OPINION ARTICLE

REVISED Response heterogeneity: Challenges for personalised medicine and big data approaches in psychiatry and chronic pain [version 2; referees: 3 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

v2 **First published:** 15 Jan 2018, **7**:55 (doi: 10.12688/f1000research.13723.1)
Latest published: 01 Mar 2018, **7**:55 (doi: 10.12688/f1000research.13723.2)

Abstract

Response rates to available treatments for psychological and chronic pain disorders are poor, and there is a substantial 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.

Keywords

personalised medicine, big data, machine learning, psychiatry, chronic pain, individual differences, response heterogeneity, clinical trial design

This article is included in the INCF gateway.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees		
	1	**2**	**3**
REVISED version 2 published 01 Mar 2018		✓ report	
		↑	
version 1 published 15 Jan 2018	✓ report	? report	✓ report

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

2 **William G. Hopkins**, Victoria University, Australia

3 **Ricardo Silva**, The Alan Turing Institute, 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 2; referees: 3 approved]** *F1000Research* 2018, **7**:55 (doi: 10.12688/f1000research.13723.2)

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)

REVISED **Amendments from Version 1**

- We have added reference to proposals that recommend determining whether clinically important response heterogeneity exists prior to quantifying it.
- We have made it more explicit that although the underlying RCTs (designed to determine treatment effectiveness at the group level) are not single arm, machine learning algorithms concerned with predicting individual differences in treatment response are usually only trained on the active treatment arm data (i.e. without reference to symptom changes in the control arm data).
- We have added reference to the use of ANCOVA under a traditional statistical framework as a way of guarding against regression to the mean and mathematical coupling artefacts – with discussion of equivalent techniques (or their absence) in the machine learning literature.
- We have added reference to the possibility that appropriately formulated data-driven models could be used to predict probability of harm (symptom increase above a clinically significant threshold), as well as probability of successful treatment response (symptom decrease above some clinically significant threshold).
- We have updated the manuscript to touch upon how prediction of differential treatment responses in future patients (i.e., out-of-sample data) can be framed as a causal inference problem (with respect to current datasets) - with brief discussion of some pertinent issues, including representativeness of training data samples, and the requirement of explicit assumptions about the causal consequences of any confounding variables.
- Some changes to grammar have been made throughout the manuscript in order to increase clarity.

See referee reports

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 for 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 pursue a personalised medicine approach have not 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 validity 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, i.e., 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 inference 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 draw this inference only 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, it is hard to 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 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].

Although the underlying RCT datasets almost always consist of at least two arms (e.g. active treatment *vs* placebo), machine learning algorithms employed to predict psychiatric treatment response are usually trained on active treatment arm data alone – without reference to control arm symptom changes (e.g. 13, 14). Unless sufficient care is taken, these kinds of predictive models may therefore be the inferential equivalent of single arm trials, and the resultant categorisation of symptom change scores may be unduly influenced by sources of variance causally unrelated to true treatment response.

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. Although self-reported symptom levels are considered the gold standard outcome measure for both psychiatric disorders and chronic pain conditions[15], reliability is limited by factors such as cognitive capacity and level of insight for patient-rated measures (e.g. 16), and by interviewer skill and inter-rater agreement for clinician-rated measures (e.g. 17–19). Further, 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[20]). The greater the variation in outcome due to these sources, the harder to it will be to detect true individual differences in treatment response, under a conventional RCT design.

A further problem in predicting true response heterogeneity is susceptibility of symptom change data to regression to the mean and mathematical coupling artefacts[21,22]. 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 baseline and change scores (i.e., that they are mathematically coupled) results in further inflation of this relationship (see 22). 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 based on psychiatric datasets, baseline severity score is usually included among the features used to train response prediction algorithms (e.g. 13, 14, 23).

