


Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach

Digital Health
Volume 7: 1–11
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076211060659
journals.sagepub.com/home/dhj


Silvan Hornstein^{1,2} , Valerie Forman-Hoffman¹, Albert Nazander¹,
Kristian Ranta¹ and Kevin Hilbert²

Abstract

Objective: Predicting the outcomes of individual participants for treatment interventions appears central to making mental healthcare more tailored and effective. However, little work has been done to investigate the performance of machine learning-based predictions within digital mental health interventions. Therefore, this study evaluates the performance of machine learning in predicting treatment response in a digital mental health intervention designed for treating depression and anxiety.

Methods: Several algorithms were trained based on the data of 970 participants to predict a significant reduction in depression and anxiety symptoms using clinical and sociodemographic variables. As a random forest classifier performed best over cross-validation, it was used to predict the outcomes of 279 new participants.

Results: The random forest achieved an accuracy of 0.71 for the test set (base rate: 0.67, area under curve (AUC): 0.60, $p = 0.001$, balanced accuracy: 0.60). Additionally, predicted non-responders showed less average reduction of their Patient Health Questionnaire-9 (PHQ-9) (-2.7 , $p = 0.004$) and General Anxiety Disorder Screener-7 values (-3.7 , $p < 0.001$) compared to responders. Besides pre-treatment Patient Health Questionnaire-9 and General Anxiety Disorder Screener-7 values, the self-reported motivation, type of referral into the programme (self vs. healthcare provider) as well as Work Productivity and Activity Impairment Questionnaire items contributed most to the predictions.

Conclusions: This study provides evidence that social-demographic and clinical variables can be used for machine learning to predict therapy outcomes within the context of a therapist-supported digital mental health intervention. Despite the overall moderate performance, this appears promising as these predictions can potentially improve the outcomes of non-responders by monitoring their progress or by offering alternative or additional treatment.

Keywords

Digital mental health, machine learning, outcome prediction, precision psychiatry, depression, general anxiety disorder

Submission date: 13 July 2021; Acceptance date: 30 October 2021

Introduction

Machine learning (ML), the ability of computers to learn without being explicitly programmed,¹ has turned into a central source of innovation in healthcare.² As computational power constantly increases and big datasets become available, ML approaches try to outperform human actors in diverse tasks by extracting information from large

¹Meru Health Inc, Palo Alto, CA, USA

²Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

Corresponding author:

Silvan Hornstein, Meru Health Inc, 470 Ramona Street, Palo Alto, CA 94301, USA.

Email: silvan.hornstein@meruhealth.com



amounts of data.^{3,4} In healthcare, this allows making clinical decisions regarding treatment and diagnosis based on much more data than what was previously possible, thus tailoring treatment to patients.⁵ This idea, frequently labelled as ‘precision medicine’,⁶ is particularly relevant for mental healthcare. First, mental illness is one of the biggest threats to well-being in general⁷ and has large economical consequences as well.⁸ Innovation in this field could therefore have a massive positive impact on public health. Second, ‘one-treatment-fits-all’ approaches seem not to work sufficiently well. For example, a majority of psychiatric disorders have meta-analytic response rates of 50% and below for cognitive behavioural therapy,⁹ which is thought of as a standard of care.¹⁰ The application of ML to large datasets of patients in mental health treatment might yield clues about what types of patients are likely to benefit from particular treatments. At the current time, patients not experiencing a reduction of their symptoms through standard of care interventions could potentially benefit most from individualization. Identifying these patients early on in treatment might help these patients achieve better outcomes by either offering them alternative or extended forms of treatment.

Indeed, numerous authors have trained machine learning models to predict therapy outcomes for all kinds of disorders. For depression, a review by Lee and colleagues¹¹ reported 26 such studies that used diverse data sources such as functional magnetic resonance imaging (fMRI),¹² electroencephalogram (EEG),¹³ genetic¹⁴ or phenomenological data,¹⁵ where ML algorithms achieved an overall accuracy of 0.83 for therapy outcomes.

However, this review also highlights a few of the limitations of the field that seem to be major obstacles to actually implementing ML-based outcome prediction in mental healthcare. On one hand, half of all reported studies had sample sizes of less than 100. This is related to the high amount of fMRI studies in the aforementioned review, which is problematic because of its difficulty in actually implementing in a real-life healthcare setting.¹⁶ This also shows that while ‘big data’ is available for ML applications in other fields, these datasets are only recently emerging in the field of mental health. On the other hand, most of the studies were done retrospectively and just a subset of all studies evaluated an algorithm’s performance on a test set. Therefore, the given accuracies will likely be overestimated. As an example for performance differences between cross-validation and test sets, Checkroud and colleagues¹⁷ predicted the outcome of antidepressant treatment with a notably big dataset of more than 4000 patients, using clinical and sociodemographic data. The model reached an accuracy of 0.65 in cross-validation but just 0.60 and 0.51% on two independent test sets. Similarly, Hilbert and colleagues predicted the outcome of cognitive behavioural therapy (CBT) treatment in a test set with a balanced accuracy of 0.59 in a sample containing diverse disorders of

which a majority had anxiety or depression.¹⁸ Interestingly, while there are increasing efforts to provide treatments for mental health problems via digital channels, such as apps or online programmes,¹⁹ there is little work on ML-based outcome prediction for such digital mental health (DMH) interventions, particularly for depression and anxiety. This is surprising because DMH solutions have been shown to be quite effective.²⁰ Integrating ML in these solutions, however, appears relevant for addressing the aforementioned shortfalls of the field. First, implementation is easier, as the path from ideation to production is arguably shorter in a DMH product. Second, in DMH, it is easier to collect large amounts of data, frequently even just as a side product of day-to-day business. Third, due to the continuous collection of new data, designing a realistic evaluation of performance on new test data are much easier. Finally, outcome prediction for DMH interventions has revealed promising results for other mental disorders such as pediatric obsessive compulsive disorder. Here, clinical baseline variables were used to predict outcomes in an Internet-delivered CBT treatment, reaching accuracies of more than 75% for the sample of 61 adolescents.²¹ This approach yielded predictive performance levels similar to non-DMH interventions.²²

Therefore, this paper will present a machine learning approach that predicts outcomes for participants of a therapist-supported DMH intervention for depression and anxiety. It is hypothesized that such an approach can reproduce the accuracies achieved by studies using similar data and a test set^{17,18} while having the benefits of being much closer to actual implementation.

