Patterns of Pre-Treatment Reward Task Brain Activation Predict Individual Antidepressant Response

Supplemental Information

APPENDIX I: SUPPLEMENTAL METHODS

I.1. Study design

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 participants. Participants were antidepressant-naïve in the current episode, must have met Structured Clinical Interview for the DSM-IV (SCID) criteria for Major Depressive Disorder (MDD), and scored ≥ 14 on the Quick Inventory of Depressive Symptomatology (QIDS-SR). Additionally, to reduce heterogeneity, participants must have had early onset (before age 30), chronic (episode duration > 2 years), or recurrent (2+ episodes) disease. 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, participants must not have had a lifetime history of psychosis, bipolar disorder, or epilepsy and must not be receiving depression-specific psychotherapy or somatic treatments.

Antidepressant treatments began at a fixed dosage, which were increased at each visit depending on participant tolerance and response. Sertraline dosage began at 50 mg daily and was titrated up to 200 mg or maximum tolerated dose, or until response. Bupropion (Wellbutrin XL, extended release formulation) dosage began at 150 mg daily and was titrated up to 450 mg or maximum tolerated dose, or until response.

I.2. Reward task paradigm

The monetary reward task (**Fig. S2**) is motivated by differential reactivity to reward anticipation and prediction error, depending on brain region, which has been identified between healthy and depressed participants. Each trial of the task begins with the *response phase*, during which the participant guesses whether an upcoming number, with possible values of 1-9, will be greater or less than 5. During the *anticipation phase*, the participant is informed about the possible outcome of the current trial. Trials can be “possible win”, where the participant wins $1 for a correct guess and loses nothing for a wrong guess, or “possible loss”, where the participant 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 participant has won money, lost money, or did not win or lose any money. This is followed by a fixation period (*baseline phase*) before the next trial. A total of 24 trials were conducted, with 12 “possible win” and 12 “possible loss” trials. All participants received a fixed monetary reward after the task regardless of outcome.

I.3. 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 functional magnetic resonance imaging (fMRI) acquisitions. Data augmentation aims to improve the accuracy of deep learning models by applying random transformations to real data samples to simulate additional samples. Such techniques have been used in both non-medical and medical image applications where they achieve substantial improvements in predictive model performance (1; 2). This study employed the BLENDS method of fMRI data augmentation, which applies random yet anatomically-realistic perturbations to brain shape (3). This simulates the situation in which a new subject was included in training that had the same functional activity as the original subject but possesses a different brain shape. First, each of the existing fMRI was nonlinearly coregistered to the MNI152 brain template using the symmetric normalization (SyN) algorithm in ANTs. This creates a set of template-to-image nonlinear warps which represent the distribution of brain morphologies present in the dataset. Next, to simulate new fMRI samples, each original fMRI was coregistered and transformed to MNI152 template space. Four template-to-image warps were randomly selected and spatially blended to create a new warp. This warp was then applied to the original fMRI to simulate a new sample. This process was repeated 10 times for each original fMRI to simulate 10 new samples per original sample. The clinical features and treatment outcome (ΔHAMD) were unmodified in the new simulated samples.

This augmentation increased the effective sample sizes by 10X, providing a total of 1060 samples for sertraline, 1160 for placebo, and 370 for bupropion. Importantly, this augmented data was used only during model training and not during evaluation. **Table S4** compares the predictive performance achieved with and without data augmentation and demonstrates the benefit of augmentation.

I.4. MRI preprocessing

All original and augmented data was preprocessed as follows. Structural MRI (sMRI) were first processed with the ROBEX tool (4) 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 SyN registrations in ANTs. This registration method was selected as it has been shown to outperform other registration methods (5; 6). The normalized sMRI was then segmented into gray matter, white matter, and cerebrospinal fluid (CSF) with FSL FAST. Functional MRI were corrected for frame-to-frame head motion with FSL MCFLIRT, and outlier frames were selected using Nipype's RapidArt to be regressed out during GLM analyses. Frames were marked as outliers if intensity was > 3 standard deviations from the mean or if the magnitude of head motion was > 1.0 mm. This threshold was selected based on published recommendations for task-based fMRI (7), however even more vigorous motion suppression was also tested (see Supplement section: Accounting for motion**)**. Brain extraction was performed using the EPI brain extraction method from fMRIPrep, which applies FSL BET and AFNI 3dAutomask and takes the intersection of the two segmentations (8). Next, spatial normalization was conducted using a direct EPI-based normalization, where the mean functional image frame was directly registered to the Montreal Neurological Institute (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 (9; 10).