These factors may help explain why previous attempts to apply machine learning approaches to outcome prediction in psychological disorders 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 drug achieved only ~60% classification accuracy when the model was applied in external validation datasets[14]. 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 identify differential response to particular treatments in different individuals can be achieved by replication of observations 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]). This approach could also be used to predict individual probability of *harm* (worsening of outcome measure above some minimum clinically important threshold) in response to a particular treatment. Some researchers have suggested that comparing the variances of symptom change data between active treatment and control arms, in order to detect whether there is clinically significant heterogeneity in response to a particular treatment in the first place, should be a pre-requisite for these kind of analyses[10].

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. Under a traditional statistical framework, an effective way of dealing with these artefacts is to include baseline scores as covariates in models of symptom change data (i.e., conduct an ANCOVA). This approach can then be used to test if a given between-subjects variable is a significant modifier of treatment effect by adding it to the model (providing measurement error is sufficiently low[24]). An interesting issue is that in machine learning, there is not really an equivalent concept to 'covariates of no interest' – rather, model features are usually selected purely on the basis of their predictive capacity. One recent paper that explicitly addresses this problem comes from Rao and colleagues, who propose a method for removing known confounds from predictive models based on functional imaging data. Rao *et al.* suggest that one solution is to first fit linear models to each image feature using the confound variables as predictors, then consider the residuals of this model to be 'adjusted' data – suitable to be used as input features for a confound-controlled predictive model[26]. There are also statistical methods that have proposed to correct for regression to the mean when simply correlating t_2-t_1 symptom changes with initial severity level that could be applied to training data (see 22). However, these may require additional measurements (e.g. multiple estimates of t_1 value, in order to estimate measurement reliability).

Counterfactual probabilistic modelling and other causal inference methods

When a particular experiment is not feasible, an alternative is to train models on 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[27]. The CGP is trained on observational symptom trajectory data 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 order to infer the existence of *causal* relationships between the two. This feature allows the 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,[27,28]).

Thus far, the CGP has only been empirically tested as a clinical decision support tool on the same subjects from whom model training data was derived[27]. However, it can be argued that prediction of the response of *future* patients to particular treatment options is an inference problem that does not necessarily involve counterfactual reasoning. Under these circumstances, we require a model that can infer causes that are likely to be active for *out-of-sample* data (individuals with certain features, who may not have received any treatment yet), as opposed to *in-sample* data (individuals whose clinical data a particular model was trained on, who might have showed a different response to different treatment strategies)[29].

This perspective raises the issue of how representative the individuals who make up a particular training dataset are of the general population (from whom future patients will be drawn). Importantly, concerns have previously been raised as to the effects of various sources of selection bias on the representativeness of participants in RCTs compared to the population at large[30]. Specific forms of selection bias that have been identified in RCTs for psychological disorders include exclusion of individuals with comorbidities (which for many some conditions may be more common than ‘pure’ presentation), selection of less severe cases (e.g. in psychotic disorders, where ability to consent and treatment compliance may be of heightened concern), or, conversely, application of minimum severity thresholds (e.g. in mood disorders, to reduce the likelihood of spontaneous remission over the trial period)[31,32]. Further, methods of recruitment to RCTs (particularly the requirement to self-select into trials) may influence the distribution of various psychological traits in trial participants, prior to any further eligibility criteria being applied (e.g. 33). Although there are methods designed to mitigate the effects of sample selection bias when transferring predictive models to a different test set (see 34), it remains an open question as to whether these are sufficiently robust for successful out-of-sample treatment prediction at the individual level.

The success of causal inference modelling approaches to response prediction may therefore depend upon availability of different kinds of data to that derived from traditional RCTs – involving semi-continuous measurement of the relevant clinical outcome (both pre- and post- intervention), and gathered from more representative sources than some previous RCT datasets. 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. 28). 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. 35).

A further important feature of predictive models derived from observational data is that they depend on explicit assumptions about the existence and causal consequences of any confounding variables present in the dataset. For example, the CGP approach requires both 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[27]. It will therefore be necessary to carefully consider how well such assumptions are met when considering applying these kinds of models to psychological and chronic pain symptom data.

