Distinct Predictive Patterns of Antidepressant Response to Sertraline, Placebo and Bupropion using Pre-Treatment Reward Processing as examined through Functional MRI and Deep Learning:

Key Results from the EMBARC Study

Kevin P. Nguyen, Cherise Chin Fatt, Alex Treacher, Cooper Mellema, Crystal Cooper, Manish Jha, Benji Kurian, Maurizio Fava, Patrick J. McGrath, Phil Adams, Myrna Weissman, Mary L. Phillips, Madhukar H. Trivedi* and Albert Montillo*

Address Correspondence to:

Albert Montillo, PhD
Director, Deep Learning for Precision Health Laboratory
Lyda Hill Department of Bioinformatics
University of Texas Southwestern Medical Center
albert.montillo@utsouthwestern.edu

Madhukar H. Trivedi, M.D.
Professor of Psychiatry
Julie K. Hersh Chair for Depression Research and Clinical Care
Betty Jo Hay Distinguished Chair in Mental Health
Director, Center for Depression Research and Clinical Care
University of Texas Southwestern Medical Center
madhukar.trivedi@utsouthwestern.edu

***Drs. Montillo and Trivedi contributed equally as senior authors to this work**

ABSTRACT

Uncertainty in selecting the appropriate antidepressant for each patient is a major challenge in treatment of major depressive disorder (MDD). No biologically driven markers are currently available to improve precision in treatment selection, thus leading to a trial-and-error process and prolonged morbidity for most patients. This study developed deep learning models that accurately predict treatment outcomes for sertraline, bupropion and placebo.

Models were trained on data from the EMBARC study, in which 223 un-medicated subjects with MDD underwent pre-treatment reward task fMRI and received 8 weeks of treatment with sertraline, bupropion, or placebo. These models integrate fMRI and clinical measures and they explain up to 37% of the variance in ∆HAMD, classify remitters with NNT of 2.3-4.3, and classify responders with NNT of 3.2-4.9. Findings reveal new regions predictive of treatment outcome such as the hippocampus and paracentral lobule, while additional regions implicated in existing research are corroborated. Distinct models were identified for each treatment and provide substantial evidence of their potential to improve precision in treatment selection for MDD.

INTRODUCTION

With an estimated lifetime prevalence of 16%, major depressive disorder (MDD) constitutes one of the most common disabling illnesses worldwide[1]. The direct cost of depression-related healthcare in addition to indirect costs from suicide and loss of productivity sums to over $200 billion in the United States alone[2]. Despite the widespread impact of MDD, effective management of the disease is limited due to a lack of available biomarkers to enhance precision in treatment selection. While antidepressants remain the mainstay of treatment, an individual's response to each antidepressant is heterogeneous and no single drug has been established as a universal first-line treatment[3]. Consequently, the current standard of care is an iterative process to identify the optimal treatment regimen for each patient. The clinician identifies a candidate treatment for the patient, the patient tests the treatment for 2 to 3 months to reach full pharmacological effect, and symptomatic improvement is assessed. If improvement is not satisfactory or if adverse effects are intolerable, the clinician selects the next treatment to test. However, about 33% of patients fail to achieve remission even after four of these trials over the period of one year without satisfactory treatment[4]. Given the substantial morbidity that can occur during this time, including loss of work productivity, employment, marriage, or even life, there is clearly a need for tools to facilitate appropriate treatment selection. Such tools should aid clinicians in prioritizing antidepressants most likely to work for each patient based on pre-treatment biomarkers and enable patients to more quickly attain remission.

Functional neuroimaging has shown promise for providing such predictive biomarkers through noninvasive measurements of individual brain activation patterns. The serotonergic emotion circuit and the dopaminergic reward circuit are the two most commonly implicated neural pathways in MDD[5]. Consequently, functional magnetic resonance imaging (fMRI) acquired during task paradigms designed to stimulate these specific circuits have proven effective at identifying abnormal function in depressed brains. Several task-based fMRI studies have measured altered blood oxygen level dependent (BOLD) activation in response to emotional and reward stimuli in MDD subjects compared to healthy control subjects, particularly in brain regions associated with these two circuits[5–7]. There is also substantial evidence for using fMRI to identify pre-treatment biomarkers of treatment outcome. Analysis of resting-state fMRI has correlated default mode network and hippocampal connectivity with outcomes for sertraline treatment[8]. In task-based fMRI with an emotional regulation task paradigm, greater BOLD activation in the anterior cingulate cortex has been associated with increased response to venlafaxine and fluoxetine[9,10]. Activation in the amygdala during emotional stimuli has similarly been connected to improvement in symptoms, though not to a specific antidepressant[11]. Reward task-based fMRI has been less studied for treatment outcome prediction, however one recent study has identified abnormal reward expectancy and prediction error-related activity in the ventral striatum as a predictor of better response to sertraline over placebo[12].

These existing studies motivate the use of task-based fMRI for developing predictors of treatment outcome, while also presenting an opportunity for the following primary contributions of this work:

First, predictive models are developed to predict an individual's response to three different treatments *from pre-treatment reward task-based fMRI*, which can be used to identify

individuals most likely to respond to these treatments. While there is previous work on using emotion task-based fMRI, the use of reward processing tasks is currently understudied. Depression treatments investigated include bupropion, which primarily modulates the neurotransmitter dopamine, sertraline, a selective serotonin reuptake inhibitor (SSRI), and a placebo treatment. Imaging and behavioral biomarkers of sertraline response have been previously studied, though bupropion has been relatively less studied. At the time of writing, only one study of bupropion has been conducted using emotion task-based fMRI[13], which principally probes serotonergic circuitry rather than the dopaminergic reward circuitry which is believed to be most directly modulated by bupropion[14].

The second major contribution is the novel application of deep learning in building predictive models of antidepressant outcome. Because the association between task-based fMRI measurements and treatment outcome is likely to be complex, a model with higher statistical complexity may be more apt to solve this predictive problem. Therefore, this work employs deep learning, which is well-suited to capture the putative complex relationship between the brain and antidepressant response. Deep learning has not been extensively explored for psychiatric prognostic problems, in part because such models typically require large amounts of data to be trained well. However, recent advancements in regularization and data augmentation have made deep learning even in the presence of limited data much more tractable. These advancements in regularization and data augmentation are combined in this study. Deep learning models also exhibit high capacity to scale to large numbers of input features (with appropriate regularization), while automatically learning the most informative features. This capacity is exploited to integrate imaging features with pre-treatment clinical measurements. Additionally, by taking a machine learning approach to this data analysis, the focus is placed on developing models that will make accurate predictions on unseen data, rather than only the dataset used herein. Consequently, this work employs extensive validation using held-out data to form an unbiased estimate of real world performance, in contrast to traditional statistical analyses that may fit models to the entire dataset.

Finally, the best predictive models are examined to identify the most important features learned by the models to predict treatment outcome. Comparing these learned features with existing knowledge about MDD neurophysiology and antidepressant mechanisms of action provides additional evidence that the models have learned true predictive information. The combination of the important features from each of the three treatments also forms a candidate moderating biomarker of treatment response. The discovered biomarkers contribute to understanding of MDD neurophysiology and the physiological mechanisms of each treatment. These results corroborate previous knowledge of particular brain regions with abnormal activation in MDD and with moderating effects on treatment outcome. Additionally, the findings suggest new moderating roles for several brain regions not previously implicated in MDD or antidepressant response.

RESULTS

Table 1. Demographic, pre-treatment clinical characteristics, and 8-week treatment outcomes for bupropion, sertraline, and placebo treatment groups. *MDE*: major depressive episode. *5 bupropion subjects, 11 sertraline subjects, and 11 placebo subjects had too many MDEs to count.

	Sertraline		Placebo		Bupropion	
Total subjects	106		116		37	
Demographics						
Female	73	69%	73	63%	26	70%
Race						
White	72	68%	83	72%	24	65%
African American	20	19%	17	15%	8	22%
Asian	5	5%	8	7%	3	8%
Other	9	8%	8	6%	2	5%
Hispanic	19	18%	22	19%	6	16%
Employed	61	58%	69	59%	20	54%
Age	38.38 ± 13.95		37.40 ± 12.80		37.51 ± 14.32	
Clinical characteristics						
Age of first MDE	16.16 ± 5.86		16.57 ± 5.91		16.11 ± 5.77	
Number of MDE*	8.55 ± 17.47		8.31 ± 12.28		8.94 ± 17.51	
Pre-treatment HAMD	18.60 ± 4.45		18.57 ± 4.26		18.00 ± 3.96	
Treatment outcomes						
ΔHAMD (Week 8 – pre-treatment)	7.89 ± 7.16		6.70 ± 6.93		5.46 ± 5.57	
Remission	41	39%	38	33%	12	32%
Response	57	54%	41	35%	15	41%

Data acquisition

In the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) study[14], subjects underwent pre-treatment fMRI acquired during a block-design monetary reward task shown to probe reactivity to reward and punishment[7,15,16] and known to be altered in individuals with MDD[7]. They were then randomly assigned to either 8 weeks of treatment with sertraline or placebo. Depression severity was tracked using the 17-item version of the Hamilton Rating Scale for Depression (HAMD), which has possible total scores ranging from 0 to 52 points[17]. The main treatment outcome of interest was the 8-week change in HAMD was ΔHAMD, defined as the difference of week 8 minus pre-treatment HAMD. Additionally, these measurements were used to identify whether or not each subject had achieved clinically-significant *response* (decrease in HAMD ≥ 50% from pre-treatment) or *remission* (week 8 HAMD ≤ 7). Mean ΔHAMD, response rates, and remission rates for each treatment arm of this study are enumerated in **Table 1**. Subjects for whom pre-treatment reward task fMRI and both pre-treatment and week 8 HAMD scores were available were selected for analysis, and this included 106 subjects from the sertraline arm and 116 subjects from the placebo arm.