Motion-related artifacts were suppressed using ICA-AROMA (11), and mean CSF and white matter signals were regressed out to reduce physiological noise. Finally, the images were spatially smoothed with a 6 mm Gaussian filter.

I.5. Contrast map computation

Individual-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 (12; 13). 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 expectancy and prediction error. The reward expectancy 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 applied to mask out unimportant voxels from the analysis and avoid the influence of physiological noise in the measured values. The GLM was fitted:

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

Where $\boldsymbol{Y}$ is the time × voxels data matrix containing the voxel timeseries, $\boldsymbol{X}$ is the time × regressors design matrix containing the regressor timeseries, $\boldsymbol{\beta}$ is the regressors × 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 expectancy and prediction error contrast maps were simply $\beta_{\text{reward expectancy}}$ and $\beta_{\text{prediction error}}$ respectively.

I.6. Computation of regional contrast values with a study-specific atlas

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 (Schaefer 2018) generated from healthy participants (14). A study-specific brain atlas with 200 regions-of-interest (ROIs) was generated from pre-treatment resting-state fMRI images of 283 MDD participants using the spatially-constrained spectral clustering method successfully developed by Craddock et al (15). The anatomical label for each ROI was determined by finding the corresponding anatomical structure with the greatest Dice overlap in the widely-used Automated Anatomical Labeling atlas (16). For each contrast map, including anticipation, reward expectancy, 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 participant.

I.7. Site effect correction

Treatment outcomes were found to differ among the 4 study sites. In particular, the mean 8-week ΔHAMD (over all treatment arms) was significantly different among the sites (one-way ANOVA, F = 5.848, p = 0.001) with one site (CU) showing a larger symptomatic change

(**Fig. S3a**) than the remaining sites. Response (F = 18.633, p = 0.0003) and remission (F = 21.029, p = 0.0001) rates also differed among sites (**Fig. S3b**). Due to these differences, explicit steps were taken to quantify and mitigate any confounding effects of site in the predictive models. Including confounding variables directly as inputs to a machine learning model can cause overfitting (17), and consequently a retrospective approach was taken to check the models for confounding effects. First, the predictive features presented in the main text (**Fig. 1, 2, and 3**), which were the 30 most important predictive features for each model, were examined for confounding effect of site. *It was found that none of predictive features showed significant differences among sites* (one-way ANOVA) at p = 0.05. Next, the ComBat confound suppression method was applied to remove any confounding effect of site in the regional contrast imaging features. ComBat has been demonstrated to correct site effects in fMRI data while preserving associations of interest and has been previously and extensively validated on EMBARC (18). ComBat was applied with the recommended parameters and age, gender, baseline HAMD, and ΔHAMD were specified as the preserved covariates. One-way ANOVAs were used to test for site effects in the imaging features before and after ComBat (**Fig. S3c**). The predictive models described in the main text were re-evaluated on this ComBat-corrected data. *It was found that performance did not significantly change (p > 0.05) compared to the non-corrected data: R^2 was 33.2% for sertraline, 25.9% for placebo, and 41.9% for bupropion.* These results demonstrate that 1) the most important predictive features learned by these models were not confounded by site and 2) any minor site effect removed by ComBat did not play a role in model predictive accuracy.

I.8. Accounting for motion

As described in the previous sections, several forms of motion artifact correction were incorporated into the preprocessing pipeline (ICA-AROMA and white matter and CSF signal regression) and into the GLM (head motion parameters). The following analysis was performed to determine whether motion-related signals impacted the construction of predictive models. *There was no association found between treatment outcome and head motion,* with motion quantified using the mean framewise displacement (mFD) metric defined by Power et al (19). Correlation between ΔHAMD and mFD was insignificant (Pearson $r = 0.042$, p = 0.524). *There was no significant difference in mean mFD between remitters and non-remitters (T = -0.109, p = 0.913) nor between responders and non-responders (T = 0.228, p = 0.820)* (**Fig. S4a-b**). Examining the top predictive features presented in the main text, only one feature learned by the sertraline model (**Fig. 1**) was correlated with mFD. Anticipation activation in the left superior frontal gyrus was significantly correlated with mFD (p = 0.049). However, given there was no association between head motion and treatment outcome, this did not contribute to the model's predictions.