Methods

Intervention

This study used anonymous participant data from Meru Health, a healthcare provider with an evidence-based, therapist-supported DMH intervention for depression and anxiety.²³ The Meru Health Program (MHP) is a DMH treatment delivered via smartphone for depression and anxiety, consisting of mindfulness, cognitive behavioural therapy techniques, psychoeducation and heart rate variability biofeedback (HRVB) using an ear-worn HearthMath® Bluetooth photoplethysmography (PPG) sensor. The intervention also includes support by a licensed clinical therapist that reviews participants’ engagement and outcomes and interacts with the participants through asynchronous chat messages on a regular basis. The intervention takes between 8 and 12 weeks, with every week having a dedicated topic such as sleep that is addressed in psychoeducative videos as well as exercises that are suggested to participants on a daily basis. For example, the week titled ‘Mind on Autopilot’ begins by introducing the concept of mindfulness in a video. Later in this week, there is a practice

on mindful eating, as well as an exercise where participants are guided to count their breaths. Participants can also share their progress within an anonymous peer group of other participants. Outcomes of the intervention are measured with biweekly Patient Health Questionnaire-9 (PHQ-9)²⁴ and General Anxiety Disorder Screener-7 (GAD-7)²⁵ questionnaires. Prior studies have shown a significant decrease in both depression and anxiety symptoms at the end of the treatment,²⁶ as well as at 12-month follow-up.²⁷ The MHP is usually provided as a healthcare benefit through an employer to its employees, with the main customers in the past being from Finland and the USA. Participants that entered the MHP through their employer by signing up on their own are referred to as ‘self-referrals’. Other participants were referred to the MHP by healthcare professionals. These are referred to as ‘healthcare-professional-referrals’.

Sample

An unbiased selection of participants that entered the MHP between September 2017 and September 2020 was used as a sample for this study. All participants provided informed consent for their anonymized data to be used for research purposes before they started with the intervention. Participants who never started the programme, as measured by activity in the app (55 participants), as well as those without a PHQ-9 and GAD-7 value for the beginning of the programme (130 participants), were excluded. Finally, those that did not score at least 5 points on either the PHQ-9 or the GAD-7 were excluded as well (24 participants), which lead to an overall sample size of 1249 participants. Data were available for 39 potentially relevant variables, including all PHQ-9 and GAD-7 items, as well as the items of the Work Productivity and Activity Impairment (WPAI) questionnaire,²⁸ a single-item burnout score,²⁹ sociodemographic variables, several variables regarding the history of the patient assessed by the therapist in the intake call (psychiatric hospitalizations, previous depressive episodes...), patients’ self-report of currently taking medication and patients way into the programme (whether they referred themselves, or were referred by a health care professional), as well as payment for service (free vs. having a co-pay for the service). The used data sources as well as preprocessing steps can be found in Table 1. Sample characteristics can be found in Table 2. For training the algorithm, data from all participants that started the programme before 29 June 2020 were used, which resulted in a train set of 970 participants. The remaining 279 participants that started the programme after that were used as a test set. Compared to the train set, the new participants from the test set had slightly higher GAD-7 values at the beginning of the programme (mean overall GAD-7 score of 12.2 in the test and 11.0 in the train set, $p < 0.001$). Also, while some of the participants from the train set were located in Finland (43%), the

whole test set consisted of participants in the US. This study was reviewed by the Pearl Institutional Review Board, which granted an exemption for analyses of previously collected and de-identified data.

Outcome variable

The primary outcome was defined as a considerable response to treatment (‘responders’, binary: yes or no), defined as having a clinical significant reduction in anxiety or depression symptoms. The cut-off for the outcome was either at least a 5-point reduction at the PHQ-9 or at least 4 points reduction of the GAD-7, which are both representations of the minimal values for significant improvement of symptoms.^{30,31} This dichotomization of the outcome was chosen to align with past approaches that also classified outcomes.^{17,18} Also, as clinical decision-making is mostly binary (e.g. contact a patient or not), a binary prediction seems most usable from a practitioner’s point of view. For those who did not finish a GAD-7/PHQ-9 questionnaire at the end of the programme, the last observation was carried forward to calculate symptom reduction, which was the case for 37% of the participants. We used this approach based on the assumption that those disengaging with the programme and not completing the final questionnaire did not experience any additional improvements. 58.9% of the participants in the train set and 67.1% of the participants in the test set were responders following the given definition. Figure 1 visualizes the distribution of PHQ-9 and GAD-7 scores at the beginning as well as the end of the programme and of the changes in the overall score. Sample characteristics by response/no-response can be found in Table 2 as well.

