OPINION ARTICLE

Response heterogeneity: Challenges for personalised medicine and big data approaches in psychiatry and chronic pain [version 1; referees: 1 approved]

Agnes Norbury [1], Ben Seymour[1,2]

[1]Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK
[2]Center for Information and Neural Networks, National Institute of Information and Communications Technology, Osaka, 565-0871, Japan

v1 **First published:** 15 Jan 2018, **7**:55 (doi: 10.12688/f1000research.13723.1)
Latest published: 15 Jan 2018, **7**:55 (doi: 10.12688/f1000research.13723.1)

Abstract

Response rates to available treatments for psychological and chronic pain disorders are poor, and there is a considerable burden of suffering and disability for patients, who often cycle through several rounds of ineffective treatment. As individuals presenting to the clinic with symptoms of these disorders are likely to be heterogeneous, there is considerable interest in the possibility that different constellations of signs could be used to identify subgroups of patients that might preferentially benefit from particular kinds of treatment. To this end, there has been a recent focus on the application of machine learning methods to attempt to identify sets of predictor variables (demographic, genetic, etc.) that could be used to target individuals towards treatments that are more likely to work for them in the first instance. Importantly, the training of such models generally relies on datasets where groups of individual predictor variables are labelled with a binary outcome category – usually 'responder' or 'non-responder' (to a particular treatment). However, as previously highlighted in other areas of medicine, there is a basic statistical problem in classifying *individuals* as 'responding' to a particular treatment on the basis of data from conventional randomized controlled trials. Specifically, insufficient information on the partition of variance components in individual symptom changes mean that it is inappropriate to consider data from the active treatment arm alone in this way. This may be particularly problematic in the case of psychiatric and chronic pain symptom data, where both within-subject variability and measurement error are likely to be high. Here, we outline some possible solutions to this problem in terms of dataset design and machine learning methodology, and conclude that it is important to carefully consider the kind of inferences that particular training data are able to afford, especially in arenas where the potential clinical benefit is so large.

incf This article is included in the INCF gateway.

Open Peer Review

Referee Status: ✓

	Invited Referees 1
version 1 published 15 Jan 2018	✓ report

1 **Greg Atkinson**, Teesside University, UK
Philip J. Williamson, Teesside University, UK

Discuss this article

Comments (0)

Corresponding author: Agnes Norbury (aen31@cam.ac.uk)

Author roles: Norbury A: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Seymour B**: Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Norbury A and Seymour B. **Response heterogeneity: Challenges for personalised medicine and big data approaches in psychiatry and chronic pain [version 1; referees: 1 approved]** *F1000Research* 2018, **7**:55 (doi: 10.12688/f1000research.13723.1)

Grant information: AN and BS are supported by the Wellcome Trust (grant number 097490/Z/11/A to BS).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 15 Jan 2018, **7**:55 (doi: 10.12688/f1000research.13723.1)

Introduction

The proportion of patients who respond to available treatments for psychological and chronic pain disorders is often low. For example, in major depression, roughly 40% of individuals experience a 'clinically significant' response (decrease in symptom severity score above some minimum value) over the course of treatment (e.g. 1,2). Similarly, a recent meta-analysis of available pharmacotherapies for neuropathic pain found estimates of 'number needed to treat' (number of patients needed to be treated to prevent one additional adverse clinical outcome) for *effective* treatments ranged from 4–10, indicating poor response rates[3]. For patients, this often means a lengthy process of cycling through different treatment options, in a sequence that may be significantly influenced by non-clinical concerns (e.g. relative drug cost, therapist availability, local health authority guidelines), and where there may be inadequate data on the safety and effectiveness of switching regimes (e.g. 4). For psychological conditions, this process can be particularly lengthy, given the significant period of time before common pharmacological treatments are expected to take effect (e.g. 4–6 weeks to conclude a particular drug treatment is ineffective,[4]). Together, this results in a substantial burden of suffering and disability to individuals with a diagnosis of these disorders, before (if) an effective treatment option can be found.