Prediction of Sertraline Treatment Outcome

For the sertraline treatment arm (n = 106), the mean ΔHAMD was 7.89 ± 7.16, remission rate was 39%, and response rate was 54% (**Table 1**). Imaging features were derived from the fMRI after performing data augmentation and image preprocessing. These features comprised measurements of an individual's BOLD activation in response to the *anticipation*, *reward*

expectation, and *prediction error* conditions of the reward task paradigm in each of 200 regions-of-interest (ROIs). The imaging features were combined with clinical measurements and demographic variables and used to train the deep learning model, a feed-forward neural network, to predict individual ΔHAMD. Using nested cross-validation to evaluate predictive performance on unseen data, this model achieved a substantial R^2 of 35%, with a 95% confidence interval of 20-50%, and an RMSE of 5.75. Permutation testing confirmed these results to be statistically significant with $p < 10^{-3}$ (**Table 2,** 1st row). Notably, this sertraline model achieved the best performance when trained on the combination of imaging and clinical/demographic features versus using imaging features or clinical features alone (**Tables S4, S5**).

The most important features learned by the predictive model were identified by computing for each feature, the permutation feature *importance* score (not to be confused with permutation *testing*)[18]. This approach measures the decrease in R^2 incurred when each feature is permuted among the subjects, ablating any information in the feature. The top 30 features ranked by importance are visualized in **Figure 1**. Remaining features had substantially lower importance. The anatomical label for each ROI in **Figure 1a** was determined by Dice overlap with the widely-used Automated Anatomical Labeling atlas[19]. Of these 30 most important features, 24 were imaging features and 6 were clinical and demographic features (**Fig. 1**), with the clinical features ranking most highly. Of the clinical features, higher pre-treatment symptomatic severity (17-item and 24-item HAMD total), body mass index (BMI), psychomotor retardation reported on the SCID[a] assessment, and higher NEO[b] Neuroticism score were predictive of <u>resistance</u> to treatment. Psychomotor agitation reported on the SCID assessment was predictive of <u>remission</u>.

From the imaging features, regions whose higher BOLD **response to reward expectation** were predictive of <u>remission</u> included the *medial orbitofrontal cortex*, *calcarine sulcus*, *precentral gyrus*, *insula*, *superior occipital lobe*, *cerebellar vermis*, and *paracentral lobule* (**Figure 1a**, in orange). Similarly, higher BOLD response in the *parahippocampal gyrus*, *hippocampus*, and *thalamus* were predictive of <u>resistance</u>. Regions whose higher BOLD **response to prediction error** were predictive of <u>remission</u> included the *precentral gyrus*, *medial superior frontal gyrus*, *inferior orbitofrontal cortex*, and *hippocampus*, while higher BOLD response in the *supplemental motor area* was predictive of <u>resistance</u> (**Figure 1a**, in blue). Regions whose higher BOLD **response to anticipation** were predictive of <u>remission</u> were the *supplemental motor area*, *superior frontal gyrus*, and *thalamus*, while the *precentral gyrus*, part of the *anterior cerebellum*, *postcentral gyrus*, and a different area of the *superior frontal gyrus* were predictive of <u>resistance</u> (**Figure 1a**, in green).

[a] Structured Clinical Interview for DSM-IV
[b] Neuroticism-Extraversion-Openness Inventory

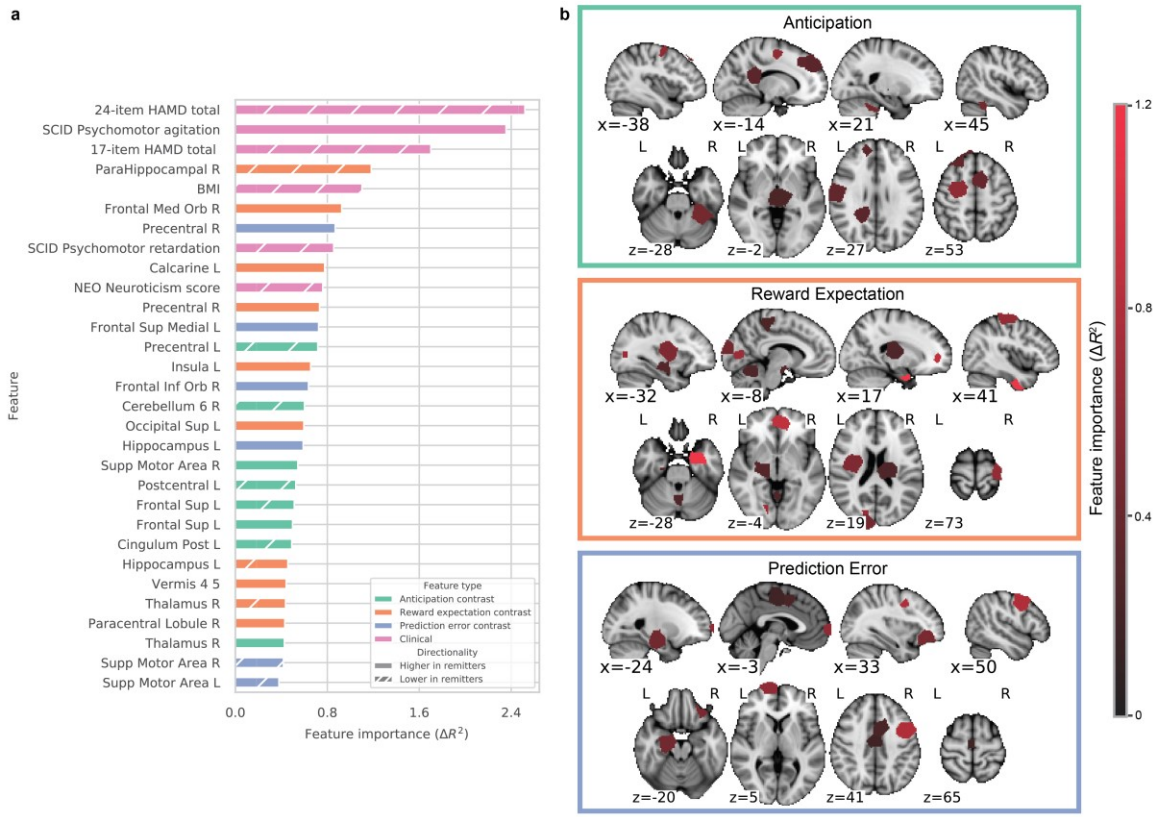

Figure 1. The thirty most important features learned by the sertraline outcome prediction model, as determined by permutation feature importance. **a)** Features are ranked by descending importance and colored by feature type. The directionality of each feature was computed by comparing the mean feature values of remitters and non-remitters; features with lower values in remitters have hashed bars. **b)** The imaging features are visualized as colored ROIs in the study-specific brain atlas and overlaid on the MNI brain template. Brighter red indicates greater feature importance.

Prediction of Placebo Treatment Outcome

In the placebo treatment arm (n = 116), the mean ΔHAMD was 6.70 ± 6.93, a slightly smaller clinical improvement than that of the sertraline arm (p = 0.11). Remission and response rates were 33% and 35% respectively, lower than those of sertraline (**Table 1**). Similar to the sertraline model, the highest predictive performance was achieved when the deep learning model was trained on the combination of imaging and clinical/demographic features (**Tables S4, S5**). The placebo model attained an R^2 of 23%, with a 95% confidence interval of 11-37%, and an RMSE of 6.06 for predicting ΔHAMD on held-out test data. This was statistically significant ($p < 10^{-3}$) upon permutation testing (**Table 2**, 2nd row).

Compared to the sertraline model, the placebo model relied more heavily on clinical and demographic features. Of the 30 most important features learned by the model (**Figure. 2**), 15 were imaging features and 15 were clinical and demographic features. Features predictive of resistance included: concurrent anxious distress and panic disorder (from SCID), greater anhedonia (MASQ[c] anhedonic depression score and SHAPS[d] ordinal total), number of comorbidities (SCQ total), Caucasian race, age, the presence of melancholic, atypical, or catatonic depression (*SCID current episode specifier*), number of children, current alcoholism, and psychomotor agitation. Remission was predicted by Asian race, longest period without dysphoria, depression-associated symptomatology (CAST[e] total), and concurrent psychotic symptoms.

The imaging features were distinct from those learned by the sertraline model. For the placebo model, higher BOLD **response to prediction error** in the *hippocampus* was highly important for predicting resistance (**Figure 3a**, in blue), and this was the only prediction error feature ranked in the top 30 learned features. Regions whose higher BOLD **response to anticipation** were associated with resistance included regions near the central sulcus (*supramarginal gyrus*, *postcentral gyrus*, *precentral gyrus*) as well as the *fusiform gyrus*, *middle frontal gyrus*, *cuneus*, and *medial orbitofrontal cortex* (**Figure 3a**, in green). In a different ROI within the *postcentral gyrus*, higher BOLD response to anticipation was predictive of remission. Finally, regions where higher BOLD **response to reward expectation** was predictive of remission included the *medial superior frontal gyrus* and *postcentral gyrus*, and regions associated with resistance were the ROIs near the central sulcus (*paracentral lobule*, *postcentral gyrus),* the *triangular part of the inferior frontal gyrus*, and *cerebellar crus* (**Figure 3a**, in orange).