Data preprocessing and training

While some basic data cleaning, such as dichotomizing variables as the medication, was done before this (Table 1), most preprocessing and the actual training of algorithms was performed with scikit-learn 0.22.2.³² The decision on algorithm and feature selection, hyperparameter tuning and preprocessing steps to be used on the incoming test set was made by evaluating performance in the train set over a 10-fold cross-validation. This was done to reserve the test set for a one-time final evaluation of the performance of the algorithm on new data, to avoid overfitting on the test data. Additionally, this resembles the actual process of implementing and using a ML approach, as for this the specifications of the algorithm have to be decided before predictions are made and used for any further interventions. Out of the algorithms implemented in scikit-learn for supervised classification, the performance of logistic regression, a support vector machine, naive Bayes and a random forest (RF) was evaluated. Because all non-outcome variables included missing data of up to around

Table 1. Data sources, preprocessing steps and missing data.

Variable(s)	Source	Preprocessing	Missing data (%)
Sex	Participants self-disclosure	As 'others' was not chosen at all, sex was dichotomized.	10%
Age	Participants self-disclosure	Calculated as (Year of Sign-up - Birth Year).	2%
Referral	Entered by the care coordinator when checking eligibility of participants.	Dichotomized to 'self-referral' and 'healthcare professional referral'.	3%
Payment	Entered by the care coordinator when checking eligibility of participants.	Dichotomized to 'Free' and '(Co)-Pay'.	14%
Motivation	Participants self-disclosure in intake call.	-	5%
Medication	Participants self-disclosure, clarified in intake call.	Dichotomized to 'Yes/No'.	20%
PHQ-9 baseline	Questionnaire presented in the app before first call with the therapist.	Single items added up for overall score.	Score: 0%. Individual items: 12% ^a
GAD-7 baseline	Questionnaire presented in the app before the first call with the therapist.	Single items added up for overall score.	Score: 0%. Individual items: 10% ^a
WPAI	Questionnaire presented in the app before the programme starts.	Scores calculated out of the items as suggested. ³⁰	18-40%
Burnout score	Participants self-disclosure	-	36%
History of trauma	Participants self-disclosure, clarified in intake call.	Dichotomized to 'Yes'/'No', 'Unknown' was set NA.	33%
Major depressive episodes	Participants self-disclosure, clarified in intake call.	-	36%
Psychiatric hospitalizations	Participants self-disclosure, clarified in intake call.	-	29%
Suicide attempts	Participants self-disclosure, clarified in intake call.	-	29%

^aAll items were answered by all participants, but due to technical reasons for these participants, just the overall sum score but not the individual scores was available.

GAD-7: General Anxiety Disorder Screener-7; PHQ-9: Patient Health Questionnaire-9; WPAI: Work Productivity and Activity Impairment.

40%, imputations were needed. Here, simple mean imputation was compared with the iterative imputer implemented in sklearn, as part of the cross-validation procedure. This imputer is an implementation of multivariate imputations using chained equations.³³ For each algorithm, the best performing, as measured by the receiver operating characteristics area under curve (ROC AUC score) and imputation method were chosen and this performance was compared between the algorithms. For the most promising algorithm, hyperparameter tuning was performed using exhaustive grid search, as well as univariate feature selection using

sklearn SelectKBest. For the RF, the following hyperparameters were tested: number of trees (from 200 to 2000), the maximum depth of the trees (from 10 to 100), maximum number of features (auto, sqrt), minimum number of samples required to split a node (2 to 20) and required number of samples in a leaf (1 to 4). The whole train-/test set procedure is visualized in Figure 2. Finally, a paired Student's *t*-test was performed comparing the best performing classifier with the logistic regression model as a baseline. This was intended to ensure that the gains in performance are above chance level. The correction

Table 2. Full sample characteristics, as well as for responders and non-responders. (Mean values, Standard Deviation in Brackets).

	Full sample (n = 1236)	Response (n = 751)	No response (n = 485)
Female sex	76.3%	79.2%	71.9%
Age (years)	38.9 (11.3)	39.4 (11.5)	38.4 (11.0)
PHQ-9 baseline score	12.0 (5.5)	12.5 (5.6)	11.3 (5.4)
GAD-7 baseline score	11.3 (4.6)	12.0 (4.4)	10.1 (4.6)
PHQ-9 final score	7.8 (5.7)	5.6 (4.1)	11.3 (5.9)
GAD-7 final score	7.2 (4.8)	5.2 (3.4)	10.1 (5.1)
PHQ-9 change over programme	-4.2 (5.3)	-6.9 (4.8)	-0.0 (2.9)
GAD change over programme	-4.1 (4.9)	-6.8 (4.3)	-0.0 (2.1)
WPAI absenteeism	10.7 (22.9)	11.1 (22.8)	10.1 (20.7)
WPAI presenteeism	45.7 (26.4)	46.1 (25.8)	44.9 (27.5)
WPAI work productivity loss	49.8 (28.2)	50.5 (27.5)	48.4 (29.2)
WPAI activity impairment	51.6 (25.6)	50.1 (25.6)	52.7 (25.6)
Burnout score	3.0 (1.0)	3.0 (1.1)	2.9 (1.1)
Motivation score	8.5 (1.2)	8.6 (1.2)	8.3 (1.2)
History of major trauma	39.8%	39.8%	39.6%
Episodes of major depression	2.0 (2.7)	1.9 (2.6)	2.1 (2.8)
Psychiatric hospitalizations	0.1 (0.4)	0.07 (0.4)	0.05 (0.3)
Suicide attempts	0.1 (0.4)	0.09 (0.4)	0.06 (0.4)
Medication (Yes/No)	42.5%	41.8%	43.6%
Type of referral (self vs. healthcare professional)	71.7%	74.3 %	67.5%

GAD-7: General Anxiety Disorder Screener-7; PHQ-9: Patient Health Questionnaire-9; WPAI: Work Productivity and Activity Impairment.

recommended by Nadeau and Bengio³⁴ was used to handle the violated assumption of independence when testing differences of classifiers over cross-validation. Following the recommendation of Bouckaert and Frank,³⁵ the test was calculated over a 10 times repeated 10-fold cross-validation procedure to ensure reproducibility of the results.

