PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Emergence of digital biomarkers to predict and modify treatment efficacy: a machine learning study
AUTHORS	Guthrie, Nicole; Carpenter, Jason; Edwards, Katherine; Appelbaum, Kevin; Dey, Sourav; Eisenberg, David; Katz, David; Berman, Mark

VERSION 1 - REVIEW

REVIEWER	Colin Simpson Victoria University of Wellington, Wellington, New Zealand
REVIEW RETURNED	09-May-2019

GENERAL COMMENTS	This is an interesting paper exploring the use of machine learning techniques to create a digital biomarker from data produced by a digital therapeutic. Data from these systems are used to modify behaviours - I am unclear however whether these devices have proven efficacy (beyond the preliminary studies that demonstrate potential). The inclusion of supportive studies on their efficacy would be helpful to the readers. In Strengths and Limitations, the first bullet point is not a strength or limitation. Please expand on the limitations. For Page 7 Line 10 sentence 2. A reference for hyperparameter optimization should be provided. The methods section contains detailed explanations of study design including ML techniques used. A section on subjects is included, but this should have its own section. The results section has some discussion of the results, which should be moved to the discussion section e.g. page 11 lines 17-27. The low sensitivity found for the models should be discussed – what are the implications for a biomarker predicting treatment response.

REVIEWER	Kit Huckvale Black Dog Institute, UNSW Sydney Australia
REVIEW RETURNED	13-May-2019

GENERAL COMMENTS	This clearly written, proof-of-concept paper describes the development and initial evaluation of a supervised machine learning approach for predicting treatment response from early behavioural/utilisation data amongst individuals receiving a digital behavioural intervention targeting blood pressure. Although none of the methods described are new, and although the study is limited in size/predictive power of the resultant models, the application of an explainer approach (SHAP algorithm) is novel in this area and is likely to be of interest to other readers also considering the application of machine learning techniques in clinical medicine. Major Comments ---------- 1. Measurement of blood pressure is notoriously error prone. The 2017 ACC/AHA guidelines recommend multiple measurement occasions and specific data discard protocols for both clinician and home BP monitoring. In the UK, NICE now recommends that at least 14 measurements be taken prior to a diagnosis of hypertension. a) Given the critical importance of ground truth labels in supervised machine learning, how did you ensure measurement validity? b) Please provide clear detail about how and by whom the measurements that dictated entry into the study and 7-14 week change were performed. c) In particular, please describe how, if multiple measurements occurred between 7-14 weeks, the values used to compute the BP delta were selected. There is an apparent risk of investigator bias if values could be selectively chosen. 2. Given that participants self-selected into the study (Methods, Page 4, Line 47) how did you control for the possibility of treatment stage acting to confound the effects of the digital intervention (and inferences about behavioural predictors)? Individuals with a new diagnosis of hypertension may be highly motivated to try new strategies for blood pressure control (such as a digital therapeutic) but are likely *also* to be titrating with antihypertensives which might reasonably be expected to drive BP reductions - particularly in the short term. Please describe treatment stage information, if available, and account for this potential risk. 3. You propose that Shapley values could be used to prioritise behavioural 'gaps' for individuals to subsequently target (Results, Page 11, Lines 17-27). a) In interpreting a Shapley value assigned to a given feature how do you know that this is, in fact, modifiable as a behavioural target? Is there a way to evaluate the potential value range for each feature? This matters because if a value is essentially fixed then it does not provided the basis for prioritised decision making that you argue. b) You note the risk of confusing correlation and causality in Discussion (Page 13, Lines 10-16) and state that this 'does not preclude the automated use of explanatory variables to guide [...] personalization.' Why? Given that each behavioural strategy comes with concrete costs to individuals, is it not a highly relevant ethical and QoL concern if there are doubts as to whether specific markers provide a genuine rational basis for action? c) How do you counter the concern that, since the ingredients in any multi-component digital behavioural intervention are expectedly evidence-based (and pre-selected for impact on health

