JAMA Network Open

Original Investigation | Psychiatry

Machine Learning Model for Response to Internet-Delivered CBT vs Antidepressant Medication

Chi Tak Lee, PhD; Derek Richards, PhD; Jakob Heinzle, PhD; Anna K. Hanlon, MSc; Kevin Lynch, MSc; Siobhan Harty, PhD; Nathalie Claus, PhD; Veronica O'Keane, MD, PhD; Klaas Enno Stephan, MD, PhD; Robert Whelan, PhD; Claire M. Gillan, PhD

Abstract

IMPORTANCE Many treatments exist for depression, yet none are universally effective. Multivariable predictive models support personalized treatment selection.

OBJECTIVE To develop a model predicting response to internet-delivered cognitive behavioral therapy (iCBT) and test its treatment specificity against antidepressant medications.

DESIGN, SETTING, AND PARTICIPANTS The Precision in Psychiatry prognostic study was a 4-week study collecting extensive baseline self-report and cognitive data online to predict early iCBT response, from February 2019 to May 2022. Patients in the iCBT group were recruited via an Irish mental health charity and a UK NHS Talking Therapies clinic. A separate antidepressant group was recruited globally online and via print advertisements. Participants were aged 18 to 70 years, fluent in English, had computer access, started treatment within 2 days of enrollment, and scored at least 10 on the Work and Social Adjustment Scale. Analysis was completed in December 2024.

EXPOSURES Low-intensity, clinician-guided iCBT with multimedia psychoeducation. Patients receiving antidepressants primarily received selective serotonin-reuptake inhibitors or serotonin-norepinephrine reuptake inhibitors.

MAIN OUTCOMES AND MEASURES Machine learning models were trained using the iCBT sample to predict change in depression severity (16-item Quick Inventory of Depressive Symptomatology–Self Report) at week 4. The best model was tested on holdout iCBT and antidepressant data. A separate model was trained on patients receiving iCBT only to assess treatment-specificity.

RESULTS Of 2674 patients screened, 883 completed baseline and final assessments, with 776 patients receiving iCBT (mean [SD] age, 31.8 [11.0] years; 600 [77.5%] female) and 107 patients receiving antidepressant medication (mean [SD] age, 30.1 [10.4] years; 78 [72.9%] female). Both samples had some treatment overlap (24% and 34%, respectively). Elastic net regression with 27 predictors best explained the variance in depression change (R^2 = 14%; SD, 0.8%; 95% CI, 13.8%-14.2%). Key predictors included baseline depression, treatment expectation, transdiagnostic symptoms, and, less strongly, cognitive variables. The model performed well on holdout iCBT (R^2 = 18.8%; root mean square error [RMSE], 0.88) and antidepressant (R^2 = 17.9%; RMSE, 1.10) data. Retraining on 181 patients who received iCBT only increased treatment specificity in predictions (R^2 = 19.3%; RMSE, 0.89) vs 71 patients who received antidepressants only (R^2 = 10.8%; RMSE, 1.17).

CONCLUSIONS AND RELEVANCE This prognostic study in a naturalistic setting found that self-reported data predicted iCBT response better than cognitive data. Model predictions generalized to

(continued)

Key Points

Question Can a machine learning model predict response to internet-based cognitive behavioral therapy (iCBT) but not antidepressant medication?

Findings In this prognostic study of 883 patients receiving iCBT or antidepressant medications, an elastic net regression with 27 mostly self-reported predictors explained 19% of the variance in depression improvement in unseen data. The model generalized well to patients receiving antidepressants when the sample contained individuals also receiving psychotherapy, but retraining the model on patients receiving only iCBT delivered a treatment-specific model.

Meaning These findings suggest that clinical and cognitive data are predictive of iCBT response in a naturalistic clinical setting, with evidence of treatment specificity.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

patients receiving antidepressants, some of whom also received psychotherapy. Training models on single-treatment cohorts may yield more treatment-specific predictions.

JAMA Network Open. 2025;8(11):e2541639. doi:10.1001/jamanetworkopen.2025.41639

Introduction

Depression is a leading cause of disability worldwide, affecting millions of individuals annually.[1] Effective treatments, like psychotherapy (eg, cognitive behavioral therapy [CBT]) and antidepressant medication, exist,[2] but approximately one-half of patients do not respond, and only one-third achieve remission.[3-5] One way to improve this is by developing decision-assistive machine learning models that can personalize treatment selection.[6,7]

Studies of this sort have increased in popularity,[6-10] but most previous models have been developed with small samples that risk overfitting, with predictions often not validated in independent data. A recent review showed only 14 of 44 studies using machine learning to predict psychotherapy response included more than 200 patients, with as few as 3 studies externally validating their models.[8] As methods have improved, there are some early signs of success. For example, a model using baseline demographic and clinical data achieved 60% accuracy in predicting antidepressant response,[11] and similar performance has been observed in internet-delivered CBT (iCBT).[12] More recent work incorporating multimodal data, eg, genetics, survey meta-data, and more diverse self-reports, has also shown enhanced prediction performance.[13,14]

Despite progress, a significant gap remains concerning the generalizability and treatment specificity of these models and an overreliance on clinical trial data that may not translate to naturalistic settings. Generalizability refers to a model's ability to perform well on unseen data from various settings (eg, different countries). Recent work has questioned this property in schizophrenia treatment prediction models.[15] Treatment specificity indicates the extent to which a model's prediction for one treatment is distinct from another. While some evidence showed that models can be drug specific,[13] it is unknown whether models that predict psychotherapy response also apply to antidepressants. This comparison is important, as practitioners frequently choose between psychotherapy or pharmacotherapy for treating patients with depression in primary care.[16] Ultimately, both the model's generalizability and treatment specificity are crucial for assessing its future clinical utility. This study aimed to develop a multivariable predictive model for iCBT that was trained and tested on large-scale data. We evaluated the model's generalizability and treatment specificity, while identifying its most predictive features.

