Psychological Medicine

cambridge.org/psm

Select or adjust? How information from early treatment stages boosts the prediction of non-response in internet-based depression treatment

Leona Hammelrath[1], Kevin Hilbert[2], Manuel Heinrich[1], Pavle Zagorscak[1] and Christine Knaevelsrud[1]

[1]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany and [2]Department of Psychology, Health and Medical University Erfurt, Erfurt, Germany

Original Article

This article incorporates a secondary data analysis from a study funded by Techniker Krankenkasse, a German health insurance company, in collaboration with Christine Knaevelsrud, Professor of Clinical Psychology, Department of Clinical-Psychological Intervention, Freie Universität Berlin, Berlin, Germany.

Cite this article: Hammelrath L, Hilbert K, Heinrich M, Zagorscak P, Knaevelsrud C (2023). Select or adjust? How information from early treatment stages boosts the prediction of non-response in internet-based depression treatment. *Psychological Medicine* 1–10. https://doi.org/10.1017/S0033291723003537

Received: 28 December 2022
Revised: 31 July 2023
Accepted: 13 November 2023

Keywords:
big data; depression; e-mental-health; early change; health informatics; machine learning; precision therapy

Corresponding author:
Leona Hammelrath;
Email: leona.hammelrath@fu-berlin.de

Abstract

Background. Internet-based interventions produce comparable effectiveness rates as face-to-face therapy in treating depression. Still, more than half of patients do not respond to treatment. Machine learning (ML) methods could help to overcome these low response rates by predicting therapy outcomes on an individual level and tailoring treatment accordingly. Few studies implemented ML algorithms in internet-based depression treatment using baseline self-report data, but differing results hinder inferences on clinical practicability. This work compares algorithms using features gathered at baseline or early in treatment in their capability to predict non-response to a 6-week online program targeting depression.

Methods. Our training and test sample encompassed 1270 and 318 individuals, respectively. We trained random forest algorithms on self-report and process features gathered at baseline and after 2 weeks of treatment. Non-responders were defined as participants not fulfilling the criteria for reliable and clinically significant change on PHQ-9 post-treatment. Our benchmark models were logistic regressions trained on baseline PHQ-9 sum or PHQ-9 early change, using 100 iterations of randomly sampled 80/20 train-test-splits.

Results. Best performances were reached by our models involving early treatment characteristics (recall: 0.75–0.76; AUC: 0.71–0.77). Therapeutic alliance and early symptom change constituted the most important predictors. Models trained on baseline data were not significantly better than our benchmark.

Conclusions. Fair accuracies were only attainable by involving information from early treatment stages. In-treatment adaptation, instead of a priori selection, might constitute a more feasible approach for improving response when relying on easily accessible self-report features. Implementation trials are needed to determine clinical usefulness.

In recent years, internet-based interventions (IBI) for mental health disorders have become an integral part of research and practice, with the ongoing pandemic reinforcing this development (Lange, 2021; Mahoney et al., 2021). For patients suffering from mild to moderate depression, IBIs are generally effective and achieve similar effect sizes as their face-to-face counterparts (Andrews et al., 2018; Carlbring, Andersson, Cuijpers, Riper, & Hedman-Lagerlöf, 2018). However, a recent meta-analysis revealed that only 37% of participants showed a reliable response (Cuijpers et al., 2021). Thus, 73% of participants show no meaningful improvements. The large rates of non-response are highly problematic both from a patient's and an economic perspective.

One approach to improve response rates is to optimize treatment selection through data-informed personalization of mental health care, also called precision therapy (Chekroud et al., 2021; Salazar de Pablo et al., 2021). The idea behind it is to overcome the current 'trial and error'-approach to treatment selection by identifying the ideal treatment for an individual based on what helped other individuals with similar characteristics in the past. To ensure an accurate fit (i.e. identify the treatment with the highest probability of effectiveness), these characteristics should ideally cover an exhaustive range of information – one reason why precision therapy goes hand-in-hand with machine learning methods.

Machine learning (ML) generally refers to the application of algorithms to large datasets to automatically learn patterns and identify relevant features for predicting the outcome of interest (here: non-response). These algorithms vary in their complexity and thus their interpretability – from simple linear models to highly complex multi-level approaches like convolutional neural networks. They can be used to predict an unobserved target variable (supervised learning), e.g. predicting antidepressant treatment response (Zhdanov et al.,

CAMBRIDGE UNIVERSITY PRESS

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

2020), or find clusters in unlabeled data (unsupervised learning), e.g. identify different types of treatment engagement (Chien et al., 2020). Compared to conventional statistical methods, most ML models can depict complex interactions and nonlinearities in heterogeneous or noisy datasets.

First studies in face-to-face treatment contexts already demonstrated that ML models trained on patient characteristics can outperform human experts in selecting suitable treatments or predicting responses (Koutsouleris et al., 2018; van Bronswijk, Lemmens, Huibers, & Peeters, 2021), pointing out its potential usefulness. Internet-based interventions might be particularly well-suited for applying ML approaches. First, their high scalability makes it easier to gather large amounts of data necessary to train those predictive algorithms. Second, common and specific factors (e.g. applied techniques, communication between patient and therapist) are often highly standardized, allowing a more precise matching of patient features to treatment components. Moreover, it is easily possible to gather variables depicting the therapeutic process (e.g. time of assessment completion, number of logins) that might represent a fruitful extension to the set of self-report features.

To date, only a few studies have attempted to develop prognostic models in IBI targeting depressive symptoms. They achieved mixed results, and comparisons are hampered by differing definitions of treatment outcomes, sample sizes, and reported metrics of predictive capability. Wallert et al. (2022) tested three ML algorithms to predict remission following a 12-week internet-based CBT program for depression. Their best classifier correctly identified 66% of individuals as remitters or non-remitters, exploiting self-report, process, and genetic variables collected pre-treatment. The area under the curve (AUC) was 0.69, which can be considered close to *fair,* according to Bone et al. (2021). Nemesure, Heinz, McFadden, and Jacobson (2021) report an AUC of 0.75 predicting response to a 9-week internet-based physical activity intervention for major depressive disorder with demographic, symptom-related, and healthcare utilization data. However, the results should be interpreted cautiously, given the small training sample of $n = 24$ participants. Finally, Pearson, Pisner, Meyer, Shumake, and Beevers (2019) investigated whether an ensemble of ML algorithms trained on baseline self-report measures, usage data, and environmental context variables (e.g. access to mental healthcare providers) explain more variance in depressive symptoms after 8-weeks of internet-based CBT than a linear regression model using only pre-treatment symptom scores. The explained variance increased only marginally from 0.17 to 0.25.