It is generally assumed that differential response to a particular treatment across individuals can be at least partially explained by patient heterogeneity within a certain diagnostic category – i.e. that individuals who present to the clinic with similar sets of symptoms may have different underlying pathologies. This seems a particularly reasonable assumption in the case of both mental health disorders and chronic pain, as diagnosis is often made purely on the basis of self-reported symptom checklists, and our lack of knowledge into the aetiology of these conditions means we have little opportunity for differential diagnosis. Indeed, in the case of psychiatric disorders, such as depression, diagnosis can often be made on the basis of directly contradictory symptom reports (e.g. sleeping too much *vs* sleeping too little), and there may be many different ways to meet diagnostic criteria (e.g. 227 possible symptom combinations for major depressive disorder, according to DSM-IV[5]). Similarly, even patients with a diagnosis of a particular pain condition are likely to have distinct patterns of nervous system damage, involving multiple pathways (e.g. 6), and definitions of chronic pain itself can vary dramatically across research groups and clinical centres[7].

Even if we lack insight into pathological mechanisms, it seems likely that if we are able to use some kind of predictive method to direct individuals towards treatments that are likely to be more effective for them – then even a small increase in the resulting response rate could potentially have a large effect on disease burden for individual patients. There has therefore recently been great interest in doing just this for psychiatric data, via application of supervised learning methods to large datasets of individual clinical predictors and treatment response data (see 8 for an excellent recent review of potential clinical advantages and best methodological practice in this area).

The current gold standard approach is firstly to define a set of features and targets for various machine learning algorithms to train on. In this context, features are individual difference variables that may potentially relate to future treatment outcome (clinical, demographic, physiological, genetic, behavioural, etc. information). The target variable (that the algorithm must learn to predict) is usually a binary category label, such as 'responder' or 'non-responder' (whether or not an individual has exhibited symptom improvement above some threshold level, following a particular course of treatment). Various supervised learning algorithms can then be trained on this labelled dataset (ideally using a rigorous cross-validated approach), and assessed in terms of their predictive accuracy on independent 'unseen' (during model training) data. Finally, the best model can be brought forward to a randomised controlled trial framework, where treatment allocation by current clinical guidelines could be compared to algorithm-assisted treatment assignment[8].

This approach is highly attractive, as the potential clinical gains from even a small increase in likelihood of treatment response for a particular individual are large. However, across the field of medicine in general, attempts to make such clinical gains via a personalised medicine approach have not often fulfilled their initial promise – with relatively few reaching the clinic (e.g. 9). Here, we explore a basic statistical issue that may limit the effectiveness of this process – i.e. the reliability of distinguishing between treatment 'responders' and 'non-responders' in the first place. We further discuss the reasons why this problem may be particularly acute in the case of available data regarding psychiatric disorders and chronic pain conditions, and some potential solutions.

The problem of response heterogeneity

The problem of properly identifying response heterogeneity, or, more simply, reliably distinguishing between responders and non-responders to a particular treatment, on the basis of randomised controlled trial (RCT) data, has previously been highlighted across various fields of medicine[10–12]. If not properly addressed, this constitutes an absolute limit on the effectiveness of predictive models at the level of input or training data, thereby limiting their future clinical usefulness.

The issue is best illustrated by considering the nature of data collected during RCTs, and the kind of inferences this process affords. The foundation of an RCT is that the mean effect of an intervention (e.g. active drug treatment) is derived by comparing what happened, on average, to the (randomly allocated) participants in the intervention group to what happened, on average, to participants in the control (e.g. placebo) arm. The random allocation of participants to the intervention *vs* control arms allows the control group to function as an illustration of what we might have expected to occur in the intervention group, had they *not* received the active treatment – in turn allowing us to draw conclusions about the overall (average) effects of the treatment itself[12]. Crucially, we can only draw this inference by direct comparison to the control arm data.

This basis of an RCT means that we cannot identify responders and non-responders by considering individuals in the intervention group *alone*. In other words, we cannot legitimately label an individual who received a particular active treatment as a 'responder' (or not) because we do not know what would have happened to

that particular individual if they had been in the comparator (or placebo) arm[10]. This kind of information is very hard to obtain at the individual (*cf* the group) level, as there is no good way to obtain a control observation. Formally, to properly infer whether a particular participant responded or didn't respond to a particular treatment, we would require knowledge of what would have happened if a key event (treatment administration) both *did* and *did not* occur (a form of counterfactual reasoning), which is not possible in the real world[11].

A particularly acute issue for psychiatric and chronic pain datasets?