[c] Mood and Anxiety Symptom Questionnaire
[d] Snaith-Hamilton Pleasure Scale
[e] Concise Associated Symptoms Tracking

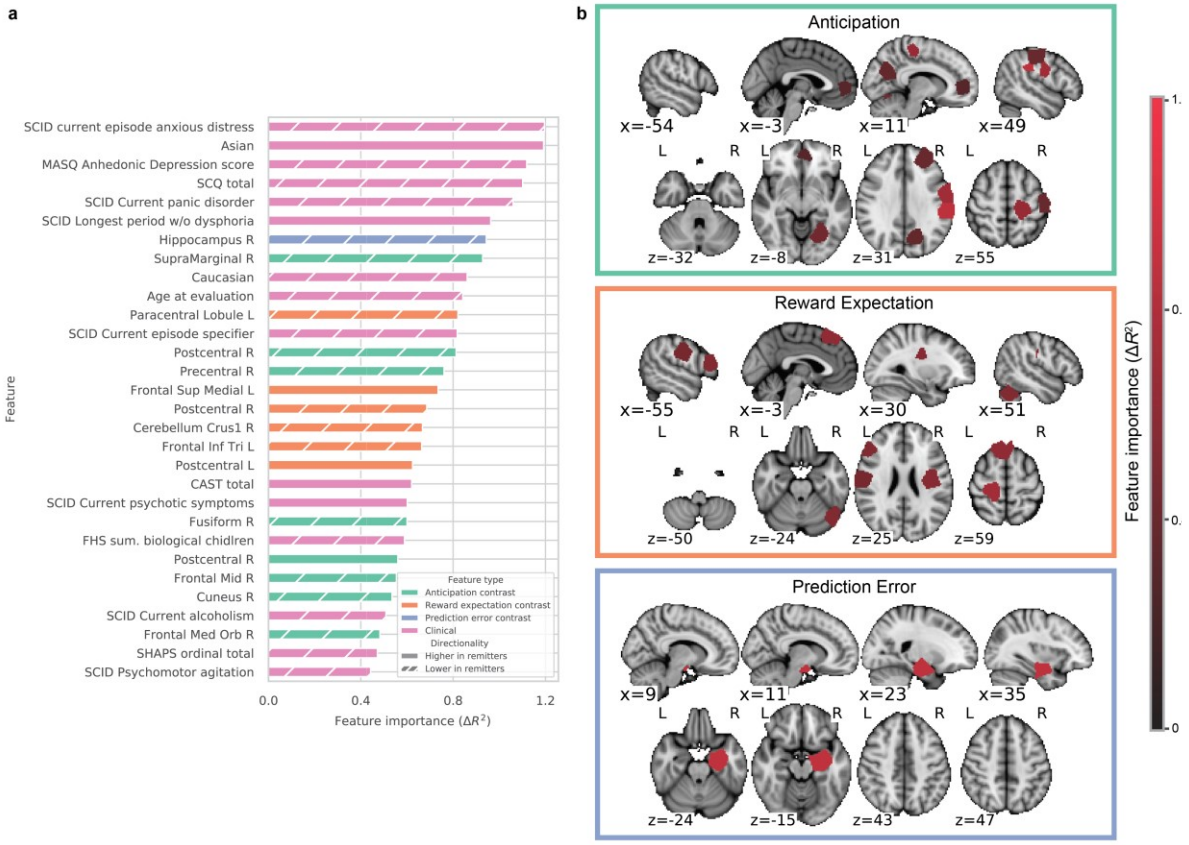

Figure 2. The thirty most important features learned by the placebo outcome prediction model, as determined by permutation feature importance. **a)** Features are ranked by descending importance and colored by feature type. The directionality of each feature was performed by comparing the mean feature values of remitters and non-remitters; features with lower values in remitters have hashed bars. **b)** The imaging features are visualized as colored ROIs in the study-specific brain atlas and overlaid on the MNI brain template. Brighter red indicates greater feature importance.

Prediction of Bupropion Treatment Outcome

At the conclusion of the first 8-week treatment period, sertraline-treated subjects who did not experience substantial improvement (Clinical Global Improvement score < "much improved") were crossed over to bupropion treatment (n = 37). These subjects underwent a similar 8-week treatment period and clinical severity was again tracked through administration of the HAMD. Mean ΔHAMD was 5.46 ± 5.57 remission rate was 32%, and response rate was 41% for this group of subjects (**Table 1**). As with the other two treatment arms, the pre-treatment fMRI from the start of the study was used to train the predictive model. In contrast to the sertraline and placebo models, the bupropion model achieved the highest performance in predicting ΔHAMD using imaging features alone; the addition of clinical and demographic features was found to decrease performance (**Table S4**). This model achieved an R^2 of 37%, with a 95% confidence interval of 12-61%, and a root mean squared error (RMSE) of 4.36 in predicting ΔHAMD on held-out data. Permutation testing confirmed the statistical significance of this performance ($p < 10^{-3}$) (**Table 2**, 3rd row).

The important imaging features of the bupropion model (**Figure 3**) were distinct from those learned by the sertraline and placebo models. Regions where higher BOLD **response to reward expectation** was predictive of <u>remission</u> included the *posterior cingulate cortex, fusiform gyrus,* and *parahippocampal gyrus*, while higher BOLD response in the *frontal operculum* and *inferior parietal gyrus* were associated with <u>resistance</u> (**Figure 2a**, in orange). Higher regional BOLD **response to prediction error** in the *fusiform gyrus, inferior temporal gyrus*, and *temporal pole* were predictive of <u>remission</u>, and higher BOLD response in the *lingual gyrus*, *postcentral gyrus*, *superior parietal gyrus*, *superior and middle frontal gyri*, *rolandic operculum*, *supplemental motor area*, *rectus*, and *thalamus* was predictive of <u>resistance</u> (**Figure 2a**, in blue). Finally, regions where higher BOLD **response to anticipation** was predictive of <u>remission</u> included the striatum (*caudate* and *putamen*), regions near the insula (*rolandic operculum* and *superior temporal pole*), medial frontal cortex (*paracentral lobule* and *orbital part of the superior frontal gyrus*), *inferior parietal gyrus*, *postcentral gyrus*, *anterior cingulate cortex*, and *superior occipital gyrus* (**Figure 2a**, in green).

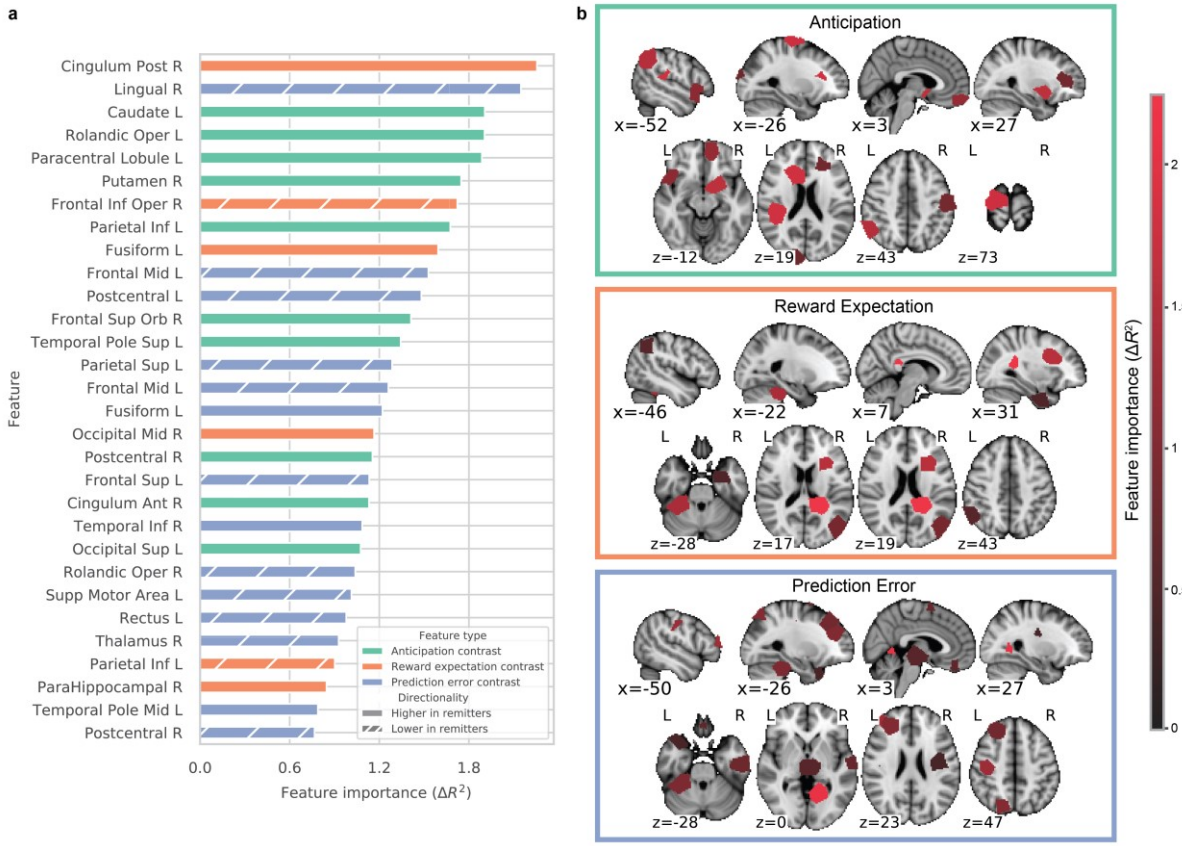

Figure 3. The thirty most important features learned by the bupropion outcome prediction model, as determined by permutation feature importance. **a)** Features are ranked by descending importance and colored by contrast. The directionality of each feature was computed by comparing the mean feature values of subjects who did and did not experience remission; features with lower values in remitters have hatched bars. **b)** The same features are visualized as colored ROIs in the study-specific brain atlas and overlaid on the MNI brain template. Brighter red indicates greater feature importance.

Table 2. Outcome prediction performance for the three treatments investigated. Models were trained to predict 8-week ΔHAMD. Performance metrics for this target include the coefficient of determination (R^2) and root mean squared error (RMSE). To obtain predictions of remission and response, which are binary variables, model outputs were thresholded post-hoc using the HAMD criteria for remission (HAMD ≤ 7 at week 8) and response (decrease in HAMD ≥ 50%). Performance metrics for remission and response are number-needed-to-treat (NNT), positive predictive value (PPV) and area under the receiver operating characteristic curve (AUROC). Statistical significance of these performance measurements over chance accuracy was measured using permutation testing, and the *p*-values are presented here.