To summarize, it remains unclear whether the prediction of non-response beyond moderate accuracy levels can be achieved with patient intake characteristics alone in IBIs for depression. However, integrating features from the early stages of treatment might provide incremental information. For example, early symptom changes were predictive of non-response in IBI targeting depression, anxiety, and panic disorder (Beard & Delgadillo, 2019; Forsell et al., 2020; Schibbye et al., 2014). Bone et al. (2021) found that gradually (i.e. on a weekly base) incorporating patient information improved therapy outcome prediction of different ML models, especially in the early phases of treatment.

Identifying non-responders before treatment initiation would be ideal for preventing treatment failures and increasing cost-effectiveness. Still, it was shown that the treatment adaptation for patients identified as at risk of non-response after 3 weeks of internet-based CBT for insomnia could still significantly improve therapy effects (Forsell et al., 2019). If baseline features are insufficient to develop clinically trustworthy models, incorporating features gathered during the first weeks of therapy to increase predictive performance and potentially adjust treatment could still benefit patients and providers.

The current study uses data from a 6-week CBT-based IBI for adults with mild to moderate depression. All participants filled out a range of questionnaires at intake, covering empirically corroborated predictors like treatment expectancy, self-efficacy, and symptom severity. They further provided weekly self-reports of depressive symptoms, cognitive distortions, and therapeutic alliance, resulting in a rich set of candidate predictors. We aimed to examine if the prediction of non-response can be achieved using easily accessible process and self-report data gathered (a) at baseline or (b) in an early stage of treatment. We hypothesized that integrating self-report data from early treatment stages would significantly improve predictive performance compared to the models trained on baseline features only.

Methods

Dataset

The data we used for model training was obtained as part of a Germany-wide study analyzing the effects of different treatment sequences in a guided, CBT-based IBI for mild to moderate depression (Brose et al., 2023a). The intervention was developed in cooperation with a German public healthcare provider. It contains seven standardized modules, covering established cognitive-behavioral methods like psychoeducation and expressive writing, and established CBT methods such as behavioral activation (e.g. daily planner, diary of positive events) and cognitive restructuring (e.g. negativity bias training, positive imagery). Study eligibility was verified using the participant's medical records, online assessments, and the structured clinical interview for DSM-IV (SCID-I, sections A through F; Wittchen, Zaudig, and Fydrich, 1997) conducted via telephone by trained interviewers. Inclusion criteria comprised a 14–28 on Beck's Depression Inventory-II (BDI-II; Hautzinger, Keller, and Kühner, 2006), indicating mild to moderate depression and computer-based internet access. Exclusion criteria were (1) a current mania or hypomania, (2) psychotic symptoms (lifetime), and (3) risk of suicide (score of 1 on BDI-II item 9). Participants assessed eligible were randomly assigned to one of two study conditions, that differed in the order of presented modules (they either started with positive behavioral activation or cognitive restructuring) and allocated to a trained counselor. For more details on study onboarding and experimental conditions, see Brose et al. (2023a). Participants completed approximately one module per week over the course of 6–8 weeks, with each module consisting of (1) a feedback letter from their respective counselor, (2) psychoeducation, (3) introduction to an exercise or homework, and (4) the introduced online exercise. To measure depressive symptomatology on a weekly basis, the Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001) was assessed before every intervention module. The PHQ-9 is a validated instrument to measure depression severity, consisting of nine items that can be scored from zero to three. Only participants who completed the intervention and filled out the post-assessment were included in this analysis. All subjects provided their written informed consent for data collection and analysis. the study protocol was approved by the Ethics Committee of *Freie Universität Berlin* prior to recruitment start (processing sign: 125/ 2016).

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

Features and outcome

The full set of features is listed in Online Supplementary Table 1.

Before treatment started, participants completed a comprehensive set of questionnaires. These covered demographics, disorder-related clinical symptom scales (e.g. cognitive distortions, self-efficacy beliefs), (psycho-) social and functional circumstances (e.g. social support, healthcare usage), life aims and values, as well as treatment expectations. Questionnaire features were included on a single-item level and as aggregated scores. Beyond that, the presence of a current or remitted depressive or dysthymic episode and changes in medication within the last 6 weeks before starting to work with the intervention, as assessed by the SCID-I interview, were included as predictors.

At the beginning of week 2 (i.e. M3), participants filled out the PHQ-9, the *Cognitive Style Assessment* measuring cognitive distortions (COSTA; Bohn et al., 2022) and the *Scale for the Multiperspective Assessment of General Change Mechanisms in Psychotherapy* assessing the therapeutic relationship (SEWIP; Mander et al., 2013). These were added to the set of baseline features as single items and sum scores. In addition, we implemented early change scores for PHQ-9 and COSTA by subtracting baseline sum scores from week 2 sum scores.

Process features encompassed the registration year, study variant, and if treatment overlapped with the first wave of infections within the global SARS-CoV-2 pandemic (yes/no).

As our outcome variable, we implemented the binary criterion of reliable and clinically significant change on PHQ-9 (yes/no), in accordance with Jacobson and Truax criteria (Jacobson & Truax, 1992). Accordingly, we defined individuals with an improvement of ≥5 points and a sub-clinical post-PHQ-9 score of <10 as responders (group = 0) and everyone else as non-responders (group = 1).

Data preparation and partitioning

Data preparation was done in Jupyter Notebook and Visual Studio Code, using Python *v. 3.7.4* and the Python packages *pandas v. 0.25.1* and *scikit-learn v 0.0.24.1*. All advanced analyses were performed using the packages *scikit-learn v 0.0.24.1* and *numpy v. 1.19.5*.