Testing of algorithms performance and feature importance

The best performing algorithm over the training procedure was used to predict the expected treatment response for the 279 participants from the test set. The decision for such a time-based train-test split was made as this resembles how the algorithm would actually be used in practice (using old data to make predictions for newly incoming participants) and therefore adding to the external validity of the approach. Performance was evaluated by randomly shuffling the labels of the train data and the test data for 5000 times and afterwards comparing the accuracy reached on the permuted data with the accuracy on the original data. Such permutation-based approaches to the significance of algorithm performance allow to check whether a classifier actually found a signal in the data.³⁶ The same procedure was repeated for other performance metrics as well. Permutations were also used to better understand how the chosen algorithm works. As multicollinearity is a major threat to most feature importance approaches,³⁷ strongly correlated features were grouped together and randomly shuffled at once, to calculate an approximation of permutation importance, which measures the feature importance by calculating the decrease in performance when shuffling a predictor.³⁸ Based on the intercorrelations, PHQ-9 items, GAD-7 items and WPAI items were grouped together. Also, suicide attempts and past psychiatric hospitalizations were grouped together. The remaining variables showed clearly lower intercorrelations (below 0.4) and were therefore not grouped together. The random shuffling of the columns was repeated 500 times to eliminate randomness in the results, then mean decrease in accuracy was measured. This approach is an adaption of the permutation importance implementation in scikit-learn, with the grouping of variables added to deal with the high amount of collinearity in the data.

Results

Cross-validation

A RF classifier was chosen as the best performing algorithm over the cross-validation procedure (Figure 2). RF reached a mean ROC AUC of 0.64 (SD = 0.04). The support vector machine scored 0.63 (SD = 0.06), the logistic regression 0.61 (SD = 0.03) and the naïve Bayes 0.60 (SD = 0.04). These performance metrics were reached while using

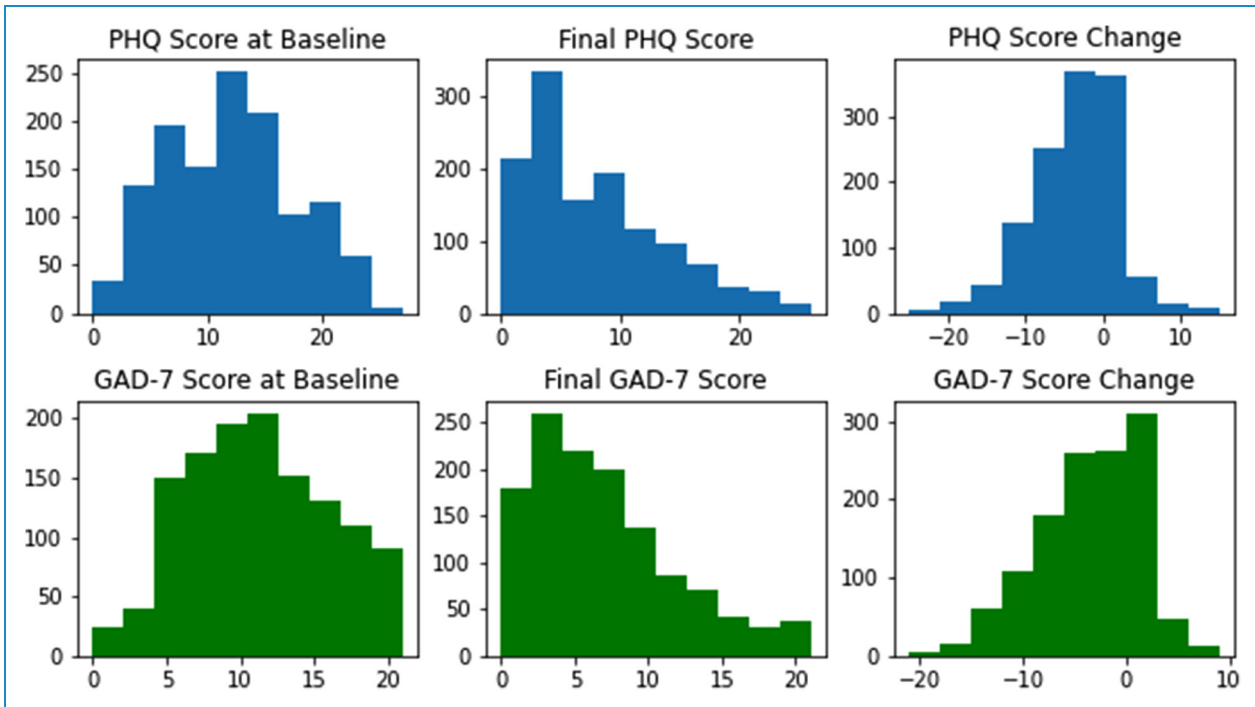


Figure 1. Distribution of PHQ-9 and GAD-7 values on the baseline (left) of the final score (middle) and of the change (right). GAD-7: General Anxiety Disorder Screener-7; PHQ-9: Patient Health Questionnaire-9.

scikit-learn's simple imputer. The use of the iterative imputer increased the performance slightly for the logistic regression (0.62, SD = 0.03), but did not influence the performance of the naïve Bayes and decreased the performance of the SVM (0.62, SD = 0.05) and the RF (0.63, SD = 0.05). Therefore, the simple imputer was used for the final predictions on the test set. Hyperparameter tuning and feature

selection was evaluated for the RF as best performing algorithm. Reducing the number of variables did not increase the performance but, in contrast, strongly lowering the number of predictors had a negative impact on ROC AUC. For example, using 16 variables selected by univariate feature selection resulted in a mean ROC AUC of 0.61. As none of the explored combination of parameters lead to

