JAMA Network Open

Original Investigation | Psychiatry

Generalizability of Treatment Outcome Prediction Across Antidepressant Treatment Trials in Depression

Peter Zhukovsky, PhD; Madhukar H. Trivedi, MD; Myrna Weissman, PhD; Ramin Parsey, MD, PhD; Sidney Kennedy, MD; Diego A. Pizzagalli, PhD

Abstract

IMPORTANCE Although several predictive models for response to antidepressant treatment have emerged on the basis of individual clinical trials, it is unclear whether such models generalize to different clinical and geographical contexts.

OBJECTIVE To assess whether neuroimaging and clinical features predict response to sertraline and escitalopram in patients with major depressive disorder (MDD) across 2 multisite studies using machine learning and to predict change in depression severity in 2 independent studies.

DESIGN, SETTING, AND PARTICIPANTS This prognostic study included structural and functional resting-state magnetic resonance imaging and clinical and demographic data from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) randomized clinical trial (RCT), which administered sertraline (in stage 1 and stage 2) and placebo, and the Canadian Biomarker Integration Network in Depression (CANBIND-1) RCT, which administered escitalopram. EMBARC recruited participants with MDD (aged 18-65 years) at 4 academic sites across the US between August 2011 and December 2015. CANBIND-1 recruited participants with MDD from 6 outpatient centers across Canada between August 2013 and December 2016. Data were analyzed from October 2023 to May 2024.

MAIN OUTCOMES AND MEASURES Prediction performance for treatment response was assessed using balanced classification accuracy and area under the curve (AUC). In secondary analyses, prediction performance was assessed using observed vs predicted correlations between change in depression severity.

RESULTS In 363 adult patients (225 from EMBARC and 138 from CANBIND-1; mean [SD] age, 36.6 [13.1] years; 235 women [64.7%]), the best-performing models using pretreatment clinical features and functional connectivity of the dorsal anterior cingulate had moderate cross-trial generalizability for antidepressant treatment (trained on CANBIND-1 and tested on EMBARC, AUC = 0.62 for stage 1 and AUC = 0.67 for stage 2; trained on EMBARC stage 1 and tested on CANBIND-1, AUC = 0.66). The addition of neuroimaging features improved the prediction performance of antidepressant response compared with clinical features only. The use of early-treatment (week 2) instead of pretreatment depression severity scores resulted in the best generalization performance, comparable to within-trial performance. Multivariate regressions showed substantial cross-trial generalizability in change in depression severity (predicted vs observed r ranging from 0.31 to 0.39).

CONCLUSIONS AND RELEVANCE In this prognostic study of depression outcomes, models predicting response to antidepressants show substantial generalizability across different RCTs of adult MDD.

JAMA Network Open. 2025;8(3):e251310. doi:10.1001/jamanetworkopen.2025.1310

Key Points

Question Can neuroimaging and clinical features predict response to sertraline and escitalopram in patients with major depressive disorder across 2 large multisite studies?

Findings In this prognostic study of depression outcomes, among 363 patients in 2 trials, the best-performing models using pretreatment clinical features and functional connectivity of the dorsal anterior cingulate showed substantial cross-trial generalizability. The addition of neuroimaging features significantly improved prediction performance of antidepressant response compared with models including only clinical features.

Meaning Promising generalizability of depression response markers emerged across 2 independent clinical trials of adults with major depressive disorder; future studies with improved predictive models are needed to optimize treatment outcomes.

+ Supplemental content

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

Introduction

Treatment of psychiatric conditions, including major depressive disorder (MDD) often fails, with more than one-half of patients with MDD not responding to first-line antidepressant treatment.[1] Leveraging machine learning in prediction of treatment response promises to accelerate symptom reduction. However, a recent study[2] of clinical markers predicting treatment outcomes has highlighted the challenge of identifying markers that generalize across trials. Although several studies[3-7] have identified promising markers of antidepressant response, it is unclear whether findings generalize across trials and, thus, to future patients.

Clinical trials featuring biomarkers alongside in-depth clinical assessments are scarce and expensive, and different trials often use distinct assessments and imaging protocols to evaluate potential biomarkers.[2,8] As a result, there are no markers of treatment response that cut across different MDD treatment trials,[8] although higher pretreatment depression severity and early change in depression scores have been linked to treatment response.[4]

Functional connectivity (FC) is an important neuroimaging predictor of antidepressant response in MDD.[8-11] Systematic reviews have identified a number of regions whose FC may be associated with response to pharmacological and neurostimulation treatments, including dorsolateral and ventrolateral prefrontal cortex, lateral parietal areas, and the anterior cingulate cortex (ACC).[8,9] Response to selective serotonin reuptake inhibitors (SSRIs) in particular has been associated with lower connectivity between the ACC with the dorsolateral prefrontal cortex (dlPFC) and insula.[6] Among structural markers of response, reduced gray matter volume in cortical regions[12] and hippocampus[13] predicted treatment outcomes, especially in late-life depression. Critically, recent studies[14,15] linking magnetic resonance imaging (MRI) data with maps of gene expression and receptor binding can help improve interpretation of molecular correlates of MRI-derived biomarkers.

Clinical and pathophysiological heterogeneity may explain variability in MDD biomarkers.[9] To address heterogeneity, unique biomarkers need to be identified in independent samples. Accordingly, studies combining clinical trials of different samples, similar to Chekroud et al,[2] are needed to test the generalizability and robustness of clinical and biological markers of treatment outcomes. Accordingly, we investigated the generalizability of models featuring clinical and functional MRI (fMRI) features across 2 MDD trials, the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) and Canadian Biomarker Integration Network in Depression (CANBIND-1). These trials administered SSRIs—sertraline in EMBARC and escitalopram in CANBIND-1—to adults with non–treatment-resistant MDD. We expected early-treatment models to outperform pretreatment models and the addition of ACC connectivity features to improve model performance.

Methods

Study Design

In this prognostic study, we used clinical, demographic, and neuroimaging data from 2 MDD randomized clinical trials: EMBARC and CANBIND-1. EMBARC is a 2-stage trial that recruited participants with MDD (aged 18-65 years) at 4 academic sites across the US between August 2011 and December 2015; participants were randomized to sertraline or placebo in stage 1. After 8 weeks, in stage 2, nonresponders to sertraline were switched to bupropion, nonresponders to placebo were switched to sertraline, and responders to sertraline or placebo continued with their treatment. Similarly, CANBIND-1 is a 2-step trial that recruited participants with MDD from 6 outpatient centers across Canada between August 2013 and December 2016; participants received escitalopram in stage 1 for 8 weeks. In stage 2, nonresponders were offered an augmentation treatment of escitalopram with aripiprazole, whereas responders continued with their treatment. Detailed descriptions of the EMBARC[16] and CANBIND-1[17,18] design have been published elsewhere. Ethical approval for EMBARC was obtained from the institutional review board at each site. Approval for

CANBIND-1 was obtained from research ethics boards at each site. Participants provided written, informed consent for all study procedures. The methods of the current study follow Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines.

Participants

Among 296 unmedicated outpatients in EMBARC, we defined 3 subgroups: individuals treated with sertraline in stage 1 with complete data, a different group of individuals treated with sertraline in stage 2 after not responding to placebo, and individuals receiving placebo in stage 1. The stage 2 sertraline group is, thus, a subset of the stage 1 placebo group, who did not respond to placebo. Among 144 patients with MDD in CANBIND-1, we included those who had completed at least 6 weeks of escitalopram treatment and had complete baseline prediction data.

Clinical Data

In addition to age, sex, employment status, and body mass index (BMI; calculated as weight in kilograms divided by height in meters squared), we used baseline depression severity (CANBIND-1, Montgomery Asberg Depression Rating Scale [MADRS][19] and EMBARC, 17-item Hamilton Depression Rating Scale [HDRS][20]), and anhedonia (Snaith Hamilton Rating Scale [SHAPS][21]) as baseline predictors. MADRS scores were converted to HDRS in CANBIND-1 following a validated mapping[22] (eAppendix 1 in Supplement 1). Early treatment models included the change in depression scores between week 2 and baseline.

Treatment Outcomes

Primary treatment outcome was treatment response, defined as a 50% or greater reduction in depression severity (EMBARC, HDRS scores; CANBIND-1, converted MADRS scores). In EMBARC, HDRS was assessed at weeks 8 and 16 in stages 1 and 2, respectively, whereas MADRS outcomes were collected at week 8 in CANBIND-1. When no data were available at week 8 in CANBIND-1, the closest assessment (ie, week 6) data were used instead. We also present analyses predicting change in depression severity in eAppendix 2 in Supplement 1.

MRI Data

We preprocessed structural and resting-state fMRI data in EMBARC using fMRIPrep software version 22.1.1 and in CANBIND-1 using fMRIPrep software version 23.0.2 (The fMRIPrep Developers), using fixed confound regression for denoising and applied a 6-mm smoothing kernel. FC matrices were calculated for the Human Connectome Project cortical parcellation.[23] Global cortical FC was obtained by averaging the rows of these connectivity matrices; we also obtained seed-based FC of the dorsal ACC (dACC) and rostral ACC (rACC), respectively.

Statistical Analysis

Data were analyzed from October 2023 to May 2024. Following recent studies,[2] we used elastic net logistic regressions with regularization (lassoglm in Matlab R2022a[24]; MathWorks) to predict treatment outcomes in the 4 datasets. We tested 5 sets of models using baseline depression severity: (1) a clinical model including age, sex, employment, baseline HRSD, SHAPS, and BMI; (2) a clinical plus global FC model; (3) a clinical plus dACC FC model; (4) a clinical plus rACC FC model; and (5) a clinical plus cortical thickness model. We then tested 5 analogous models that included week 2 instead of baseline depression severity scores. We provide more details on predictive modeling in eAppendix 2 in Supplement 1. The baseline models included the following: (1) response predicted by age plus sex plus employment plus baseline HDRS plus SHAPS plus BMI; (2) response predicted by age plus sex plus employment plus baseline HDRS plus SHAPS plus BMI plus global FC; (3) response predicted by age plus sex plus employment plus baseline HDRS plus SHAPS plus BMI plus dACC FC; (4) response predicted by age plus sex plus employment plus baseline HDRS plus SHAPS

plus BMI plus rACC FC; and (5) response predicted by age plus sex plus employment plus baseline HDRS plus SHAPS plus BMI plus brain structure

First, we tested the prediction performance of models within CANBIND-1 and EMBARC stage 1 by creating 100 random training and test data splits, training the elastic net model with 10-fold cross-validation on the training dataset, and predicting outcomes in the test dataset on each iteration. Second, we evaluated the prediction performance of models trained on CANBIND-1 and tested in EMBARC stage 1, EMBARC stage 2, and EMBARC placebo. Area under the curve (AUC) and balanced accuracy[2] were used to assess prediction performance. Across all models, we used thresholds derived from simulation studies to assess whether balanced accuracy was significantly higher than chance ($P < .05$).[3,25] Second, for a select subset of models (clinical features only and clinical plus dACC models), bootstrapping was used to test whether AUCs were significantly higher than chance and to compare the most promising models with each other (1-tailed $P < .05$).

Next, in secondary analyses we used a multivariate partial least squares regression (PLS-R) to predict change in depression severity scores after treatment (eAppendix 2 in Supplement 1). PLS-R predictors included 360 dACC connectivity features, age, sex, employment, baseline HDRS-17, SHAPS, and BMI. We used permutation testing (n = 5000) to assess model significance (permutation $P < .05$) and bootstrapping (n = 5000; $|Z| > 3$) to identify robust features. After training the model on CANBIND-1, we applied the regression weights to predict change in depression severity in EMBARC. Similarly, after training the model on EMBARC stage 1 sertraline, we applied the regression weights to CANBIND-1, EMBARC stage 2 sertraline, and EMBARC stage 1 placebo.