The original sample consisted of 2304 participants. We first removed all participants with missing values on one of the PHQ-post (28.7% from total) or PHQ-M3 (11.8% from total) items, leaving a sample of 1591 records for model training and validation. Missing values on the remaining features amounted to max. 2.8% and were imputed by either mean or mode depending on the respective data type. Categorical features with no clear ordering (e.g. assigned counselor, recruitment strategy) were one-hot encoded, creating binary variables that indicate the presence or absence of a certain category. Aggregating categories reduced high feature cardinality. Features representing clinical symptom scale items were reverse-scored and/or aggregated if indicated in the respective manual. Finally, continuous features like age and minutes of daily internet usage were centered and scaled. The final datasets contained 213 (baseline) and 260 (early change) features, respectively.

For cross-validation and to avoid overfitting and bias, a train-test split of 80/20 was employed. Due to a small class imbalance favoring responders, we performed down sampling of the majority class, leaving 1270 records for training and 318 records for model validation. Since generalization performance is strongly dependent on the respective train-test partitioning (Orrù, Monaro, Conversano, Gemignani, & Sartori, 2020), we performed 100 iterations of our ML pipeline with independent train-test splits per iteration as done by Hilbert et al. (2021). Model performance is therefore reported as mean across all 100 iterations, including range and standard deviations. Since we were mainly interested in correctly identifying non-responders, our main performance measure, also used for feature selection and hyperparameter tuning, is recall (also known as sensitivity, i.e. the proportion of correctly identified non-responders). To allow comparison with other studies, we further report balanced accuracy (i.e. the arithmetic mean of recall/sensitivity and specificity), AUC (i.e. the probability that the model will correctly distinguish between true negatives and true positives), and f1 scores (i.e. the harmonic mean of recall/sensitivity and precision).

Model versions and machine learning pipeline

We chose a random forests (RF) classifier to develop our predictive models, as it can automatically deal with nonlinearities and higher-order interaction effects, has been demonstrated to be robust against bias, and is a commonly used model in studies predicting therapy outcomes (Breiman, 2001). Four versions of random forest classifiers were compared against two simpler benchmark models, a linear main-effects logistic regression (C = 1.0) predicting non-response using (1) baseline PHQ-9 sum scores or (2) the PHQ-9 early change score.

To determine the impact of including early treatment information, we contrasted models including baseline characteristics only or baseline plus early treatment characteristics. Further, we compared different model pipelines by applying either no feature selection or hyperparameter tuning or applying both feature selection using Elastic Net (L1-penalty = 0.5, max_iter = 1000) and random search hyperparameter tuning (Bergstra & Bengio, 2012) using nested cross-validation with five folds and 100 iterations. This resulted in the following models:

1. RF with baseline features; no feature selection or hyperparameter tuning.
2. RF with baseline features; feature selection and hyperparameter tuning.
3. RF with baseline and early treatment features; no feature selection or hyperparameter tuning.
4. RF with baseline and early treatment features; feature selection and hyperparameter tuning.

To compare those models with our benchmark, we implemented corrected resampled *t*-tests (Nadeau & Bengio, 2003).

We report the mean number of features chosen by automatic feature selection, as well as the 10 most important features ranked by their Gini impurity index (or mean decrease impurity: measures the weighted average of uncertainty reduction achieved by the respective feature across trees) and number of occurrences across 100 iterations in case of feature selection.

Results

Sample characteristics

Sociodemographic, clinical, and process features per outcome group (reliable and clinically significant change from pre-to post-treatment yes/no) are depicted in Table 1. Among participants

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

Table 1. Patient summary characteristics stratified by treatment outcome

		Overall	Responder	Non-responder	*p*-Value
n		1591	797	794	
Age, mean (s.d.)		43.5 (12.8)	43.5 (12.8)	43.4 (12.9)	0.895[a]
Sex, *n* (%)	Male	610 (38.3)	306 (38.4)	304 (38.3)	0.999[b]
	Female	981 (61.7)	491 (61.6)	490 (61.7)	
Highest educational level, *n* (%)					
	Certificate of secondary education or no school leaving certificate	68 (4.3)	40 (5.0)	28 (3.5)	0.111[b]
	General certificate of secondary education	283 (17.8)	155 (19.4)	128 (16.1)	
	Higher education entrance qualification	372 (23.4)	180 (22.6)	192 (24.2)	
	Polytechnic school degree	189 (11.9)	100 (12.5)	89 (11.2)	
	University degree	679 (42.7)	322 (40.4)	357 (45.0)	
Living arrangement, *n* (%)					
	Alone	412 (25.9)	205 (25.7)	207 (26.1)	0.422[b]
	With partner only	536 (33.7)	263 (33.0)	273 (34.4)	
	With partner and children	424 (26.6)	226 (28.4)	198 (24.9)	
	With other people or with children only	219 (13.8)	103 (12.9)	116 (14.6)	
Marital status, *n* (%)					
	Single	688 (43.2)	342 (42.9)	346 (43.6)	0.738[b]
	Married	754 (47.4)	384 (48.2)	370 (46.6)	
	Divorced or widowed	149 (9.4)	71 (8.9)	78 (9.8)	
Employment status, *n* (%)					
	Employed worker	1045 (65.7)	538 (67.5)	507 (63.9)	0.111[b]
	Student, pupil or trainee	160 (10.1)	85 (10.7)	75 (9.4)	
	Retired or currently unemployed	217 (13.6)	102 (12.8)	115 (14.5)	
	Self-employed or other	169 (10.6)	72 (9.0)	97 (12.2)	
Residence size					
	Big city	529 (33.2)	252 (31.6)	277 (34.9)	0.148[b]
	Outskirts or suburb of a big city	292 (18.4)	159 (19.9)	133 (16.8)	
	Medium or small town	287 (18.0)	153 (19.2)	134 (16.9)	
	Village, farmstead or detached house	483 (30.4)	233 (29.2)	250 (31.5)	
BMI, median[Q1, Q2]		24.1 [21.5,27.4]	24.0 [21.7,27.1]	24.2 [21.5,27.5]	0.697[c]
Internet usage, median [Q1,Q3]		180.0 [90.0240.0]	180.0 [90.0240.0]	180.0 [90.0240.0]	0.795[c]
Treatment during corona, *n* (%)					
	No	1363 (85.7)	680 (85.3)	683 (86.0)	0.744[b]
	Yes	228 (14.3)	117 (14.7)	111 (14.0)	
Registration year, *n* (%)					
	2016	87 (5.5)	43 (5.4)	44 (5.5)	0.301[b]
	2017	607 (38.2)	284 (35.6)	323 (40.7)	
	2018	316 (19.9)	169 (21.2)	147 (18.5)	
	2019	329 (20.7)	172 (21.6)	157 (19.8)	
	2020	252 (15.8)	129 (16.2)	123 (15.5)	
Study version, *n* (%)[e]					
	PAF	773 (48.6)	406 (50.9)	367 (46.2)	0.067[b]
	CRF	818 (51.4)	391 (49.1)	427 (53.8)	

