment – a case study from a replicate cross-over study. *Stat Methods Med Res* **20**, 657–666.

Smith TB & Hopkins WG (2011). Variability and predictability of finals times of elite rowers. *Med Sci Sports Exerc* **43**, 2155–2160.

Taylor CE, Jones H, Zaregarizi M, Cable NT, George KP & Atkinson G (2010). Blood pressure status and post-exercise hypotension: an example of a spurious correlation in hypertension research? *J Hum Hypertens* **24**, 585–592.

Vickers AJ & Altman DG (2001). Analysing controlled trials with baseline and follow up measurements. *BMJ* **323**, 1123–1124.

Zar JH (1999). *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

Additional information

Competing interests

None declared.

Author contributions

Conception and format of the paper was made by GA. Analysis and interpretation of data was made by GA and AMB. GA drafted the article and AMB revised it critically for important intellectual content. Both authors approved the final version of the manuscript, all persons designated as authors qualify for authorship, and all those who qualify for authorship are listed.

Funding

None declared.

Supporting information

Data S1. The data for the intervention participants used to plot Figures 2 and 3.