Variability of change (e.g. $t_2 - t_1$ symptom score) in the intervention arm is not a true estimate of variability in treatment response, because it includes components of within-subject variation and measurement error[10]. Even if measurement error is small (i.e. we can precisely measure the outcome variable of interest), for many medical interventions, the outcome variable will depend on a complex interplay of biological factors (e.g. time of day, stress level, etc.), and so within-subject variability will be relatively high. This means that the reliability of within-subject measurements across time points can be somewhat poor, and large variation in changes between study time points may be evident – even where there is no true individual difference in treatment response.

Unfortunately, for psychiatric and chronic pain symptom data, both measurement error and within-subject variation are likely to be high. Measurement error may be higher than other areas of medicine, as the main tools used to assess clinical outcomes are patient or clinician-completed questionnaire measures, which are relatively low precision tools. Further, although self-reported symptom levels are considered the gold standard outcome measure for both psychiatric disorders and chronic pain conditions[13], reliability is limited by factors such as cognitive capacity and level of insight for patient-rated measures (e.g. 14), and by interviewer skill and inter-rater agreement for clinician-rated measures (e.g. 15–17). Finally, these classes of disorders represent episodic, chronically relapsing conditions, which will likely contribute to large within-subject variation, particularly at typical RCT follow-up timescales (often around 6 months–1 year; *cf* e.g. median duration of a depressive episode of ~20 weeks,[18]). If the variation in outcome due to these sources is greater than that due to any true individual differences in treatment response, it will be very hard to detect the latter under a conventional RCT framework.

A further problem in predicting true response heterogeneity is susceptibility of symptom change data to regression to the mean and mathematical coupling artefacts[19,20]. Regression to the mean refers to the phenomenon whereby if an individual is selected on the basis of having an extreme measurement value at time point one, their second measurement value will, on average, be closer to the mean of the population distribution (due to the influences of measurement error and normal within-subject variation). A corollary of this effect is that t_1 severity is often a significant covariate of change in symptom score between t_1 and t_2, – meaning that individuals with higher initial scores may appear to show the greatest improvement in symptom levels at follow-up, even when the true magnitude of change does not vary across individuals (see 10 for a worked example). The fact the t_1 score is used to calculate both quantities (i.e. they are mathematically coupled) results in further inflation of this relationship (see 20). Care should therefore be taken when key predictors in response algorithms closely index t_1 severity, as this may result in a poorly generalising model. However, in previous studies in psychiatric datasets, baseline severity score is usually included among the features used to train response prediction algorithms (e.g. 21–23).

These factors may help explain why previous attempts to apply machine learning approaches to outcome prediction in psychiatric datasets have thus far had limited success in terms of out-of-sample (unseen data) classification. For example, a recent methodologically rigorous trial aiming to predict significant response (remission) following treatment with a particular antidepressant achieved only ~60% classification accuracy when the model was applied in external validation datasets[23]. However, as previously noted, tools with only modest true predictive value may still have reasonably high clinical utility compared to current best practice[8]; therefore this is still an approach very much worth pursuing.

Potential solutions

Clinical trial design

The problem of identifying true response heterogeneity is a problem of appropriately partitioning variance components in observed outcomes[11]. The ability to properly identify differential response to a particular treatment in different individuals requires replication at the level at which the differential response is claimed (i.e., that particular treatment in that particular individual). Differential treatment response (i.e. identification of patient by treatment interactions) can therefore be identified by use of repeated period cross-over designs – a form of trial where each participant receives both placebo and active treatments more than once[11]. However, in practice, these designs are rare, as they are likely to be impractical (prohibitively lengthy and expensive) and/or unethical. This kind of design also assumes that treatments wash out fully between administrations, which might not be reasonable for some interventions (e.g. psychological therapies)[24].

Training data definition and selection

An alternative approach is to improve the way data from existing RCTs is used to train predictive models. For example, it has been suggested that the uncertainty in each individual's 'response' (change in symptom score in the active treatment group) could be expressed as a confidence interval by reference to the standard deviation of the change scores in the control (placebo) group multiplied by the appropriate value from the *t* distribution (e.g. individual change score ± 1.96*SD of control arm changes for a 95% CI, see 24). The probability that any given individual in the intervention group is a true responder (true change score is greater than the minimum clinically significant change) can then be derived from individual CIs using a Bayesian approach[10]. Appropriate supervised learning algorithms could then be trained to predict (continuous) treatment response probability, as opposed to dividing individuals into binary response categories (e.g. using Gaussian process regression,[25]).