In sensitivity analyses, first, we reanalyzed the data while correcting for batch effects in the resting-state fMRI data and the gray matter brain structure using ComBat.[26] Second, we bootstrapped elastic net models with the full set of 366 predictors. All code is available elsewhere.[27]

Results

An overview of the demographic and clinical features of the participant groups is provided in **Table 1**. Briefly, of the 363 participants (225 from EMBARC and 138 from CANBIND-1; mean [SD] age, 36.6

Table 1. Demographic and Clinical Sample Characteristics

Characteristic	Participants, No. (%)			
	CANBIND-1 escitalopram (n = 138)	EMBARC stage 1 sertraline (n = 110)	EMBARC stage 2 sertraline (n = 60)[a]	EMBARC stage 1 placebo (n = 115)
Sex				
Female	89 (64.5)	75 (68.2)	34 (56.7)	71 (61.7)
Male	49 (35.5)	35 (31.8)	26 (43.3)	44 (38.3)
Employed	65 (47.1)	62 (56.4)	30 (50.0)	63 (54.8)
Race				
African American or Black	5 (3.6)	22 (20.0)	10 (16.7)	18 (15.7)
Asian	19 (13.8)	5 (4.5)	2 (3.3)	8 (6.9)
White	106 (76.8)	73 (66.4)	43 (71.7)	80 (69.7)
Other[b]	12 (8.7)	10 (9.1)	5 (8.3)	9 (7.8)
Ethnicity				
Hispanic	9 (6.5)	20 (18.2)	10 (16.7)	20 (17.4)
Non-Hispanic	129 (93.5)	90 (81.8)	50 (83.3)	95 (82.6)
Response	60 (43.5)	58 (52.7)	36 (60.0)	42 (36.5)
Age, mean (SD), y	34.8 (12.4)	38.2 (14.0)	39.2 (13.3)	37.3 (13.1)
Body mass index, mean (SD)[c]	26.3 (5.9)	28.7 (8.1)	27.5 (5.9)	28.2 (6.8)
Years of education, mean (SD)	16.9 (2.1)	15.1 (2.6)	15.3 (2.5)	15.4 (2.4)
Pretreatment depression severity, mean (SD)[d]	22.1 (4.3)	18.7 (4.4)	18.6 (3.8)	18.7 (4.3)
Snaith Hamilton Rating Scale, mean (SD)	35.6 (6.0)	33.5 (5.4)	33.2 (5.4)	33.2 (5.8)

Abbreviations: CANBIND-1, Canadian Biomarker Integration Network in Depression; EBARC, Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care.

[a] Participants who received sertraline in EMBARC stage 2 had not responded to placebo in stage 1.

[b] Other race refers to multiracial or unknown. In CANBIND-1 participants could indicate multiple races at the same time.

[c] Body mass index is calculated as weight in kilograms divided by height in meters squared.

[d] Pretreatment depression severity was assessed using the 17-item Hamilton Depression Rating Scale in EMBARC and scores converted from the Montgomery Asberg Depression Rating Scale to the Hamilton Depression Rating Scale scores in CANBIND-1.

[13.1] years), 235 (64.7%) were women, and they showed response rates ranging between 40% and 60%. A detailed breakdown of participant drop-out and missing data is provided in eFigure 1 in Supplement 1. We evaluated the out-of-trial performance of 10 machine learning models in 2 training scenarios: first, we trained the models in CANBIND-1 and tested them in EMBARC stage 1 sertraline, stage 2 sertraline, and stage 1 placebo conditions; second, we trained the models in EMBARC stage 1 sertraline and tested them in CANBIND-1 and EMBARC stage 2 sertraline, as well as EMBARC stage 1 placebo conditions.

Pretreatment Models of Response

AUC and balanced accuracy values for pretreatment models are summarized in **Table 2**. We found that the clinical data model and the model using dACC-to-cortex connectivity alongside clinical data performed best (trained on CANBIND-1 and tested on EMBARC, AUC = 0.62 for stage 1 and AUC = 0.67 for stage 2; trained on EMBARC stage 1 and tested on CANBIND-1, AUC = 0.66). Although the clinical model reached out-of-trial AUCs of 0.58 to 0.61 and balanced accuracy of 59% to 61% when trained and tested on CANBIND-1 and EMBARC antidepressant groups, the addition of dACC connectivity features (clinical plus dACC) improved pairwise out-of-trial model performance to AUCs of 0.61 to 0.68 and balanced accuracy of 61% to 71%. Although the clinical and rACC connectivity model (clinical plus rACC) and the clinical and brain structure (clinical plus cortical thickness) model also achieved good performance when trained in CANBIND-1, they did not generalize well when trained in EMBARC stage 1. The addition of global FC features (clinical plus global FC) did not improve model performance, with worse AUC values across all training and testing setups for groups given SSRIs. Bootstrapping the models showed that the addition of dACC connectivity data significantly improved model performance for EMBARC stage 2 when the models were trained on CANBIND-1, with a trend in improvement in model performance for EMBARC stage 1 sertraline. When mapping

Table 2. Summary of Out-of-Trial Model Performance for Models Trained in the CANBIND-1 and EMBARC Clinical Trials[a]

	Models trained on CANBIND-1						Models trained on EMBARC stage 1 sertraline					
	Tested on EMBARC stage 1 sertraline		Tested on EMBARC stage 2 sertraline		Tested on EMBARC stage 1 placebo		Tested on CANBIND-1 escitalopram		Tested on EMBARC stage 2 sertraline		Tested on EMBARC stage 1 placebo	
Models of response	AUC	BA	AUC	BA	AUC	BA	AUC	BA	AUC	BA	AUC	BA
Pretreatment												
Clinical model[b]	0.58	0.61[c]	0.58	0.60[c]	0.63	0.59[c]	0.59	0.59[c]	0.61	0.60[c]	0.52	0.55
Clinical plus global FC[d]	0.56	0.59[c]	0.50	0.51	0.62	0.59[c]	0.52	0.56	0.60	0.59	0.49	0.57
Clinical plus dACC FC[d]	0.62	0.63[c]	0.67	0.65[c]	0.57	0.55	0.66	0.64[c]	0.70	0.71[c]	0.62	0.61[c]
Clinical plus rACC FC[d]	0.59	0.60[c]	0.63	0.65[c]	0.68	0.64[c]	0.51	0.52	0.57	0.56	0.44	0.46
Clinical plus CT[e]	0.58	0.56	0.61	0.63[c]	0.63	0.63[c]	0.56	0.52	0.62	0.62[c]	0.48	0.51
Early treatment												
Clinical model[f]	0.68	0.69[c]	0.73	0.66[c]	0.71	0.66[c]	0.66	0.69[c]	0.68	0.66[c]	0.63	0.66[c]
Clinical plus global FC[d]	0.64	0.64[c]	0.65	0.60[c]	0.69	0.64[c]	NV[g]	NV[g]	NV[g]	NV[g]	NV[g]	NV[g]
Clinical plus dACC FC[d]	0.68	0.64[c]	0.79	0.73[c]	0.70	0.69[c]	0.59	0.58[c]	0.71	0.70[c]	0.64	0.63[c]
Clinical plus rACC FC[d]	0.66	0.67[c]	0.74	0.67[c]	0.73	0.69[c]	NV[g]	NV[g]	NV[g]	NV[g]	NV[g]	NV[g]
Clinical plus CT[e]	0.69	0.69[c]	0.73	0.70[c]	0.73	0.69[c]	0.57	0.57	0.66	0.61[c]	0.56	0.51

Abbreviations: AUC, area under the curve; BA, balanced accuracy; dACC, dorsal anterior cingulate cortex; FC, functional connectivity; CT, cortical thickness; NV, no variables; rACC, rostral anterior cingulate cortex.

[a] We analyzed data from 138 participants in CANBIND-1, 110 participants who received sertraline in stage 1 of EMBARC, 115 participants who received placebo in stage 1 of EMBARC, and 60 participants who received sertraline in EMBARC stage 2 after not responding to placebo in EMBARC stage 1.

[b] Model includes age, sex, Snaith Hamilton Rating Scale score, employment, body mass index, and baseline Hamilton Depression Rating Scale and Montgomery Asberg Depression Rating Scale scores.

[c] Balanced accuracy values were significantly higher than chance ($P < .05$, not correcting for the number of models) based on prior simulations.[25]

[d] FC models trained on EMBARC stage 1 used variables derived from the CANBIND-1 models.

[e] CT models trained on EMBARC stage 1 used variables derived from the CANBIND-1 models.

[f] Model includes age, sex, Snaith Hamilton Rating Scale score, employment, body mass index, baseline Hamilton Depression Rating Scale and Montgomery Asberg Depression Rating Scale scores, and change in Hamilton Depression Rating Scale and Montgomery Asberg Depression Rating Scale scores at week 2.

[g] No variables survive regularization.

the global FC predictors of response in CANBIND-1 and EMBARC (eFigure 2 in Supplement 1), we found a different pattern of connectivity that did not generalize across trials or within EMBARC. However, we found that lower connectivity of the dACC with dlPFC, medial temporal and parietal areas and higher dACC connectivity with the posterior cingulate (**Figure 1**A and **Figure 2**A) were predictive of response to antidepressants, generalizing across trials (Figure 1B and Figure 2B).

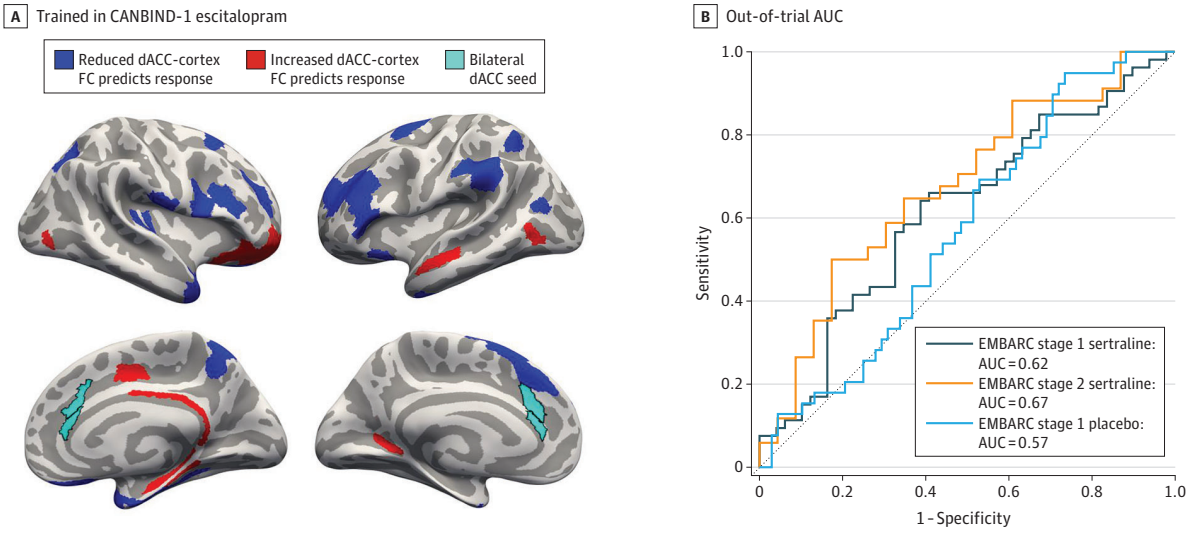
Early-Treatment Models of Response

AUC and balanced accuracy values for early treatment models are summarized in Table 2, and resampling results are shown in eFigure 3 and eFigure 4 in Supplement 1. We found that the early-treatment models performed the best overall, outperforming the clinical plus dACC pretreatment models. The clinical model reached out-of-trial AUCs of 0.66 to 0.73 and balanced accuracy of 66% to 69% when trained on CANBIND-1 and EMBARC data and tested on samples who received SSRIs rather than placebo (eTable in Supplement 1). Although the addition of dACC connectivity features improved the model fit in EMBARC stage 2, this improvement was not significant in bootstrapping analyses (**Figure 3** and eFigure 5 in Supplement 1).