More aggressive outlier frame scrubbing was also tested, with the motion threshold decreased from 1.0 mm to 0.5 mm. Model performance on this data was not significantly different (p < 0.05) from the results presented in the main text using the 1.0 mm threshold: R^2 was 32.7% for sertraline, 20.1% for placebo, and 22.6% for bupropion.

I.9. Deep learning model training and hyperparameter optimization

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 (**Fig. S5**), such as number of layers, number of neurons per layer, learning rate, regularization strength, and dropout rate were optimized using Bayesian optimization (BO) (20). This was implemented in Ray Tune using the Scikit-optimize backend. The BO was allowed to evaluate 100 candidate hyperparameter configurations within each cross-validation fold. The predefined hyperparameter ranges are given in **Table S3**. 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 using the Nadam optimizer, with learning rate as a hyperparameter during BO, to minimize the mean squared error loss. Input features were normalized and scaled to zero mean unit variance.

The predictive performance of each candidate model was validated using 20x20 nested cross-validation (21; 22). The data was first split into 20 outer cross-validation folds, stratified by 8-week change in Hamilton Rating Scale for Depression score (ΔHAMD) to ensure representative distributions of participants in each fold. The training data of each fold was then split again into 20 inner cross-validation folds, which were used to evaluate the performance of each candidate hyperparameter configuration during BO. For each outer fold, the hyperparameter configuration with the lowest root mean squared error (RMSE) across the inner folds was selected, retrained on all inner-fold data of that outer fold, and used to predict on the held-out outer fold data. These predictions on held-out data not seen during training or BO were used to compute the final performance reported in the results.

In addition to the 20x20 nested cross-validation used in the main results, a 20x20 Monte Carlo cross-validation strategy was also tested to further validate the performance estimates. A similar cross-validation strategy was used by Wu et al. to validate their EEG-based treatment outcome prediction models (23). The data was split into 20 cross-validation folds stratified by ΔHAMD. BO was conducted and the hyperparameter configuration with the lowest RMSE on validation data across the folds was selected. This was repeated 20 times with different random shuffling of the data before splitting. Mean performance across the 20 repetitions is reported in **Table S5**. A similar performance was observed to that achieved with the nested cross validation approach used in the main text.

I.10. 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 participants 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:

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

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. The typical antidepressant response rate in clinical practice is estimated to be about 45% (24). This can be used to define

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

where the control event rate is now

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

NNT_{clin} for the 3 predictive models is reported in Appendix II.

I.11. Permutation testing

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 (25). 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 100 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.

I.12. Feature importance

The importance of each feature in forming predictions was quantified by computing the Jacobian, i.e. the partial derivative of the model output with respect to each model input (26; 27). The magnitude of this importance measure indicates the sensitivity of the model output to changes in a particular feature's values, while the sign indicates the direction in which the model output changes when the feature's value increases. For ease of interpretability in this analysis, the signs were negated such that positive importance indicates greater predicted improvement in HAMD with higher feature values and negative importance indicates lesser predicted improvement in HAMD with higher feature values.

These importance measures were computed for the final trained model in each outer fold. The mean importance of each feature over the outer folds is presented in the main text results (**Figs. 1-3**).

APPENDIX II: SUPPLEMENTAL RESULTS

II.1. Ablation experiments

To determine whether 1) the combination of imaging with clinical/demographic features and 2) fMRI data augmentation were both necessary to achieve the observed performance, ablation experiments were conducted (**Table S4**). For each ablation, the same hyperparameter optimization, model training, and cross-validation were used as in the main results. With only augmented imaging features and no clinical features, R^2 reduced to 12% (one-tailed test of correlations, p = 0.0002) for sertraline and 18% for placebo (p = 0.153). For bupropion, R^2 increased slightly to 38% but the difference was not significant (p = 0.585). Removing data augmentation or using only clinical or only non-augmented imaging features resulted in poor performance ($R^2 < 0$).