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 symptom change scores are not adjusted for other important sources of variation. This might be particularly important when considering patients with episodic, chronically-relapsing disorders, as within-subject variability is likely to be high (and symptom measurement itself may be 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 training data selection and predictive model design, but changes in the way symptom data is collected and monitored may still be required in the future in order to maximise the clinical utility of model-aided treatment selection approaches.

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. 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

14. 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

15. 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

16. 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

17. Kobak KA, Feiger AD, Lipsitz JD: **Interview quality and signal detection in clinical trials.** *Am J Psychiatry.* 2005; **162**(3): 628.
PubMed Abstract | Publisher Full Text

18. 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

19. 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

20. 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

21. 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

22. 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

23. 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

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. Rao A, Monteiro JM, Mourao-Miranda J, *et al.*: **Predictive modelling using neuroimaging data in the presence of confounds.** *Neuroimage.* 2017; **150**: 23–49.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

29. Dawid AP: **Causal Inference without Counterfactuals.** *J Am Stat Assoc.* 2000; **95**(450): 407–424.
Publisher Full Text

30. Rothwell PM: **External validity of randomised controlled trials: "to whom do the results of this trial apply?"** *Lancet.* 2005; **365**(9453): 82–93.
PubMed Abstract | Publisher Full Text

31. Hofer A, Hummer M, Huber R, *et al.*: **Selection bias in clinical trials with antipsychotics.** *J Clin Psychopharmacol.* 2000; **20**(6): 699–702.
PubMed Abstract

32. Fava M, Evins AE, Dorer DJ, *et al.*: **The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach.** *Psychother Psychosom.* 2003; **72**(3): 115–27.
PubMed Abstract | Publisher Full Text

33. Almeida L, Kashdan TB, Nunes T, *et al.*: **Who volunteers for phase I clinical trials? Influences of anxiety, social anxiety and depressive symptoms on self-selection and the reporting of adverse events.** *Eur J Clin Pharmacol.* 2008; **64**(6): 575–582.
PubMed Abstract | Publisher Full Text

34. Bareinboim E, Pearl J: **Causal inference and the data-fusion problem.** *Proc Natl Acad Sci U S A.* 2016; **113**(27): 7345–7352.
PubMed Abstract | Publisher Full Text | Free Full Text

35. 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 2

Referee Report 16 March 2018

doi:10.5256/f1000research.15388.r31390

✓ **William G. Hopkins**
Institute for Health and Sport, Victoria University, Melbourne, VIC, Australia

There has been no attempt to use my suggestions for making the abstract clearer, and you ignored many of my suggestions for improvement in grammar and sense and my requests for clarification in the rest of the manuscript. Never mind.

I don't agree with your argument about repeatability. In your response to my previous critique, you wrote: "Regarding the issue of 'repeatability', we would argue that since this perspective deals with the ability to predict the responses of *future* patients (i.e., individuals with similar characteristics to past 'responders'), then repeatability of these responses is not only relevant, but vital. Although we would not deny the real benefit to an individual patient of any significant improvement in clinical outcome, if this benefit is not 'repeatable' in that it is reliably casually related to treatment administration (e.g. score improvement is largely due to fluctuations in symptoms that would have occurred otherwise), then it should not be taken as evidence for use of that treatment in future (similar) individuals." In other words, you think a treatment has to reproduce a benefit on a second administration in the same subject if it is to produce benefit on the first administration in a similar subject. But if the treatment produced a benefit on the first administration in one kind of subject, then there is no reason to assume it would not produce benefit on the first administration in a similar subject. Besides, we are often talking about treatments that are supposed to produce a permanent cure, in which case the question of repeatability is not an issue. But if the effect of the treatment wears off, a repeat of the same treatment could fail acutely for the same reason it failed chronically the first time (a change in the subject resulting in desensitization), so you can't extrapolate from the effect of the second treatment in the first subject to the expected effect of the treatment the first time it is administered to a similar subject.