It also may be important to think carefully about the nature of the predictors (features) included in supervised learning model training data – as those that reference initial clinical severity may be vulnerable to regression to the mean-related artefacts. There are statistical methods that have proposed to correct for regression to the mean when correlating t_2-t_1 symptom changes with initial severity level (see 20). However, these may require additional measurements (e.g. multiple estimates of t_1 value, in order to control for effects of measurement error).

Counterfactual probabilistic modelling

When a particular experiment is not feasible, one alternative is to train models from observational (non-experimental) data that are able to make counterfactual predictions – i.e. of the outcomes that would have been observed, had we run that particular experiment. For example, Saria and colleagues have recently developed a counterfactual Gaussian process (CGP) approach to modelling clinical outcome data[26]. The CGP is trained on observational (non-experimental) time series data, in order to form a model of clinical outcomes under a series of treatments in continuous time. Crucially, the CGP is trained using a joint maximum likelihood objective, which parses dependencies between observed actions (e.g. treatments) and outcomes in the data. This feature allows prediction of how future trajectories (symptom levels) may change in response to different treatment interventions, and has previously been shown to successfully predict real clinical data (renal health markers following different kinds of dialysis,[26,27]).

This modelling approach requires datasets with semi-continuous measurement of the relevant clinical outcome (both pre- and post- intervention), in order to generate hypothetical treatment response traces – a kind of data that is not usually available from existing RCTs. Given sufficient attention to patient confidentiality and other ethical concerns, it may be possible to obtain appropriate training data from health service clinical records; however, frequency and consistency of symptom reporting may pose analytical problems (e.g. 27). The use of personal devices such as smartphones or other wearable technology to regularly self-record symptom levels may be a potential source of this kind of data in the future, given sufficient insight and patient compliance (e.g. 28). The CGP approach also rests on two key mathematical assumptions: that there will be a consistency of outcomes between training observations and future outcomes, given a particular treatment; and that there are no important confounding variables missing from the dataset[26]. It may require careful consideration as to whether these are reasonable assumptions for modelling psychological and chronic pain symptomatology.

Conclusions

The issues discussed above underline the importance of focusing on where data comes from when considering strategies for personalised medicine. In particular, it is problematic to designate individual data points from a conventional RCT design as 'responders' or 'non-responders' to a particular treatment, as this is in effect a single-arm (no control) study not adjusted for other important sources of response variation. This might be particularly important when considering patients with episodic, chronically-relapsing disorders as control variability is likely to be high (and symptom measurement itself is often imprecise). One solution to this problem is to use data derived from repeated cross-over design clinical trials, although in practice these can be prohibitively difficult and/or ethically problematic. It may be possible to alleviate these issues with careful model design, but this may still require changes to the way data is collected and monitored in the future in order to maximise potential clinical utility.

Competing interests

No competing interests were disclosed.

Grant information

AN and BS are supported by the Wellcome Trust (grant number 097490/Z/11/A to BS).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Rush AJ, Trivedi MH, Wisniewski SR, *et al.*: **Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report.** *Am J Psychiatry.* 2006; **163**(11): 1905–1917.
PubMed Abstract | Publisher Full Text

2. Pigott HE, Leventhal AM, Alter GS, *et al.*: **Efficacy and effectiveness of antidepressants: current status of research.** *Psychother Psychosom.* 2010; **79**(5): 267–279.
PubMed Abstract | Publisher Full Text

3. Finnerup NB, Attal N, Haroutounian S, *et al.*: **Pharmacotherapy for neuropathic pain in adults: a systematic review and meta-analysis.** *Lancet Neurol.* 2015; **14**(2): 162–173.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Cleare A, Pariante CM, Young AH, *et al.*: **Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines.** *J Psychopharmacol.* 2015; **29**(5): 459–525.
PubMed Abstract | Publisher Full Text

5. Zimmerman M, Ellison W, Young D, *et al.*: **How many different ways do patients meet the diagnostic criteria for major depressive disorder?** *Compr Psychiatry.* 2015; **56**: 29–34.
PubMed Abstract | Publisher Full Text

6. Smith SM, Dworkin RH, Turk DC, *et al.*: **The Potential Role of Sensory Testing, Skin Biopsy, and Functional Brain Imaging as Biomarkers in Chronic Pain Clinical Trials: IMMPACT Considerations.** *J Pain.* 2017; **18**(7): 757–777.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Steingrímsdóttir ÓA, Landmark T, Macfarlane GJ, *et al.*: **Defining chronic pain in epidemiological studies: a systematic review and meta-analysis.** *Pain.* 2017; **158**(11): 2092–2107.
PubMed Abstract | Publisher Full Text