	outcomes), there is no mystery in relation to gaps in self-management behaviour that will drive outcomes: an individual simply needs to do the parts of the intervention that they are not doing already. In other words, does the explainer approach really add much over and above that which could be ascertained from simple utilisation/adherence data? If it does, please explain why. Minor Comments ---------- 4. Is it realistic that clinicians would consider stopping antihypertensive medication in the short term on the basis of predictors identified within a digital intervention (Methods, Page 6, Lines 6-7)? Although a case for rationalising therapy has been made recently (e.g. doi: 10.1097/HJH.0000000000001405) only a quarter of people appear to be able to sustain BP improvements without medication and neither the ACC/AHA (despite the citation on Page 6, Line 7) nor NICE guidelines include protocols for titrating down medication. Although not central to your subsequent argument (in Discussion) about the value of the approach, please consider revising this statement. 5. Abstract, Page 2, Line 19. 'Ad libitum' is not accessible to an international readership whose first language is not Latin. Please rephrase. 6. Methods, Page 7, Line 13. Please describe your approach to hyperparameter optimization. 7. Discussion, Page 11, Line 54. Is this approach really correctly characterised as being analogous to a credit score? A credit score is intended to support inferences about future behaviour on the basis that past borrowing behaviour is likely to dictate future profligacy. In the context of health behaviour change, it is not clear that success in achieving one lifestyle modification (e.g. sodium reduction) dictates a) future ability to maintain this behaviour or b) success in achieving any other behavioural modification. 8. Figure 3, Page 20. To aid interpretation, consider annotating Figure 3 to highlight the direction on the Y axis that tends to increase the likelihood that the overall prediction will favour a successful reduction in BP, and using 'probability of successful reduction in BP' (or similar) rather than 'output value' (assuming that this interpretation is correct.)

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1. Data from these systems are used to modify behaviours - I am unclear however whether these devices have proven efficacy (beyond the preliminary studies that demonstrate potential). The inclusion of supportive studies on their efficacy would be helpful to the readers.

There is a growing literature on the effect of digital therapeutics on established clinical biomarkers. We referenced several articles demonstrating clinically meaningful improvements in biomarkers with the use of digital therapies that focus on self monitoring as the primary intervention (references 4-6)

and we added a recently published RCT to the reference list (Wang 2018). Since most of the literature consists of short-term studies in small cohorts, we qualified the evidence as “preliminary” and as having “potential” cost-efficacy in end of the first paragraph of the introduction section.

2. In Strengths and Limitations, the first bullet point is not a strength or limitation. Please expand on the limitations.

Thank you for pointing out that the first bullet point does not fit in the list of strengths and limitations. We removed that point, but since it added context to the list we modified the first bullet to indicate the practical application of the biomarkers. We also added two limitations of the current study which are included in the discussion section:

● Use of additional explanatory variables to develop the biomarkers may enhance the accuracy of predictions.

● Generalizability of the biomarkers is unknown and may be limited by the demographics of the training dataset.

3. For Page 7 Line 10 sentence 2. A reference for hyperparameter optimization should be provided.

We have added a reference for the hyperparameter optimization methods (https://scikit-learn.org/stable/modules/grid_search.html)

4. The methods section contains detailed explanations of study design including ML techniques used. A section on subjects is included, but this should have its own section.

We appreciate that a seperate section for participants can improve a readers’ ability to scan the paper for key information. We pulled out the description of participants to a distinct section.

5. The results section has some discussion of the results, which should be moved to the discussion section e.g. page 11 lines 17-27.

Within the results section, there is some narrative that is intended to help readers better understand the results, for example the description of results from the SHAP algorithm. We suspect that many or most readers will be seeing these kind of results for the first time. We acknowledge that these descriptions contain some interpretations but prefer to leave this text in place to support the understanding of results. We feel it is most helpful to the reader to see these explanations as close to the visualizations as possible. If it’s felt that readers are well versed in machine learning models and outputs, we would be happy to move these expanded descriptions to the discussion section.

6. The low sensitivity found for the models should be discussed – what are the implications for a biomarker predicting treatment response.

This is an excellent question, which highlights the trade-offs between specificity and sensitivity that typically occurs when developing a biomarker. We acknowledge this trade-off in the 3rd paragraph of the model validation section as well as in the 3rd paragraph of the discussion section. In the limitations section, we added a paragraph to highlight the clinical implications of a biomarker with lower sensitivity.

Reviewer: 2

Major Comments ----------

1. Measurement of blood pressure is notoriously error prone. The 2017 ACC/AHA guidelines recommend multiple measurement occasions and specific data discard protocols for both clinician and home BP monitoring. In the UK, NICE now recommends that at least 14 measurements be taken prior to a diagnosis of hypertension.

1.1 Given the critical importance of ground truth labels in supervised machine learning, how did you ensure measurement validity?

We acknowledge the large variance endemic to blood pressure data and used averages over 6-day intervals for the baseline and end-values (ground truth) to help address this concern. A description of how the averages were calculated have been added to the manuscript in the methods sections paragraphs 1 and 2. Further, while efforts have been taken to ensure measurement validity, the type of error described is non-systematic error, or “noise”, which, if present, tends to bias analysis to the null hypothesis. So, if this type of error is present, it would tend to cause us to underestimate treatment effects, thereby reducing the predictive power of our algorithms.