II.2. Value of deep learning methods over traditional statistical and classical machine learning approaches

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 the deep learning approach described above.

To further evaluate the need for deep learning models, several other multivariate regression methods were compared to the recommended deep learning approach. An elastic net model was tested, serving as a baseline linear model. Hyperparameters were optimized with a random search over 100 configurations and performance on held-out data was evaluated using the same approach that was used on the deep learning models. The same 10x data augmentation was also applied. No models were able to explain any variance (positive R^2) for any treatment group. Other classical machine learning models including K-nearest neighbors, support vector machine, and random forest were also tested with similar results. Compared to the deep learning models, these models were unable to learn to predict treatment outcome from the data with high accuracy.

II.3. Alternate computation of Number-Needed-to-Treat

In the main text, the reported NNT values are computed relative to the actual remission or response rates in each treatment group of the study (see Appendix I, Computation of Number-needed-to-treat). Because treatment assignment was randomized in this study, this NNT may be less relevant to real-world clinical practice. A second metric, $\mathrm{NNT}_{\mathrm{clin}}$, was 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 (24), $\mathrm{NNT}_{\mathrm{clin}}$ was 4.35 for sertraline and 1.82 for bupropion. This indicates that a clinician would need to screen about 4 individuals using the predictive models to identify one additional individual who could be treated with sertraline or bupropion and achieve response, compared to current clinician decision-making.

II.4. Examination of 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, were unable to provide predictive power (**Table S4**). Additionally, using imaging features alone provided low predictive performance for sertraline and placebo, even with data augmentation (R^2 12-18%, **Table S4**).

For the sertraline model, several clinical measurements were found to be highly important features for predicting treatment outcome (**Fig. 1** in Main Text). Psychomotor agitation was the most important predictor of improvement learned by the model. 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 participants (28; 29). Pre-treatment HAMD score was also highly important, with a higher total score on the either the 17-item or 24-item versions predicting greater improvement. Family history of suicide, comorbidities (SCQ total score) and older age of first dysphoric or depressive episode predicted less improvement.

Different clinical and demographic features were learned by the placebo model (**Fig. 2** in Main Text). Concurrent panic disorder, hypersomnia, and older age at evaluation predicted less improvement. Older age has previously been connected with lower remission rates, though this is believed to be due to medical comorbidities rather than age itself (30). Separated marital status and Asian race were both learned as predictors of greater improvement, but this may be an artifactual finding given that only 3% of the placebo group (3 participants) were separated and 7% (8 participants) were Asian.

Examining the bupropion model (**Fig. 3** in Main Text), higher education level was a top predictor of greater improvement. This association has been previously reported, though not specifically for bupropion (31; 32). Family history of mental illness also predicted greater improvement. Anxious distress predicted less improvement, which mirrors previous findings on other antidepressants (33; 34).