		Prediction target							
		ΔHAMD		**Remission**			**Response**		
Treatment	**Features used**	R^2	RMSE	NNT	PPV	AUROC	NNT	PPV	AUROC
Sertraline	Imaging, clinical	35% $p < 10^{-3}$	5.75 $p < 10^{-3}$	4.31 $p < 10^{-3}$	0.62 $p < 10^{-3}$	0.60 p=0.017	4.88 $p < 10^{-3}$	0.68 $p < 10^{-3}$	0.63 $p < 10^{-3}$
Placebo	Imaging, clinical	23% $p < 10^{-3}$	6.06 $p < 10^{-3}$	2.78 $p < 10^{-3}$	0.69 $p < 10^{-3}$	0.61 p=0.048	3.19 $p < 10^{-3}$	0.67 $p < 10^{-3}$	0.64 $p < 10^{-3}$
Bupropion	Imaging	37% $p < 10^{-3}$	4.36 $p < 10^{-3}$	2.35 p=0.018	0.75 p=0.044	0.71 p=0.196	3.24 p=0.002	0.71 p=0.002	0.62 p=0.006

Comparison to Traditional Statistical Analysis and Classical Machine Learning

A traditional voxel-wise analysis using statistical parametric mapping was performed to identify any group differences in reward-related activation between treatment responders and non-responders. The following group-level comparisons were conducted: responders vs. non-responders, remitters vs. non-remitters, and top quartile of ΔHAMD vs. bottom quartile of ΔHAMD. None of these comparisons identified significant group differences after false discovery rate correction at $p < 0.05$. These results underscore the importance of using a more statistically powerful analysis such as those described in the preceding sections.

Additionally, a number of classical machine learning methods were tested to compare the current deep learning approach to other multivariate regression models. The model types tested included multiple linear regression, elastic net regression, K-nearest neighbors, support vector machine, and random forest. Hyperparameter models were optimized and performance on held-out data was evaluated using the same approach that was used on the deep learning models. The same data augmentation was applied and models were tested with imaging features alone and with both imaging and clinical features. The best performance achieved with these classical machine learning models was an R^2 of 11% and RMSE of 6.54 in predicting ΔHAMD for sertraline, and results for bupropion and placebo were poor. Compared to the deep learning models, these models were unable to learn to predict treatment outcome from the data with high accuracy.

Clinical Significance

To evaluate the ability of the models to detect clinically significant outcomes, the predictions of ΔHAMD on held-out data were thresholded post-hoc using the clinical criteria for remission (HAMD ≤ 7 at week 8 of treatment) and response (change in HAMD of ≥ 50% from pre-treatment). These predictions of remission and response were compared to the ground truth labels in the dataset to compute number-needed-to-treat (NNT), which is a common and intuitive metric for assessing the clinical value of a proposed screening tool. For prediction of remission, NNT was, 4.31 for sertraline, 2.35 for bupropion, and 2.78 for placebo (**Table 2**). This indicates, for example, that about 4 individuals would need to be screened with the sertraline model to identify one additional individual who would experience remission when prescribed sertraline, compared to fortuitous randomized treatment assignment in this study. Positive predictive values (PPV) for predicting remission were 62% for sertraline, 75% for bupropion, and 69% for placebo. For prediction of response, performance was only slightly smaller with NNTs of 3.19 for bupropion, 3.24 for placebo, and 4.88 for sertraline. PPVs for predicting response were 71% for bupropion, 68% for sertraline, and 67% for placebo. All NNTs and PPVs measured were statistically significant ($p < 0.05$) upon permutation testing.

A second metric, NNT_{clin}, can be computed to compare the performance of these models to clinician performance for the same antidepressant selection decisions. Using an estimated medication response rate of 45% in clinical practice[20], NNT_{clin} was 4.35 for sertraline and 3.85 for bupropion. This indicates that a clinician would need to examine about 4 individuals using the proposed predictive models to identify one additional individual who could be treated well with sertraline or bupropion compared to current clinical practice without using the proposed models.

Treatment Specificity of the Predictive Models

To test whether the models had identified composite predictive biomarkers that are *specific to each treatment*, the model for each treatment was evaluated on subjects from the other two treatment arms. In each case, the predictive performance was low (negative R^2), confirming that each had learned predictive features *specific* to each treatment.

DISCUSSION

These results, to the best of our knowledge, describe for the first time specific prediction model-based results with strong certainty for three treatment options (sertraline, placebo and bupropion). Deep learning predictive models were developed to forecast individual outcome for treatment with these three treatment arms: sertraline, placebo, and bupropion. All models were capable of explaining a substantial proportion of the variance in ΔHAMD and were accurate in predicting remission and response after 8 weeks of treatment. The highest performance was achieved by the bupropion model, which had an NNT of 2.35 for predicting remission, and yet even the lowest performing model (sertraline) had an NNT of 4.31, which is well within what is widely described as strongly clinically meaningful[20]. The higher performance of the bupropion model may be accounted for by the fact that, while the ΔHAMD observed with sertraline had a significant placebo component (given the non-significant difference from the ΔHAMD observed with placebo), the ΔHAMD reported with bupropion was less likely to result from a placebo effect and was therefore more likely the reflection of a true drug effect.

Additionally, these results provide evidence of stronger prediction compared with other proposed predictors of antidepressant treatment outcome. A predictor using measurements from behavioral tests of cognitive and emotional function achieved NNT of 3.8 for predicting remission in cognitively-impaired subjects treated with escitalopram[21]. A later analysis of the same data predicted remission for escitalopram- and venlafaxine-treated subjects with NNT of 2.7 and 4.6, respectively[22]. The authors also attained NNT of 3.5 for sertraline, but PPV was 43%, lower than the 62% of the sertraline model developed in this current study. Neither of these previous studies examined prediction of bupropion or placebo outcome. For the more difficult task of predicting response, the models proposed in this study demonstrated relatively lower yet still clinically relevant performance with NNTs ranging from 3-5.

Interactions between sertraline and the reward processing task

In this study, the reward processing task yielded biomarkers predictive of bupropion outcome. This finding is consistent with the drug's mechanism of action because bupropion modulates dopamine, the principal neurotransmitter in reward processing neural circuitry. However, the finding that sertraline outcome could also be predicted with high R^2 from the same task paradigm is unexpected given that sertraline is a selectively serotonergic antidepressant. It should be noted that the sertraline model used both imaging and clinical features, while using either imaging or clinical features alone produced poorer performance. Therefore, the sertraline model learned synergistic information from both modalities of data, and some of its predictive power must be attributed to the use of clinical features.

Regarding the capability of the reward processing task fMRI to predict sertraline outcome, one possible explanation is that serotonin does play a role in reward processing

circuitry, and consequently sertraline may modulate this circuitry to some extent. Accordingly, rodent studies of reward circuitry have demonstrated a dependence on serotonin for normal function[23]. Knock-out of the serotonin transporter causes behavioral changes and altered reward circuit connectivity[24]. Sertraline, but not the SSRIs fluvoxamine and paroxetine, increases dopamine levels in the nucleus accumbens and striatum[25]. A second possible explanation is that the reward processing task stimulates the serotonergic emotion circuitry to some degree. Given the shared anatomical regions between the reward and emotion processing circuits, such as the anterior cingulate cortex, medial frontal cortex and amygdala, and components of the limbic system, there is likely an interaction between the reward processing task and the serotonergic circuits modulated by sertraline[5]. A third explanation is that a combination of these factors are involved.

Examination of learned composite neuroimaging biomarkers

Feature importance measurements of the trained predictive models revealed the imaging, clinical, and demographic features that most facilitated the models' prediction of ΔHAMD. The imaging features can be compared to existing studies of reward processing neural circuitry, MDD pathophysiology, and antidepressant response biomarkers in order to identify features that have a previously known neurobiological basis and those that are novel and could warrant further investigation (**Table 3**).

The predictive model for sertraline (**Table 3**, 3rd column) identified regions that have been previously found to have altered BOLD activity during reward processing in MDD, particularly the medial frontal cortex (MFC), orbitofrontal cortex (OFC), insula, and thalamus[5,6]. While most of these regions were also learned by the bupropion model, the specific predictive contrasts in these regions are distinct between the antidepressants; e.g. prediction error-related activation in the thalamus is predictive of resistance for bupropion while anticipation-related activation in the same region is predictive of remission for sertraline. The placebo model (**Table 3**, 3rd column) also learned the MFC and OFC as important regions, though the predictive contrasts in these regions were different from the other treatments. While the OFC was learned by all three models, higher anticipation-related activation predicted remission for bupropion but resistance for placebo, and higher reward expectation and prediction error-related activation predicted remission for sertraline.

The predictive model for bupropion (**Table 3**, 4th column) also learned several regions that have been previously implicated in altered reward processing in MDD, including the anterior cingulate cortex (ACC), OFC, dorsal striatum (caudate and putamen), insula, and thalamus[5,6,26]. Activation in the ACC in response to emotional stimuli has been shown to predict faster improvement in response to fluoxetine, an SSRI[10], and the current results suggest an additional role in dopaminergic drug response. Also, connectivity between the ACC and the insula was predictive of treatment-independent clinical improvement in an analysis of EEG data from this same study (EMBARC)[27]. Reward task activity in the ventral striatum has been previously identified as a moderator of antidepressant outcome[12], and a similar moderating relationship is seen here, where higher dorsal striatal activation predicted remission in bupropion-treated subjects.