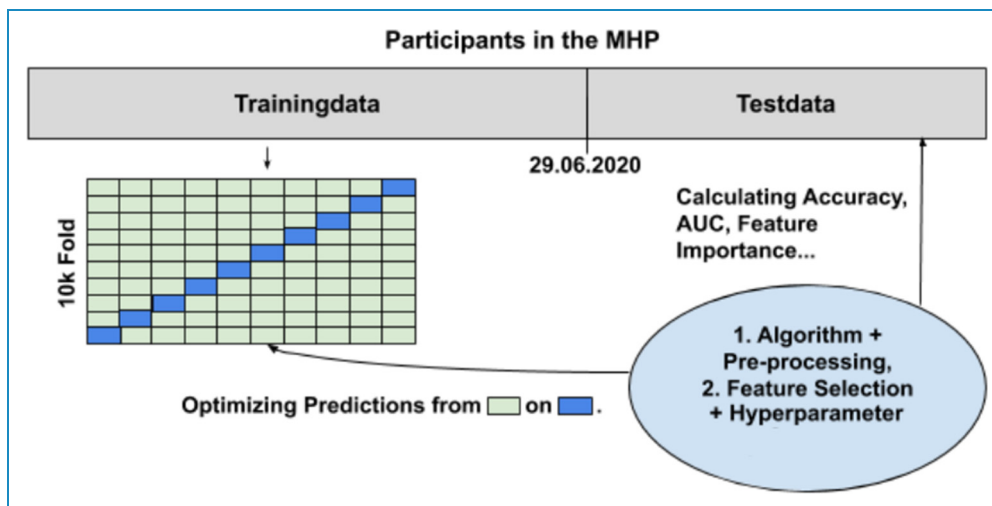


Figure 2. Visualization of the training of the machine learning (ML) algorithm. The best performing solution over a 10-fold cross-validation (CV) is used to predict the outcomes for the test data. Algorithm and preprocessing were selected first, and feature selection and hyperparameter tuning were evaluated afterwards.

clearly improved performance, the default parameters were used for predicting the treatment response of the test set. Finally, the performance of the logistic regression was compared with the performance of the best performing algorithm, the RF, revealing a significant difference in their ROC AUC score ($p = 0.04$). Figure 3 compares the performance of the four different algorithms by plotting their ROC AUC curves.

Test set performance

Applied to the test set, this led to an unbalanced accuracy of 0.71, which was significantly above the base rate of 0.67 ($p = 0.001$), an AUC of 0.60 ($p = 0.001$) and, for the highest accuracy, a sensitivity of 0.72 ($p = 0.04$) and specificity of 0.60 ($p = 0.19$), respectively. The balanced accuracy was 0.60 ($p = 0.004$). Also, comparing the absolute values of symptom reduction reveals that those participants that were predicted to experience response on average had a symptom decrease of 5.1 on the PHQ-9 and 5.5 on the GAD-7, while those predicted as having no response had an average decrease of 2.4 and 1.8 respectively, which is significantly less compared to the responders ($p < 0.001$ for GAD-7, $p = 0.004$ for PHQ-9; independent samples t -test).

Out of the groups of predictors, the PHQ-9 items were the most important, as randomly shuffling them would have decreased the test set accuracy by 4.6%. This was followed by the GAD-7 values (4% decrease), the motivation score (2.75% decrease), the sex of the participants (2.2% decrease), type of referral to the programme (2.1% decrease) and WPAI values (2% decrease). Finally, shuffling the burnout score and the history of major trauma item decreased the performance as well (1.0% and 0.7% decrease, respectively), while the remaining variables left the accuracy nearly unchanged. Therefore, participants with high PHQ-9 and GAD-7 values, high motivation scores, female sex and who referred themselves into the programme had higher chances of good outcomes. It should be noted that these values are not the same as the actual drop in performance when going through the whole procedure of model building and test set predictions without these variables and should be read as an overall estimation of each variable's importance for the final prediction.⁴⁰

Discussion

Predicting therapy outcomes by utilizing ML has been done rarely within a DMH context. Therefore, this paper investigated the predictive performance of ML algorithms for the outcomes of a therapist-supported DMH intervention targeting depression and anxiety. This was evaluated using an independent test set of new participants, to avoid overfitting and reach more realistic performance estimates in a real-world setting. The selected algorithm predicted response to the treatment robustly above chance level

with a comparable performance to similar designed studies from a non-DMH context. This, as well as clearly lower PHQ-9/GAD-7 decreases in the predicted non-responders, provides hope for implementing ML approaches in DMH. Besides PHQ-9 and GAD-7 values, a self-report of motivation, the WPAI values, as well as the type of referral into the programme added most to the predictive power of the algorithm.

As hypothesized, performance metrics were lower than several studies from a less application-heavy context,¹¹ but comparable to those reported from studies with a similar aim, methodology and context.^{17,18} Therefore, this paper successfully demonstrated the applicability of an ML-based outcome prediction approach in DMH, while having a considerable sample size and high external validity, due to data coming from an actual real-world DMH programme. However, for potential applications of predicted outcomes, predictive performance would have to improve clearly. A potential advantage of ML approaches in DMH is that smartphone-based interventions can collect data that is not available in a traditional treatment setting. As data such as app usage has been demonstrated to be informative in predicting mental-health-related outcomes,⁴¹ making use of these data sources appears to be a promising path towards improved predictive accuracy. Also, as outcome prediction studies using neurobiological data regularly outperform the solutions using solely phenomenological data,¹¹ such data could improve performance as well. As some DMH interventions, including the one whose data were used for this study, actually collect biological data as heart rate variability, this data could be integrated into a predictive model quite easily.