1.2 Please provide clear detail about how and by whom the measurements that dictated entry into the study and 7-14 week change were performed.

1.3 In particular, please describe how, if multiple measurements occurred between 7-14 weeks, the values used to compute the BP delta were selected. There is an apparent risk of investigator bias if values could be selectively chosen.

In response to both points 1.2 and 1.3, we appreciate the risk of manually selecting data to include in analysis. To remove that risk the measurements included in the blood pressure values used to calculate the change over the study interval were determined by the defined 6-day intervals as described above (response to 1.1). All values that fell into the defined interval were included in the average for baseline and end program. In addition, we have added the average number of

observations and 95% CI of both baseline and end-intervention values in the end of the first paragraph in the Results - Dataset section.

2. Given that participants self-selected into the study (Methods, Page 4, Line 47) how did you control for the possibility of treatment stage acting to confound the effects of the digital intervention (and inferences about behavioural predictors)? Individuals with a new diagnosis of hypertension may be highly motivated to try new strategies for blood pressure control (such as a digital therapeutic) but are likely *also* to be titrating with antihypertensives which might reasonably be expected to drive BP reductions - particularly in the short term. Please describe treatment stage information, if available, and account for this potential risk.

We agree that time since diagnosis can modulate the level of motivation for making lifestyle changes and the stability of an individual's medication regime. Unfortunately, in this retrospective analysis, we did not have complete data on time since diagnosis for all participants so did not feel it appropriate to include it in our manuscript. For those with data available, the average time since diagnosis was 12.4 years (95% CI 8.9 to 15.9) which is consistent with values found in our other cohort studies and makes it less likely that time since diagnosis had a major impact.

Nonetheless, the lack of data on the time of diagnosis and changes to medications over the study interval are known limitations of this study. The lack of data on changes in medications is currently acknowledged in the limitations section of the manuscript and we added the limitation of not including the time since diagnosis as an explanatory variable in the model to third paragraph.

3. You propose that Shapley values could be used to prioritise behavioural 'gaps' for individuals to subsequently target (Results, Page 11, Lines 17-27).

3.1 In interpreting a Shapley value assigned to a given feature how do you know that this is, in fact, modifiable as a behavioural target? Is there a way to evaluate the potential value range for each feature? This matters because if a value is essentially fixed then it does not provided the basis for prioritised decision making that you argue.

We agree that a fixed value such as baseline blood pressure or baseline BMI do not provide opportunities for treatment adjustments based on their Shapley values. This is why the model intentionally also includes modifiable explanatory variables such as the number of plant-based meals consumed or exercise minutes achieved. In translating Shapley values to a participant user interface/user experience (UI/UX), we would have to flag variables that are theoretically modifiable or not so they are appropriately displayed to motivate changes or add context. We added mention of this methodology to the Making Use of Explainable AI section i the second paragraph. Since the biomarker is trained only on actual range of values achieved by prior participants, it makes it more likely Shapley values accurately represent plausible modifications to explanatory variables.

3.2 You note the risk of confusing correlation and causality in Discussion (Page 13, Lines 10-16) and state that this 'does not preclude the automated use of explanatory variables to guide [...] personalization.' Why? Given that each behavioural strategy comes with concrete costs to individuals, is it not a highly relevant ethical and QoL concern if there are doubts as to whether specific markers provide a genuine rational basis for action?

Thank you for these excellent questions. The nature of predictive models is that explanatory variables are correlated with the prediction made. It is true, as discussed in the manuscript that this does not mean they are causative. However, at the same time, it does not mean they are not causative. In fact, some explanatory variables were chosen for inclusion because our hypothesis is that they are causative. Thus, there is a real probability that using explanatory variables to guide personalization may enhance outcomes. This is one reason why we use the term hypothesis-driven approach to describe our methodology. However, a cost-benefit analysis of adhering to personalizations derived from explanatory variables can not be known from an internal validation study. Instead, predictive models need to go through three stages of evidence: internal validation, external validation and finally, an impact analysis which could determine whether implementing the biomarker (and corresponding personalization of behavioral therapy) causes an improvement of outcomes and at what cost. These are important constructs, so we added mention of them, as well as three additional references that detail these methods, in the Future Work section.

3.3 How do you counter the concern that, since the ingredients in any multi-component digital behavioural intervention are expectedly evidence-based (and pre-selected for impact on health outcomes), there is no mystery in relation to gaps in self-management behaviour that will drive outcomes: an individual simply needs to do the parts of the intervention that they are not doing already. In other words, does the explainer approach really add much over and above that which could be ascertained from simple utilisation/adherence data? If it does, please explain why.