Multivariate Regression Predicting Change in Depression Severity

A PLS-R model predicting change in depression severity in CANBIND-1 explained significantly more variance than expected by chance, whereas a similar PLS-R model explained significantly more variance than expected by chance in EMBARC stage 1. Models trained and tested on the same data showed very high levels of performance (eFigure 6 in Supplement 1); performance decreased when models were trained on one trial and tested on a different trial, with out-of-trial predicted vs observed correlations for SSRI-to-SSRI generalization ranging between 0.31 and 0.39. In sensitivity analyses, reanalyzing the data while applying batch harmonization (using ComBat) within trials reduced out-of-trial performance slightly, but did not alter the overall results (eAppendix 3 and eTable in Supplement 1).

Figure 1. Functional Connectivity (FC) Predictors of Treatment Response in Canadian Biomarker Integration Network in Depression (CANBIND-1) and the Out-of-Trial Generalization Performance



We trained models on CANBIND-1 escitalopram data and then tested them on Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) stage 1 sertraline, EMBARC stage 2 sertraline, and EMBARC stage 1 placebo groups. We show the seed-based dorsal anterior cingulate (dACC) connectivity maps predicting response in CANBIND-1 (A) alongside the respective out-of-trial area under the receiver operator curve (AUC) analyses (B). The dACC seed highlighted in light green (A) was selected on the basis of the overlap between the global FC maps in CANBIND-1 (eFigure 2 in Supplement 1) and prior literature on the ACC involvement in major depression.

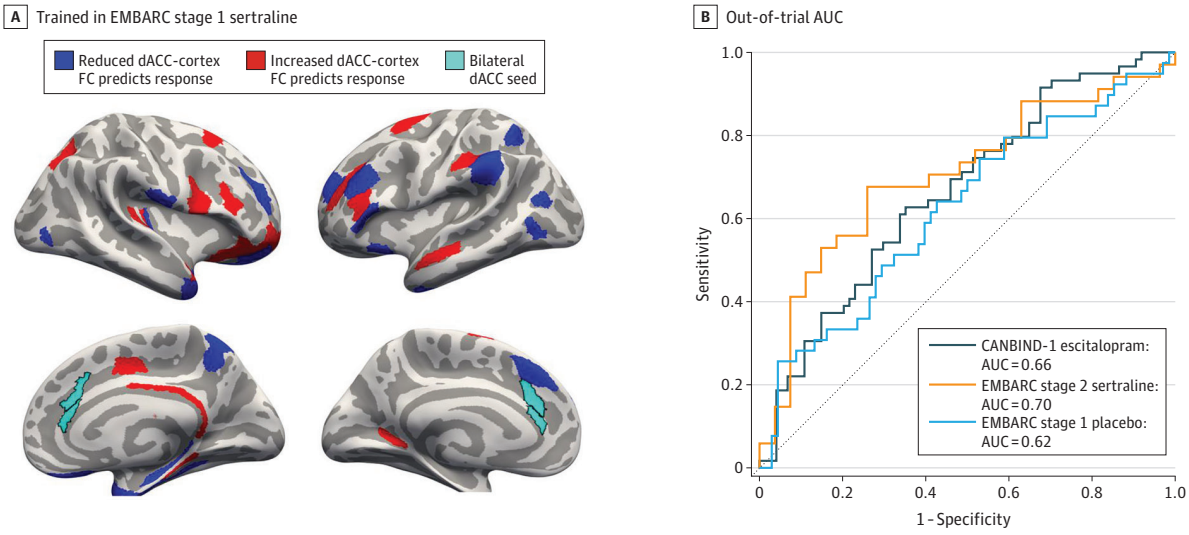
Discussion

The findings of this prognostic study identified promising biological and clinical markers of antidepressant treatment response. Overall cross-trial model prediction performance for SSRIs was encouraging, with baseline models including dACC connectivity features achieving cross-trial balanced accuracy of 63% to 71% in 2 multisite trials from Canada and the US. Importantly, early (2-week) depression severity produced model performance that was on par or better than performance of a combination of baseline clinical and fMRI FC features. Notably, neuroimaging predictors included lower connectivity of the dACC with dlPFC, which is consistent with transcranial magnetic stimulation (rTMS) studies for MDD.

Although our model performance was moderate, it was better than expected based on similar studies of schizophrenia treatment markers.[2] This may be due to shared clinical features of the 2 trials considered here. Testing across trials featuring patient groups who vary widely in levels of severity, chronicity, and age groups[2] may result in lower generalization. Differences in underlying cause likely underpin patient heterogeneity and may require context-dependent models that identify markers in more homogeneous patient populations (eg, treatment-resistant late-life depression vs nonresistant MDD in adults). We also harmonized clinical and neuroimaging data across trials, using the same fMRI processing streams and predictive features, and our results were relatively robust to batch harmonization. Nevertheless, the moderate performance likely reflects MDD heterogeneity.[28-30]

The circuit identified as predictive of treatment response included the anticorrelation between dACC and dlPFC as well as the angular gyrus. The dACC seed included posterior Brodmann area 24 and the anterior Brodmann area 32 prime. It is thus at the intersection between the hot and cold subdivisions of the dACC.[31] A similar connectivity map has been previously identified as predicting treatment outcomes in CANBIND-1,[6] despite differences in fMRI processing and statistical models. In addition, rTMS trials use the most anticorrelated portion of the dlPFC with the subgenual ACC as a target stimulation region to maximize effectiveness,[10,32] implicating this circuit in both pharmacotherapy and rTMS response.[9] These findings fit theories proposing that top-down

Figure 2. Functional Connectivity (FC) Predictors of Treatment Response in Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) and the Out-of-Trial Generalization Performance

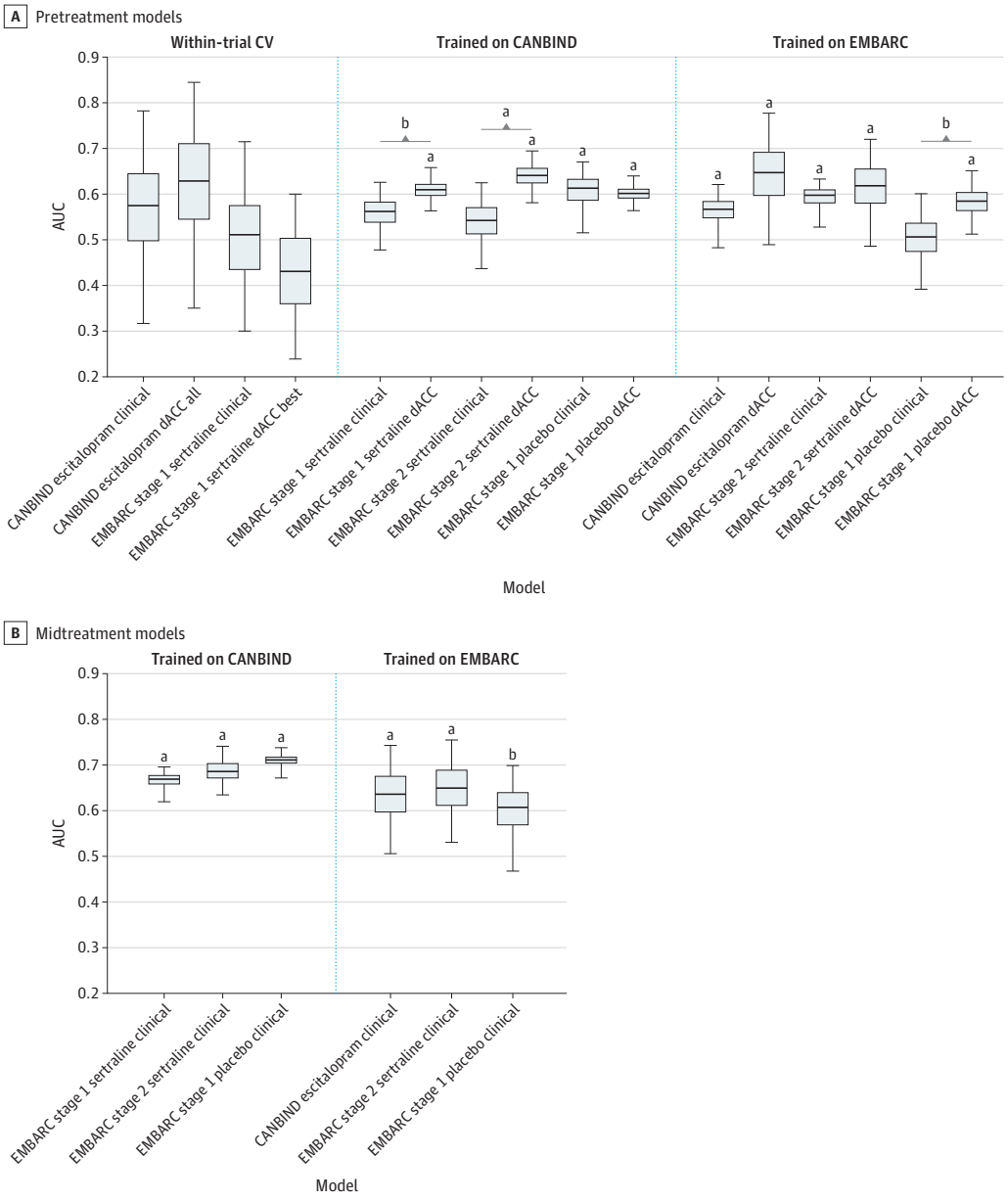


We trained models on the EMBARC stage 1 sertraline sample and tested the resulting models on Canadian Biomarker Integration Network in Depression (CANBIND-1) escitalopram, EMBARC stage 2 sertraline, and EMBARC stage 1 placebo samples. We show the seed-based dorsal anterior cingulate (dACC) connectivity maps (A) predicting response in EMBARC alongside the respective out-of-trial area under the receiver-operator curve (AUC) analyses (B). The dACC seed highlighted in light green (A) was selected on the basis of the overlap between the global FC maps in CANBIND-1 (eFigure 2 in Supplement 1) and prior literature on the ACC involvement in major depression.

emotional regulation is achieved by prefrontal regions regulating the ACC and amygdala activity, regions typically recruited by emotional stimuli.[33-35]

The inclusion of global cortical connectivity and gray matter structure did not improve prediction performance, consistent with previous studies of CANBIND-1 cortical thickness data.[36] These biomarkers may be different in distinct patient populations, however. Finally, early treatment

Figure 3. Area Under the Curve (AUC) for Models Predicting Treatment Response Across Best-Performing Models Derived From Bootstrapping Analyses



Models trained on Canadian Biomarker Integration Network in Depression (CANBIND-1) data were tested on Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) stage 1 sertraline, EMBARC stage 2 sertraline, and EMBARC stage 1 placebo data. Similarly, models trained on EMBARC stage 1 sertraline were tested on all other groups. Error bars represent 95% CIs (2.5%-97.5%), not adjusted for multiple comparisons. We compared dACC models with their respective clinical counterparts. Within-trial model performance was tested using repeated 10-fold cross-validation (CV). For dACC models, within-trial cross-validation was conducted on a full set of predictors; conversely, cross-trial bootstrapping was conducted on a smaller set of predictors (52 clinical and dACC variables) that survived regularization in CANBIND-1. Bootstrapping of the full predictor set can be found in eFigure 4 in Supplement 1.