In the paragraph on regression to the mean, you state "The fact the t_1 score is used to calculate both baseline and change scores (i.e., that they are mathematically coupled) results in further inflation of this relationship (see 22)." No, it doesn't "further inflate" regression to the mean. You get exactly the same artefact whether you use t1 to predict t2 or t2-t1. The benefit of predicting the change score is that the artefact is more obvious in scatterplots.

Some of the new material is beyond my expertise. It reads OK.

Find "which for many some conditions" and remove either "many" or "some". In the same paragraph, "31,32" needs to be superscripted. Whether 33 and 34 need superscripting is a problem with that method of referencing.

"availability of different kinds of data to that derived " should be "availability of kinds of data different from those derived"

"may be a potential" is another double doubtful you will probably not bother to fix.

Competing Interests: No competing interests were disclosed.

Referee Expertise: Research design and analysis, with special reference to physical activity, sport and lifestyle.

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

Version 1

Referee Report 20 February 2018

doi:10.5256/f1000research.14907.r30529

✓ **Ricardo Silva**
The Alan Turing Institute, London, UK

I would like to thank the authors for the informative review of the many difficulties pertinent to modeling the treatment of psychiatric disorders. My main take will be on the statistical and machine learning aspects, causal inference in particular, as I do not have a clinical background.

With the availability of larger sources of data, it is reasonable to ask what can be leveraged in order to better predict how individual patients will respond to particular treatments. This raises questions of causal inference, as treatments are meant to be interventions that will (or are expected to) change the condition of a patient. The article properly addresses factors that make this challenging, including measurement error. I do not have much to comment on measurement error as this goes into the specifics of the domain, but I would like to second the authors' warning on how important this issue is. Even different ways of executing the data collection protocol in different environments (e.g., the way questionnaires are applied) can have an influence on what response distribution we obtain in the end.

I will focus instead on what it means to perform counterfactual reasoning, and which challenges in performing it are warranted. I share much of the view of Dawid [1], that many problems of causal inference are decision-theoretical rather than genuinely counterfactual. In fact, counterfactuals are incompatible with out-of-sample problems, if only because it is impossible to be "contrary to a fact" that has not happened yet and can still happen. In the classical statistical approach for causal inference, the motivation is intrinsically an in-sample problem, where the randomness is all in the treatment assignment: we are dealing with what would have happened to a particular group of people, at a particular time, at a particular environment, had treatment assignments been different [2]. In the questions raised by the authors, we are interested in what *will* happen to a patient given a treatment coming from a set of two or more choices (to follow one psychiatric treatment, or an alternative, or none etc.), not what *would have happened* had things been different. This is a predictive policy question, not unlike what is found in contextual bandits (but notice that, unlike contextual bandits, we are not necessarily interested in exploration-exploitation trade-offs, but on the assessment of a policy that maps features of an individual to an action). This boils down to evaluating hypothetical actions, and picking the most beneficial one according to some risk/utility function. This is still a causal inference question as it concerns the effects of actions, but it does not need to be framed as counterfactual reasoning as the notion of rewinding the clock to apply a different treatment to an individual is never on the cards when assessing out-of-sample treatment recommendation. Put simply, the estimand does not involve counterfactuals.

How to obtain such a model is however a nontrivial question and requires much work, which I briefly discuss below in the context of the article. As a final remark before continuing, I will say that, unlike Dawid, I have no issues on using counterfactual models to address hypothetical questions. By the end of the day, there are many equivalent languages to express causal assumptions, and we should be free to choose our "syntactic sugar" as it is seen fit, as long as we know what the limitations of our models are. The following is agnostic to whether counterfactuals have motivated the model or some other predictive counterfactual-free causal approach has been used instead.