(*Continued*)

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

Table 1. (*Continued.*)

		Overall	Responder	Non-responder	*p*-Value
Previous psychotherapy, *n* (%)					
	No	623 (39.2)	328 (41.2)	295 (37.2)	0.113[b]
	Yes	968 (60.8)	469 (58.8)	499 (62.8)	
Other prof. support, *n* (%)					
	No	1104 (69.4)	552 (69.3)	552 (69.5)	0.953[b]
	Yes	487 (30.6)	245 (30.7)	242 (30.5)	
Prefers conventional therapy, *n* (%)					
	No	392 (24.6)	181 (22.7)	211 (26.6)	0.084[b]
	Yes	1199 (75.4)	616 (77.3)	583 (73.4)	
Serious illness, *n* (%)					
	No	962 (60.5)	464 (58.2)	498 (62.7)	0.074[b]
	Yes	629 (39.5)	333 (41.8)	296 (37.3)	
Sick leave, *n* (%)					
	No	968 (60.8)	484 (60.7)	484 (61.0)	**0.030**[b]
	Yes	384 (24.1)	209 (26.2)	175 (22.0)	
	Unemployed	239 (15.0)	104 (13.0)	135 (17.0)	
Physician visits, *n* (%)					
	No	764 (48.0)	358 (44.9)	406 (51.1)	**0.015**[b]
	Yes	827 (52.0)	439 (55.1)	388 (48.9)	
Neurologist visits, *n* (%)					
	No	1366 (85.9)	686 (86.1)	680 (85.6)	0.862[b]
	Yes	225 (14.1)	111 (13.9)	114 (14.4)	
Counseling, *n* (%)					
	No	1505 (94.6)	757 (95.0)	748 (94.2)	0.567[b]
	Yes	86 (5.4)	40 (5.0)	46 (5.8)	
Psychotherapy, *n* (%)					
	No	1498 (94.2)	759 (95.2)	739 (93.1)	0.084[b]
	Yes	93 (5.8)	38 (4.8)	55 (6.9)	
Medication intake, *n* (%)					
	No	821 (51.6)	411 (51.6)	410 (51.6)	0.999[b]
	Yes	770 (48.4)	386 (48.4)	384 (48.4)	
TI[d] sum score, median [Q1,Q3]		17.0 [13.0,20.0]	17.0 [13.0,20.0]	17.0 [13.0,20.0]	0.460[c]
TI[d] MDE, mean (s.d.)					
	No	1089 (68.4)	549 (68.9)	540 (68.0)	0.748[b]
	Yes	502 (31.6)	248 (31.1)	254 (32.0)	
TI[d] dysthymia, mean (s.d.)					
	No	1431 (89.9)	733 (92.0)	698 (87.9)	**0.009**[b]
	Yes	160 (10.1)	64 (8.0)	96 (12.1)	
TI[d]MDE fully remitted, mean (s.d.)					
	No	1175 (73.9)	578 (72.5)	597 (75.2)	0.249[b]
	Yes	416 (26.1)	219 (27.5)	197 (24.8)	
TI[d] MDE part. remitted, mean (s.d.)					
	No	1357 (85.3)	681 (85.4)	676 (85.1)	0.919[b]
	Yes	234 (14.7)	116 (14.6)	118 (14.9)	

(*Continued*)

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

Table 1. (*Continued.*)

	Overall	Responder	Non-responder	*p*-Value
TI[d] change in medication, mean (s.d.)				
No	1420 (89.3)	701 (88.0)	719 (90.6)	0.111[b]
Yes	171 (10.7)	96 (12.0)	75 (9.4)	
PHQ-9 pre, median [Q1,Q3]	11.0 [9.0,14.0]	12.0 [10.0,14.0]	10.0 [8.0,13.0]	**<0.001**[c]
BDI, median [Q1,Q3]	22.0 [18.0, 25.0]	23.0 [19.0, 25.0]	21.0 [17.0, 24.0]	**<0.001**[c]
PHQ-S, mean (s.d.)	8.9 (3.3)	9.1 (3.3)	8.8 (3.4)	0.078[a]
EUROHIS, mean (s.d.)	24.9 (4.0)	24.7 (4.1)	25.0 (3.9)	0.228[a]
IMET, median [Q1,Q3]	34.0 [26.0,43.0]	33.0 [26.0,43.0]	34.0 [26.0,43.0]	0.654[c]
GAD, median [Q1,Q3]	9.0 [7.0,12.0]	10.0 [8.0,13.0]	9.0 [7.0,11.0]	**<0.001**[c]
IPQR, median [Q1,Q3]	65.0 [61.0,68.0]	65.0 [62.0,68.0]	65.0 [61.0,68.0]	0.366[c]
GPSE, mean (s.d.)	24.5 (4.7)	24.6 (4.7)	24.4 (4.8)	0.409[a]
PATHEV hope, median [Q1,Q3]	15.0 [13.0,17.0]	16.0 [13.0,17.0]	15.0 [13.0,17.0]	**0.033**[c]
PATHEV FUR, median [Q1,Q3]	4.0 [3.0,5.0]	4.0 [3.0,5.0]	4.0 [3.0,5.0]	0.844[c]
PATHEV PAS, mean (s.d.)	13.5 (2.4)	13.5 (2.4)	13.4 (2.3)	0.298[a]
COSTA, mean (s.d.)	29.5 (9.9)	29.5 (9.9)	29.5 (9.9)	0.876[a]
PHQ-9 early, median [Q1,Q3]	9.0 [6.0,12.0]	8.0 [5.0,10.0]	10.0 [7.0,13.0]	**<0.001**[c]
COSTA early, median [Q1,Q3]	29.0 [23.0,35.0]	28.0 [22.0,34.0]	30.0 [24.0,36.0]	**<0.001**[c]
SEWIP early, median [Q1,Q3]	48.0 [36.0,58.0]	50.0 [37.0,60.0]	46.0 [36.0,55.0]	**<0.001**[c]