[a] Denotes models whose performance was significantly higher than chance (1-tailed $P < .05$).

[b] Denotes significant differences between models ($P < .10$ from bootstrapping the differences in AUC).

outcome data from week 2 depression severity scores were very informative as early improvements in depressive symptoms predicted treatment response at the end of a full course of treatment.

In addition to the main analyses of a binary response outcome variable, we also found substantial cross-trial generalizability of multivariate models predicting change in depression severity, with out-of-trial predicted vs observed correlations between 0.31 and 0.39. Overall, although our model performance is moderate, it shows that adding early treatment response information, as well as resting-state biomarkers, improves our ability to predict response after a full course of treatment. Integrating neuroimaging markers with other modalities, such as cognitive and psychological assessments, may further boost prediction performance.[3]

If the FC and clinical markers identified here are replicated, prospective trials with biomarker-guided treatment assignment will be needed to test the markers' utility. Although using neuroimaging and cognitive markers to help assign treatments is promising, this research avenue is not without challenges, limited by the training data and generalizability of models as well as harmonization and similarity between different samples.[37]

Limitations

Our study has some limitations. First, we included only 2 clinical trials, which limited our sample size. Lack of preregistration of the analytic approach is an additional limitation, although our methods follow previously published modeling approaches.[2] Furthermore, data harmonization across trials can be challenging, because no approaches for prospective harmonization of data from new, previously unseen participants in novel biomarker-guided trials exist. Future work should combine data across trials for more robust biomarkers and reveal more mechanistic insights into treatment response. In addition, we focused on SSRIs, and future studies will be needed to develop robust biomarkers for different therapies, including pharmacological medications and rTMS.

Conclusions

In conclusion, the current cross-trial generalization results represent an important step toward biomarkers of antidepressant response. Leveraging data to identify robust biomarkers that generalize across patient populations in different geographic locations will allow us to test such biomarkers in prospective randomized clinical trials and hopefully help connect patients with treatments that work best for them.

ARTICLE INFORMATION

Accepted for Publication: January 17, 2025.

Published: March 20, 2025. doi:10.1001/jamanetworkopen.2025.1310

Corresponding Author: Diego A. Pizzagalli, PhD, Center for Depression, Anxiety and Stress Research, Department of Psychiatry, McLean Hospital, Harvard Medical School, 115 Mill St, Belmont, MA 02478-9106 (dap@mclean.harvard.edu).

Author Affiliations: Center for Depression, Anxiety and Stress Research, Department of Psychiatry, McLean Hospital, Harvard Medical School, Belmont, Massachusetts (Zhukovsky, Pizzagalli); Department of Psychiatry, University of Texas, Southwestern Medical Center, Dallas (Trivedi); Department of Psychiatry, New York State Psychiatric Institute, New York, New York (Weissman); Columbia University Vagelos College of Physicians and Surgeons, New York, New York (Weissman); Department of Psychiatry, Stony Brook University, Stony Brook, New York (Parsey); Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada (Kennedy); Centre for Depression and Suicide Studies, Unity Health Toronto, Toronto, Ontario, Canada (Kennedy).

Author Contributions: Dr Zhukovsky had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: All authors.

Acquisition, analysis, or interpretation of data: Zhukovsky, Trivedi, Parsey, Kennedy, Pizzagalli.

Drafting of the manuscript: Zhukovsky, Kennedy.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Zhukovsky.

Obtained funding: Weissman, Parsey, Pizzagalli.

Administrative, technical, or material support: Kennedy, Pizzagalli.

Supervision: Pizzagalli.

Conflict of Interest Disclosures: Dr Trivedi reported receiving personal fees from Acadia Pharmaceuticals, Alkermes, Alto Neuroscience, Axsome Therapeutics, BasePoint Health Management, Biogen MA, Cerebral, Circular Genomics, Compass Pathfinder Limited, Daiichi Sankyo, GH Research, GreenLight VitalSign6, Heading Health, Janssen Pharmaceutical, Legion Health, Merck Sharp & Dohme, Mind Medicine, Myriad Neuroscience, Naki Health, Neurocrine Biosciences, Noema Pharma AG, Orexo US, Otsuka America Pharmaceutical, Otsuka Europe, Otsuka Pharmaceutical Development & Commercialization, Praxis Precision Medicines, PureTech LYT, Relmada Therapeutics, Sage Therapeutics, Seaport Therapeutics, Signant Health, Sparian Biosciences, Titan Pharmaceuticals, Takeda Pharmaceuticals, and WebMD; grants from the National Institute of Mental Health (NIMH), National Institute on Drug Abuse, National Center for Advancing Translational Sciences, American Foundation for Suicide Prevention, Patient-Centered Outcomes Research Institute, Blue Cross Blue Shield of Texas, Substance Abuse and Mental Health Services Administration, and Department of Defense; and editorial compensation from Elsevier and Oxford University Press outside the submitted work. Dr Kennedy reported receiving grants from Brain Canada, CIHR, Janssen, Lundbeck, Neurocrine, Ontario Brain Institute, Otsuka, and SPOR; and funding for consulting or speaking engagements from Abbvie, Boehringer-Ingelheim, Brain Canada, Janssen, Lundbeck, Otsuka, Pfizer, Sanofi, Sunovion, and Servier outside the submitted work. Dr Pizzagalli reported receiving personal fees from Boehringer Ingelheim, Compass Pathways, Engrail Therapeutics, Neumora Therapeutics, Neurocrine Biosciences, Neuroscience Software, Sage Therapeutics, Alkermes, American Psychological Association, Psychonomic Society, and Springer; grants from Millennium Pharmaceuticals; and stock options from Compass Pathways, Engrail Therapeutics, Neumora Therapeutics, and Neuroscience Software outside the submitted work. No other disclosures were reported.

Funding/Support: The EMBARC study was supported by NIMH grants U01MH092221 and U01MH092250. CANBIND-1 is an Integrated Discovery Program carried out in partnership with, and financial support from, the Ontario Brain Institute, an independent nonprofit corporation, the Brain-CODE platform, and was funded partially by the Ontario government. There is no other funding to report.

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.

Disclaimer: The opinions, results and conclusions are those of the authors and no endorsement by the Ontario Brain Institute is intended or should be inferred.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. 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

2. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science*. 2024;383(6697):164-167. doi:10.1126/science.adg8538

3. Poirot MG, Ruhe HG, Mutsaerts HMM, et al. Treatment response prediction in major depressive disorder using multimodal MRI and clinical data: secondary analysis of a randomized clinical trial. *Am J Psychiatry*. 2024;181(3):223-233. doi:10.1176/appi.ajp.20230206

4. Webb CA, Trivedi MH, Cohen ZD, et al. Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychol Med*. 2019;49(7):1118-1127. doi:10.1017/S0033291718001708

5. Ang YS, Kaiser R, Deckersbach T, et al. Pretreatment reward sensitivity and frontostriatal resting-state functional connectivity are associated with response to bupropion after sertraline nonresponse. *Biol Psychiatry*. 2020;88(8):657-667. doi:10.1016/j.biopsych.2020.04.009

6. van der Wijk G, Harris JK, Hassel S, et al. Baseline functional connectivity in resting state networks associated with depression and remission status after 16 weeks of pharmacotherapy: a CAN-BIND report. *Cereb Cortex*. 2022;32(6):1223-1243. doi:10.1093/cercor/bhab286

7. Harris JK, Hassel S, Davis AD, et al. Predicting escitalopram treatment response from pre-treatment and early response resting state fMRI in a multi-site sample: a CAN-BIND-1 report. *Neuroimage Clin*. 2022;35(July):103120. doi:10.1016/j.nicl.2022.103120

8. Taylor JJ, Kurt HG, Anand A. Resting state functional connectivity biomarkers of treatment response in mood disorders: a review. *Front Psychiatry*. 2021;12(March):565136. doi:10.3389/fpsyt.2021.565136

9. Tura A, Goya-Maldonado R. Brain connectivity in major depressive disorder: a precision component of treatment modalities? *Transl Psychiatry*. 2023;13(1):196. doi:10.1038/s41398-023-02499-y

10. Klooster D, Voetterl H, Baeken C, Arns M. Evaluating robustness of brain stimulation biomarkers for depression: a systematic review of magnetic resonance imaging and electroencephalography studies. *Biol Psychiatry*. 2024;95(6):553-563. doi:10.1016/j.biopsych.2023.09.009

11. Roalf DR, Figee M, Oathes DJ. Elevating the field for applying neuroimaging to individual patients in psychiatry. *Transl Psychiatry*. 2024;14(1):87. doi:10.1038/s41398-024-02781-7

12. Grehl MM, Hameed S, Murrough JW. Brain features of treatment-resistant depression: a review of structural and functional connectivity magnetic resonance imaging studies. *Psychiatr Clin North Am*. 2023;46(2):391-401. doi:10.1016/j.psc.2023.02.009

13. Gerlach AR, Karim HT, Peciña M, et al. MRI predictors of pharmacotherapy response in major depressive disorder. *Neuroimage Clin*. 2022;36(April):103157. doi:10.1016/j.nicl.2022.103157

14. Arnatkeviciute A, Markello RD, Fulcher BD, Misic B, Fornito A. Toward best practices for imaging transcriptomics of the human brain. *Biol Psychiatry*. 2023;93(5):391-404. doi:10.1016/j.biopsych.2022.10.016

15. Arnatkeviciute A, Fulcher BD, Bellgrove MA, Fornito A. Imaging transcriptomics of brain disorders. *Biol Psychiatry Glob Open Sci*. 2021;2(4):319-331. doi:10.1016/j.bpsgos.2021.10.002

16. Trivedi MH, McGrath PJ, Fava M, et al. Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): rationale and design. *J Psychiatr Res*. 2016;78:11-23. doi:10.1016/j.jpsychires.2016.03.001

17. Lam RW, Milev R, Rotzinger S, et al; CAN-BIND Investigator Team. Discovering biomarkers for antidepressant response: protocol from the Canadian biomarker integration network in depression (CAN-BIND) and clinical characteristics of the first patient cohort. *BMC Psychiatry*. 2016;16(1):105. doi:10.1186/s12888-016-0785-x

18. Kennedy SH, Lam RW, Rotzinger S, et al; CAN-BIND Investigator Team. Symptomatic and functional outcomes and early prediction of response to escitalopram monotherapy and sequential adjunctive aripiprazole therapy in patients with major depressive disorder: a CAN-BIND-1 Report. *J Clin Psychiatry*. 2019;80(2):18m12202. doi:10.4088/JCP.18m12202

19. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979;134:382-389. doi:10.1192/bjp.134.4.382

20. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967;6(4):278-296. doi:10.1111/j.2044-8260.1967.tb00530.x

21. Snaith RP, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P. A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. *Br J Psychiatry*. 1995;167(1):99-103. doi:10.1192/bjp.167.1.99

22. Carmody TJ, Rush AJ, Bernstein I, et al. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol*. 2006;16(8):601-611. doi:10.1016/j.euroneuro.2006.04.008