So, what are the challenges of causal inference for heterogeneous effects? It is still the case that RCTs are much limited in this scenario: it is one thing to use a RCT so show that there exists a group of people which at particular time and at a particular environment would have shown different responses had treatment assignment been different. It is a different ballgame to extract meaningful predictive power of a dataset if the sample is not representative of the target population. This sample selection bias, where volunteers of a RCT are very likely not to be representative, is pervasive in many sciences. To the best of my understanding, this is also the case in psychiatric research. Some mitigation can be done to transfer some conclusions to the test set of interest [3] by tapping into some aspects of the process that remain invariant out-of-sample. Unfortunately, this may not be enough, and attending to the needs of many individuals of interest may require unwarranted extrapolations from the training set. Although RCTs are extremely desirable for causal inference, as a well-designed trial will remove unmeasured confounding, the selection bias that is natural in studies with human subjects may be too strong for its conclusions to be applicable to a large fraction of out-of-sample personalised treatments. I'm particularly skeptical of putting any effort on cross-over designs: those have the shortcomings of RCTs (sample not being representative), while adding assumptions such as "treatments wash out fully between administrations" which seem unbelievable to me in the context of psychiatric research, while the motivation (estimating counterfactuals) is unnecessary for the ultimate goal of deciding among the initial treatment options for out-of-sample patients.

A promising venue is to exploit observational data, under the provision that we understand its many limitations. We need causal assumptions about which sources of confounding exist and how to remove them, and to which extent a RCT may provide information on the degree of confounding for patient profiles observed in the general population (via the observational data), but which are far from those found in the RCT sample. While the machine learning literature has done much in terms of providing ways of combining data from a set of experiments and observations [4, 5], they put much emphasis on finding causal networks as opposed to reliably estimating heterogeneous effects. This remains open, and with the added difficulty of dealing with assumptions about measurement error. Hopefully we will see this raising research questions for those up to the challenge, and with the motivation of improving the lives of the many who rely on our better understanding of psychiatric treatments and their effectiveness.

References

1. Dawid A: Causal Inference Without Counterfactuals. *Journal of the American Statistical Association*. 2000; **95** (450). Publisher Full Text

2. Rosenbaum P: Observation and Experiment. *Harvard University Press*. 2017.
3. Bareinboim E, Pearl J: Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*. 2016; **113** (27): 7345-7352 Publisher Full Text
4. Cooper GF, Yoo C: Causal discovery from a mixture of experimental and observational data. *Morgan Kaufmann Publishers Inc.(San Francisco, CA)*. 1999. 116-125
5. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP: Causal protein-signaling networks derived from multiparameter single-cell data.*Science*. 2005; **308** (5721): 523-9 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?
Yes

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?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Machine learning

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

Author Response (*Member of the F1000 Faculty*) 27 Feb 2018

Agnes Norbury, Department of Engineering, University of Cambridge, UK

Thank you for taking the time to review this manuscript, and for providing a thoughtful meditation on individual response prediction from a casual inference perspective. We have updated the manuscript to touch upon some of the issues you have raised (please see below):

“Counterfactual probabilistic modelling **and other causal inference methods**

When a particular experiment is not feasible, an alternative is to train models on 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 [27] . The CGP is trained on observational symptom trajectory data 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 order to infer the existence of *causal* relationships between the two**. This feature allows the 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, [27, 28]).

Thus far, the CGP has only been empirically tested as a clinical decision support tool on the same subjects from whom model training data was derived [27]. However, it can be argued that prediction of the response of *future* patients to particular treatment options is an inference problem that does not necessarily involve counterfactual reasoning. Under these circumstances, we require a model that can infer causes that are likely to be active for *out-of-sample* data (individuals with certain features, who may not have received any treatment yet), as opposed to *in-sample* data (individuals whose clinical data a particular model was trained on, who might have showed a different response to different treatment strategies) [29].