Notes. [a]Two-sample *t* test, [b]Chi-squared test, [c]Kruskal–Wallis test, [d]TI = expert rating from telephone interview, [e]Study version: PAF = positive activities module first; CRF: cognitive restructuring module first.
p-values meeting the criterion of $p < 0.05$ are highlighted in bold.

with complete PHQ-9 scores, 50.09% (797/1591) fulfilled the criteria for reliable and clinically significant improvement from pre- to post-treatment. Participants categorized as non-responders reported higher unemployment rates but fewer days of sick leave, as well as doctor visits within the last 4 weeks before treatment. Concerning clinical features, participants fulfilling criteria for reliable and clinically significant change scored significantly higher on baseline sum scores of PHQ-9, BDI-II, GAD, and PATHEV – hope. Furthermore, they were less likely to fulfill the criteria for dysthymia as determined in the SCID interview pre-treatment. In turn, they show significantly lower week two sum scores of PHQ-9 and COSTA and significantly higher therapeutic alliances as indicated by higher SEWIP sum scores at week 2. Apart from that, there were no significant differences between outcome groups.

Model performance

All ML models predicted non-response with above-chance accuracies (see Fig. 1). Models incorporating early symptom developments (RF early) performed significantly better than models trained solely on baseline features and then both benchmark models. They achieved recall scores of 0.76 (s.d. = 0.04) and 0.75 (s.d. = 0.04) correctly identified non-responders, using the full or an automatically reduced set of features respectively (for an overview of evaluation metrics see Table 2).

With recall scores of 0.59 and 0.60, respectively our baseline models performed descriptively worse than our benchmark trained on baseline PHQ-9 sum scores (recall = 0.62, s.d. = 0.03) and significantly worse than the benchmark using early change scores (recall = 0.69, s.d. = 0.03, $p < 0.001$).

We observed similar predictive performances between models using the same set of features, indicating no disadvantages for the models involving automatic feature selection and thus, only around a third of the available variables. Figure 2 compares ROC curves from the RF early model including feature selection and hyperparameter tuning to the mean ROC of the benchmark using early change.

The 10 most important features, ranked by their Gini impurity index, are depicted in Fig. 3 topped by information on their occurrence across the 100 iterations. Among the most important features were early treatment information, like therapeutic alliance (SEWIP items 1 and 4) and early change on PHQ-9. Beyond that, baseline information on cognitive distortions (COSTA items 1, 5, and 6), anxiety (GAD items 4 and 5), and symptom severity (PHQ-9 sum) ranked comparatively important.

Discussion

The present paper tested an ML-based approach to predict non-response in participants of a CBT-based IBI for depression. We found that fair prediction (e.g. accuracies >0.70) could only be reached when information on early treatment stages was included.

Models relying exclusively on information assessed before the intervention reached only moderate performance that even a benchmark model using only baseline PHQ-9 sum scores exceeded. This moderate predictive performance lags behind other studies predicting IBI outcomes using pre-treatment features only, like those of Nemesure et al. (2021) and Wallert

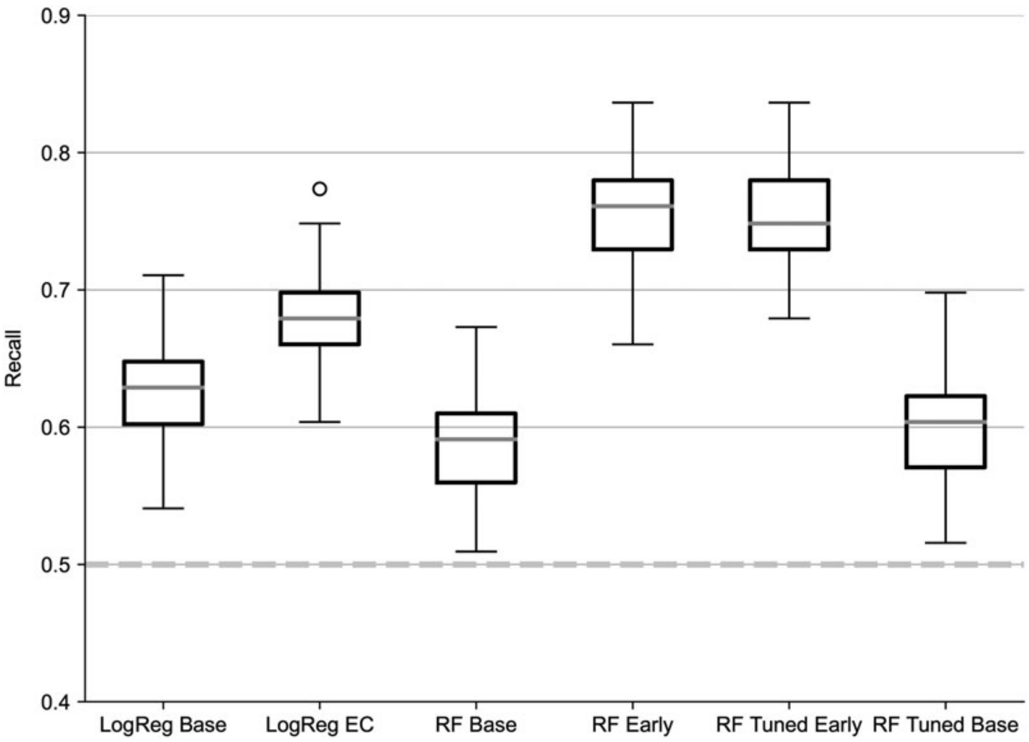

Figure 1. Comparison of model performance in identifying non-response.