Also, whether knowing about those predicted to be non-responders allows efficient improvement of future treatment results depends not just on the accuracy of the algorithm but also on how the predictions are actually used. First studies suggested that monitoring patients, for whom low outcomes are predicted, can help to improve their outcomes.⁴² This appears as a promising use case also for a DMH setting, as the monitoring would be easy to implement and not too cost-intensive. Additionally, studies that allocate patients to treatments based on outcome predictions demonstrated superior outcomes⁴³ or lower costs⁴⁴ and thus highlight another group of interventions that could be built on top of the presented algorithm. However, other studies failed to show the superiority of tailored treatment,⁴⁵ therefore, more research is needed to evaluate individualized treatment procedures based on predicted outcomes before definite conclusions can be made. Finally, it should be noted that besides the technical ability to predict outcomes, there are also numerous ethical challenges that need to be considered for using outcome predictions, such as potential biases of the used algorithm⁴⁶ or potential negative

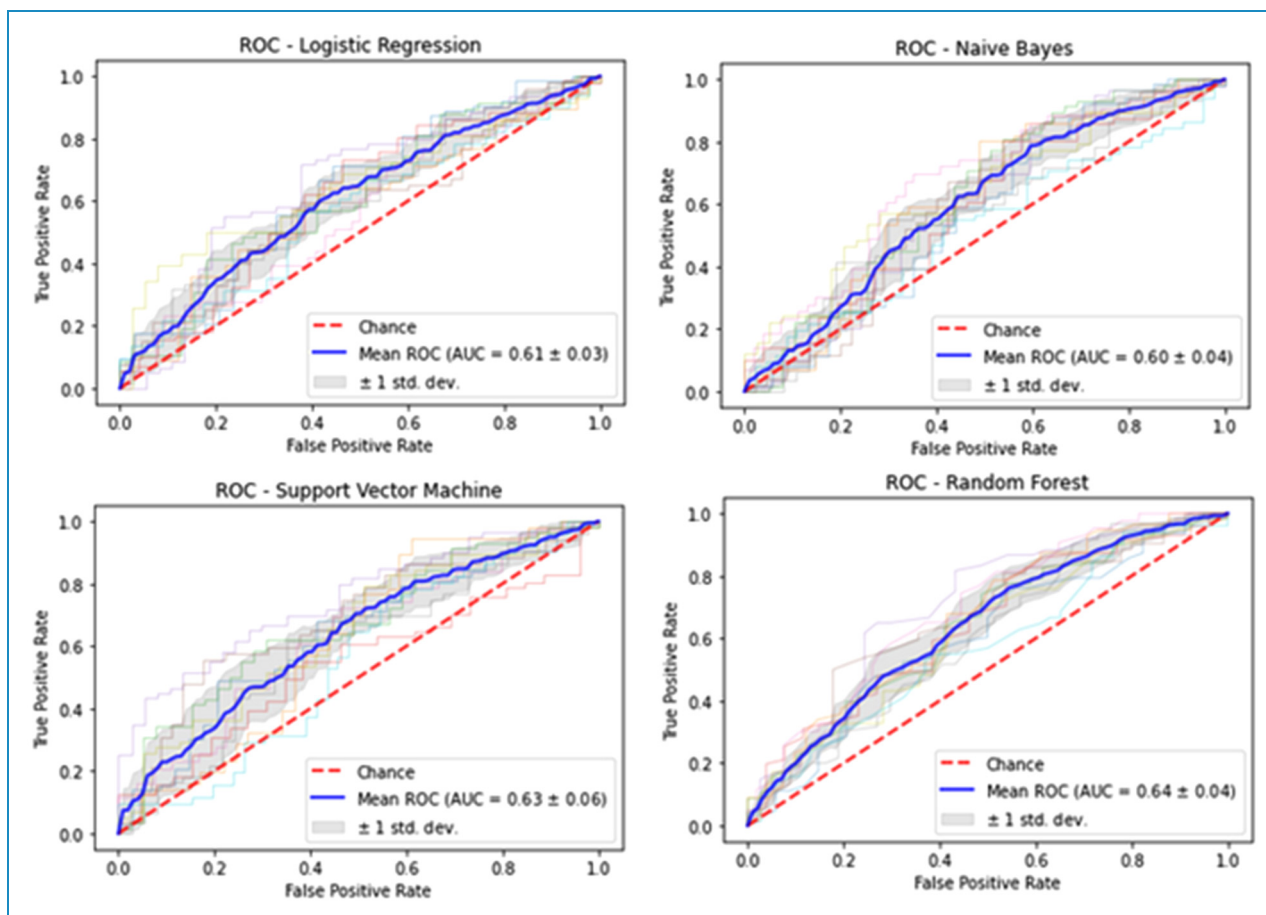


Figure 3. Comparison of the performance of four algorithms over the cross-validation procedure.³⁹ Thin lines represent performance per cross-fold. Simple mean imputations and normalized variables were used.

side effects on other variables than outcomes.⁴⁷ For example, offering additional treatment based on predicted outcomes might increase costs in an unscalable way.