23. Glasser MF, Coalson TS, Robinson EC, et al. A multi-modal parcellation of human cerebral cortex. *Nature*. 2016;536(7615):171-178. doi:10.1038/nature18933

24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

25. Combrisson E, Jerbi K. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods*. 2015;250:126-136. doi:10.1016/j.jneumeth.2015.01.010

26. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037

27. Zhukovsky P. MDD response prediction. Accessed February 7, 2025. https://github.com/peterzhukovsky/MDD_response_prediction

28. Zhukovsky P, Anderson JAE, Coughlan G, Mulsant BH, Cipriani A, Voineskos AN. Coordinate-based network mapping of brain structure in major depressive disorder in younger and older adults: a systematic review and meta-analysis. *Am J Psychiatry*. 2021;178(12):1119-1128. doi:10.1176/appi.ajp.2021.21010088

29. Buch AM, Liston C. Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics. *Neuropsychopharmacology*. 2021;46(1):156-175. doi:10.1038/s41386-020-00789-3

30. Marawi T, Zhukovsky P, Rashidi-Ranjbar N, et al; PACt-MD Study Group. Brain-cognition associations in older patients with remitted major depressive disorder or mild cognitive impairment: a multivariate analysis of gray and white matter integrity. *Biol Psychiatry*. 2023;94(12):913-923. doi:10.1016/j.biopsych.2023.05.018

31. Pizzagalli DA. Frontocingulate dysfunction in depression: toward biomarkers of treatment response. *Neuropsychopharmacology*. 2011;36(1):183-206. doi:10.1038/npp.2010.166

32. Cole EJ, Stimpson KH, Bentzley BS, et al. Stanford accelerated intelligent neuromodulation therapy for treatment-resistant depression. *Am J Psychiatry*. 2020;177(8):716-726. doi:10.1176/appi.ajp.2019.19070720

33. Park C, Rosenblat JD, Lee Y, et al. The neural systems of emotion regulation and abnormalities in major depressive disorder. *Behav Brain Res*. 2019;367(April):181-188. doi:10.1016/j.bbr.2019.04.002

34. Kober H, Ochsner KN. Regulation of emotion in major depressive disorder. *Biol Psychiatry*. 2011;70(10):910-911. doi:10.1016/j.biopsych.2011.09.019

35. Ochsner KN, Ray RR, Hughes B, et al. Bottom-up and top-down processes in emotion generation: common and distinct neural mechanisms. *Psychol Sci*. 2009;20(11):1322-1331. doi:10.1111/j.1467-9280.2009.02459.x

36. Suh JS, Minuzzi L, Raamana PR, et al. An investigation of cortical thickness and antidepressant response in major depressive disorder: a CAN-BIND study report. *Neuroimage Clin*. 2020;25:102178. doi:10.1016/j.nicl.2020.102178

37. Kelley ME, Choi KS, Rajendra JK, et al. Establishing evidence for clinical utility of a neuroimaging biomarker in major depressive disorder: prospective testing and implementation challenges. *Biol Psychiatry*. 2021;90(4):236-242. doi:10.1016/j.biopsych.2021.02.966

SUPPLEMENT 1.

eAppendix 1. Demographic and clinical information
eFigure 1. Overview of the participant flow in CANBIND and EMBARC
eAppendix 2. Supplemental methods
eAppendix 3. Supplemental results
eFigure 2. Global functional connectivity (Global FC) predictors of treatment response in CANBIND-1 and EMBARC and the out-of-trial generalization performance
eFigure 3. Receiver operating characteristic curves (A) and area-under the curve (AUC) histogram (B) for predicting response to sertraline and placebo in EMBARC after repeated 10-fold cross-validation training on CANBIND data for the clinical+dACC model
eTable. Summary of out-of-trial model performance for models trained in the CANBIND and EMBARC clinical trials following ComBat batch harmonization
eFigure 4. Area under the curve (AUC) for models predicting treatment response in clinical+dACC models with all 366 features derived from bootstrapping analyses
eFigure 5. Difference in areas under the curve (ΔAUC) for dACC and clinical models predicting treatment response derived from bootstrapping analyses
eFigure 6. Scatterplots of predicted vs observed change in HDRS-17 depression severity scores between the last clinical assessment and baseline
eReferences

SUPPLEMENT 2.

Data Sharing Statement

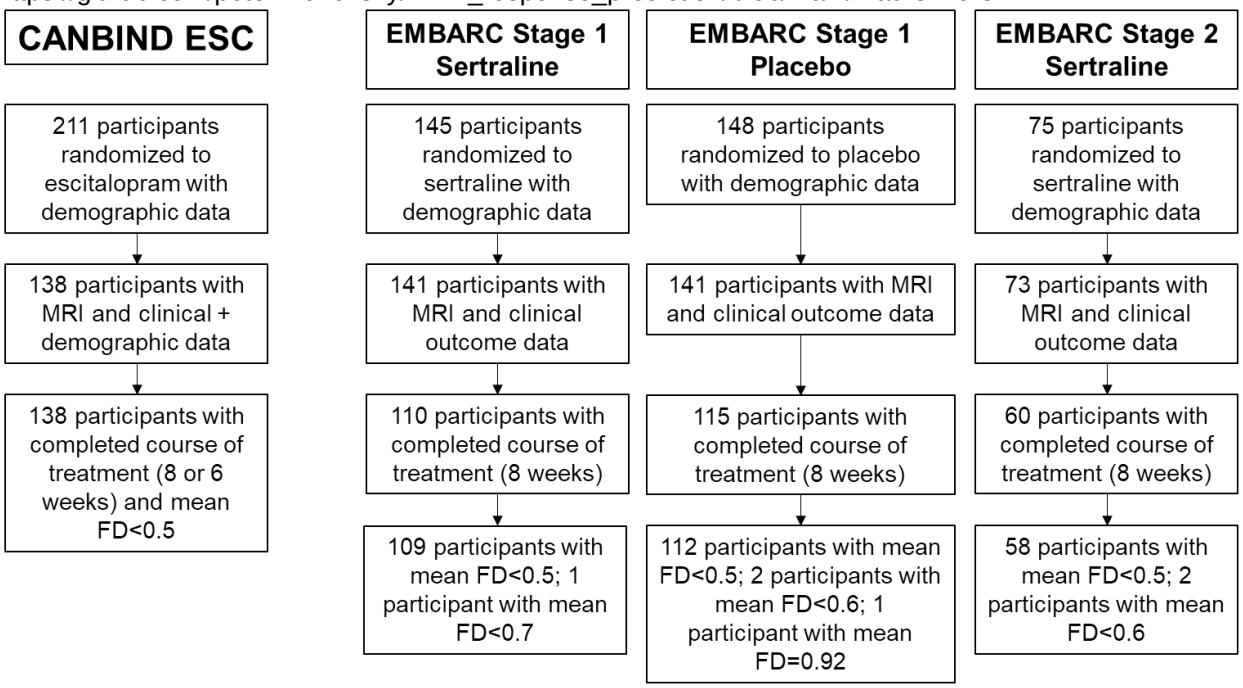
Supplemental Online Content

Zhukovsky P, Trivedi MH, Weissman M, Parsey R, Kennedy S, Pizzagalli DA. Generalizability of treatment outcome prediction across antidepressant treatment trials in depression. *JAMA Netw Open*. 2025;8(3):e251310. doi:10.1001/jamanetworkopen.2025.1310

This supplemental material has been provided by the authors to give readers additional information about their work.

eAppendix 1. Demographic and clinical information

We provide more details on the demographics of the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) and the Canadian Biomarker Integration Network in Depression (CANBIND) samples in Table 1. An overview of the participant flow is shown in Supplementary Figure 1. Employment status was determined as full or part-time employment in EMBARC; and 'working now' in CANBIND. The Body Mass Index (BMI) data in EMBARC included some extreme outliers, hence we excluded BMI data for those with BMI<13 and BMI>55. We used linear regression to impute BMI from waist circumference for those participants with available data. Pre-treatment depression severity was measured using the Montgomery Asberg Depression Rating Scale (MADRS)[1] in CANBIND and Hamilton Depression Rating Scale (HDRS) [2] in EMBARC. We converted MADRS scores to HDRS-17 scores as part of data harmonization using a previously described approach[3] (see also https://mood-disorders.co.uk/ASSETS/FILES/QIDS-MADRS-HRSD-conversion-table-pdf.pdf). The conversion approach used a previously published table mapping between HDRS-17 and MADRS scores. This approach has been extensively validated using item response theory analyses[3]. While the mapping is mostly linear, sometimes a range of scores from one questionnaire corresponds to one score on the other questionnaire. For instance, a MADRS score of 8 or 9 corresponds to an HDRS-17 score of 7 while a MADRS score of 10 corresponds to a HDRS-17 score of 8. We provide the code used for this mapping in a public github repository:
https://github.com/peterzhukovsky/MDD_response_prediction/blob/main/madrs2hdrs17.m



eFigure 1. Overview of the participant flow in CANBIND and EMBARC. MRI data refer to baseline MRI only.

eAppendix 2. Supplemental Methods

Participants. EMBARC included 296 outpatients 18–65 years of age recruited from four academic sites in the US (Columbia University, Massachusetts General Hospital, University of Michigan, UT Southwestern Medical Center) who were diagnosed as having recurrent or chronic MDD and were not taking medication for MDD; participants were not required to be medication naive. They took part in an MRI session that included both resting state fMRI and structural MRI imaging. To test our hypotheses, we defined three population subgroups, with the first group being treated with sertraline in Stage 1 (n=110 with complete data), the second group being treated with sertraline in Stage 2 after not responding to placebo in Stage 1 (i.e., a different set of n=64) and the third group receiving placebo in Stage 1 (n=115). CANBIND included 144 MDD patients 18-61 years of age recruited from six sites in Canada (Toronto General Hospital (TGH); CAMH (CAM); McMaster University (MCU);University of Calgary (UCA); University of British Columbia (UBC); Queens University (QNS)) who completed a resting state fMRI and structural MRI scan before starting escitalopram treatment. Among those participants, 138 had completed at least six weeks of escitalopram treatment and had complete baseline prediction data.

Clinical data. Participants provided information on their age, sex, employment status (binarized as unemployed vs. partially or fully employed here) and overlapping clinical data were available on baseline depression severity (MADRS [1] in CANBIND and 17-item HDRS [2] in EMBARC) and anhedonia (Snaith Hamilton Rating Scale (SHAPS) [4]). BMI scores were also available in both datasets. BMI outlier scores (<15 or >55) were removed in EMBARC; we imputed BMI using regression of waist circumference scores for participants who had those data available. No outliers were present in CANBIND. Finally, to harmonize across prediction models, we converted MADRS to HRSD scores in CANBIND [3].

MRI Data. We preprocessed structural and resting-state fMRI data in EMBARC using fMRIPrep v22.1.1 and in CANBIND using fMRIPrep v23.0.2[5]. For denoising, we regressed out 24 fixed confounds: 6 motion parameters, average signal from the white matter and cerebrospinal fluid and their first and second-order temporal derivatives. Next, we applied a 6-mm smoothing kernel. We registered fMRIPrep outputs from the MNI152 (Nlin6) space to the FreeSurfer fsaverage space (mri_vol2surf) and extracted fMRI timeseries for 360 cortical regions in the Human Connectome Project (HCP) parcellation [6]. We then calculated global cortical FC measures by averaging the rows of a 360 x 360 connectivity matrix, excluding the values along the diagonal. In addition to the global cortical FC measures, due to a priori hypotheses, we also calculated seed-based FC of the dorsal anterior cingulate (dACC) and rostral anterior cingulate (rACC), respectively. Seeds were selected based on the bilateral dACC and rACC regions that survived regularization in predicting treatment response in CANBIND. The dACC seed comprised bilateral p24 and a32pr regions, while the rACC seed comprised bilateral a24 and p32 regions from the HCP parcellation (Figure 1B, 1D). Structural data were processed using FreeSurfer as part of the fMRIPrep pipeline, producing cortical thickness outputs in the aparc parcellation[7], and subcortical volumes in the aseg parcellation. Subcortical volumes were divided by the total intracranial volumes to provide normalized values.