This perspective raises the issue of how representative the individuals who make up a particular training dataset are of the general population (from whom future patients will be drawn). Importantly, concerns have previously been raised as to the effects of various sources of selection bias on the representativeness of participants in RCTs compared to the population at large [30]. Specific forms of selection bias that have been identified in RCTs for psychological disorders include exclusion of individuals with comorbidities (which for many some conditions may be more common than 'pure' presentation), selection of less severe cases (e.g. in psychotic disorders, where ability to consent and treatment compliance may be of heightened concern), or, conversely, application of minimum severity thresholds (e.g. in mood disorders, to reduce the likelihood of spontaneous remission over the trial period) [31,32]. Further, methods of recruitment to RCTs (particularly the requirement to self-select into trials) may influence the distribution of various psychological traits in trial participants, prior to any further eligibility criteria being applied (e.g. [33]). Although there are methods designed to mitigate the effects of sample selection bias when transferring predictive models to a different test set (see [34]), it remains an open question as to whether these are sufficiently robust for successful out-of-sample treatment prediction at the individual level.

The success of causal inference modelling approaches to response prediction may therefore depend upon availability of different kinds of data to that derived from traditional RCTs – involving semi-continuous measurement of the relevant clinical outcome (both pre- and post- intervention), and gathered from more representative sources than some previous RCT datasets. 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. 28). 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. 35).

A further important feature of predictive models derived from observational data is that they depend on explicit assumptions about the existence and causal consequences of any confounding variables present in the dataset. For example, the CGP approach requires both 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 [27]. It will therefore be necessary to carefully consider how well such assumptions are met when considering applying these kinds of models to psychological and chronic pain symptom data."

Competing Interests: No competing interests were disclosed.

Referee Report 02 February 2018

doi:10.5256/f1000research.14907.r29962

?

William G. Hopkins

Institute for Health and Sport, Victoria University, Melbourne, VIC, Australia

This article represents a valuable contribution to the developing literature on quantification of individual responses to treatments and identification of patients who respond positively to treatments. Perhaps there should be some attention to the issue of identifying negative responders, since it is more important to avoid harming individual patients than to miss out on benefitting them. By "harm" I don't mean side effects. I also point out in my review that prediction models could be developed from identified modifiers of the treatment effect in controlled trials, something that you should also mention. I also have a strong view about repeatability of individual responses to treatments, something that you will need to address by refuting my claim or accommodating it. Otherwise there are only minor points for you to consider. These were all made on a first and only read-through, so although you explain some points further on, it is important to avoid any confusion in the first place.

"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." I don't understand inclusion of "within-subject variability"? Are you referring to real changes of individual subjects pre to post the treatment that occur even in the absence of an active treatment? I normally think about that as another kind of measurement error. Perhaps you need to clarify by stating "both within-subject variability over the period of the treatment and short-term measurement error " Also, solitary "this" is a grammatical error known as an ambiguous antecedent. There are a few other instances in the manuscript that need to be fixed.

"especially in arenas where the potential clinical benefit is so large" I don't understand the use of "so". Are you referring to the arenas of psych disorders and chronic pain? Are the "potential clinical benefits" in these arenas any larger than in any other arenas of pathology? And you haven't established (in the Abstract, anyway) that there is the potential for large benefit when individual responders have been characterised. Maybe something along the lines of "trustworthy identification of characteristics of positive responders to treatments could result in substantial clinical benefit in psychological disorders, chronic pain, and other pathologies."

In the Introduction, you make it clear–at some length–that non-responders to one kind of treatment could be responders to another, and it may therefore be possible to improve the health of the majority of patients by targeting specific patients with specific treatments, where the pathology has a range of underlying causes and a range of treatments is available. Maybe you need to make that clearer in the Abstract.

"variables that may potentially relate" is a "double doubtful." Remove either "may" or "potentially".

"i.e. the reliability of distinguishing between treatment 'responders' and 'non-responders'". I think validity would be a better word than reliability. Also no need for quote marks.

e.g. and i.e. normally have commas after them and are used only in parentheses.

"Crucially, we can only draw this inference by direct comparison…" Make it "Crucially, we can draw this inference only by direct comparison…" Check for any other instances of this misplaced modifier.

You tend to use too many parenthetical asides. Remove parentheses from as many as you can.

"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." I found this sentence confusing. What is counterfactual reasoning? If I understand this sentence correctly, I disagree with it. In crossovers it is possible to determine the outcome with an individual who received the active and the control treatment. You go on to state that yourself.