In addition to the overall moderate accuracy of the model derived in the current investigation, the chosen approach had some other limitations. First, due to the dataset coming from a naturalistic setting, there was a considerable amount of missing data. Also, missingness was mostly not random, which is an assumption of several advanced imputation techniques,⁴⁸ but due to changes in the data collected during the intervention. This might have lowered the predictive performance, but also biased the variable importance. For example, the burnout score might have been a more informative predictor if there were not nearly 40% of the data missing. Second, there were considerable differences between the test set and the data the algorithm was trained on originally. This occurred, as the intervention was rolled out into a new population with partly different population characteristics. This might have lowered the predictive performance as well, though the ability to make predictions across diverse populations might also be seen as a strength, proofing external validity of the chosen approach.

Third, comparable to similar studies,⁴⁹ the chosen approach lacks a realistic baseline. Particularly in approaches where easily interpretable data such as questionnaires are used, it might be possible that human clinicians could predict therapy outcomes above chance level as well, and potentially even better than an algorithm. This was rarely explicitly tested, although an approach showing that ML predictions failed to outperform human therapists in forecasting outcomes in alcohol addiction⁵⁰ highlights that collecting and testing against realistic baselines is quite relevant. Fourth, this paper took a very conservative approach on hyperparameter tuning and feature selection, acknowledging the risk that every decision made on a train set increases the risk of overfitting the algorithm, even when using cross-validation methods.⁵¹ Though hyperparameter tuning undeniably can improve an algorithm's performance,¹¹ particularly with increasing sample size and the potential integration of behavioural and neurobiological data sources, future approaches would likely benefit from some additional decisions regarding the algorithms specification. Making these decisions over time and evaluating them on new incoming data

could help to minimize the risk of overfitting. Fifth, the use of LOCF for handling missing data are appropriate under the assumption that those disengaging from the programme did not experience any additional improvements but may introduce bias if participants experience worsening of symptoms after dropping out. Additionally, LOCF is not robust when data are not missing at random (MNAR). However, suitable alternatives within this predictive modelling framework are lacking. Finally, the importance of the strongly correlated predictors was likely overestimated in the feature importance approach.³⁷The final approach in this paper did not use feature selection as this did not reveal significant improvements in performance over the cross-validation procedure. However, as evident in the limitations in interpreting the chosen feature importance approach, feature selection can also increase the interpretability of ML approaches. Interpretability appears particularly relevant for novel applications of ML in real-world settings, as it might benefit the trust of those using it.⁵² Therefore, future similar minded approaches might choose to integrate feature selection more willingly, not just to benefit model's performance but also its interpretability.

This paper predicted the response to a DMH intervention addressing depression and anxiety. While the reached performance was similar to studies from other settings, this performance was only moderate and would need to increase to monitor outcomes or influence the treatment decisions. Integrating behavioural and neurobiological data, as regularly collected from DMH interventions, might help to improve predictive accuracy and get closer towards more precise, individualized mental healthcare.

Acknowledgements: The authors would like to thank Meru Health for the possibility to conduct this study.

Author's contribution: SH and KH performed data analysis. SH, KH and VFH did literature research. VFH gained ethical approval, SH, VFH, AN, KR and VFH reviewed edited the manuscript and approved the final version of the manuscript.

Declaration of Conflicting Interests: The author(s) declared the following potential conflicts of interest with respect to the research, authorship and/or publication of this article: Xxxxxxx. Mr Hornstein is employed as a Data Scientist at Meru Health, Inc, receives salary from the company and owns options of the company. Dr Forman-Hoffman is employed as the Head of Research at Meru Health, Inc, receives salary from the company and owns options of the company. Mr Nazander is the Chief Technology Officer (CTO) and one of the founders at Meru Health, Inc, receives salary from the company and owns options of the company. Mr Ranta is the Chief Executive Officer (CEO) and one of the founders at Meru Health, Inc, receives salary from the company and owns options of the company. Dr Hilbert does not have any competing interests.

Ethical approval: The ethics committee Pearl IRB approved this study (Protocol #20-MERU-109).

Funding: The author(s) received no financial support for the research, authorship and/or publication of this article.

Guarantor: SH.

Informed consent: All participants provided informed consent for their anonymized data to be used for research purposes before they started with the intervention.

ORCID ID: Silvan Hornstein  <https://orcid.org/0000-0002-0398-7096>

Trial registration: Not applicable, because this article does not contain any clinical trials.

References

1. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev* 1959; 3: 210–229.
2. Beam AL and Kohane IS. Big data and machine learning in health care. *JAMA* 2018; 319: 1317–1318. PMID: 29532063.
3. Jordan MI and Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015; 349: 255–260. PMID: 26185243.
4. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; 577: 89–94. PMID: 31894144.
5. Rajkomar A, Dean J and Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380: 1347–1358. PMID: 30943338.
6. Collins FS and Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015; 372: 793–795. PMID: 25635347.
7. Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet* 2012; 380: 2163–2196. PMID: 23245607.
8. Trautmann S, Rehm J and Wittchen HU. The economic costs of mental disorders: do our societies react appropriately to the burden of mental disorders? *EMBO Rep* 2016; 17: 1245–1249. PMID: 27491723.
9. Hofmann SG, Asnaani A, Vonk IJ, et al. The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cognit Ther Res* 2012; 36: 427–440. PMID: 23459093.
10. David D, Cristea I and Hofmann SG. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Front Psychiatry* 2018; 9: 4. PMID: 29434552.
11. Lee Y, Raguett RM, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J Affect Disord* 2018; 241: 519–532. PMID: 30153635.
12. Costafreda SG, Chu C, Ashburner J, et al. Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One* 2009; 4: e6353. PMID: 19633718.