Methods

Study Design and Participants

The Precision in Psychiatry (PIP) study is a prognostic, longitudinal study approved by Trinity College Dublin School of Psychology Research Ethics Committee, the NHS Northwest-Greater Manchester West Research Ethics Committee, Health Research Authority, and Health and Care Research Wales. All participants provided informed consent online. This study is reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline.

The PIP study aimed to use machine learning to identify baseline self-reported and cognitive predictors of early iCBT response in depression. Using a fully internet-based protocol,[17] we recruited patients initiating clinician-guided iCBT from a mental health charity in Dublin, Ireland, and an NHS Talking Therapies clinic in Berkshire, UK, of whom 24% also reported using antidepressant medications. We further recruited a second group of patients initiating antidepressants. Similarly, 34% of these patients reported concurrently receiving psychotherapy. Eligible patients were aged 18 to 70 years, fluent in English, had computer access, started treatment withing 2 days of study enrollment, and scored at least 10 on the Work and Social Adjustment Scale,[18] a transdiagnostic measure of mental health–related functional impairment. Participants were sent an individualized schedule for completing the baseline assessment, 3 weekly check-ins, and the final assessment (eFigure 1 in Supplement 1).

Outcome Measure

Our primary outcome is the 16-item Quick Inventory of Depressive Symptomatology–Self Report (QIDS-SR), a widely used measure of major depressive disorder symptoms based on *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition) (*DSM-IV*) criteria with high internal consistency and strong concurrent validity.[19] The QIDS-SR total score classifies severity into none (0-5), mild (6-10), moderate (11-15), severe (16-20), and very severe (21-27). Early treatment response was defined as the change in QIDS-SR from baseline to final assessment. We included baseline QIDS-SR in our models which, in the context of regression, transforms change score models into residualized change models with equivalent interpretation for effects of interest[20] while retaining clinical interpretability of the outcome variable.

With the final assessment (approximately 1.5 hours) being much longer than weekly check-ins (10 minutes), attrition was high from week 3 to 4 (74 participants in the iCBT group; 10 participants in the antidepressant group). To maximize sample sizes for model training and validation, we imputed missing final QIDS-SR scores in the training sample by regressing week 4 QIDS-SR on week 3 QIDS-SR ($b = 0.88$; SE, 0.03; $P < .001$; where a 1-point increase at week 3 predicted a 0.88-point increase at week 4), before applying the resulting coefficient to impute missing week 4 QIDS-SR in the test samples. This increased the final sample from 799 participants to 883 participants. Sensitivity analyses using multiple imputation and complete-case analysis produced similar results consistent with the main analyses (eAppendix 1 and eAppendix 2 in Supplement 1).

Predictors of Interest

At baseline, participants completed self-report assessments and gamified cognitive tasks. Predictors of interest encompassed 6 categories: sociodemographics, psychosocial, physical health and lifestyle, clinical, treatment, and cognitive data. eAppendix 3 in Supplement 1 details the data processing and transformation procedures we undertook for modeling. The predictive efficacy of features varies based on the algorithm used and the level of information they offer. Therefore, we used an exploratory strategy for feature selection in our training set, testing different combinations of feature sets with varying granularity based on an item-level approach (357 predictors), where each item serves as an individual predictor; a subscale score approach (100 predictors) using psychometric subscale scores (total scores if unavailable) as predictors; a total score approach (78 predictors) using psychometric total scores as predictors; and a filter-based approach using the random forest variable importance calculation method applied to all items to circumvent human bias in feature selection. These approaches rely on the same input data but differ in how they are aggregated, meaning that participant burden is identical across methods. Notably, the total and subscale feature sets incorporated 9 symptom-specific QIDS-SR items, whereas the item-level and filter-based feature sets incorporated all 16 items. A comprehensive list of study variables is available elsewhere.[17]

Model Training and Cross-Validation

All data processing and analyses were carried out using R software version 4.2.2 (R Project for Statistical Computing). **Figure 1** illustrates the modeling pipeline for the study. We first trained a variety of linear and nonlinear machine learning models using 70% of the iCBT sample (543 patients) as the training data. These included elastic net regression, random forest, extreme gradient boosting machine (XGBoost) in linear and tree form to explore the optimal model fit with regards to diversity in model complexity, interpretability, and feature selection. To internally evaluate model performance in predicting QIDS-SR change, we carried out a 10-fold nested repeated cross validation (nrCV) for each model. nrCV is a resampling technique that separates hyperparameter optimization and model selection (inner folds) from performance evaluation (outer folds) to minimize overfitting and optimistic bias in performance estimation[21] (eAppendix 4 in Supplement 1). We evaluated the models based on their root mean square error (RMSE) and selected the model with the lowest RMSE. The model was then subjected to a last round of 10-fold cross-validation performed on the entire training dataset to determine the optimal hyperparameters used for fitting the final model for prediction with external data.

We applied nrCV to each algorithm with each feature set separately. In the filter-based approach, the random forest–based method ranks predictors by the mean decrease in prediction accuracy when one is shuffled (while others remained unchanged), where larger decreases indicate greater predictor importance. For each algorithm type, we estimated a benchmark model comprising baseline depression severity, age, and sex as predictors. This allowed us to quantify the improvement in prediction from additional predictors by comparing their performance to a minimal set of variables. Altogether, we trained and compared 20 models based on nrCV performance estimates. We used the coefficient of determination R^2, which quantifies the proportion of variance in QIDS-SR change explained by the model, to determine its expected predictive utility in unseen data.