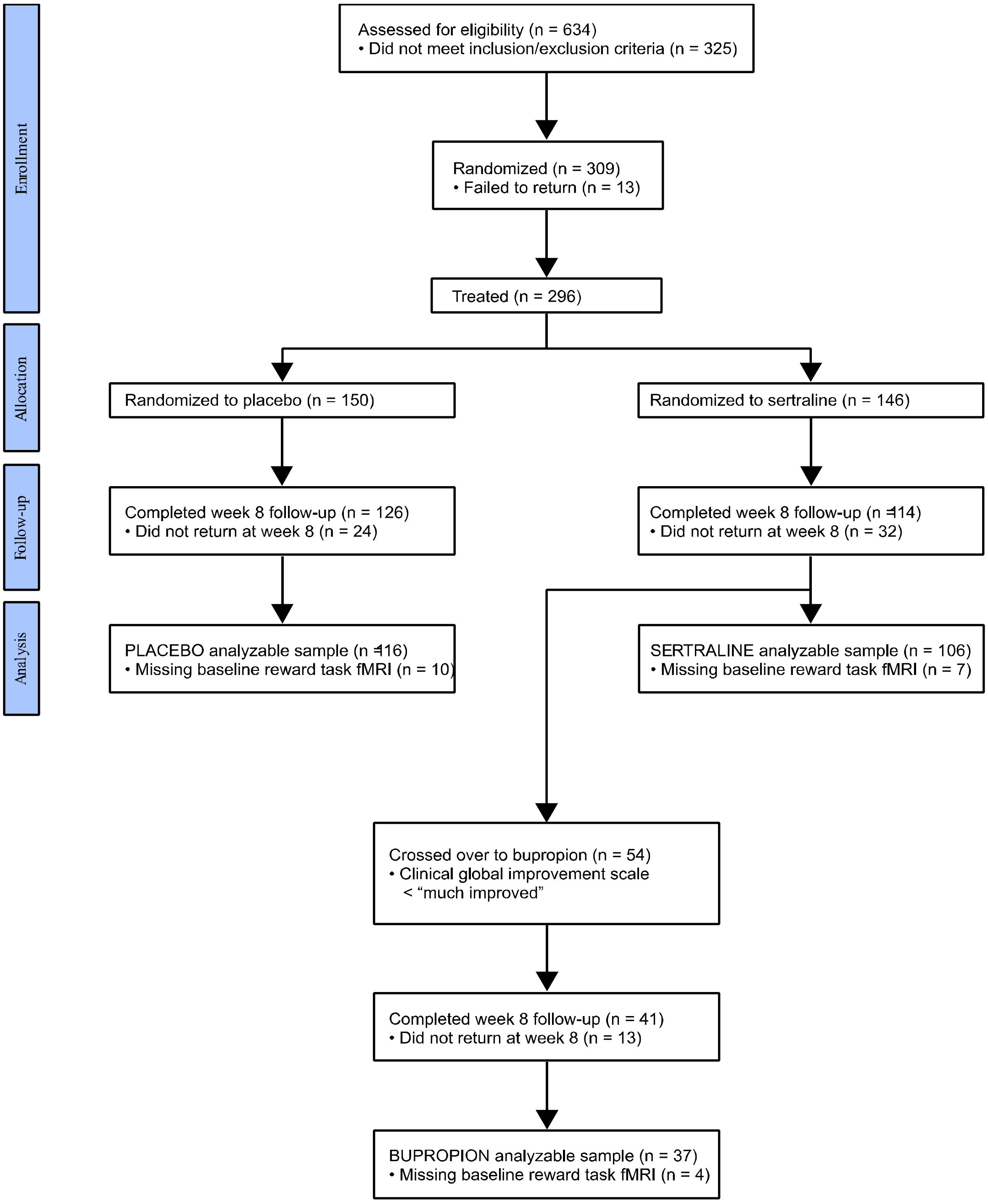
Enrollment
Allocation
Follow-up
Analysis
Assessed for eligibility (n = 634)
• Did not meet inclusion/exclusion criteria (n = 325)
Randomized (n = 309)
• Failed to return (n = 13)
Treated (n = 296)
Randomized to placebo (n = 150)
Randomized to sertraline (n = 146)
Completed week 8 follow-up (n = 126)
• Did not return at week 8 (n = 24)
Completed week 8 follow-up (n =114)
• Did not return at week 8 (n = 32)
PLACEBO analyzable sample (n =116)
• Missing baseline reward task fMRI (n = 10)
SERTRALINE analyzable sample (n = 106)
• Missing baseline reward task fMRI (n = 7)
Crossed over to bupropion (n = 54)
• Clinical global improvement scale
< "much improved"
Completed week 8 follow-up (n = 41)
• Did not return at week 8 (n = 13)
BUPROPION analyzable sample (n = 37)
• Missing baseline reward task fMRI (n = 4)

Figure S1. CONSORT diagram of study participants.