While EMBARC also includes arterial spin labeling data [8,9] and both EMBARC [10,11] and CANBIND [12,13] have task-based fMRI with different tasks (e.g., emotion conflict monitoring in EMBARC and Go/NoGo and incentive delay in CANBIND), we focus on imaging data that are common across both EMBARC and CANBIND, namely structural MRI and resting-state functional connectivity. Future studies should also examine other imaging modalities such (e.g. ASL) as candidate biomarkers of treatment response.

Predictive modeling approach. We used elastic net logistic regressions with regularization (*lassoglm*, Matlab R2022a [16]) to predict treatment outcomes in the four datasets. While various machine learning approaches with non-linear prediction exist, recent evidence from large-scale brain-behavior studies suggests that linear models perform on par with some of the more complex prediction models [17]. Given the

moderate size of the imaging datasets analyzed here, and following recent studies of treatment outcome prediction [18], we use elastic net logistic models aiming to maintain a higher feature-to-observation ratio [19]. Regularization also allows us to identify a smaller, most salient set of predictors that could be tested in future studies. There were five sets of models tested using baseline depression severity: (1) a clinical model including age, sex, employment, baseline HDRS, SHAPS, and BMI; (2) a clinical + global FC model that included all features from the clinical model and added 360 global FC features; (3) a clinical + dACC FC model that added 360 seed-based dACC features to the clinical model; (4) a clinical + rACC FC model that added 360 seed-based rACC features to the clinical model; and (5) a clinical + cortical thickness (CT) model that added 74 gray matter features to the clinical model. We then tested five analogous models that included week 2 instead of baseline depression severity scores. We provide an overview of the models below:

(1) response ~ age + sex + employment + baseline HDRS + SHAPS + BMI
(2) response ~ age + sex + employment + baseline HDRS + SHAPS + BMI + global FC
(3) response ~ age + sex + employment + baseline HDRS + SHAPS + BMI + dACC FC
(4) response ~ age + sex + employment + baseline HDRS + SHAPS + BMI + rACC FC
(5) response ~ age + sex + employment + baseline HDRS + SHAPS + BMI + brain structure

First, we tested the prediction performance of models within CANBIND and EMBARC Stage 1 by creating 100 random training and test data splits, training the elastic net model with 10-fold cross-validation on the training dataset, and predicting outcomes in the test dataset on each iteration. Second, we evaluated the prediction performance of models trained on CANBIND and tested in EMBARC Stage 1, EMBARC Stage 2, and EMBARC placebo. Training models used 10-fold cross-validation to optimize regression weights. This process resulted in point-estimates of AUC and balanced accuracy for out-of-trial prediction performance, reported in Table 2. Balanced accuracy was calculated as the mean of sensitivity and specificity [18].

We next used bootstrapping (bootfun in Matlab) to assess whether AUCs were significantly higher than chance and to compare the most promising models with each other. We conducted two sets of bootstrapping analyses: first, across 1,000 bootstraps, we sampled from CANBIND and tested the out-of-trial performance of models with clinical features only and those with clinical and dACC FC features that survive regularization in the EMBARC datasets. Second, we sampled from EMBARC Stage 1 and tested the out-of-sample performance of models with clinical features only and those with clinical and dACC FC features that survive regularization (n=1,000 iterations) in CANBIND and in EMBARC Stage 2 and EMBARC placebo data. A limitation of this approach is that while the models are trained on EMBARC Stage 1 data and tested on CANBIND and EMBARC Stage 2 data, they only include features that survive regularization in original CANBIND training. While this process makes feature selection circular, feature tuning or coefficient selection is not. Models were deemed to perform significantly higher than chance by comparing the AUCs from the bootstrap distribution to AUC=0.5. We calculated the p-values by dividing the number of bootstraps with AUC below 0.5 by 1,000 with a one-tailed p-value threshold of 0.05 for significance (50 out of 1,000 iterations). When comparing between models, we calculated the difference in AUC for pairs of models (i.e., clinical models vs. clinical and dACC-to-cortex FC models) on each bootstrap and calculated the p-values by dividing the number of bootstraps with ΔAUC<0 by 1,000, with a one-tailed p-value threshold of 0.05. We removed models leaving no variables after regularization, primarily among clinical-only models, affecting up to 10% of bootstraps. In those cases, following previous studies [18], we did not include the AUC in our plots and analyses. In our elastic net models, we used the hyperparameter α=0.01 for models trained on CANBIND data and α=0.001 for models trained on EMBARC data. One-tailed p-value thresholds were chosen given that we tested whether models performed significantly better than chance and whether more complex models including connectivity features performed better than clinical models.

Hyperparameter optimization. The elastic net models used the default hyperparameter optimization options (*lassoglm.m*). We used the following alpha hyperparameters: α=0.3 for clinical+fMRI models trained in CANBIND; α=0.1 for structural MRI models trained in CANBIND; and α=0.001 for models trained on clinical models or models trained on EMBARC data.

We tested a range of α thresholds when training the model within the CANBIND or EMBARC data α=[0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6], and selected the above values as they produced a reasonable set of predictors at the training stage. Models with higher α criteria resulted in more conservative regularization and often did not return any predictors surviving regularization. Conversely, models with very low α criteria returned regression weights for nearly every predictor variable. Since we expected only some imaging features to be contributing to predictions, we selected α criteria that would prune over 50% of the functional connectivity features or brain structure features. We also reran the models with the best performing α values several times and selected the α value for which the model returned the nearly same number of features every time.

When reporting balanced accuracy values, we preferentially select balanced accuracy values based on both sensitivity and specificity being larger then 0.55; if that was not possible, we selected balanced accuracy values based on both sensitivity and specificity being larger then 0.50. Sometimes, higher balanced accuracy is possible at the trade-off cost of having either low sensitivity and high specificity or vice versa; however, we were aiming to balance both sensitivity and specificity performance.

Data harmonization. We repeated the analyses described above twice. First, we report findings without batch harmonization. Current batch harmonization tools require full datasets to estimate site- and confound-specific biases, posing the challenge of prospective harmonization to a new patient from a new site. To test generalizability in context of potential clinical trials, we wanted to test model performance with the currently available tools. However, we also report results of predicting modelling after batch harmonization in ComBat (see Supplementary Section 3.2).

Sensitivity analyses. First, we reanalyzed the data while correcting for batch effects in the rs-fMRI data and the gray matter brain structure using ComBat [20] https://github.com/Jfortin1/ComBatHarmonization/ in the Supplementary Information.

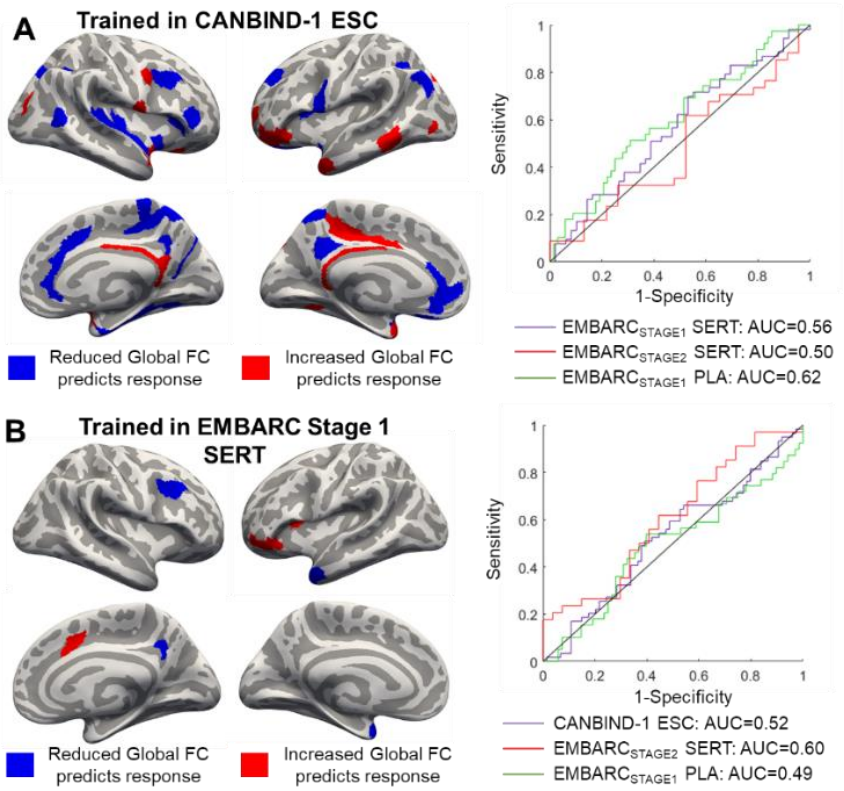
To test the robustness of our findings we conducted several sensitivity analyses. First, while we report in the main results the bootstrapped performance of the most parsimonious model featuring 52 predictors, we also bootstrapped elastic net models with the full set of 366 predictors. This analysis allowed the elastic net models to prune any predictors via regularization on each bootstrap iteration in addition to fitting regression weights to a specific set of predictors. We found a similar bootstrapping performance in this analysis to that of the 52-predictor model (Supplementary Section 3.3).

Multivariate regression predicting change in depression severity. In this secondary analysis, we aimed to test model performance predicting change in HDRS-17 scores in EMBARC and the MADRS scores transformed to HDRS-17 scores in CANBIND. Absolute difference between 8-week and baseline was used as an outcome. Whenever 8-week outcomes were not available, 6-week scores were used in CANBIND. In EMBARC Stage 2, 12-week scores were available while in EMBARC Stage 1, 6-week scores were available. Given that a 4-week course of treatment is short, and we aimed for consistency within each trial we only include EMBARC data with a full 8-week course of treatment. Partial least squares models with 360 dACC connectivity features, age, sex, employment, baseline HDRS-17, SHAPS, and BMI were run. These were the same predictors as those used in the main analyses. We used permutation testing (n=5,000) and bootstrapping (n=5,000, $Z > 3$ and $Z < -3$) to assess model significance and to identify robust features. We trained the model on CANBIND and then applied the regression weights from the resulting model to predict change in depression severity in EMBARC; similarly, we trained the model on EMBARC

Stage 1 sertraline and applied the regression weights to predict change in depression scores in the other datasets.