Model Testing and External Validation

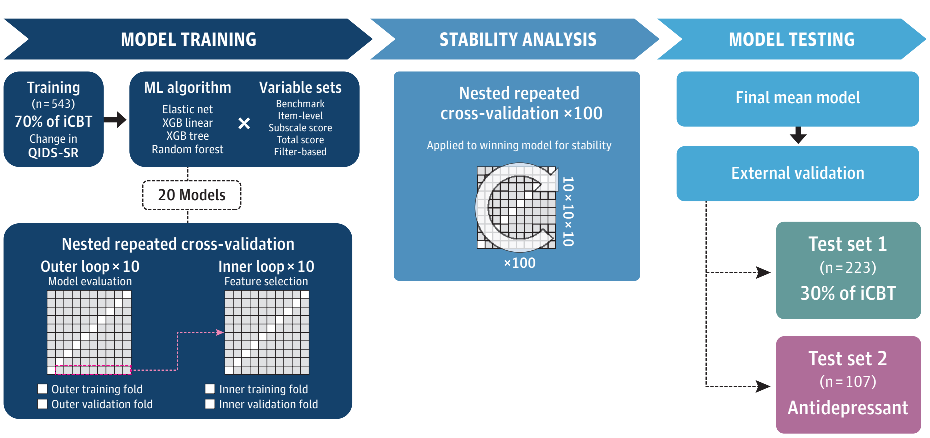
Stability Analyses

To ensure robustness against random data partitioning, we repeated the nrCV analyses 100 times for the final model. We calculated the means of feature importance metrics, such as coefficients from each iteration, to identify the most consistently retained predictors for external validation.

External Validation

We tested the final mean model's performance on 2 holdout datasets. The first dataset included the remaining 30% of the iCBT sample (233 patients). The 70:30 random split of the iCBT sample for model training and testing was stratified to balance recruitment site (Ireland or UK), age, sex, baseline, and final QIDS-SR. The second dataset comprised the antidepressant sample (107 patients), which allowed us to assess the model's generalizability and treatment specificity. Both samples differed in their geographic dispersion, recruitment method, and various demographic and clinical characteristics.

Figure 1. Predictive Modeling Pipeline



iCBT indicates internet-delivered cognitive behavioral therapy; ML, machine learning; QIDS-SR, Quick Inventory of Depressive Symptomatology–Self Report; XGB, extreme gradient boosting machine.

Treatment Specificity

Given the mixed nature of the samples, we further trained a model following the same analytical procedures as our winning model using patients receiving iCBT only (no medication), and tested on holdout data that were similarly parsed (181 patients receiving iCBT only; 71 patients receiving antidepressants only).

Predictor Importance

To understand the factors driving the final model's predictions, we assessed variable importance using Shapley additive explanation values.[22] Shapley additive explanation values determine the mean marginal contribution of each predictor to the model's performance across all possible combinations of predictors in a model-agnostic approach.

Statistical Analysis

In the holdout samples, we conducted permutation tests to assess whether the final model performed significantly different to the benchmark model including only age, sex, and baseline severity. We created a null distribution of R^2 differences between the final and the benchmark model by randomly switching each prediction entry from the 2 models with a probability of 0.5 before comparing their predictive power. We repeated this 1000 times and derived a *P* value based on the overlap of the true R^2 difference between the models and this null distribution. Data analyses were completed in December 2024. Statistical significance was set at 2-sided α = .05.

Results

The study included 883 patients, with 776 receiving iCBT (mean [SD] age, 31.8 [11.0] years; 600 [77.5%] female) and 107 receiving antidepressants (mean [SD] age, 30.1 [10.4] years; 78 [72.9%] female) (**Table**; eFigure 1, and eAppendix 5 in Supplement 1). Both samples were typically married, employed, third-level–educated, based in Ireland or the UK, and self-rated as middle social class. Most patients had 1 or more mental health diagnoses, most commonly depression. While most patients reported no family history of mental illness, they had typically experienced at least 2 lifetime episodes, beginning in adolescence or adulthood, with the current episode lasting a median (range) of 199 (0-19 490) days in the iCBT group and 200 (2-15 904) days in the antidepressants group. Most patients had some history of mental health treatment. Participants reported a moderate level of expectation that their treatment would work on a scale from 1 to 10, with 10 indicating greater expectation (mean [SD] score, 5.06 [2.05] in the iCBT group; 4.77 [1.95] in the antidepressant group). Treatment overlap occurred in 185 patients (23.8%) of the iCBT group and 36 patients (33.6%) of the antidepressant group.

nrCV Performance and Final Model Derivation

The performance of the 20 models assessed during nrCV analyses of the training dataset are shown in **Figure 2** and eTable 1 in Supplement 1. The best-performing model was the elastic net regression with the total score feature set, which had the lowest prediction error (RMSE, 0.92; mean absolute error, 0.736) and the highest variance explained (R^2 = 14.6%). Stability analyses, repeating the nrCV 100 times, revealed that the elastic net regression with total score features predicted a mean R^2 of 14% (SD, 0.8%; 95% CI, 13.8%-14.2%) in QIDS-SR change (mean RMSE, 0.93; SD, 0.004; 95% CI, 0.926-0.927).