Notes: The dashed line indicates chance level. RF, random forest; base, pre-treatment features only; early, features from the beginning of week 2 were incorporated; LogReg base, logistic regression using the baseline PHQ sum; LogReg EC, logistic regression using PHQ difference from baseline to week 2 (early change)

Table 2. Outcomes by model type averaged across 100 iterations

Model		*n* features	Recall	Accuracy	AUC	F1	*p-phq-sum*[c]	*p-phq-ec*[d]
Benchmark[a] phq-sum	Mean (s.d.)	1	0.62 (0.03)	0.62 (0.02)	0.65 (0.03)	0.62 (0.02)		
	Min, max		0.54, 0.71	0.56, 0.68	0.60, 0.72	0.57, 0.68		
Benchmark[b] phq-ec	Mean (s.d.)	1	0.69 (0.03)	0.68 (0.02)	0.74 (0.02)	0.68 (0.02)		
	Min, max		0.60, 0.77	0.61, 0.72	0.68, 0.80	0.61, 0.74		
RF base	Mean (s.d.)	213	0.59 (0.03)	0.63 (0.03)	0.68 (0.03)	0.62 (0.03)	0.043	<0.001
	Min, max		0.51, 0.67	0.57, 0.69	0.63, 0.74	0.54, 0.68		
RF early	Mean (s.d.)	260	0.76 (0.03)	0.70 (0.02)	0.77 (0.02)	0.72 (0.02)	0.004	<0.001
	Min, max		0.66, 0.84	0.65, 0.75	0.72, 0.83	0.66, 0.76		
p base – early			<0.001	<0.001	<0.001	<0.001		
RF tuned base	Mean (s.d.)	70.32 (4.98)	0.60 (0.04)	0.63 (0.03)	0.68 (0.03)	0.62 (0.02)	0.469	0.003
	Min, max	59, 84	0.52, 0.70	0.57, 0.69	0.62, 0.74	0.54, 0.69		
RF tuned early	Mean (s.d.)	103.91 (5.00)	0.75 (0.04)	0.70 (0.02)	0.77 (0.02)	0.72 (0.02)	0.009	0.003
	Min, max	90, 116	0.68, 0.84	0.64, 0.75	0.71, 0.82	0.66, 0.76		
p tuned base – early			<0.001	<0.001	<0.001	<0.001		

Notes: [a]A logistic regression predicting non-response using baseline PHQ-9 sum scores; [b]A logistic regression predicting non-response using PHQ-9 early change; [c]Statisical comparison of recall scores with the baseline PHQ sum benchmark using corrected resampled *t* tests; [d]Statisical comparison of recall scores with the PHQ early change benchmark using corrected resampled *t* tests; SD, standard deviation; RF, random forest.

et al. (2022), which may in part be explained by our more basic Random Forest algorithm – both studies implemented advanced ensembling approaches like XGBoost (Chen & Guestrin, 2016). On the other hand, studies often find similar performances for algorithms of varying complexity, especially in low-dimensional data like self-report variables (e.g. Hilbert et al., 2021). Further, larger sample sizes tend to produce more robust and generalizable ML models (e.g. Luedtke, Sadikova, & Kessler, 2019). Finally, studies from face-to-face therapy settings with comparable study protocols (i.e. baseline features, comprehensive samples

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

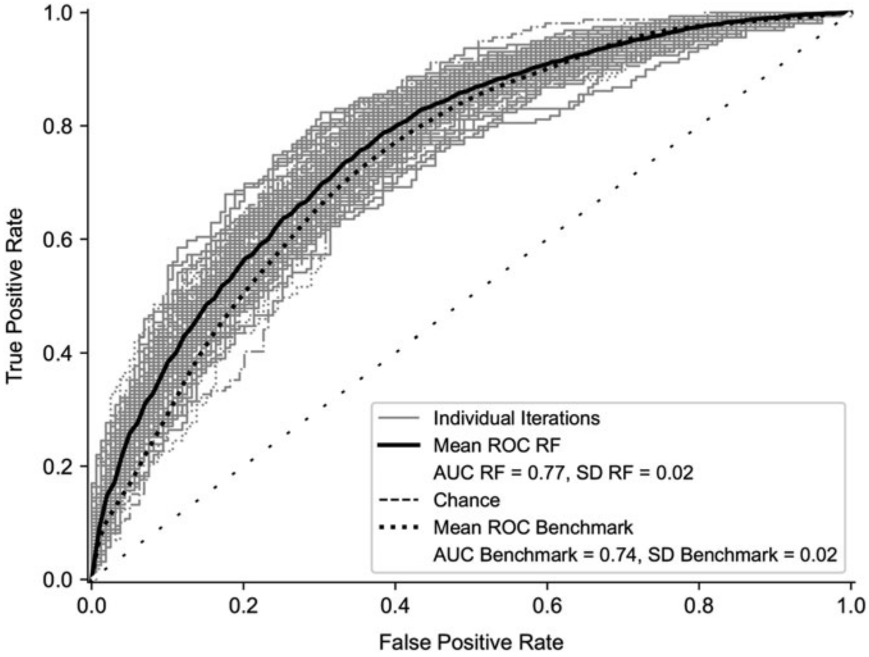

Figure 2. ROC curves for the random forest model involving early treatment features, hyperparameter tuning and automatic feature selection.
Notes: The dashed line indicates chance level. The bold line indicates the averaged ROC across 100 iterations for the random forest model. The bold dotted line indicated the averages ROC across 100 iterations for the benchmark model trained on PHQ early-change. RF, random forest; AUC, area under the curve; SD, standard deviation.