This question nicely highlights the distinction between a health-education intervention and behavioral therapy. If people did not struggle to adhere to recommended behaviors, there would be no need for behavioral therapy. But in practice, simply telling people what they should do is not extremely effective and it is highly unlikely that participants will achieve maximal adherence to all behavioral strategies that are recommended. This is why personalization of recommendations and feedback are important strategies for achieving behavior change. Of course, the extent of value offered by an explainer approach can not be fully known without an impact analysis, as added per our response to 3.2.

Minor Comments ----------

1. Is it realistic that clinicians would consider stopping antihypertensive medication in the short term on the basis of predictors identified within a digital intervention (Methods, Page 6, Lines 6-7)? Although a case for rationalising therapy has been made recently (e.g. doi: 10.1097/HJH.0000000000001405) only a quarter of people appear to be able to sustain BP improvements without medication and neither the ACC/AHA (despite the citation on Page 6, Line 7) nor NICE guidelines include protocols for titrating down medication. Although not central to your

subsequent argument (in Discussion) about the value of the approach, please consider revising this statement.

This is an excellent discussion, although we believe it to be beyond the scope of this manuscript. It is very true that there are insufficient deprescription guidelines in place, which is one reason we began this line of work. However, in the US and many parts of the world, there is increasing recognition by patients, providers and health systems of the over-reliance of pharmacotherapy for cardiometabolic conditions that are caused by poor behaviors. If we were able to help a quarter of this population reduce reliance on pharmacotherapy, it would be a major public health achievement and greatly lower the cost of health care. From our perspective, and to your point, because not everyone will be able to rely fully on behavioral therapy alone it increases the value of developing biomarkers to help determine who will succeed or not.

2. Abstract, Page 2, Line 19. 'Ad libitum' is not accessible to an international readership whose first language is not Latin. Please rephrase.

Thank you for pointing out the use of a phrase that may not be understood by all readers. Ad libitum is a commonly used term of art in the biomedical literature and we believe it to be well understood. For the moment, we have left the term as is, but would be more than willing to change it to “at will” at the discretion of the editor.

3. Methods, Page 7, Line 13. Please describe your approach to hyperparameter optimization.

We used a cross validated grid search method for the hyperparameter optimization. Specifically we used the parallelized hyperparameter optimization package to run multiple points in the grid in parallel to speed up the search (https://github.com/jmcarpenter2/parfit). We chose to use this traditional method for this low dimensionality data set because it is straightforward to implement and fast. The reference for this method has been added to the manuscript on page 7.

4. Discussion, Page 11, Line 54. Is this approach really correctly characterised as being analogous to a credit score? A credit score is intended to support inferences about future behaviour on the basis that past borrowing behaviour is likely to dictate future profligacy. In the context of health behaviour change, it is not clear that success in achieving one lifestyle modification (e.g. sodium reduction) dictates a) future ability to maintain this behaviour or b) success in achieving any other behavioural modification.

We appreciate this question and admit the credit score is not a perfect analogy but one we felt was known widely and illustrative. Like a credit score, the approach outlined is also supporting inferences about past behavior as it pertains to future behavior, which is then validated by a change in outcome. Like a credit score, the inputs are multifactorial and each component may carry different weights for individuals, and the output can be displayed on numeric scale. It is true that the biomarkers were not

designed to predict long-term ability to maintain these behaviors.That could certainly be the subject of additional work!

5. Figure 3, Page 20. To aid interpretation, consider annotating Figure 3 to highlight the direction on the Y axis that tends to increase the likelihood that the overall prediction will favour a successful reduction in BP, and using 'probability of successful reduction in BP' (or similar) rather than 'output value' (assuming that this interpretation is correct.)

Thank you for this suggestion. We agree that relabeling ‘output value’ to the more descriptive title of ‘probability of successful BP reduction’.

VERSION 2 – REVIEW

REVIEWER	Colin Simpson Victoria University of Wellington
REVIEW RETURNED	20-Jun-2019

GENERAL COMMENTS	My points have been addressed.

REVIEWER	Kit Huckvale Black Dog Institute, UNSW Sydney
REVIEW RETURNED	01-Jul-2019

GENERAL COMMENTS	Thank you for the opportunity to re-review this revised manuscript and to the authors for their thoughtful responses to peer review. In my original review I highlighted three major questions/concerns which the authors have addressed effectively in their revised manuscript: 1) Relating to participant selection/blood pressure measurement (addressed by detailing the BP averaging procedure in Methods and coverage statistics in Results.) 2) Concerning self-selection effects (addressed in Limitations.) 3) Concerning Shapley values and their interpretation/utility (addressed in Future Work and in responses to reviewers.)