External Validation

Bringing forward the final model for external validation, we found the model explained more variance in QIDS-SR change than in nrCV when applied to the iCBT holdout dataset (R^2 = 18.8%; RMSE, 0.88) and the antidepressant dataset (R^2 = 17.9%; RMSE, 1.10). For comparison, we additionally tested the elastic net regression benchmark model on holdout datasets (iCBT: R^2 = 11.1%; RMSE, 0.92;

Table. Baseline Characteristics of Study Completers

Characteristic	Patients, No. (%)		t/χ^2 (*df*)	*P* value
	iCBT[a]	Antidepressant[b]		
Sex				
Female	600 (77.5)	78 (72.9)	2.03 (2)	.36
Male	168 (21.7)	27 (25.2)		
Other	6 (0.8)	2 (1.9)		
Country				
United Kingdom	648 (83.7)	42 (39.2)	124.09 (2)	<.001
Ireland	109 (14.1)	46 (43.0)		
Other	17 (2.2)	19 (17.8)		
Age, mean (SD), y	31.8 (11.0)	30.1 (10.4)	1.52 (877)	.13
Marital status				
Single	303 (39.1)	47 (43.9)	1.60 (5)	.90
In a relationship	226 (29.2)	31 (29.0)		
Married	196 (25.3)	23 (21.5)		
Divorced	28 (3.6)	3 (2.8)		
Separated	19 (2.5)	3 (2.8)		
Widowed	2 (0.3)	0		
Education				
<Third Level	181 (23.4)	15 (14.0)	4.81 (2)	.09
Some/complete third level	412 (53.2)	63 (58.9)		
>Third Level	181 (23.4)	29 (27.1)		
Employment				
Employed	555 (71.7)	57 (53.3)	15.38 (2)	<.001
Unemployed	213 (27.5)	48 (44.8)		
Retired	6 (0.8)	2 (1.9)		
Subjective social status, mean (SD)[c]	4.17 (1.68)	4.40 (1.98)	−1.30 (879)	.19
Current diagnoses, No.				
None	263 (34.0)	8 (7.5)	34.11 (2)	<.001
1	276 (35.6)	45 (42.0)		
>1	235 (30.4)	54 (50.5)		
Types of diagnoses[d]				
None	263 (34.0)	8 (7.5)	9.84 (5)	.08
Depression	367 (47.4)	81 (75.5)		
GAD	300 (38.8)	58 (54.2)		
Panic disorder	32 (4.1)	5 (4.2)		
PTSD	28 (3.6)	14 (13.1)		
OCD	36 (4.7)	4 (3.7)		
Others	55 (7.1)	13 (12.1)		
Family members with mental disorders, No.				
None	320 (41.3)	37 (34.6)	3.44 (3)	.33
1	238 (30.8)	33 (30.8)		
2	122 (15.8)	18 (16.8)		
≥3	94 (12.1)	19 (17.8)		
Lifetime episodes, No.				
<2	81 (10.6)	8 (7.6)	7.54 (2)	.02
2-5	371 (48.7)	40 (37.7)		
>5	310 (40.7)	58 (54.7)		
Age of onset				
Childhood	132 (17.4)	27 (25.7)	10.04 (2)	.007
Adolescence	315 (41.6)	51 (48.6)		
Adulthood	311 (41.0)	27 (25.7)		
Current episode duration, median (range), d[e]	199 (0-19 490)	200 (2-15 904)	0.19 (797)	.85

(continued)

Table. Baseline Characteristics of Study Completers (continued)

Characteristic	Patients, No. (%) iCBT[a]	Patients, No. (%) Antidepressant[b]	t/χ^2 (*df*)	*P* value
History of past treatment				
Never	353 (45.6)	34 (31.8)	10.11 (3)	.02
Psychotherapy and medication	170 (22.0)	34 (31.8)		
Medication only	123 (15.9)	15 (14.9)		
Psychotherapy only	128 (16.5)	24 (22.4)		
Concurrent treatment[f]	185 (23.8)	36 (33.6)	14.95 (1)	<.001
Treatment expectation, mean (SD)[g]	5.06 (2.05)	4.77 (1.93)	1.40 (879)	.16

Abbreviations: GAD, generalized anxiety disorder; iCBT, internet-delivered cognitive behavioral therapy; OCD, obsessive compulsive disorder; PTSD, posttraumatic stress disorder.

[a] Excluded missing data for descriptive analyses (2 participants for sex, country, marital status, education, employment, subjective social status, number of diagnoses, types of diagnoses, family with mental health disorders, treatment history, and treatment expectations; 3 participants for age; 14 participants for number of lifetime episodes; 18 participants for age of onset; and 76 participants for current episode duration).

[b] Excluded missing data for descriptive analyses (1 participant for age, 1 participant for number of lifetime episodes, 2 participants for age of onset, and 8 participants for current episode duration).

[c] Range, 0 to 10; higher score indicates higher perceived social status.

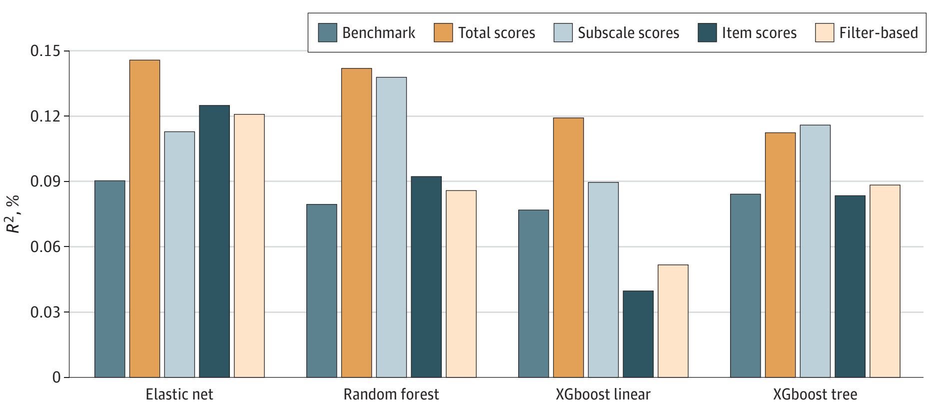
[d] The total number of diagnoses type exceeds the sample size of completers (ie, participants have the option to pick >1 diagnosis).

[e] Median and range reported instead of mean and SD due to extremely skewed data.