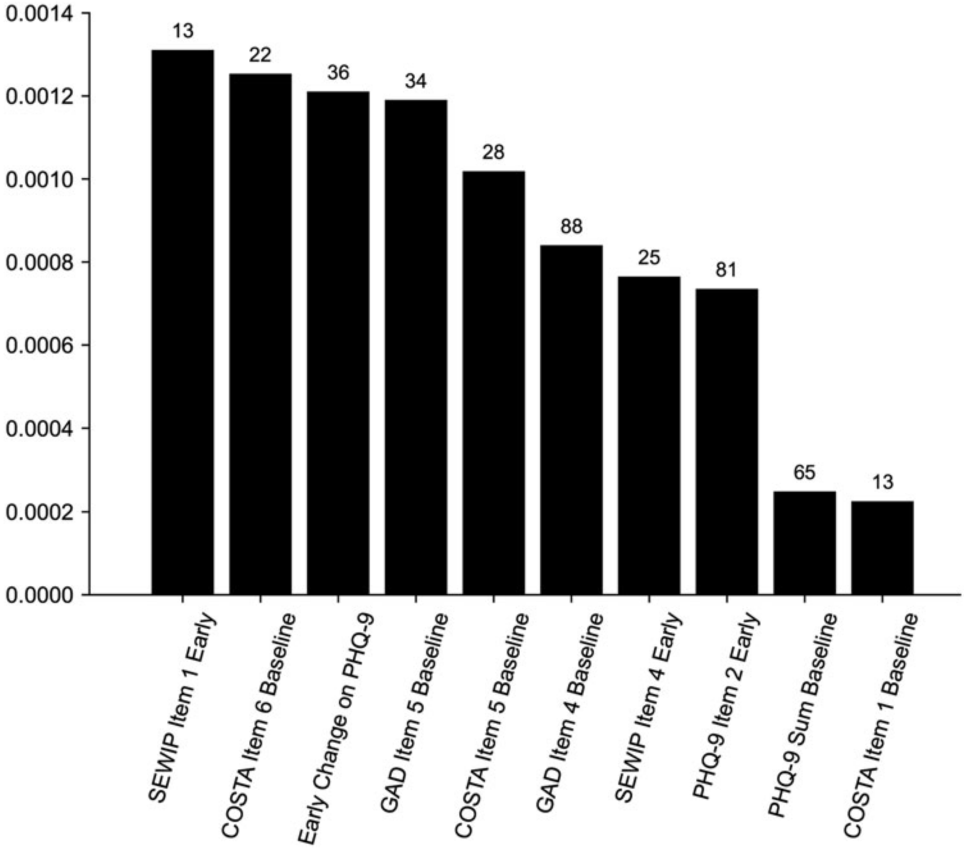

Figure 3. The 10 most important features for the random forest involving early treatment features, hyperparameter tuning and feature selection.
Notes: Importance is computed by the Gini impurity index averaged across iterations. The numbers above the bars indicate the amount of rounds the feature has been selected by automatic feature selection. There were 100 iterations in total.

of patients, ML models of varying complexity) also fail to exceed moderate thresholds of outcome prediction (Hilbert et al., 2021).

In line with our hypothesis, both models incorporating information from the early stages of therapy (e.g. depressive symptoms, therapeutic relationship, and cognitive distortions following

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

2 weeks of treatment) achieved the best accuracy in identifying non-responders. The tuned model, involving hyperparameter tuning and automatic feature selection, performed only slightly worse using around a third of the original feature stack, making it the preferred choice for implementation. Our findings align well with the study by Bone et al. (2021): they repeatedly trained machine learning algorithms on weekly symptom measures to predict response to psychotherapy for depression and anxiety. Predictive performance was moderate at baseline and improved with each passing week, with a particularly prominent rise in the early phase (i.e. the first 2–3 weeks of therapy). Beyond that, Brose et al. (2023b), using the same data set, found that early symptom change and symptom variability are related to changes in BDI-II scores from pre- to post-test.

When it comes to integrating predictive algorithms into routine care, information from early phases of treatment may be necessary to reach beneficial and trustworthy accuracies for both patients and therapists. Thus, treatment adaptation – instead of a priori selection – may constitute a promising avenue forward. This could take the form of a stepped care approach: patients start with a low-threshold approach like the (un-)guided internet-based interventions. Then, after a certain part of the treatment has been completed, an outcome prognosis is made based on pre-treatment assessments and information concerning the treatment progress. Consequently, treatment is either adapted to increase the chance of a beneficial outcome or continued consistently when a beneficial outcome is likely.

The improving access to psychological therapy programs in England followed a similar approach by implementing progressive care (i.e. all patients start with a low-intensity treatment, and intensity is increased if necessary; Boyd, Baker, & Reilly, 2019). They evaluated that this approach led to higher recovery rates than a stratified model of care (i.e. the therapist selects treatment form based on pre-treatment symptom severity). To increase the usefulness of such an adaptive approach, future studies should also focus on modifiable predictors, providing therapists instructions on how to proceed in case of imminent non-response. Beyond that, one could include ecological momentary assessments and passive sensing as features to enhance informative density while keeping costs and expenses low (Zarate, Stavropoulos, Ball, de Sena Collier, & Jacobson, 2022).

Our study has several limitations. First, we had to remove around 30% of our original sample due to missing data in the outcome variable. Since we do not know the mechanism of missingness (i.e. the proportion of data points not missing at random) this significant minority might influence the generalizability of our algorithm. As we were conducting secondary data analysis, we used all available features instead of selecting predictors based on domain knowledge. Carefully selecting relevant predictors based on domain knowledge may help to improve prediction accuracy (Salazar de Pablo et al., 2021). Further, we could not externally validate our algorithms in another IBI for depression. Thus, it is unclear whether our results only hold for this specific program or generalize to other IBI and therapy formats. Finally, clinical usefulness must be more thoroughly answered before applying these prediction models in real-world practice. Here, trials comparing 'precision therapy' (i.e. therapists following algorithm-supported decision tools) against treatment as usual regarding patient-related risks and benefits and health-economic aspects are urgently needed.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291723003537.

Competing interests. The authors declare none.

References

Andrews, G., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., English, C. L., & Newby, J. M. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *Journal of Anxiety Disorders*, *55*, 70–78. https://doi.org/10.1016/j.janxdis.2018.01.001.

Beard, J. I. L., & Delgadillo, J. (2019). Early response to psychological therapy as a predictor of depression and anxiety treatment outcomes: A systematic review and meta-analysis. *Depression and Anxiety*, *36*(9), 866–878. https://doi.org/10.1002/da.22931.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, *13*(null), 281–305.

Bohn, J., Heinrich, M., Brose, A., Knaevelsrud, C., & Zagorscak, P. (2022). Measuring cognitive distortions in depression: Development of the Cognitive Styles Assessment (COSTA) questionnaire [Manuscript in preparation].

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., … Delgadillo, J. (2021). Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health*, *3*(4), e231–e240. https://doi.org/10.1016/S2589-7500(21)00018-2.

Boyd, L., Baker, E., & Reilly, J. (2019). Impact of a progressive stepped care approach in an improving access to psychological therapies service: An observational study. *PLOS ONE*, *14*(4), e0214715. https://doi.org/10.1371/journal.pone.0214715.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brose, A., Heinrich, M., Bohn, J., Kampisiou, C., Zagorscak, P., & Knaevelsrud, C. (2023a). Sequencing effects of behavioral activation and cognitive restructuring in an internet-based intervention for depressed adults are negligible: Results from a randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *91*(3), 122–138. https://doi.org/10.1037/ccp0000789.