Several regions were learned by the models that have not previously associated with altered reward processing in MDD (**Table 3**, bottom). The superior frontal gyrus was learned by both the sertraline and bupropion models, and this region is known to be stimulated by the reward processing task, and consequently would contain a strong signal to be learned by the models[16]. Similarly, the hippocampus, learned by the sertraline and placebo models, and the nearby temporal pole, learned by the bupropion model, are constituents of the limbic system and would likely be activated by reward stimuli[28,29]. The paracentral lobule was learned by all three models and the precentral gyrus by the sertraline and placebo models. These regions do not have a clear role in the canonical reward processing circuitry but have demonstrated altered activation during memory encoding and retrieval tasks in MDD subjects[30]. This could suggest possible altered reward processing function in these regions. Additionally, decreased perfusion in the paracentral lobule on arterial spin labelling imaging has been related to antidepressant resistance[31].

Examining the bupropion model, several novel regions were learned uniquely for this treatment. The inferior parietal gyrus and fusiform gyrus are plausible predictive regions as they are known to activate during the reward processing task[16]. The posterior cingulate cortex (PCC) and parahippocampal gyrus have not previously been examined in the context of reward processing and MDD, but activation during emotional stimuli in these areas has been shown to be modulated by bupropion[13].

Examination of learned clinical and demographic biomarkers

For the sertraline and placebo models, clinical and demographic features were found to be complementary with imaging features for achieving high predictive performance. Clinical features alone, however, yielded poor predictive performance with R^2 of 1.9% for sertraline and 1.8% for placebo (**Table S4**). Additionally, using imaging features alone also provided low predictive performance (R^2 < 10%, **Table S5**). For the sertraline model, a small number of clinical measurements were found to be highly important features for predicting treatment outcome (6 out of the top 30 features, **Figure 1a**). Pre-treatment HAMD score was the most important feature overall, with a higher total score on the either the 17-item or 24-item version predicting treatment resistance. Higher body mass index (BMI) also predicted resistance, which is corroborated by other studies showing that obesity correlates with poorer outcomes with SSRIs[32,33]. A notable finding was that concurrent psychomotor agitation predicted remission. Sertraline is known to effectively treat psychomotor agitation, compared to other SSRIs such as fluoxetine, and a prior study saw a non-significantly higher response rate to sertraline vs. nortriptyline in agitated subjects[34,35].

Regarding the placebo model, several associated symptoms and psychiatric comorbidities were important features predictive of resistance, such as concurrent anxious distress, concurrent panic disorder, and anhedonia. As with sertraline, pre-treatment comorbidity score was a highly important feature for predicting resistance. Temporal features also appeared to be predictive, including older age at time of study being predictive of resistance and a longer period of time without dysphoria being predictive of remission. Asian and Caucasian race were both learned as predictive features, with Asian subjects being more

Table 3. Brain regions identified by the models as containing reward task BOLD activation features predictive of treatment outcome. The regions are separated into two categories. The first category contains regions previously implicated in the context of altered reward processing in MDD or treatment outcome prediction, or known to be activated by the reward processing task used in this study. The second category contains regions that have not been directly associated with reward processing in MDD and that may constitute novel candidate biomarkers of treatment outcome. (↑) indicates that BOLD activation in the region was associated with remission, (↓) indicates association with resistance to treatment, and (↕) indicates that the directionality of the association varied among contrasts. Superscripts indicate which reward task contrast feature was learned by the model in a particular region; *AN*: anticipation, *RE*: reward expectation, *PE*: prediction error. Region abbreviations include the anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), medial frontal cortex (MFC), and orbitofrontal cortex (OFC).

Category	Association with published results	Sertraline	Placebo	Bupropion
Regions previously implicated in reward processing in MDD	Reward processing activity altered in MDD[5–7,25]	↑MFC^{PE} ↑OFC^{RE,PE} ↑Insula^{RE} ↕Thalamus^{AN,RE}	↑MFC^{RE} ↓OFC^{AN}	↑ACC^{AN} ↑OFC^{AN} ↑Caudate^{AN} ↑Putamen^{AN} ↕Insula/Operculum^{AN,RE,PE} ↓Thalamus^{PE}
	Reward processing activity associated with treatment outcome[8]	↑MFC^{PE}	↑MFC^{RE}	↑Caudate^{AN} ↑Putamen^{AN}
Regions not previously associated with reward processing activity in MDD	Known to activate during the reward processing task[16]	↕Superior frontal gyrus^{AN}	↓Fusiform gyrus^{AN}	↓Inferior parietal gyrus^{AN,RE} ↓Superior frontal gyrus^{PE} ↑Fusiform gyrus^{RE,PE}
	Associated with the limbic system[27,28]	↕Hippocampus^{RE,PE}	↓Hippocampus^{PE}	↑Temporal pole^{PE}
	Emotion processing activation modulated by bupropion[13]	n/a	n/a	↑PCC^{RE} ↑Parahippocampal gyrus^{RE}
	Memory encoding and retrieval altered in MDD[29]	↑Paracentral lobule^{PE} ↑Precentral gyrus^{RE,PE}	↓Paracentral lobule^{RE} ↓Precentral gyrus^{AN}	↑Paracentral lobule^{AN} ↓Lingual gyrus^{PE}
	Other regions identified by the predictive model	↑Cuneus^{RE} ↓Postcentral gyrus^{AN} ↓Parahippocampal gyrus^{RE} ↑Calcarine sulcus^{RE} ↑Superior occipital gyrus^{RE} ↕Supplemental motor area^{AN,PE}	↓Middle frontal gyrus^{AN} ↑Postcentral gyrus^{AN,RE} ↓Cerebellar crus^{RE} ↓Cuneus^{AN} ↓Supramarginal gyrus^{AN}	↓Middle frontal gyrus^{PE} ↕Postcentral gyrus^{AN,PE} ↓Superior parietal gyrus^{PE} ↑Middle occipital gyrus^{RE} ↑Inferior temporal gyrus^{PE}

likely to remit, but this may be an artifactual finding given that only 7% of the placebo treatment group (8 subjects) was Asian.

Importantly, the observation that the combination of multiple data modalities produces more accurate predictions of antidepressant outcome is novel. *This is one of the first studies to synergize imaging measurements with another modality of information, namely clinical assessments, and this combination yielded stronger predictive signals.* These findings support further investigations into cross-modal composite moderators of treatment response for MDD.

Limitations

In order to most accurately estimate the generalization performance of these predictive models, i.e. the predictive accuracy on new real-world data, the models would need to be tested on an additional, independent dataset. However, such replication data is limited. EMBARC is the largest randomized controlled study of antidepressant response with fMRI to date, and other datasets with compatible imaging data are not yet available. In the meantime, the rigorous cross-validation and permutation testing employed in this work provide high confidence that the models truly learned predictive features and that the reported model performance will generalize to additional subjects.

Several other reputable studies have found significant biomarkers of antidepressant response with modest sample sizes in the range of 10-20 subjects[10,13,36], compared to this study which used 106 subjects (sertraline), and 116 subjects (placebo), and 37 subjects (bupropion). However, the 37 subjects for bupropion could be considered a limitation of this work. To address this, this work employed an innovative data augmentation approach to synthesize additional fMRI and improve the ability of the predictive models to find true associations in the data. Such data augmentation methods have been previously shown to enhance the performance of deep learning in data-limited situations[37]. Application of data augmentation increased the apparent dataset size by 500%, to a total of 222 bupropion fMRI samples. Additionally, the 95% confidence intervals were reported for each R^2 accounting for sample size, indicating that at worst, the lowest performing model explained 12% of the variance which is still a substantial effect size. It should be noted that these confidence intervals were computed using the sample size *before* data augmentation, making 12% a conservative estimate. Furthermore, permutation testing to check the significance of results also helped to ensure that findings were not spurious.

A secondary limitation for the bupropion treatment group in particular results from the nature of the crossover study design. Subjects were selected to receive bupropion after not responding (Clinical Global Improvement scale less than “much improved”) to an initial 8-week course of sertraline treatment, which may have led to a sampling bias for subjects less responsive to antidepressants. A follow-up study with a wider treatment design is warranted.

Technical limitations of this work include the following. While the current methodology uses a brain atlas to parcellate the contrast maps and compute mean regional contrast values for input into the predictive models, a predictor that uses the contrast maps directly may attain better accuracy. Additionally, the imaging and clinical features were concatenated for input into the models, which relies on the ability of the feed-forward neural networks to automatically identify the best features from this pool of information. A more sophisticated neural network architecture with separate subnetworks for imaging and clinical features may further improve performance, particularly for bupropion where clinical features have not added predictive power in the current analysis.

CONCLUSION

This work presents the following contributions.

1) To date, this is the first application of deep learning to prediction of treatment outcome in MDD. While traditional statistical parametric mapping methods were unable to identify *group*-level differences between responders and non-responders, deep learning models learned to predict *individual* outcomes for three different treatments and explained a substantial portion of the variance in outcome. Furthermore, the use of multivariate modeling with high statistical capacity allowed the integration of clinical measurements with imaging data to create *multi-modal* predictors, which has not been extensively explored in other studies of treatment moderators.

2) Imaging-based measurements of the reward processing circuitry are shown to have prognostic applications, and composite neuroimaging candidate biomarkers were identified for sertraline, bupropion, and placebo. Deeper examination of the important learned features revealed many brain regions in concordance with previous knowledge of MDD neurophysiology, progression, and treatment response. However several identified regions, such as the temporal pole and paracentral lobule, have not previously been implicated in MDD treatment response and could present new targets of research in therapeutics.

3) All three treatment outcome prediction models exhibited actionable predictive capability with low NNTs of 2-4 for predicting remission. Such models may begin to assist clinicians in treatment planning, particularly to determine whether bupropion or sertraline should be prioritized in each patient's medication trials.

Expediting the selection of appropriate antidepressants will help to reduce the need for long and subjective medication trials and to alleviate the considerable morbidity in depression.