[f] Concurrent treatment refers to brain health medication for participants receiving iCBT and psychotherapy for participants receiving antidepressants reported at any time during the study. This definition deviated from that detailed in our methodological protocol, which considered concurrent treatments reported poststudy initiation,[16] as we recovered missing data on this information reported at baseline since the publication of the protocol.

[g] Range, 0 to 9; higher score indicates higher expectations.

Figure 2. Model Performance From Nested Repeated Cross-Validation



antidepressants: R^2 = 12.7%; RMSE, 1.14) (**Figure 3**A). A permutation significance test showed that the final mean model was superior to the benchmark model in the holdout iCBT data (change in R^2 = 7.7%; P = .004), but this was not the case for the antidepressant data (change in R^2 = 5.2%; P = .10) (eFigure 2 in Supplement 1).

Treatment Specificity of Predictions

Due to the nature of our data, some patients were receiving concurrent treatments. To test whether full treatment segmentation would improve model treatment specificity, we retrained our final model using 410 patients receiving iCBT and no antidepressants. Although the model's performance during nrCV (mean from 100 runs) decreased slightly (mean R^2 = 11.4%; mean RMSE, 0.91), it still generalized well to 181 participants in the iCBT holdout dataset who did not use antidepressants (R^2 = 19.3%; RMSE, 0.89), but performed markedly worse on 71 patients using antidepressants but not receiving psychotherapy (R^2 = 10.8%; RMSE, 1.17) (Figure 3B). Indeed, permutation tests showed that the benchmark model was not significantly different from the retrained final model when tested on the antidepressant sample who were not receiving psychotherapy (change in R^2 = −2.0%; P = .73) where the benchmark performed nominally better. In contrast, in the iCBT holdout sample not using antidepressants, the final model significantly outperformed the benchmark model (R^2 = 11.7%; RMSE, 0.93; change in R^2 = 7.5%, P = .007) (eFigure 2 in Supplement 1). Altogether, this suggests that a model trained on patients receiving iCBT and not using antidepressants showed specificity in predicting iCBT response over antidepressant.

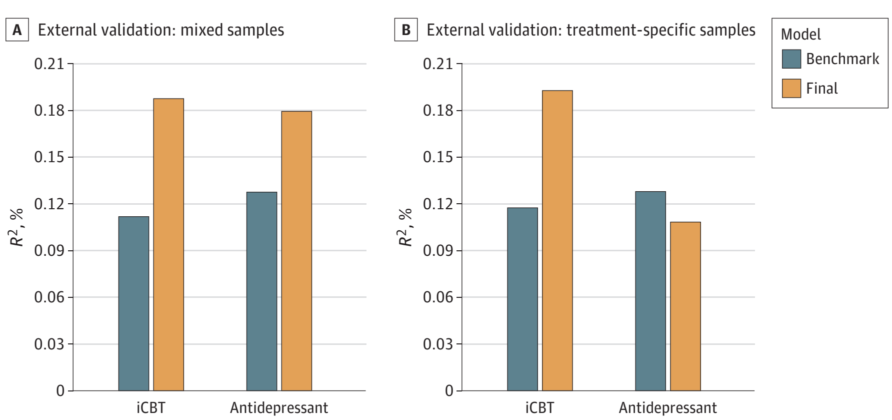
Predictor Importance

Among 78 predictors in the final mean model trained with our mixed iCBT sample, the same 27 variables were retained (ie, coefficients >0) in all 100 repeats of the nrCV (**Figure 4**A). Four additional variables were retained 61 times, which were removed from final analyses (eTable 2 and eTable 3 in Supplement 1). The top 2 predictors were 2 different baseline depression severity measures on the QIDS-SR and the Zung Self-Rating Depression (SDS) scale, albeit in opposing directions (increased QIDS-SR and decreased SDS were associated with greater reductions in QIDS-SR). Following this, treatment expectation ranked highly, with higher ratings associated with greater improvement in depression symptoms. Other predictors included mostly psychosocial and health factors (eg, stressful life events, social support, exercise, diet), depression-specific symptoms (eg, low mood, lack of interest), and transdiagnostic characteristics (eg, apathy, impulsivity). Several cognitive variables were also retained (eg, metacognition efficiency, model-based and model-free indices) along with 1 demographic variable (ie, sex) but with lower importance compared to the rest. The iCBT-only model had fewer (ie, 18) but similar predictors as the mixed-sample model (Figure 4B).

Discussion

In this prognostic study using baseline data from a large-scale naturalistic sample of patients receiving iCBT, we developed a model with 27 predictors that explained 19% of variance in depression improvement during external validation. The model outperformed a benchmark model including age, sex, and baseline depression and generalized well to a distinct sample of patients initiating antidepressants, despite notable differences in recruitment and patient characteristics. Importantly, retraining the model on patients receiving iCBT only (ie, excluding those also receiving antidepressants) resulted in predictions that failed to generalize to the sample receiving antidepressant only (ie, no psychotherapy), demonstrating treatment specificity.