8. Gillan CM, Whelan R: **What big data can do for treatment in psychiatry.** *Curr*

Opin Behav Sci. 2017; **18**: 34–42.
Publisher Full Text

9. Drucker E, Krapfenbauer K: **Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine.** *EPMA J.* 2013; **4**(1): 7.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Atkinson G, Batterham AM: **True and false interindividual differences in the physiological response to an intervention.** *Exp Physiol.* 2015; **100**(6): 577–588.
PubMed Abstract | Publisher Full Text

11. Senn S: **Mastering variation: variance components and personalised medicine.** *Stat Med.* 2016; **35**(7): 966–977.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Dahly D: **Response Heterogeneity.** 2017.
Reference Source

13. Dworkin RH, Turk DC, Farrar JT, *et al.*: **Core outcome measures for chronic pain clinical trials: IMMPACT recommendations.** *Pain.* 2005; **113**(1–2): 9–19.
PubMed Abstract | Publisher Full Text

14. Alwin DF, Krosnik JA: **The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes.** *Sociol Methods Res.* 1991; **20**: 139–181.
Publisher Full Text

15. Kobak KA, Feiger AD, Lipsitz JD: **Interview quality and signal detection in clinical trials.** *Am J Psychiatry.* 2005; **162**: 628.
PubMed Abstract | Publisher Full Text

16. Engelhardt N, Feiger AD, Cogger KO, *et al.*: **Rating the raters: assessing the quality of Hamilton rating scale for depression clinical interviews in two industry-sponsored clinical drug trials.** *J Clin Psychopharmacol.* 2006; **26**(1): 71–74.
PubMed Abstract | Publisher Full Text

17. Rothman B, Yavorsky C, De Fries A, *et al.*: **P02-88 - Quantifying rater drift on the HAM-D in a sample of standardized rater training events: Implications for reliability and sample size calculations.** *Eur Psychiatry.* 2011; **26**: 683.
Publisher Full Text

18. Solomon DA, Keller MB, Leon AC, *et al.*: **Recovery from major depression. A 10-year prospective follow-up across multiple episodes.** *Arch Gen Psychiatry.* 1997; **54**(11): 1001–1006.
PubMed Abstract | Publisher Full Text

19. Oldham PD: **A note on the analysis of repeated measurements of the same subjects.** *J Chronic Dis.* 1962; **15**(10): 969–977.
PubMed Abstract | Publisher Full Text

20. Tu YK, Gilthorpe MS: **Revisiting the relation between change and initial value: a review and evaluation.** *Stat Med.* 2007; **26**(2): 443–457.
PubMed Abstract | Publisher Full Text

21. Riedel M, Möller HJ, Obermeier M, *et al.*: **Clinical predictors of response and remission in inpatients with depressive syndromes.** *J Affect Disord.* 2011; **133**(1–2): 137–149.
PubMed Abstract | Publisher Full Text

22. Kessler RC, van Loo HM, Wardenaar KJ, *et al.*: **Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports.** *Mol Psychiatry.* 2016; **21**(10): 1366–1371.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Chekroud AM, Zotti RJ, Shehzad Z, *et al.*: **Cross-trial prediction of treatment outcome in depression: a machine learning approach.** *Lancet Psychiatry.* 2016; **3**(3): 243–250.
PubMed Abstract | Publisher Full Text

24. Hopkins WG: **Individual responses made easy.** *J Appl Physiol (1985).* 2015; **118**(12): 1444–1446.
PubMed Abstract | Publisher Full Text

25. Rasmussen CE, Williams KI: **Regression.** In *Gaussian Processes for Machine Learning.* The MIT Press; 2006.
Reference Source

26. Schulam P, Saria S: **What-If Reasoning with Counterfactual Gaussian Processes.** *ArXiv170310651 Cs Stat.* 2017.
Reference Source

27. Soleimani H, Subbaswamy A, Saria S: **Treatment-Response Models for Counterfactual Reasoning with Continuous-time, Continuous-valued Interventions.** *ArXiv170402038 Cs Stat.* 2017.
Reference Source