"(e.g. time of day, stress level, etc.)" Either e.g. or etc., but not both!

"Measurement error may be higher than [in] 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." I think this statement is false, so you'd better support it with references. The square root of the alpha reliability (which provides an upper limit to the criterion validity correlation of multi-item instruments) and short-term retest reliability ICC could well be high enough to reasonably identify responders to short-term treatments. Even VASs have high short-term ICCs. The ICCs over periods for long-term treatments are likely to be a different story, as you point out.

"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." It depends what you mean by "detect". To be on the safe side, perhaps you should state "The greater the variation in outcome due to these sources compared with that of true individual responses, the harder it will be to characterize the latter in a controlled trial." Note that I have removed superfluous words. Bottom line is that you can make up for the short- and log-term errors with a big-enough sample size, at least for characterizing the mean effect and its modifiers.

"The fact the t1 score is used to calculate both quantities…" I fully understand regression to the mean, but I re-read this paragraph and still don't know what "both quantities" refers to.

"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)." This rather obscurely worded claim has been made by others, but it is false. You can have a patient who responds individually to a treatment on one administration of the treatment. Whether that patient would respond similarly again following washout and reapplication of the treatment is irrelevant. What matters is that the patient has obtained benefit from the treatment when it was applied the first time. Period. Whether you can adequately quantify the extent of an individual patient's response to the (first) application of the treatment depends on short- and long-term errors of measurement and on the magnitude of the response in that patient, but regardless, you can certainly characterise modifiers of the treatment effect with realistic sample sizes: 4x the sample size required to characterize the mean effect (Hopkins, 2006). The identified modifiers could then be used to build courses-for-horses (treatments-for-patients) prediction models, something you haven't considered. Anyway, there is no need to confuse everyone by raising the spectre of repeatability of the treatment effect in individuals.

"However, these may require additional measurements (e.g. multiple estimates of t1 value, in order to control for effects of measurement error)." No, repeating the t1 assessment reduces but does not eliminate the effect of regression to the mean. What you need for the adjustment is the reliability ICC over the time-frame of the treatment. In fact, the control group effectively provides that: when you predict the likelihood of an individual's response in a controlled trial using a mixed model in which the pre-test (t1) is included as a modifying covariate, you have controlled for (adjusted away the effect of) regression to the mean.

I don't understand the paragraph headed Counterfactual probabilistic modelling. You will have to explain what is going on here without the jargon, for my benefit and for those who are even less statistically savvy. I see the word "trajectory" in there, which suggests to me that you are talking about clinical trials with multiple repeated measurements during the course of the treatment. That's a luxury that may not be available in many settings, and in any case, it requires an appropriate model for the time course, which is bound to be non-linear. You still need a control group, if you want to eliminate the contribution of the placebo effect.

Ah, I see you provide some explanation in the next paragraph. Please make the preceding paragraph clearer.

"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…" No. See above.

"and that there are no important confounding variables missing from the dataset." Be more explicit. If it's a properly balanced controlled trial (or randomized with a sufficiently large sample size), what's the problem?

"One solution to this problem is to use data derived from repeated cross-over design clinical trials…" Well, no, because it introduces what I called above the spectre of repeatability.

"It may be possible to alleviate these issues with careful model design…" Explain "careful". You will need to have identified the point(s) explaining "careful" previously, as this sentence is in the Conclusions.

Hopkins, WG. Estimating sample size for magnitude-based inferences. Sportscience 10, 63-70 (http://www.sportsci.org/2006/wghss.htm)

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.

Referee Expertise: Research design and analysis, with special reference to physical activity, sport and lifestyle.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response (*Member of the F1000 Faculty*) 27 Feb 2018

Agnes Norbury, Department of Engineering, University of Cambridge, UK

Thank you for taking the time to review this manuscript. Please see below for responses to your main comments, which are addressed in the revised manuscript.