Figure 3. Final Mean Model Performance During External Validation

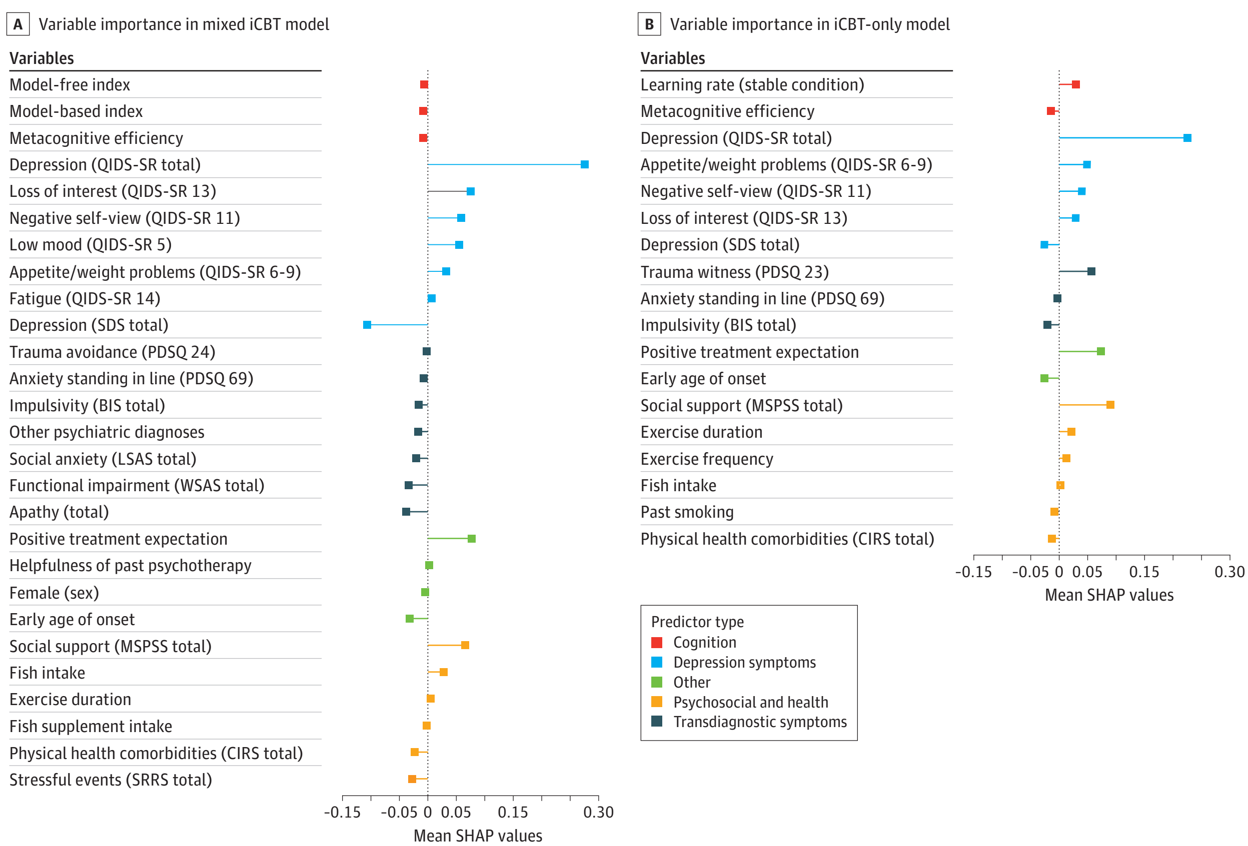


iCBT indicates internet-delivered cognitive behavioral therapy.

Our final model performed comparably to, or better than, similar studies using baseline data to predict response after full treatment courses or over longer follow-ups across a range of treatment-seeking samples and, crucially, in more tightly controlled designs.[11,13,23,24] This matters because clinical settings are much noisier, yet they are where these models are intended to be deployed. Other data types, such as genetics, could improve predictions[14] but are more costly and burdensome for participants than the digital data emphasized in this study.[17]

In our final model, baseline depression severity was the strongest predictor, consistent with numerous machine learning prediction studies,[11,14,25,26] and studies investigating iCBT.[27-29] Treatment expectation was the next most influential predictor following baseline depression severity, consistent with studies showing its significant contribution to therapeutic gains across psychotherapy.[30,31] Similar findings have been reported for invasive chronic pain treatments[32] and noninvasive brain stimulation, where a study showed participants' treatment beliefs had a greater effect than the treatment itself.[33] Cognitive features made a modest contribution to the model's prediction, ranking below most self-report variables.[6,7,13,14] This is notable, as cognitive tasks required considerably more time than questionnaires and were generally less well-tolerated by participants, suggesting their limited predictive value may not outweigh the added participant burden. Indeed, the study end point including these tasks lasted approximately 1.5 hours and coincided with a significant rise in dropout.

Figure 4. Ranked Mean Shapley Additive Explanation (SHAP) Values of Retained Predictors in the Final Mean Model



BIS indicates Barratt Impulsiveness Scale; CIRS, Cumulative Illness Rating Scale; iCBT, internet-delivered cognitive behavioral therapy; LSAS, Liebowitz Social Anxiety Scale; MPSS, Multidimensional Scale of Perceived Social Support; QIDS-SR, Quick Inventory of Depressive Symptomatology–Self Report; PDSQ, Psychiatric Diagnostic Screening Questionnaire; SDS, Zung Self-Rating Depression; SRRS, Social Readjustment Rating Scale; WSAS, Work and Social Adjustment Scale; XGB, extreme gradient boosting machine.

Limitations

This study has some limitations. Models that can predict whether a patient responds to psychotherapy or pharmacotherapy are clinically valuable for treatment selection, but an important next step is to develop models capable of predicting differential response to 2 or more available treatments.[11,13] When we trained a model on patients receiving iCBT only, we found some evidence for this treatment specificity: the final model no longer outperformed the benchmark model in participants receiving antidepressants. While indicative, these results should be interpreted with caution, as the observed difference in significance is not equivalent to a significant model × group interaction. We did not separately train an antidepressant-only model to assess its treatment specificity in predicting iCBT response due to limited sample size (ie, ≥500 patients, as suggested by simulations),[34] which constitutes another important target for future work. Moreover, the observational design of the study means that treatment assignment was nonrandom, underpinned by routine clinical decisions.[7,17] Although this limits our ability to infer causation, it is also a strength of our study, showing commensurate model performance with that of randomized trials.