eAppendix 3. Supplemental Results

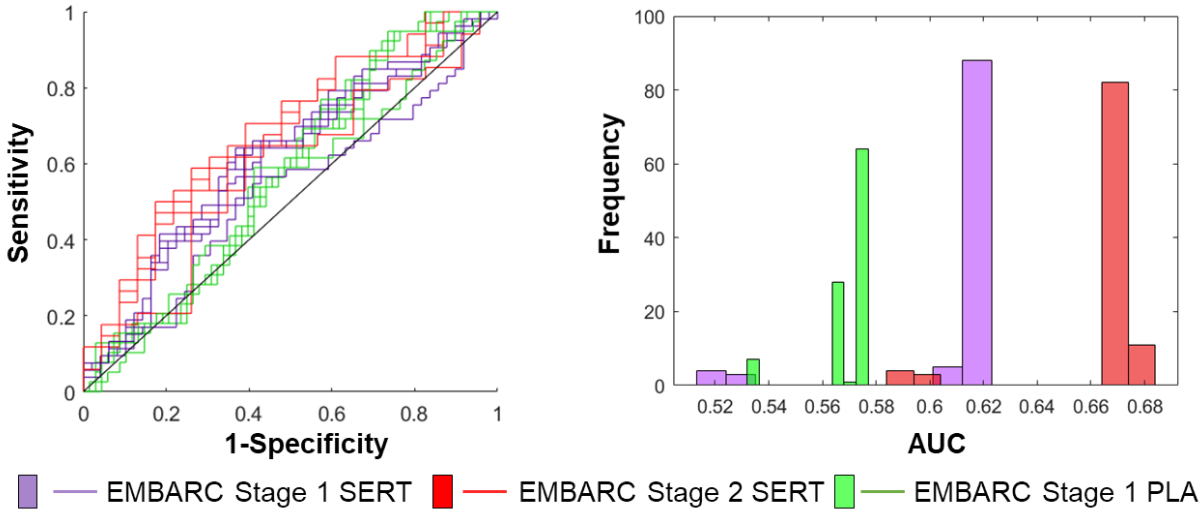
3.1. Global functional connectivity maps predicting treatment response



eFigure 2. Global functional connectivity (Global FC) predictors of treatment response in CANBIND-1 and EMBARC and the out-of-trial generalization performance. We first trained models on CANBIND-1 escitalopram data **(A)** and then tested them on EMBARC Stage 1 sertraline, EMBARC Stage 2 sertraline, and EMBARC Stage 1 placebo groups. We then trained models on the EMBARC Stage 1 sertraline sample **(B)** and tested the resulting models on CANBIND-1 escitalopram, EMBARC Stage 2 sertraline and EMBARC Stage 1 placebo samples. We show the global FC maps predicting response in CANBIND-1 **(A)** and EMBARC **(B)** alongside the respective out-of-trial receiver-operator curve (ROC-AUC) analyses. ESC: escitalopram, FC: functional connectivity, SERT: sertraline, AUC: area under the curve.

3.2. Model stability

We evaluated the stability of models following repeated training. Training the regularized elastic net models involved random data splits, which may produce slightly different models on each iteration, and we wanted to test whether model performance for the clinical+dACC (dorsal Anterior Cingulate Cortex) varied depending on the training split. We found the best performing pre-treatment model (including dACC connectivity features) to be very stable (Supplementary Figure S3).



eFigure 3. Receiver operating characteristic curves **(A)** and area-under the curve (AUC) histogram **(B)** for predicting response to sertraline and placebo in EMBARC after repeated 10-fold cross-validation training on CANBIND data for the clinical+dACC model. Models trained on CANBIND data were tested on EMBARC Stage 1 sertraline, EMBARC Stage 2 sertraline and EMBARC Stage 1 placebo data. We plot receiver-operator curves after running the same model 100 times. On every iteration, the model was trained using a different 10-fold data split, with regularization leading to slightly different model coefficients. Regularization hyperparameter was set at α=0.3 for this fMRI model. Overall, the AUCs are very similar across the iterations. SERT: sertraline; PLA: Placebo

3.3. Model performance following batch harmonization using ComBat

We applied ComBat batch harmonization to adjust for age, sex and scanning site within each trial. We then repeated the main out-of-trial prediction analyses from the main text. We found the results of this re-analysis (shown in Supplementary Table 2) to be largely consistent with the findings reported in the Table 1 of the main text. The overall out-of-trial model performance was similar, and the addition of dACC connectivity features also improved model performance in this analysis.

	Models trained on CANBIND ESC						Models trained on EMBARC Stage 1 SERT					
Pre-treatment models of response												
Tested on:	**EMBARC Stage 1 SERT**		**EMBARC Stage 2 SERT**		**EMBARC Stage 1 PLA**		**CANBIND Escitalopram**		**EMBARC Stage 2 SERT**		**EMBARC Stage 1 PLA**	
	AUC	**bACC**	**AUC**	**bACC**	**AUC**	**bACC**	**AUC**	**bACC**	**AUC**	**bACC**	**AUC**	**bACC**
Clinical Model*	0.58	**0.61**	0.58	**0.60**	0.63	**0.59**	0.59	**0.59**	0.61	**0.60**	0.52	0.55
Clinical + Global FC#	0.55	0.57	0.49	0.52	0.63	**0.63**	0.45	0.46	0.60	0.55	0.49	0.49
Clinical + dACC FC#	0.61	**0.61**	0.64	**0.62**	0.55	0.56	0.71	**0.68**	0.67	**0.66**	0.61	**0.63**
Clinical + rACC FC#	0.61	**0.64**	0.66	**0.67**	0.66	**0.62**	0.44	0.51	0.49	0.51	0.44	0.47
Clinical + CT^	0.60	0.57	0.63	0.56	0.64	**0.61**	0.60	**0.63**	0.65	**0.68**	0.51	0.56
***Age, Sex, SHAPS, Employment, BMI, baseline HDRS/MADRS**												
Early treatment models of response							**Early treatment models of response**					
Clinical Model*	0.68	**0.69**	0.73	**0.66**	0.71	**0.66**	0.66	**0.69**	0.68	**0.66**	0.63	**0.66**
Clinical + Global FC#	0.64	**0.67**	0.63	**0.60**	0.70	**0.65**	no variables survive regularization					
Clinical + dACC FC#	0.67	**0.62**	0.78	**0.74**	0.69	**0.65**	0.67	**0.62**	0.74	**0.74**	0.64	**0.65**
Clinical + rACC FC#	0.66	**0.64**	0.77	**0.73**	0.74	**0.68**	no variables survive regularization					
Clinical + CT^	0.66	**0.63**	0.80	**0.77**	0.72	**0.65**	0.56	0.56	0.68	**0.63**	0.55	0.56

***Age, Sex, SHAPS, Employment, BMI, baseline HDRS/MADRS, change in HDRS/MADRS at week 2**

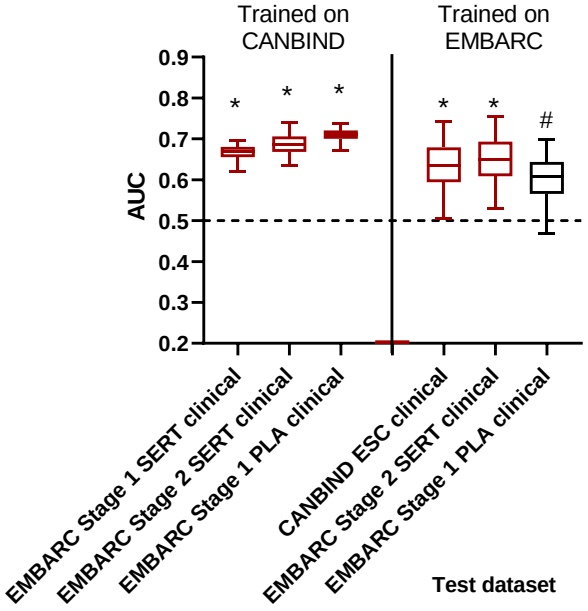
#FC models trained on EMBARC1 used variables derived from the CANBIND models

^CT models trained on EMBARC1 used variables derived from the CANBIND models

eTable. Summary of out-of-trial model performance for models trained in the CANBIND and EMBARC clinical trials following ComBat batch harmonization. AUC: area under the curve; bACC: balanced accuracy. Balanced accuracy values highlighted in bold were significantly higher than chance ($p<0.05$, not correcting for the number of models) based on prior simulations [33].

3.4. Bootstrapping the full set of dACC predictors

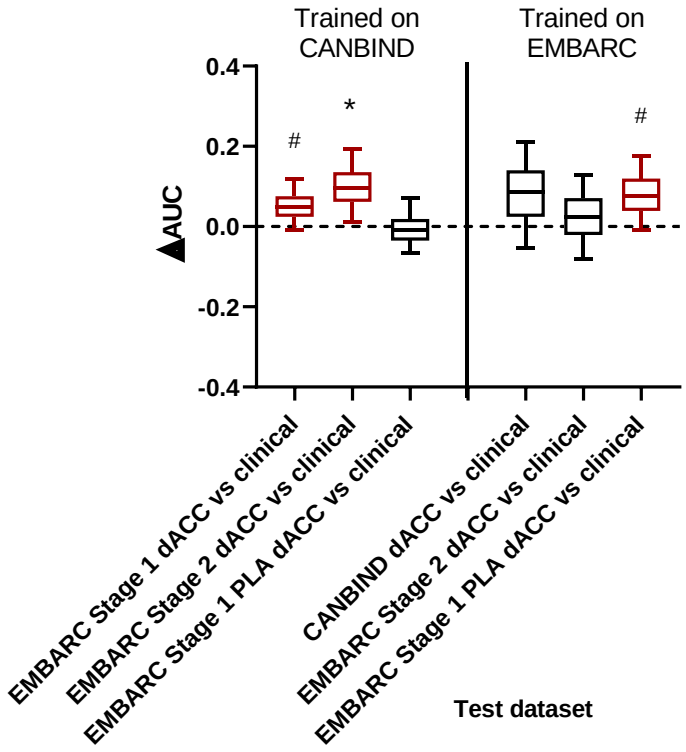
We re-ran the bootstrapping of the out-of-trial model performance in CANBIND using all potential 366 predictors, which allowed us to apply regularization (α=0.3) and fit regression weights for the surviving predictors on each bootstrap iteration. We show the results of this re-analysis in Supplementary Figure S4. Overall, we found bootstrapping performance in this analysis to be similar to the performance of the 52-predictor model reported in the main analyses.



eFigure 4. Area under the curve (AUC) for models predicting treatment response in clinical+dACC models with all 366 features derived from bootstrapping analyses. Models trained on CANBIND data were tested on EMBARC Stage 1 sertraline (EMBARC1), EMBARC Stage 2 sertraline (EMBARC2) and EMBARC Stage 1 placebo data (EMBARC1 PLA). Error bars represent 95% confidence intervals [2.5%-97.5%], not adjusted for multiple comparisons. Boxplots of models whose performance was significantly higher than chance (one-tailed *p<0.05, #p<0.1) are highlighted in red. We compared dACC models with their respective clinical counterparts, with significant differences between models highlighted with a bar (one-tailed *p<0.05; #p<0.1 from bootstrapping the differences in AUC).