METHODS

Materials

While full details of the design of the EMBARC study are available[14], the most pertinent details to the current analysis are described here. A CONSORT flow diagram can be found in the supplemental material (**Figure S1**). Institutional review board approval was obtained from each study site: Columbia University, Massachusetts General Hospital, University of Michigan, and University of Texas Southwestern Medical Center. Written informed consent was obtained from all subjects. Subjects must have met Structured Clinical Interview for the DSM-IV (SCID) criteria for MDD and scored ≥ 14 on the Quick Inventory of Depressive Symptomatology (QIDS-SR). Exclusion criteria included pregnancy, concurrent use of antipsychotics or mood stabilizers, and significant risk of suicide during the study as evaluated by study investigators. Additionally, subjects must not have had a lifetime history of psychosis, bipolar disorder, or epilepsy and must not be receiving depression-specific psychotherapy or somatic treatments. A total of 296 depressed subjects were enrolled and initiated treatment.

The treatment period was separated into two 8-week stages. In the first stage, subjects were randomized in a double-blind manner into the sertraline or placebo treatment arms. Subjects in the sertraline arm received up to 200 mg daily as tolerated. Clinical severity was

measured by clinicians using the Hamilton Rating Scale for Depression (HAMD) at pre-treatment and at week 8 of treatment. At week 8, sertraline-treated subjects who did not meet the response criteria (Clinical Global Improvement score less than “much improved”) are crossed over to bupropion. These subjects then completed another 8 weeks of treatment, with up to 450 mg daily dose as tolerated for bupropion, and HAMD was measured at week 8 of bupropion treatment.

After eliminating images with apparent signal loss and clipped field-of-view artifacts, 281 reward task-based fMRI acquisitions and 290 structural MRI (sMRI) acquisitions were obtained. Additionally, 283 resting-state fMRI acquisitions were obtained to generate the functional brain atlas, as described below. Final inclusion criteria for this analysis were completion of 8 weeks of treatment of a given treatment and availability of pre-treatment reward task-based fMRI and sMRI. This yielded 106 subjects for sertraline, and 116 for placebo, and 37 for bupropion.

MRI Acquisition

MRI was acquired on General Electric, Siemens, or Phillips 3T scanners depending on study site. Structural MRI was acquired with T1-weighted MPRAGE or the manufacturer equivalent, and functional MRI was acquired using gradient echo EPI with TR of 2 s. Resting-state fMRI was acquired for 6 minutes. While two resting-state runs were performed, before and after the task-based runs, only the first run was used for this study to avoid possible disruptive effects of the tasks. Full scanner and sequence information can be found in **Table S2**. Importantly, differences among scanner manufacturers at each study site are inherently normalized by the task-based fMRI processing stream described below. This processing generates contrast maps, which consider the relative differences in BOLD signals rather than the absolute measured values.

Reward task-based fMRI was acquired for 8 minutes during a block-design monetary reward task that has demonstrated to probe reward processing neural circuitry known to be altered in MDD (**Figure 4a**)[7,16]. Specifically, differential reactivity to reward anticipation and prediction error depending on brain region have been identified between healthy subjects and MDD subjects. Each trial of the task begins with the *response phase*, during which the subject guesses whether an upcoming number, with possible values of 1-9, will be greater or less than 5. During the *anticipation phase*, the subject is informed about the possible outcome of the current trial. Trials can be “possible win”, where the subject wins $1 for a correct guess and loses nothing for a wrong guess, or “possible loss”, where the subject loses $0.50 for a wrong guess and wins nothing for a correct guess. During the *outcome phase*, the actual number is revealed, followed by visual feedback indicating whether the subject has won money, lost money, or did not win or lose any money. This is followed by a fixation period before the next trial. A total of 24 trials were conducted, with 12 “possible win” and 12 “possible loss” trials. All subjects received a fixed monetary reward after the task regardless of outcome. The rationale for this task design is that MDD subjects commonly exhibit anhedonia and display reduced brain activation in response to rewarding stimuli compared to healthy subjects[7,26].

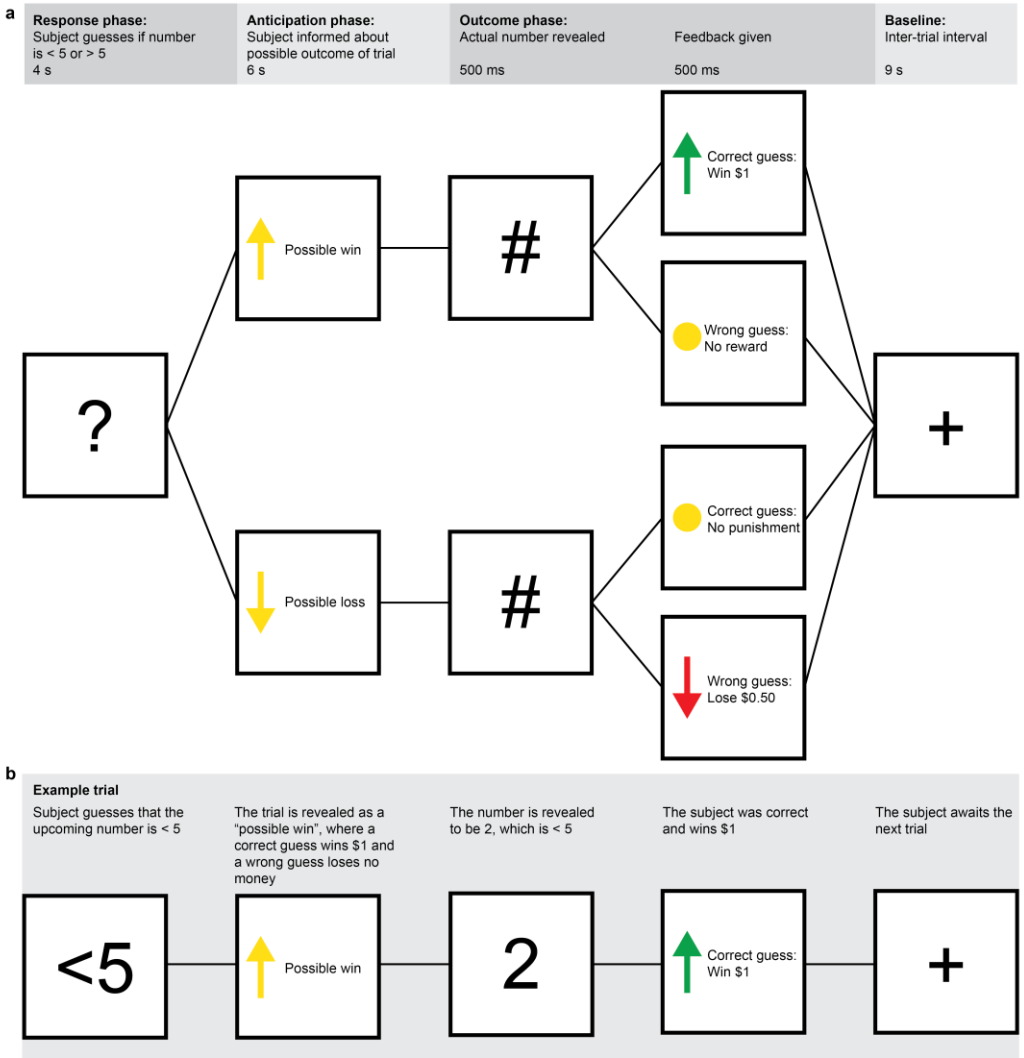

Figure 4. Block-design reward task paradigm employed in this study. The task lasts 8 minutes and includes 24 trials. **a**) Flowchart demonstrating the possible stimuli and outcomes for a single trial. In each trial, the subject guesses whether the upcoming number (1-9) is greater or less than 5. They are shown whether the trial is a "possible win" with a reward for a correct guess or a "possible loss" with a punishment for a wrong guess, and the outcome is then presented. **b**) Diagram for an example trial. In this case, the subject guesses that the upcoming number is less than 5, and the trial is a "possible win". The actual number is 2, and the subject receives $1 for a correct guess.

Anatomically Realistic fMRI Data Augmentation

To improve the performance of the deep learning models and increase their ability to learn the true association between imaging features and treatment outcome, an anatomically-informed data augmentation approach was used to simulate additional fMRI acquisitions. Similar augmentation approaches have demonstrated to dramatically increase the accuracy of deep learning models in natural (non-medical) image applications[37]. In this study, data augmentation was employed to synthesize additional fMRI data from the existing data by simulating variations in brain morphology. For each subject (the *source* subject), 5 age decile and gender-matched *target* subjects were identified from the other treatment groups. Because the target subjects are not included in the training data of each treatment-specific model, this avoids biasing the model performance results. The following registration was applied between the *source* subject fMRI and each *target* subject fMRI (**Figure 5**). First, the brain extraction was performed for the *source* sMRI and *target* sMRI using the ROBEX tool. Brain extraction was also performed for the *source* fMRI by applying FSL BET[38] and AFNI 3dAutomask[39] and taking the intersection of the two segmentations. Second, the *source* fMRI was coregistered to the *source* sMRI using the SyNQuick routine in ANTs[40] to obtain transformation T_{fs} (**Figure 5a**). Third, the *source* sMRI was coregistered to the *target* sMRI using the SyNRegistration routine in ANTs to ensure close alignment of the *source* brain to the *target* anatomy, yielding transformation T_{ss} (**Figure 5b**). Finally, the two transformations T_{fs} and T_{ss} were composed and applied to the *source* fMRI to obtain a new fMRI image containing the *source* brain activity mapped onto the *target* brain anatomy (**Figure 5c**).

This augmentation increased the effective sample sizes by 500%, providing a total of 222 fMRI images for bupropion, 978 for sertraline, and 696 for placebo. Importantly, this augmented data was used only during model training and not during evaluation, and it did not include data generated from held-out subjects during cross-validation. Additionally, target subjects were selected only from other treatment groups and do not appear in the models' held-out validation or test sets, and thus this does not bias model results. **Table S5** compares the predictive performance achieved with and without data augmentation and demonstrates the benefit of augmentation.