Our outcome measure relied solely on subjective patient ratings. While this is a pragmatic approach for clinical translation and implementation in iCBT, a clinically assigned outcome measure may improve model performance. Moreover, binary remission may be a more preferrable outcome than continuous scores for yes or no clinical decisions, but for the 4-week timescale used here, remission rates were too low. Although our study gathered an extensive range of predictors, it lacked some key variables (eg, race and ethnicity) and other variable modalities (eg, genetics). Future work might further differentiate treatments by including features conceptually linked to iCBT (eg, digital fluency). In addition, future studies should explore smartphone-based and passive-sensing data (eg, app usage, geographic location), which are increasingly accessible in digitized research and intervention.[6,7]

Conclusions

The PIP prognostic study used a fully-remote, scalable digital method in a naturalistic clinical setting. We found self-reported and cognitive characteristics predictive of early response in iCBT, with some suggestion of treatment specificity. The digital nature of this study lends itself to rapid clinical translation in the growing field of online psychotherapy.

ARTICLE INFORMATION

Accepted for Publication: September 10, 2025.

Published: November 6, 2025. doi:10.1001/jamanetworkopen.2025.41639

Corresponding Author: Chi Tak Lee, PhD (clee@tcd.ie), and Claire M. Gillan, PhD (gillancl@tcd.ie), Lloyd Institute, Trinity College Institute of Neuroscience, School of Psychology, Dublin D02 R123, Ireland.

Author Affiliations: School of Psychology, Trinity College Dublin, Dublin, Ireland (Lee, Richards, Hanlon, Lynch, Harty, Whelan, Gillan); Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland (Lee, Hanlon, Lynch, Harty, O'Keane, Whelan, Gillan); Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland (Lee, Hanlon, Lynch, Harty, Whelan, Gillan); Amwell Science, Amwell, Dublin, Ireland (Lee, Richards, Harty); Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zürich & Eidgenössische Technische Hochschule, Zürich, Switzerland (Heinzle, Stephan); Institute of Psychology, Division of Clinical Psychology and Psychological Treatment in Adults, University of Bremen, Bremen, Germany (Claus); Department of Psychiatry, Trinity College Dublin, Dublin, Ireland (O'Keane); Tallaght Hospital, Trinity Centre for Health Sciences, Tallaght University Hospital, Dublin, Ireland (O'Keane); Max Planck Institute for Metabolism Research, Cologne, Germany (Stephan).

Author Contributions: Drs Lee and Gillan had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Lee, Richards, Harty, O'Keane, Stephan, Gillan.

Acquisition, analysis, or interpretation of data: Lee, Heinzle, Hanlon, Lynch, Claus, Stephan, Whelan, Gillan.

Drafting of the manuscript: Lee, O'Keane, Gillan.

Critical review of the manuscript for important intellectual content: Lee, Richards, Heinzle, Hanlon, Lynch, Harty, Claus, Stephan, Whelan, Gillan.

Statistical analysis: Lee, Heinzle, Stephan, Whelan, Gillan.

Obtained funding: Richards, Gillan.

Administrative, technical, or material support: Lee, Hanlon, Lynch, Harty.

Supervision: Richards, Stephan, Gillan.

Conflict of Interest Disclosures: Dr Lee reported previously being employed by SilverCloud outside the submitted work. Dr Stephan reported receiving support from the René and Susanne Braginsky Foundation and the ETH Foundation. Dr Claus reported receiving support from Erasmus+ and the European Union during the conduct of this study. Dr Harty reported being a current employee of and holding shares in SilverCloud Health, Amwell during the conduct of the study. Dr Richards previously being employed by SilverCloud Health, Amwell during the conduct of the study. No other disclosures were reported.

Funding/Support: This project was funded by MQ: transforming mental health (MQ16IP13; Dr Gillan). It subsequently formed part of Dr Lee's PhD studentship, cofunded by SilverCloud Health, Amwell, and the Irish Research Council (award No. EPSPG/2020/8). Dr Gillan holds additional funding from Science Foundation Ireland's Frontiers for the Future Scheme (award No. 19/FFP/6418) and a European Research Council Starting Grant (ERC-H2020-HABIT). KES acknowledges support from the René and Susanne Braginsky Foundation and the ETH Foundation. NC acknowledges support from Erasmus+, European Union.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. Relevant funding was used solely for data collection and management.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. World Health Organization. Depression. Accessed October 6, 2025. https://www.who.int/news-room/fact-sheets/detail/depression

2. National Institute for Health and Care Excellence. *Depression in Adults: Treatment and Management*. NICE; 2022.

3. Andersson G, Carlbring P, Rozental A. Response and remission rates in internet-based cognitive behavior therapy: an individual patient data meta-analysis. *Front Psychiatry*. 2019;10:749. doi:10.3389/fpsyt.2019.00749

4. Cuijpers P, Karyotaki E, Ciharova M, Miguel C, Noma H, Furukawa TA. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: a meta-analysis. *Acta Psychiatr Scand*. 2021;144(3):288-299. doi:10.1111/acps.13335

5. Trivedi MH, Rush AJ, Wisniewski SR, et al; STAR*D Study Team. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry*. 2006;163(1):28-40. doi:10.1176/appi.ajp.163.1.28

6. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*. 2021;20(2):154-170. doi:10.1002/wps.20882

7. Rost N, Binder EB, Brückl TM. Predicting treatment outcome in depression: an introduction into current concepts and challenges. *Eur Arch Psychiatry Clin Neurosci*. 2023;273(1):113-127. doi:10.1007/s00406-022-01418-4

8. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. *Psychother Res*. 2021;31(1):92-116. doi:10.1080/10503307.2020.1808729

9. Ermers NJ, Hagoort K, Scheepers FE. The predictive validity of machine learning models in the classification and treatment of major depressive disorder: state of the art and future directions. *Front Psychiatry*. 2020;11:472. doi:10.3389/fpsyt.2020.00472

10. Sajjadian M, Lam RW, Milev R, et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. *Psychol Med*. 2021;51(16):2742-2751. doi:10.1017/S0033291721003871

11. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243-250. doi:10.1016/S2215-0366(15)00471-X

12. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. *Digit Health*. 2021;7:20552076211060659. doi:10.1177/20552076211060659

13. Iniesta R, Hodgson K, Stahl D, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Sci Rep*. 2018;8(1):5530. doi:10.1038/s41598-018-23584-z

14. Wallert J, Boberg J, Kaldo V, et al. Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. *Transl Psychiatry*. 2022;12(1):357. doi:10.1038/s41398-022-02133-3

15. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science*. 2024;383(6679):164-167. doi:10.1126/science.adg8538

16. Park LT, Zarate CA Jr. Depression in the primary care setting. *N Engl J Med*. 2019;380(6):559-568. doi:10.1056/NEJMcp1712493

17. Lee CT, Palacios J, Richards D, et al. The Precision in Psychiatry (PIP) study: testing an internet-based methodology for accelerating research in treatment prediction and personalisation. *BMC Psychiatry*. 2023;23(1):25. doi:10.1186/s12888-022-04462-5

18. Mundt JC, Marks IM, Shear MK, Greist JH. The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *Br J Psychiatry*. 2002;180(5):461-464. doi:10.1192/bjp.180.5.461

19. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*. 2003;54(5):573-583. doi:10.1016/S0006-3223(02)01866-8

20. Castro-Schilo L, Grimm KJ. Using residualized change versus difference scores for longitudinal research. *J Soc Pers Relat*. 2018;35(1):32-58. doi:10.1177/0265407517718387

21. Lewis MJ, Spiliopoulou A, Goldmann K, Pitzalis C, McKeigue P, Barnes MR. nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. *Bioinform Adv*. 2023;3(1):vbad048. doi:10.1093/bioadv/vbad048

22. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, et al. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. NeurIPS Proceedings; 2017:4768-4777.

23. Buckman JEJ, Cohen ZD, O'Driscoll C, et al. Predicting prognosis for adults with depression using individual symptom data: a comparison of modelling approaches. *Psychol Med*. 2023;53(2):408-418.

24. Jacobson NC, Nemesure MD. Using artificial intelligence to predict change in depression and anxiety symptoms in a digital intervention: evidence from a transdiagnostic randomized controlled trial. *Psychiatry Res*. 2021;295:113618. doi:10.1016/j.psychres.2020.113618

25. Pearson R, Pisner D, Meyer B, Shumake J, Beevers CG. A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychol Med*. 2019;49(14):2330-2341. doi:10.1017/S003329171800315X

26. Li F, Jörg F, Merkx MJM, Feenstra T. Early symptom change contributes to the outcome prediction of cognitive behavioral therapy for depression patients: a machine learning approach. *J Affect Disord*. 2023;334:352-357. doi:10.1016/j.jad.2023.04.111

27. Edmonds M, Hadjistavropoulos HD, Schneider LH, Dear BF, Titov N. Who benefits most from therapist-assisted internet-delivered cognitive behaviour therapy in clinical practice: predictors of symptom change and dropout. *J Anxiety Disord*. 2018;54:24-32. doi:10.1016/j.janxdis.2018.01.003

28. Karyotaki E, Efthimiou O, Miguel C, et al; Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*. 2021;78(4):361-371. doi:10.1001/jamapsychiatry.2020.4364

29. Hadjistavropoulos HD, Pugh NE, Hesser H, Andersson G. Predicting response to therapist-assisted internet-delivered cognitive behavior therapy for depression or anxiety within an open dissemination trial. *Behav Ther*. 2016;47(2):155-165. doi:10.1016/j.beth.2015.10.006

30. Mogoașe C, Cobeanu O, David O, Giosan C, Szentagotai A. Internet-based psychotherapy for adult depression: what about the mechanisms of change? *J Clin Psychol*. 2017;73(1):5-64. doi:10.1002/jclp.22326

31. Wampold BE, Minami T, Baskin TW, Callen Tierney S. A meta-(re)analysis of the effects of cognitive therapy versus 'other therapies' for depression. *J Affect Disord*. 2002;68(2-3):159-165. doi:10.1016/S0165-0327(00)00287-1

32. Müller-Schrader M, Heinzle J, Müller A, et al. Individual treatment expectations predict clinical outcome after lumbar injections against low back pain. *Pain*. 2023;164(1):132-141. doi:10.1097/j.pain.0000000000002674

33. Fassi L, Hochman S, Daskalakis ZJ, Blumberger DM, Cohen Kadosh R. The importance of individual beliefs in assessing treatment efficacy. *Elife*. 2024;12:RP88889. Published online March 28, 2024. doi:10.7554/eLife.88889.3

34. Luedtke A, Sadikova E, Kessler RC. Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin Psychol Sci*. 2019;7(3):445-461. doi:10.1177/2167702618815466

SUPPLEMENT 1.

eFigure 1. PIP Study Design Overview and Flow Chart
eTable 1. Model Performance Results From Nested Repeated Cross-Validation (nrCV) Analyses
eAppendix 1. Core analyses Without Final QIDS-SR Imputation
eAppendix 2. Impact of Multiple Imputation on nrCV Model Performance
eAppendix 3. Data Processing Procedures for Modelling
eAppendix 4. Nested Repeated Cross-Validation (nrCV)
eAppendix 5. Between-Arm Baseline Depression Severity Differences and Treatment Effects
eFigure 2. Permutation Testing Between the Benchmark and Final Averaged Model
eTable 2. Ranked Coefficients of Retained Predictors in the Final Elastic Net Regression Model Employing the Total Score Feature Set
eTable 3. Ranked Averaged Coefficients of Retained Predictors in the Final Elastic Net Regression Model Employing the Total Score Feature Set

SUPPLEMENT 2.

Data Sharing Statement