28. Faurholt-Jepsen M, Vinberg M, Christensen EM, *et al.*: **Daily electronic self-monitoring of subjective and objective symptoms in bipolar disorder--the MONARCA trial protocol (MONitoring, treAtment and pRediCtion of bipolAr disorder episodes): a randomised controlled single-blind trial.** *BMJ Open.* 2013; **3**(7): pii: e003353.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔

Version 1

Referee Report 18 January 2018

doi:10.5256/f1000research.14907.r29856

✔ **Greg Atkinson** [1], **Philip J. Williamson** [2]

[1] Health and Social Care Institute, Teesside University, Middlesbrough, UK

[2] Health and Social Care Institute, Teesside University, Middlesbrough, UK

In this manuscript, the authors discussed the concept of inter-individual differences in response to treatment interventions, particularly those focussed on psychological-related outcomes. The consideration of inter-individual responses is an important issue and the authors provide further insights previously not considered in detail within the domain of psychology. The topic is generally discussed accurately in the context of the current literature, statements are generally correct and supported by relevant citations. I have thoroughly read and considered the manuscript, which was interesting in content and constructed with a logical flow. I have only minor comments for the authors' consideration.

1. The article focuses on the prediction of response heterogeneity, especially prediction of responders/non-responders. Have you considered the roadmap that has been suggested[1] to actually confirm whether the general amount of true heterogeneity is clinically important or not BEFORE we might explore for predictors of that heterogeneity?

2. Page 3, Right hand Column, Lines 27 onwards– whilst discussing RCT trial design and highlighting that responders/non-responders cannot be identified through the analysis of intervention sample data alone, perhaps it might be appropriate to address any research making similar claims even in the total absence of any control sample data.

3. Page 4, Left hand column, Line 8 and conclusion. The arguments that you make on this point, particularly in the conclusion are, at present, unsupported by scientific literature and require justification. You allude to the fact that a lack of 'true' counterfactual information makes an RCT in effect a single-arm (no control study). It is agreed that one cannot say with 100% certainty whether the intervention group as a whole or any specific individual in the intervention group is a positive responder, as what would have happened to that person if they had been in the control group is of course unknown. This is the fundamental counterfactual basis of the RCT. Nevertheless, as the control group variability over the same time period as the intervention effectively provides our best guess of the counterfactual (what would have happened to individuals in the intervention group if they had been in the control arm), I feel that this applies to changes at both the group mean and the individual level, and that disregarding RCTs as 'single-arm studies' is unsupported. According to the previously-mentioned "roadmap" that has been presented, the analysis of the control group changes (specifically the comparison of change variance between treatment and control) can provide information as to what the general clinical importance is of "true" individual response

heterogeneity. By “true”, one knows from this comparison whether the overall amount of heterogeneity in changes surpasses the overall amount of random within-subject heterogeneity of changes in the control group. If heterogeneity of change is similar between treatment and control, it could be argued that moving on to attempts to predict treatment response variability is a somewhat meaningless exercise.

4. Page 4, Left hand column, Lines 43 – 54. Whilst discussing regression to the mean and the mathematical coupling of pre- to post change scores, the use of covariates (especially baseline values of the study outcome) in the statistical model (ANCOVA) could be suggested as a potential solution to this – a notable absence in many studies’ data analyses.

5. Pages 4 – 5. You make a number of pertinent suggestions for potential solutions to the problem, and briefly allude to the methods recently suggested [1,3]. We have suggested how this might be approached in RCTs and tied to an appropriate anchor usually a minimal clinically important difference or smallest worthwhile change [1,2]. Addressing these issues may assist the reader in applying this methodology in their applied practice and/or research environments.

References

1. Atkinson G, Batterham AM: True and false interindividual differences in the physiological response to an intervention.*Exp Physiol*. 2015; **100** (6): 577-88 PubMed Abstract | Publisher Full Text
2. Williamson PJ, Atkinson G, Batterham AM: Inter-Individual Responses of Maximal Oxygen Uptake to Exercise Training: A Critical Review.*Sports Med*. 2017; **47** (8): 1501-1513 PubMed Abstract | Publisher Full Text
3. Hopkins WG: Individual responses made easy.*J Appl Physiol (1985)*. 2015; **118** (12): 1444-6 PubMed Abstract | Publisher Full Text

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Are arguments sufficiently supported by evidence from the published literature?
Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Partly

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.