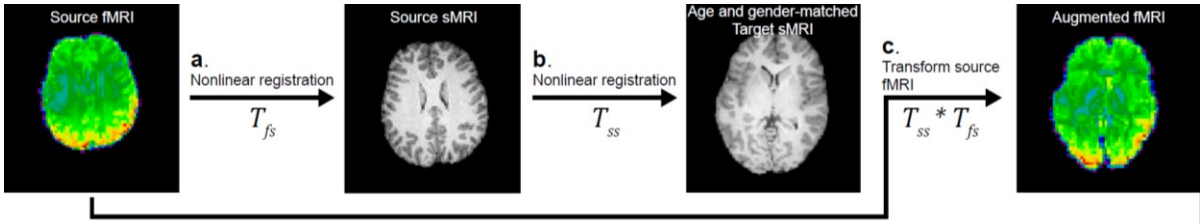

Figure 5. Data augmentation was employed to increase dataset sizes by synthesizing additional, anatomically realistic fMRI images. Each original, *source* fMRI image was spatially transformed into the brain anatomy of 5 age and gender-matched *target* subjects using a T1-based registration. **a**) The *source* fMRI is first nonlinearly registered to the *source* sMRI. **b**) The *source* sMRI is then registered using a precise SyN registration to the *target* sMRI. **c**) The transformations computed during steps **a** and **b** are combined and applied to the *source* fMRI to transform it into the *target* brain anatomy.

MRI Preprocessing

All original and augmented data was preprocessed as follows. Structural MRI were first processed with the ROBEX tool[41] to remove the skull and non-brain voxels. The image is then spatially normalized to the MNI152 T1-weighted template brain using a series of rigid body, affine, and nonlinear symmetric normalization (SyN) registrations in ANTs. This registration method was selected as it has been shown to outperform other registration methods[42]. The normalized sMRI was then segmented into gray matter, white matter, and cerebrospinal fluid with FSL FAST. Functional MRI were corrected for frame-to-frame head motion with FSL MCFLIRT, and frames where the magnitude of head motion was > 1 mm or the Z-score of the intensities was > 3 were marked as outliers to be regressed out during GLM analyses. Brain extraction was performed by applying FSL BET and AFNI 3dAutomask and taking the intersection of the two segmentations. Next, spatial normalization was conducted using a direct EPI-based normalization, where the mean functional image frame was directly registered to the MNI152 EPI brain template with ANTs. This direct normalization has been demonstrated to better correct for geometric distortions caused by EPI magnetic inhomogeneities than traditional, T1-based normalization which registers the functional to the structural image and the structural image to the template in two steps[43,44]. Finally, the normalized fMRI was spatially smoothed with a 6 mm Gaussian filter.

Generalized Linear Model Fitting

To create BOLD contrast maps, i.e. spatial maps of task-related BOLD activation, subject-level generalized linear models (GLMs) were fitted to the fMRI using the SPM12 package. Regressors were defined based on methodology used in prior analyses of this reward task fMRI data[7,12]. These included regressors for each of the *response, anticipation, outcome,* and *baseline* phases in the task paradigm. Additionally, parametrically modulated regressors were added to represent reward expectation and prediction error. The reward expectation regressor had a value of +0.5 during the *anticipation* phase of “possible win” trials and -0.25 during the *anticipation* phase of “possible loss” trials, which are the expected values of the monetary outcome of these two trial types. The prediction error regressor corresponded to the *outcome* phase and was set to the difference between the outcome and the expected value: +0.5 for a correct guess in a “possible win” trial, -0.5 for a wrong guess in a “possible win” trial, +0.25 for a correct guess in a “possible loss” trial, and -0.25 for a wrong guess in a “possible loss” trial. These 6 primary regressors, their first temporal derivatives, the head motion parameters obtained during preprocessing, and the regressors for the outlier frames were included in the GLM design matrix $\boldsymbol{X}$. White matter and cerebrospinal fluid masks from the sMRI segmentation were used to mask out unimportant voxels from the analysis. The GLM was fitted:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

Where $\boldsymbol{Y}$ is the $\text{time} \times \text{voxels}$ data matrix containing the voxel timeseries, $\boldsymbol{X}$ is the $\text{time} \times \text{regressors}$ design matrix containing the regressor timeseries, $\boldsymbol{\beta}$ is the $\text{regressors} \times \text{voxels}$ matrix containing the fitted coefficients, and $\boldsymbol{\epsilon}$ contains the residuals. The anticipation contrast map was computed as $\beta_{\text{anticipation}} - \beta_{\text{baseline}}$. The reward expectation and prediction error contrast maps were simply $\beta_{\text{reward expectation}}$ and $\beta_{\text{prediction error}}$ respectively.

Computation of Regional Contrast Values

Preliminary results showed that a study-specific functional brain atlas, generated from MDD fMRI, yielded superior predictive results when used to extract imaging features from contrast maps compared to a canonical functional brain atlas generated from healthy subjects (Schaefer 2018[45]). A MDD-specific brain atlas with 200 ROIs was generated from pre-treatment resting-state fMRI images of 283 MDD subjects using the spatially-constrained spectral clustering method developed by Craddock et al[46]. For each contrast map, including anticipation, reward expectation, and prediction error, the mean of the voxel intensities from the contrast map was computed for each ROI. Concatenation of the 200 mean regional values from each of the 3 contrast maps yielded a vector of 600 fMRI features for each subject.

Acquisition of Clinical Measurements

In addition to imaging features, 95 pre-treatment clinical measures and demographic features were also included as predictor inputs. Demographics consisted of race, ethnicity, age, education, gender, and marital status. Clinical measures included total scores and sub-scores for several subject-reported forms and psychiatrist-administered assessments (**Table S1**). These assessments include measures of pre-treatment functioning, depression symptomatology, suicide risk comorbidities, and family psychiatric history, among other factors.

Deep Learning Model Training and Hyperparameter Optimization

Feed-forward neural networks (**Figure S2**) were constructed to take parcellated contrast map values (and clinical features for sertraline and placebo) as inputs and return the predicted ΔHAMD. To mitigate overfitting, in addition to using the previously described data augmentation, the models were rigorously regularized with L1 and L2 weight regularization, batch normalization, and dropout layers. *Hyperparameters* defining the model architecture, such as number of layers, number of neurons per layer, learning rate, regularization strength, and dropout rate were optimized using a random search. Random search has shown to be an unbiased, highly efficient method for quickly identifying optimal hyperparameters with low computational overhead[47,48]. With an appropriately defined hyperparameter search space, a random search will find a high-performing model architecture without dependency on the expertise of the investigator. Five hundred hyperparameter configurations were sampled randomly from uniform distributions over predefined hyperparameter ranges (**Table S3**). These 500 configurations were then combined with each of the three brain atlas versions (100, 200, and 400 ROIs) to create 1500 total candidate models, which were each evaluated for each of the 3 treatments. The same set of 1500 candidate models was used for all treatment groups. The models were implemented in the Keras and Tensorflow packages and trained using Nvidia Tesla P100 GPUs on the BioHPC computing cluster at UT Southwestern. Models were trained

with the Nadam optimizer, with learning rate and decay included as hyperparameters in the random search. The loss function was designed to maximize R^2 as was defined as:

$$\ell = 100\big(1 - f_{R^2}(y, \hat{y})\big) + \lambda\Omega(\theta)$$

where $f_{R^2}(\cdot)$ computes R^2 between the true output values y and the predicted outputs $\hat{y}$ and $\lambda\Omega(\theta)$ is the weight regularization term. The coefficient of 100 was chosen empirically to keep the magnitudes of the first and second terms of the loss function within similar ranges.

The predictive performance of each candidate model was validated using nested cross-validation[49]. The data was first split into 3 outer cross-validation folds, stratified by ΔHAMD to ensure representative distributions of subjects in each fold. The training data of each fold was then split again into 5 inner cross-validation folds, which were used to evaluate the performance of each candidate model. For each outer fold, the model with the best R^2 across the inner folds was selected, retrained on all inner-fold data of that outer fold, and evaluated on the held-out outer fold data. The mean performance of the best model from each outer fold is reported in the results of this work.

After completing this random search, plots of performance vs. hyperparameter values were visualized to ensure that local maxima of performance had been observed. This verified that sufficiently large hyperparameter ranges had been searched to identify a high-performing model.

Permutation Testing of Model Performance

The statistical significance of the model performance results was measured using permutation testing, which tests the null hypothesis that the model did not learn the association between the data and the prediction target[50]. In this approach, a null distribution is generated by permuting the target labels, i.e. ΔHAMD in this study. Specifically, the labels were randomly permuted 500 times and the model was refit and evaluated each time. The *p*-value for each performance metric was obtained by computing the cumulative density function of the null distribution at the actual model performance.

Feature Importance

For the best performing models of each random search, important learned features were identified using permutation feature importance[18]. This method leaves the labels undisturbed while each individual feature is permuted among the subjects, ablating any useful information in that feature, and the change in model performance (R^2) was measured. This process is then repeated for each feature. Features that incur a greater decrease in performance when permuted are more important for the model's prediction[18]. Within each outer cross-validation fold, features were ranked by importance, and the mean importance rank over all folds was used to identify the most important features learned by the model. The top 30 features are reported because an inflection point was observed at 30 features, with subsequent features having much more similar feature importance.

Computation of Number-Needed-to-Treat

In this work, the number-needed-to-treat (NNT) is defined as the number of individuals that must be screened by a predictive model to identify one additional remitter or responder, compared to the overall remission or response rate of the treatment in this study:

$$NNT = \frac{1}{r_e - r_c}$$

For example, to compute the NNT for predicting remission, the experimental event rate r_e is the true remission rate in the subjects predicted by the model to remit:

$$r_e = \frac{\#\text{ true remitters}}{\#\text{ predicted remitters}}$$

And the control event rate r_c is the overall remission rate of the treatment group in the study:

$$r_c = \frac{\#\text{ remitters}}{\#\text{ subjects in treatment group}}$$

Additionally, a second NNT can be defined as the number of individuals that must be screened to identify one additional remitter or responder, relative to a clinician's performance in making the same treatment selection decisions. One source estimates the typical antidepressant response rate in clinical practice to be about 45%[20]. This can be used to define

$$NNT_{clin} = \frac{1}{r_e - r_{clin}}$$

where the control event rate is now

$$r_{clin} = 45\%$$

DATA AVAILABILITY

All data from EMBARC are publicly available from the NIMH Data Archive at: https://nda.nih.gov/edit_collection.html?id=2199.