Brose, A., Koval, P., Heinrich, M., Zagorscak, P., Bohn, J., & Knaevelsrud, C. (2023b). Depressive symptom dynamics and therapeutic outcomes over time: An integrative approach using location-scale modeling [Manuscript in preparation].

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: An updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, *47*(1), 1–18. https://doi.org/10.1080/16506073.2017.1401115.

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., … Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882.

Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785–794. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Chien, I., Enrique, A., Palacios, J., Regan, T., Keegan, D., Carter, D., … Belgrave, D. (2020). A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. *JAMA Network Open*, *3*(7), e2010791. https://doi.org/10.1001/jamanetworkopen.2020.10791.

Cuijpers, P., Karyotaki, E., Ciharova, M., Miguel, C., Noma, H., & Furukawa, T. A. (2021). The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: A meta-analysis. *Acta Psychiatrica Scandinavica*, *144*(3), 288–299. https://doi.org/10.1111/acps.13335.

Forsell, E., Isacsson, N., Blom, K., Jernelöv, S., Ben Abdesslem, F., Lindefors, N., … Kaldo, V. (2020). Predicting treatment failure in regular care internet-delivered cognitive behavior therapy for depression and anxiety using only weekly symptom measures. *Journal of Consulting and Clinical*

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press

Psychology, *88*(4), 311–321. (2019-76267-001). https://doi.org/10.1037/ccp0000462.

Forsell, E., Jernelöv, S., Blom, K., Kraepelien, M., Svanborg, C., Andersson, G., … Kaldo, V. (2019). Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: A single-blind randomized clinical trial with insomnia patients. *American Journal of Psychiatry*, *176*(4), 315–323. https://doi.org/10.1176/appi.ajp.2018.18060699.

Hautzinger, M., Keller, F., & Kühner, C. (2006). *Beck depressions-inventar* (*BDI-II*). Göttingen: Hogrefe.

Hilbert, K., Jacobi, T., Kunas, S. L., Elsner, B., Reuter, B., Lueken, U., & Kathmann, N. (2021). Identifying CBT non-response among OCD outpatients: A machine-learning approach. *Psychotherapy Research*, *31*(1), 52–62. https://doi.org/10.1080/10503307.2020.1839140.

Jacobson, N. S., & Truax, P. (1992). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. In A. E. Kazdin (Ed.), *Methodological Issues & Strategies in Clinical Research* (pp. 631–648). https://doi.org/10.1037/10109-042.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., & Dwyer, D. B., … PRONIA Consortium. (2018). Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry*, *75*(11), 1156–1172. https://doi.org/10.1001/jamapsychiatry.2018.2165.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Lange, K. W. (2021). Coronavirus disease 2019 (COVID-19) and global mental health. *Global Health Journal*, *5*(1), 31–36. https://doi.org/10.1016/j.glohj.2021.02.004.

Luedtke, A., Sadikova, E., & Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, *7*(3), 445–461. https://doi.org/10.1177/2167702618815466.

Mahoney, A. E. J., Elders, A., Li, I., David, C., Haskelberg, H., Guiney, H., & Millard, M. (2021). A tale of two countries: Increased uptake of digital mental health services during the COVID-19 pandemic in Australia and New Zealand. *Internet Interventions*, *25*, 100439. https://doi.org/10.1016/j.invent.2021.100439.

Mander, J. V., Wittorf, A., Schlarb, A., Hautzinger, M., Zipfel, S., & Sammet, I. (2013). Change mechanisms in psychotherapy: Multiperspective assessment and relation to outcome. *Psychotherapy Research*, *23*(1), 105–116. https://doi.org/10.1080/10503307.2012.744111.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/A:1024068626366.

Nemesure, M. D., Heinz, M. V., McFadden, J., & Jacobson, N. C. (2021). Predictive modeling approach to evaluate individual response to a physical activity digital intervention for subjects with major depressive disorder. PsyArXiv [Preprint]. Available online at: https://psyarxiv.com/3kjyh/. https://doi.org/10.31234/osf.io/3kjyh.

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, *10*, 2970. https://doi.org/10.3389/fpsyg.2019.02970.

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2019). A machine learning ensemble to predict treatment outcomes following an internet intervention for depression. *Psychological Medicine*, *49*(14), 2330–2341. https://doi.org/10.1017/S003329171800315X.

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., … Fusar-Poli, P. (2021). Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice. *Schizophrenia Bulletin*, *47*(2), 284–297. https://doi.org/10.1093/schbul/sbaa120.

Schibbye, P., Ghaderi, A., Ljótsson, B., Hedman, E., Lindefors, N., Rück, C., … Kaldo, V. (2014). Using early change to predict outcome in cognitive behaviour therapy: Exploring timeframe, calculation method, and differences of disorder-specific versus general measures. *PLoS ONE*, *9*(6), e100614.

van Bronswijk, S. C., Lemmens, L. H. J. M., Huibers, M. J. H., & Peeters, F. P. M. L. (2021). Selecting the optimal treatment for a depressed individual: Clinical judgment or statistical prediction? *Journal of Affective Disorders*, *279*, 149–157. https://doi.org/10.1016/j.jad.2020.09.135.

Wallert, J., Boberg, J., Kaldo, V., Mataix-Cols, D., Flygare, O., Crowley, J. J., … Rück, C. (2022). Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. *Translational Psychology*, *12*(1), 357.

Wittchen, H.-U., Zaudig, M., & Fydrich, T. (1997). SKID. Strukturiertes Klinisches Interview für DSM-IV. Achse I und II. Handanweisung. Retrieved from https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_1646481.

Zarate, D., Stavropoulos, V., Ball, M., de Sena Collier, G., & Jacobson, N. C. (2022). Exploring the digital footprint of depression: A PRISMA systematic literature review of the empirical evidence. *BMC Psychiatry*, *22*(1), 421. https://doi.org/10.1186/s12888-022-04013-y.

Zhdanov, A., Atluri, S., Wong, W., Vaghei, Y., Daskalakis, Z. J., Blumberger, D. M., … Farzan, F. (2020). Use of machine learning for predicting escitalopram treatment outcome from electroencephalography recordings in adult patients with depression. *JAMA Network Open*, *3*(1), e1918377. https://doi.org/10.1001/jamanetworkopen.2019.18377.

https://doi.org/10.1017/S0033291723003537 Published online by Cambridge University Press