1. We have added reference to the possibility that appropriately formulated data-driven models could be used to predict probability of harm (symptom increase above a clinically significant threshold), as well as probability of successful treatment response (symptom decrease above some clinically significant threshold).

2. We have added reference to methods for the identification of treatment 'modifiers', whilst controlling for other artefacts, under a traditional statistical approach – plus a discussion of an analogous approach from the field of machine learning (please see below):

"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. **Under a traditional statistical framework, an effective way of dealing with these artefacts is to include baseline scores as covariates in models of symptom change data (i.e., conduct an ANCOVA). This approach can then be used to test if a given between-subjects variable is a significant modifier of treatment effect by adding it to the model (providing measurement error is sufficiently low, [24]). An interesting issue is that in machine learning, there is not really an equivalent concept to 'covariates of no interest' – rather, model features are usually selected purely on the basis of their predictive capacity. One recent paper that explicitly addresses this problem comes from Rao and colleagues, who propose a method for removing known confounds from predictive models based on functional imaging data. Rao et al. suggest that one solution is to first fit linear models to each image feature using the *confound* variables as predictors, then consider the residuals of this model to be 'adjusted' data – suitable to be used as input features for a confound-controlled predictive model [26]**."

3. Regarding the issue of 'repeatability', we would argue that since this perspective deals with the ability to predict the responses of *future* patients (i.e., individuals with similar characteristics to past 'responders'), then repeatability of these responses is not only relevant, but vital. Although we would not deny the real benefit to an individual patient of any significant improvement in clinical outcome, if this benefit is not 'repeatable' in that it is reliably casually related to treatment administration (e.g. score improvement is largely due to fluctuations in symptoms that would have occurred otherwise), then it should not be taken as evidence for use of that treatment in future

(similar) individuals.

4. Some changes to grammar have also been made throughout the manuscript in order to increase clarity.

Competing Interests: No competing interests were disclosed.

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.

Author Response (*Member of the F1000 Faculty*) 27 Feb 2018
Agnes Norbury, Department of Engineering, University of Cambridge, UK

Thank you for taking the time to review this manuscript. Please see below for responses to your main comments, which are addressed in the revised manuscript.

1. We have added reference to the proposed ‘roadmap’ to determine whether clinically important response heterogeneity exists prior to quantifying it.

2. Regarding the reference to ‘single arm’ studies, we have made it more explicit that although the underlying RCTs (in order to determine treatment effectiveness as the group level) are not single arm, machine learning algorithms concerned with predicting individual differences in treatment response are usually only trained on the active treatment arm data (please see below):

“**Although the underlying RCT datasets almost always consist of at least two arms (e.g. active treatment *vs* placebo), machine learning algorithms employed to predict psychiatric treatment response are usually trained on active treatment arm data alone – without reference to control arm symptom changes (e.g. [13,14]). Unless sufficient care is taken, these kinds of predictive models may therefore be the inferential equivalent of single arm trials, and the resultant categorisation of symptom change scores may be unduly influenced by sources of variance causally unrelated to true treatment response.**”

3. We have added reference to the use of ANCOVA under a traditional statistical framework as a way of guarding against regression to the mean and mathematical coupling artefacts – with discussion of equivalent techniques (or their absence) in the machine learning literature (please see below):

“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. **Under a traditional statistical framework, an effective way of dealing with these artefacts is to include baseline scores as covariates in models of symptom change data (i.e., conduct an ANCOVA). An interesting issue is that in machine learning, there is not really an equivalent concept to covariates of no interest – rather, model features are selected purely on the basis of their predictive capacity. One recent paper that explicitly addresses this issue comes from Rao and colleagues, who propose a method for removing known confounds from predictive models based on functional imaging data. Rao et al. suggest that a potential solution is to first fit linear models to each image feature using the *confound* variables as predictors, then consider the residuals of this model to be ‘adjusted’ data – suitable to be used as input features for a confound-controlled predictive model [26].**

Competing Interests: No competing interests were disclosed.