3.5. Model comparison

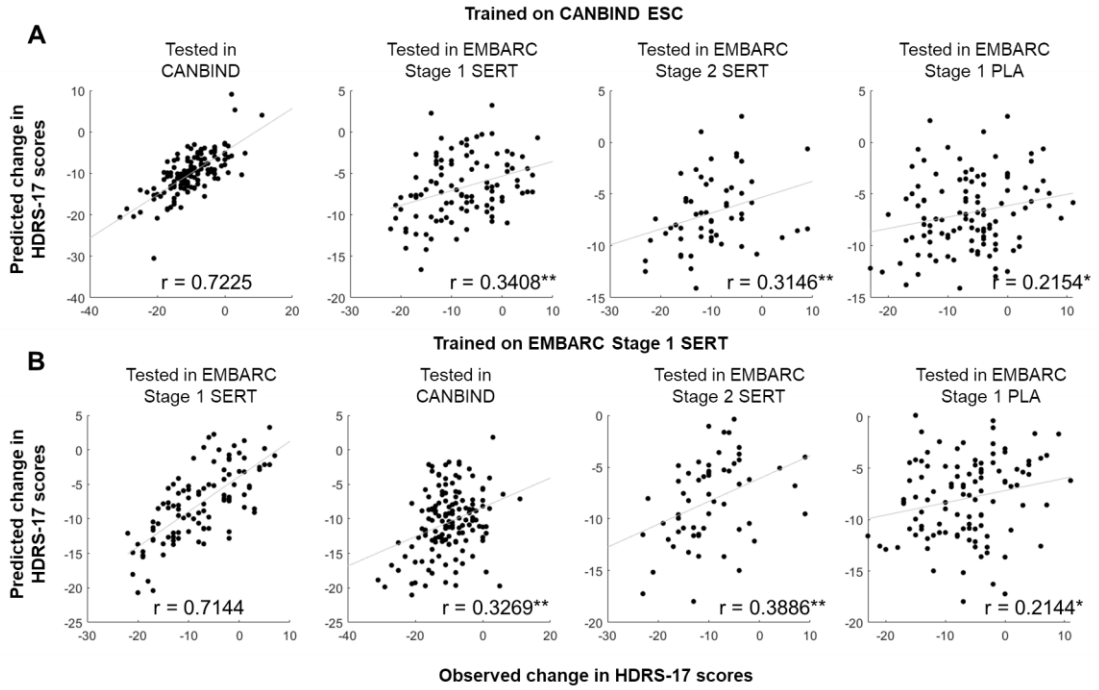
The model performance was improved by the addition of dACC features when the models were trained on CANBIND and tested on EMBARC Stage 1 SERT (p=0.0824) and on EMBARC Stage 2 SERT (p=0.033); model performance was also improved by the addition of dACC features when the models were trained on EMBARC Stage 1 SERT and tested on EMBARC Stage 1 PLA (p=0.068) as shown in Supplementary Figure S5.



eFigure 5. Difference in areas under the curve (ΔAUC) for dACC and clinical models predicting treatment response derived from bootstrapping analyses. Models trained on CANBIND data were tested on EMBARC Stage 1 sertraline, EMBARC Stage 2 sertraline and EMBARC Stage 1 placebo data. Similarly, models trained on EMBARC Stage 1 sertraline were tested on all other groups. dACC models included 52 features in total. Error bars represent 95% confidence intervals [2.5%-97.5%], not adjusted for multiple comparisons. Boxplots of dACC models that performed better than the clinical models across bootstraps (one-tailed *p<0.05; #p<0.1) are highlighted in red.

3.6. Partial Least Squares Regression predicting change in depression severity

We found the model predicting change in depression severity in CANBIND to explain significantly more variance than expected by chance (permutation p=0.004), with various dACC functional connectivity features and baseline depression severity crossing the Z > 3 or Z < -3 threshold after bootstrapping. Similarly, we found the model predicting change in depression severity in EMBARC Stage 1 to explain significantly more variance than expected by chance (permutation p=0.0298), with various dACC functional connectivity features and baseline depression severity crossing the Z > 3 or Z < -3 threshold after bootstrapping. Similar to the main analyses predicting treatment response, we found that models trained and tested on the same data showed very high levels of performance. However, performance decreased when models were trained on one trial and tested on a different trial, with out of trial predicted vs observed correlations for SSRI-to-SSRI generalization ranging between 0.3 and 0.4. Models trained on SSRI data and used to predict changes in depression severity to placebo showed predicted vs observed correlations of approximately 0.2. We show all predicted vs. observed correlations in Supplementary Figure S6.



eFigure 6. Scatterplots of predicted vs observed change in HDRS-17 depression severity scores between the last clinical assessment and baseline. Partial least squares regression models were first trained on CANBIND data and tested within CANBIND, on EMBARC Stage 1 sertraline, Stage 2 sertraline and Stage 1 placebo (A). Similarly, models were trained on EMBAR Stage 1 sertraline data and tested within EMBARC Stage 1 sertraline, on CANBIND, on Stage 2 sertraline and Stage 1 placebo (B). Pearson's correlations (r) for observed vs predicted values are shown. Uncorrected p<0.05*; uncorrected p<0.01**.

eReferences

1. Montgomery A, Asberg M. A New Depression Scale Designed to be Sensitive to Change. *Br J Psychiatry*. 1979;134:382-389.

2. Hamilton M. Development of a rating scale for depressive illness. *Br J Soc Clin Psychol*. 1967;6:278-296. doi:10.1159/000395073

3. Carmody TJ, Rush AJ, Bernstein I, et al. The Montgomery Äsberg and the Hamilton ratings of depression: A comparison of measures. *Eur Neuropsychopharmacol*. 2006;16(8):601-611. doi:10.1016/j.euroneuro.2006.04.008

4. Snaith RP, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P. A scale for the assessment of hedonic tone. The Snaith-Hamilton Pleasure Scale. *Br J Psychiatry*. 1995;167(JULY):99-103. doi:10.1192/bjp.167.1.99

5. Esteban O, Markiewicz CJ, Blair RW, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods*. 2019;16(1):111-116. doi:10.1038/s41592-018-0235-4

6. Glasser MF, Coalson TS, Robinson EC, et al. A multi-modal parcellation of human cerebral cortex. *Nature*. 2016;536(7615):171-178. doi:10.1038/nature18933

7. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 2006;31(3):968-980. doi:10.1016/j.neuroimage.2006.01.021

8. Poirot MG, Ruhe HG, Mutsaerts HJMM, et al. Treatment Response Prediction in Major Depressive Disorder Using Multimodal MRI and Clinical Data: Secondary Analysis of a Randomized Clinical Trial. *Am J Psychiatry*. 2024;181(3):223-233. doi:10.1176/appi.ajp.20230206

9. Cooper CM, Chin Fatt CR, Jha M, et al. Cerebral Blood Perfusion Predicts Response to Sertraline versus Placebo for Major Depressive Disorder in the EMBARC Trial. *EClinicalMedicine*. 2019;10(April):32-41. doi:10.1016/j.eclinm.2019.04.007

10. Nguyen KP, Chin Fatt C, Treacher A, et al. Patterns of Pretreatment Reward Task Brain Activation Predict Individual Antidepressant Response: Key Results From the EMBARC Randomized Clinical Trial. *Biol Psychiatry*. 2022;91(6):550-560. doi:10.1016/j.biopsych.2021.09.011

11. Chase HW, Fournier JC, Greenberg T, et al. Accounting for dynamic fluctuations across time when examining fMRI test-retest reliability: Analysis of a reward paradigm in the EMBARC study. *PLoS One*. 2015;10(5):1-20. doi:10.1371/journal.pone.0126326

12. Macqueen GM, Hassel S, Arnott SR, et al. The canadian biomarker integration network in depression (CAN-BIND): Magnetic resonance imaging protocols. *J Psychiatry Neurosci*. 2019;44(4):223-236. doi:10.1503/jpn.180036

13. Alders GL, Davis AD, MacQueen G, et al. Reduced accuracy accompanied by reduced neural activity during the performance of an emotional conflict task by unmedicated patients with major depression: A CAN-BIND fMRI study. *J Affect Disord*. 2019;257(April):765-773. doi:10.1016/j.jad.2019.07.037

14. Markello RD, Arnatkevičiūtė A, Poline JB, Fulcher BD, Fornito A, Misic B. Standardizing workflows in imaging transcriptomics with the Abagen toolbox. *Elife*. 2021;10:1-27. doi:10.7554/eLife.72129

15. Hansen JY, Shafiei G, Markello RD, et al. Mapping neurotransmitter systems to the structural and functional organization of the human neocortex. *Nat Neurosci*. 2022;25(November):1569–1581. doi:10.1038/s41593-022-01186-3

16. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

17. Schulz MA, Yeo BTT, Vogelstein JT, et al. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat Commun*. 2020;11(1). doi:10.1038/s41467-020-18037-z

18. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science (80-)*. 2024;383:164-167. doi:10.1126/science.adg8538

19. Helmer M, Warrington S, Mohammadi-Nejad AR, et al. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Commun Biol*. 2024;7(1). doi:10.1038/s42003-024-05869-4

20. Johnson WE, Li C. Adjusting batch effects in microarray expression data using empirical Bayes methods. Published online 2007:118-127. doi:10.1093/biostatistics/kxj037

21. Morgan SE, Seidlitz J, Whitaker KJ, et al. Cortical patterning of abnormal morphometric similarity in psychosis is associated with brain expression of schizophrenia-related genes. *Proc Natl Acad Sci*. Published online 2019:201820754. doi:10.1073/pnas.1820754116

22. Zhukovsky P, Wainberg M, Milic M, et al. Multiscale neural signatures of major depressive, anxiety, and stress-related disorders. *Proc Natl Acad Sci*. 2022;119(23):1-10. doi:10.1073/pnas.2204433119

23. Zhukovsky P, Savulich G, Morgan S, Dalley JW, Williams GB, Ersche KD. Morphometric similarity deviations in stimulant use disorder point towards abnormal brain ageing. *Brain Commun*. 2022;4(3). doi:10.1093/braincomms/fcac079

24. Morgan SE, Seidlitz J, Whitaker KJ, et al. Cortical patterning of abnormal morphometric similarity in psychosis is associated with brain expression of schizophrenia-related genes. *Proc Natl Acad Sci*. 2019;116(19):9604-9609. doi:10.1073/pnas.1820754116

25. French L, Ma T, Oh H, Tseng GC, Sibille E. Age-Related Gene Expression in the Frontal Cortex Suggests Synaptic Function Changes in Specific Inhibitory Neuron Subtypes. 2017;9(May):1-14. doi:10.3389/fnagi.2017.00162

26. Zhukovsky P, Ironside M, Duda JM, et al. Acute stress increases striatal connectivity with cortical regions enriched for μ- and κ-opioid receptors. *Biol Psychiatry*. 2024;(14):1-10. doi:10.1016/j.biopsych.2024.02.005

27. Howard DM, Adams MJ, Shirali M, et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat Commun*. 2018;9(1):1-10. doi:10.1038/s41467-018-03819-3

28. Anderson KM, Collins MA, Kong R, et al. Convergent molecular, cellular, and cortical neuroimaging signatures of major depressive disorder. *Proc Natl Acad Sci*. 2020;117(40):202008004. doi:10.1073/pnas.2008004117

29. Fabbri C, Corponi F, Souery D, et al. The Genetics of Treatment-Resistant Depression: A Critical Review and Future Perspectives. *Int J Neuropsychopharmacol*. 2019;22(2):93-104. doi:10.1093/ijnp/pyy024

30. Wigmore EM, Hafferty JD, Hall LS, et al. Genome-wide association study of antidepressant treatment resistance in a population-based cohort using health service prescription data and meta-analysis with GENDEP. *Pharmacogenomics J*. 2020;20(2):329-341. doi:10.1038/s41397-019-0067-3

31. Li QS, Tian C, Seabrook GR, Drevets WC, Narayan VA. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response. *Transl Psychiatry*. 2016;6(9):1-9. doi:10.1038/TP.2016.171

32. Fabbri C, Hagenaars SP, John C, et al. Genetic and clinical characteristics of treatment-resistant

depression using primary care records in two UK cohorts. *Mol Psychiatry*. 2021;26(7):3363-3373. doi:10.1038/s41380-021-01062-9

33. Combrisson E, Jerbi K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods*. 2015;250:126-136. doi:10.1016/j.jneumeth.2015.01.010