ACKNOWLEDGEMENTS

The EMBARC study was supported by the National Institute of Mental Health of the National Institutes of Health under award numbers U01MH092221 (to M.H.T.) and U01MH092250 (to P.J.M., R.V.P., and M.W.). Valeant Pharmaceuticals donated the Wellbutrin XL used in the study. This work was supported by the EMBARC National Coordinating Center at UT Southwestern Medical Center, with M.H.T. as Coordinating PI, and the Data Center at Columbia and Stony Brook Universities.

ClinicalTrials.gov identifier: NCT01407094.

AUTHOR CONTRIBUTIONS

K.P.N, C.C-F., M.H.T., and A.M. contributed to formulation of the research problem, experimental design, and initial manuscript writing. C.C-F., C.C., M.J., B.K., M.F., P.J.M, P.A., R.V.P., M.W., and M.P., and M.H.T conducted the clinical study design and data collection. K.P.N., C.C-F., A.T., C.M., and A.M. contributed to image preprocessing, data analysis, and deep learning methodology. All work was supervised principally by M.H.T. and A.M. All authors contributed to the completion of the final manuscript.

REFERENCES

1. Kessler, R.C. *et al.* The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* **289**, 3095–3105 (2003).

2. Greenberg, P.E., Fournier, A.-A., Sisitsky, T., Pike, C.T. & Kessler, R.C. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *The Journal of clinical psychiatry* **76**, 155–162 (2015).

3. Dupuy, J.M., Ostacher, M.J., Huffman, J., Perlis, R.H. & Nierenberg, A.A. A critical review of pharmacotherapy for major depressive disorder. *The international journal of neuropsychopharmacology* **14**, 1417–1431 (2011).

4. Rush, A.J. *et al.* Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *The American journal of psychiatry* **163**, 1905–1917 (2006).

5. Phillips, M.L. *et al.* Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: neuroimaging approaches. *The American journal of psychiatry* **172**, 124–138 (2015).

6. Lener, M.S. & Iosifescu, D.V. In pursuit of neuroimaging biomarkers to guide treatment selection in major depressive disorder: a review of the literature. *Annals of the New York Academy of Sciences* **1344**, 50–65 (2015).

7. Greenberg, T. *et al.* Moderation of the Relationship Between Reward Expectancy and Prediction Error-Related Ventral Striatal Reactivity by Anhedonia in Unmedicated Major Depressive Disorder: Findings From the EMBARC Study. *The American journal of psychiatry* **172**, 881–891 (2015).

8. Chin Fatt, C. *et al.* Effect of Intrinsic Patterns of Functional Brain Connectivity in Moderating Antidepressant Treatment Response in Major Depression. *The American journal of psychiatry* (2019).

9. Davidson, R.J., Irwin, W., Anderle, M.J. & Kalin, N.H. The neural substrates of affective processing in depressed patients treated with venlafaxine. *The American journal of psychiatry* **160**, 64–75 (2003).

10. Chen, C.-H. *et al.* Brain imaging correlates of depressive symptom severity and predictors of symptom improvement after antidepressant treatment. *Biological Psychiatry* **62**, 407–414 (2007).

11. Canli, T. *et al.* Amygdala reactivity to emotional faces predicts improvement in major depression. *Neuroreport* **16**, 1267–1270 (2005).

12. Greenberg, T. *et al.* Reward related ventral striatal activity and differential response to sertraline versus placebo in depressed individuals. *Molecular psychiatry* (2019).

13. Robertson, B. *et al.* Effect of Bupropion Extended Release on Negative Emotion Processing in Major Depressive Disorder. *J. Clin. Psychiatry* **68**, 261–267 (2007).

14. Trivedi, M.H. *et al.* Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. *Journal of psychiatric research* **78**, 11–23 (2016).

15. Chase, H.W. *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* **10**, e0126326 (2015).

16. M. R. Delgado, L. E. Nystrom, C. Fissell, D. C. Noll & and J. A. Fiez. Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum.

17. Max Hamilton. A Rating Scale for Depression. *Journal of Neurology, Neurosurgery, and Psychiatry* **23**, 56–62 (1960).

18. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics (Oxford, England)* **26**, 1340–1347 (2010).

19. Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–289 (2002).

20. Roose, S.P., Rutherford, B.R., Wall, M.M. & Thase, M.E. Practising evidence-based medicine in an era of high placebo response: number needed to treat reconsidered. *The British journal of psychiatry : the journal of mental science* **208**, 416–420 (2016).

21. Etkin, A. *et al.* A cognitive-emotional biomarker for predicting remission with antidepressant medications: a report from the iSPOT-D trial. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **40**, 1332–1342 (2015).

22. Gordon, E., Rush, A.J., Palmer, D.M., Braund, T.A. & Rekshan, W. Toward an online cognitive and emotional battery to predict treatment remission in depression. *Neuropsychiatric disease and treatment* **11**, 517–531 (2015).

23. Fischer, A.G. & Ullsperger, M. An Update on the Role of Serotonin and its Interplay with Dopamine for Reward. *Frontiers in human neuroscience* **11**, 484 (2017).

24. Bearer, E.L., Zhang, X., Janvelyan, D., Boulat, B. & Jacobs, R.E. Reward circuitry is perturbed in the absence of the serotonin transporter. *NeuroImage* **46**, 1091–1104 (2009).

25. Kitaichi, Y. *et al.* Sertraline increases extracellular levels not only of serotonin, but also of dopamine in the nucleus accumbens and striatum of rats. *European journal of pharmacology* **647**, 90–96 (2010).

26. Forbes, E.E. *et al.* Reward-related decision-making in pediatric major depressive disorder: an fMRI study. *Journal of child psychology and psychiatry, and allied disciplines* **47**, 1031–1040 (2006).

27. Whitton, A.E. *et al.* Pretreatment Rostral Anterior Cingulate Cortex Connectivity With Salience Network Predicts Depression Recovery: Findings From the EMBARC Randomized Clinical Trial. *Biological Psychiatry* **85**, 872–880 (2019).

28. Chau, D.T., Roth, R.M. & Green, A.I. The neural circuitry of reward and its relevance to psychiatric disorders. *Curr Psychiatry Rep* **6**, 391–399 (2004).

29. Morgane, P.J., Galler, J.R. & Mokler, D.J. A review of systems and networks of the limbic forebrain/limbic midbrain. *Progress in neurobiology* **75**, 143–160 (2005).

30. Werner, N.S. *et al.* Functional MRI study of memory-related brain regions in patients with depressive disorder. *Journal of affective disorders* **119**, 124–131 (2009).

31. Lui, S. *et al.* Depressive disorders: focally altered cerebral perfusion measured with arterial spin-labeling MR imaging. *Radiology* **251**, 476–484 (2009).

32. Jha, M.K. *et al.* Validating pre-treatment body mass index as moderator of antidepressant treatment outcomes: Findings from CO-MED trial. *Journal of affective disorders* **234**, 34–37 (2018).

33. Khan, A., Schwartz, K.A., Kolts, R.L. & Brown, W.A. BMI, sex, and antidepressant response. *Journal of affective disorders* **99**, 101–106 (2007).

34. Flament, Martime F., Lane, M. F. Acute antidepressant response to fluoxetine and sertraline in psychiatric outpatients with psychomotor agitation. *International journal of psychiatry in clinical practice* **5**, 103–109 (2001).

35. Bondareff, W. *et al.* Comparison of sertraline and nortriptyline in the treatment of major depressive disorder in late life. *The American journal of psychiatry* **157**, 729–736 (2000).

36. Forbes, E.E. *et al.* Reward-related brain function as a predictor of treatment response in adolescents with major depressive disorder. *Cognitive, affective & behavioral neuroscience* **10**, 107–118 (2010).

37. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V. & Le V, Q. AutoAugment: Learning Augmentation Policies from Data. In *Computer Vision and Pattern Recognition* (California, USA, 2019).

38. Smith, S.M. Fast robust automated brain extraction. *Human brain mapping* **17**, 143–155 (2002).

39. Cox, R.W. AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research* **29**, 162–173 (1996).

40. Avants, B.B. *et al.* A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration. *NeuroImage* **54**, 2033–2044 (2010).

41. Iglesias, J.E., Liu, C.-Y., Thompson, P.M. & Tu, Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging* **30**, 1617–1634 (2011).

42. Klein, A. *et al.* Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage* **46**, 786–802 (2009).

43. Calhoun, V.D. *et al.* The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. *Human brain mapping* **38**, 5331–5342 (2017).

44. Dohmatob, E., Varoquaux, G. & Thirion, B. Inter-subject Registration of Functional Images: Do We Need Anatomical Images? *Frontiers in Neuroscience* **12** (2018).

45. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral cortex (New York, N.Y. : 1991)* **28**, 3095–3114 (2018).

46. Craddock, R.C., James, G.A., Holtzheimer, P.E., Hu, X.P. & Mayberg, H.S. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping* **33**, 1914–1928 (2012).

47. Bergstra, J. & Bengio, Y. Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).

48. Li, L. & Talwalkar, A. Random Search and Reproducibility for Neural Architecture Search. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence* (eds. Globerson, A. & Silva, R.) 129 (AUAI Press, Pittsburgh, 2019).

49. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics* **7**, 91 (2006).

50. Ojala, M. & Garriga, G.C. Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research* **11**, 1833–1863 (2010).