13. Mumtaz W, Xia L, Mohd Yasin MA, et al. A wavelet-based technique to predict treatment outcome for major depressive disorder. *PLoS One* 2017; 12: e0171409. PMID: 28152063.
14. Belzeaux R, Lin CW, Ding Y, et al. Predisposition to treatment response in major depressive episode: a peripheral blood gene coexpression network analysis. *J Psychiatr Res* 2016; 81: 119–126. PMID: 27438688.
15. Kautzky A, Dold M, Bartova L, et al. Refining prediction in treatment-resistant depression: results of machine learning analyses in the TRD III sample. *J Clin Psychiatry* 2017; 79: 0–0.
16. Dwyer DB, Falkai P and Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol* 2018; 14: 91–118. PMID: 29401044.
17. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry* 2016; 3: 243–250. PMID: 26803397.
18. Hilbert K, Kunas SL, Lueken U, et al. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. *Behav Res Ther* 2020; 124: 103530
19. Tal A and Torous J. The digital mental health revolution: opportunities and risks. *Psychiatr Rehabil J* 2017; 40: 263–265. PMID: 2889165820.
20. Lattie EG, Adkins EC, Winquist N, et al. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. *J Med Internet Res* 2019; 21: e12869. PMID: **31333198**.
21. Lenhard F, Sauer S, Andersson E, et al. Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: a machine learning approach. *Int J Methods Psychiatr Res* 2018; 27: e1576.
22. Hilbert K, Jacobi T, Kunas SL, et al. Identifying CBT non-response among OCD outpatients: a machine-learning approach. *Psychother Res* 2021 Jan; 31(1): 52–62.
23. Solution. Meru Health website. Accessed Decembre 2, 2020. <https://www.meruhealth.com/solution>
24. Kroenke K and Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 2002; 32: 509–515. PMID: 11914441.
25. Spitzer RL, Kroenke K, Williams JB, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006; 166: 1092–1097. PMID: 16717171.
26. Goldin PR, Lindholm R, Ranta K, et al. Feasibility of a therapist-supported, mobile phone–delivered online intervention for depression: longitudinal observational study. *JMIR Formative Research* 2019; 3: e11509. PMID: 30682726.
27. Economides M, Ranta K, Nazander A, et al. Long-term outcomes of a therapist-supported, smartphone-based intervention for elevated symptoms of depression and anxiety: quasiexperimental, pre-postintervention study. *JMIR mHealth uHealth* 2019; 7: e14284. PMID: 31452521.
28. Reilly MC, Zbrozek AS and Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics* 1993; 4: 353–365. PMID: 10146874.
29. Dolan ED, Mohr D, Lempa M, et al. Using a single item to measure burnout in primary care staff: a psychometric evaluation. *J Gen Intern Med* 2015; 30: 582–587. PMID: 25451989.
30. Löwe B, Unützer J, Callahan CM, et al. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care* 2004; 42: 1194–1201. PMID: 15550799.
31. Toussaint A, Hüsing P, Gumz A, et al. Sensitivity to change and minimal clinically important difference of the 7-item generalized anxiety disorder questionnaire (GAD-7). *J Affect Disord* 2020; 265: 395–401. PMID: 32090765.
32. Pedregosa, et al. Scikit-learn: machine learning in Python. *JMLR* 2011; 12: 2825–2830.
33. Van Buuren S and Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011; 45: 1–67.
34. Nadeau C and Bengio Y. Inference for the generalization error. *Mach Learn* 2003; 52: 239–281.
35. Bouckaert RR and Frank E. Evaluating the replicability of significance tests for comparing learning algorithms. In: *Pacific-Asia conference on knowledge discovery and data mining*. Berlin, Heidelberg: Springer, 2004, pp.3–12.
36. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 2013; 36: 27–46.
37. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
38. Receiver Operating Characteristic (ROC) with cross validation. Accessed Decembre 2, 2020. https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html
39. Ojala M and Garriga GC. Permutation tests for studying classifier performance. *J Mach Learn Res* 2010; 11: 6.
40. Hooker G and Mentch L. Please stop permuting features: an explanation and alternatives. *arXiv preprint arXiv* 2019; 1905: 03151.
41. Mehrotra A, Hendley R and Musolesi M. Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp.1132–1138.
42. Delgado J, de Jong K, Lucock M, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry* 2018; 5: 564–572. PMID: 29937396.
43. Schwartz B, Cohen ZD, Rubel JA, et al. Personalized treatment selection in routine care: integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychother Res* 2021 Jan; 31(1): 33–51.
44. Bremer V, Becker D, Kolovos S, et al. Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: data-driven analysis. *J Med Internet Res* 2018; 20: e10275. PMID: 30131318.
45. Batterham PJ, Calear AL, Farrer L, et al. Fitmindkit: randomised controlled trial of an automatically tailored online program for mood, anxiety, substance use and suicidality. *Internet Interv* 2018; 12: 91–99. PMID: 30135773.
46. Parikh RB, Teeple S and Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019; 322: 2377–2378.

47. Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety. 2016. arXiv preprint arXiv:1606.06565.
 48. Schafer JL and Graham JW. Missing data: our view of the state of the art. *Psychol Methods* 2002; 7: 147. PMID: 1209040.
 49. DeMasi O, Kording K and Recht B. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one* 2017; 12: e0184604. PMID: 28949964.
 50. Symons M, Feeney GF, Gallagher MR, et al. Machine learning vs addiction therapists: a pilot study predicting alcohol dependence treatment outcome from patient data in behavior therapy with adjunctive medication. *J Subst Abuse Treat* 2019; 99: 156–162.
 51. Skocik M, Collins J, Callahan-Flintoft C, et al. I tried a bunch of things: the dangers of unexpected overfitting in classification. *BioRxiv* 2016; 078816, PMID: 33035522.
 52. Ribeiro MT, Singh S and Guestrin C. “ Why should I trust you?” Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp.1135–1144.
-