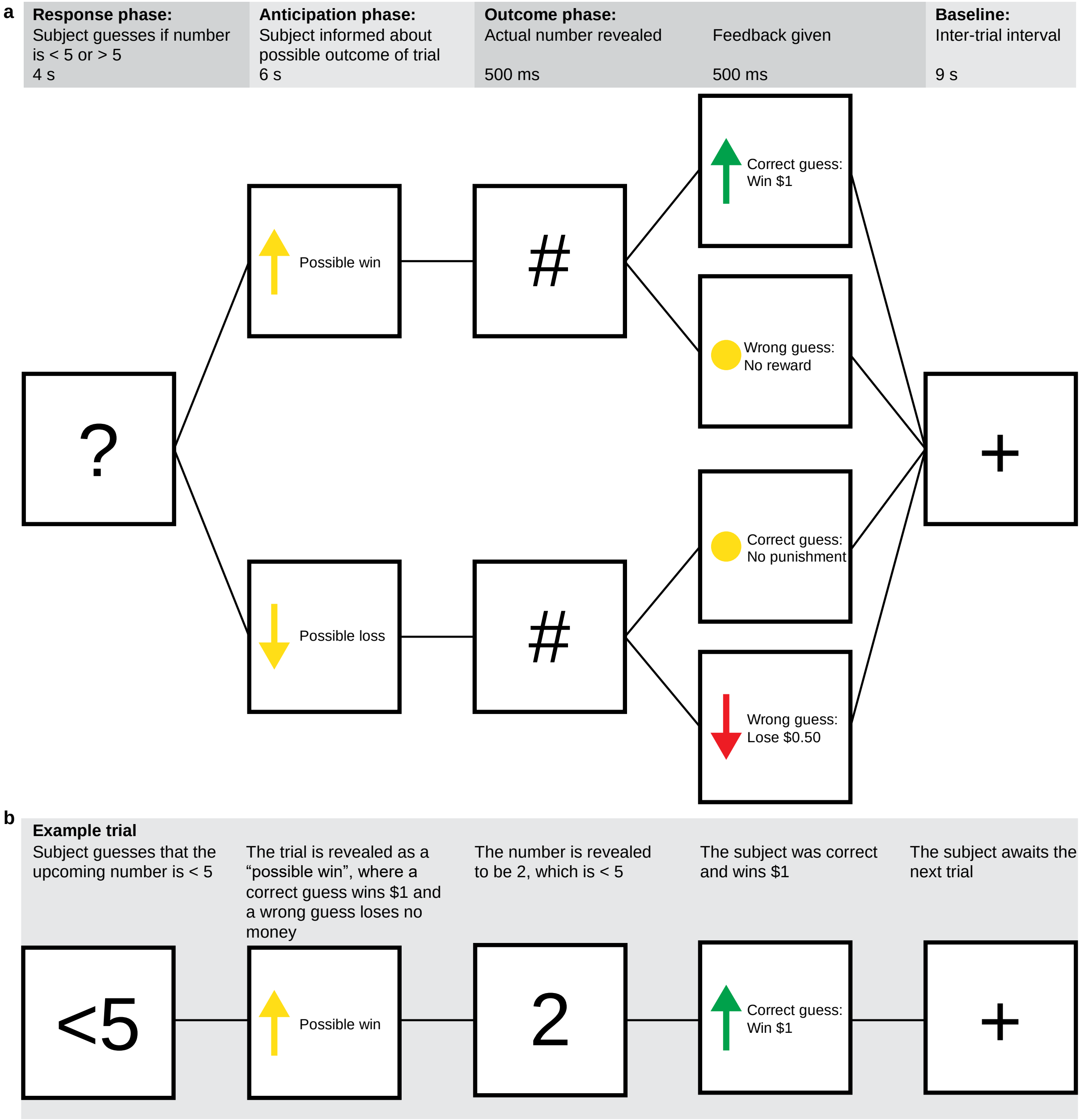
a
Response phase:
Subject guesses if number is < 5 or > 5
4 s
Anticipation phase:
Subject informed about possible outcome of trial
6 s
Outcome phase:
Actual number revealed
500 ms
Feedback given
500 ms
Baseline:
Inter-trial interval
9 s
?
Possible win
Possible loss
#
#
Correct guess: Win $1
Wrong guess: No reward
Correct guess: No punishment
Wrong guess: Lose $0.50
+
b
Example trial
Subject guesses that the upcoming number is < 5
The trial is revealed as a “possible win”, where a correct guess wins $1 and a wrong guess loses no money
The number is revealed to be 2, which is < 5
The subject was correct and wins $1
The subject awaits the next trial
<5
Possible win
2
Correct guess: Win $1
+

Figure S2. 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 participant 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 participant guesses that the upcoming number is less than 5, and the trial is a “possible win”. The actual number is 2, and the participant receives $1 for a correct guess.

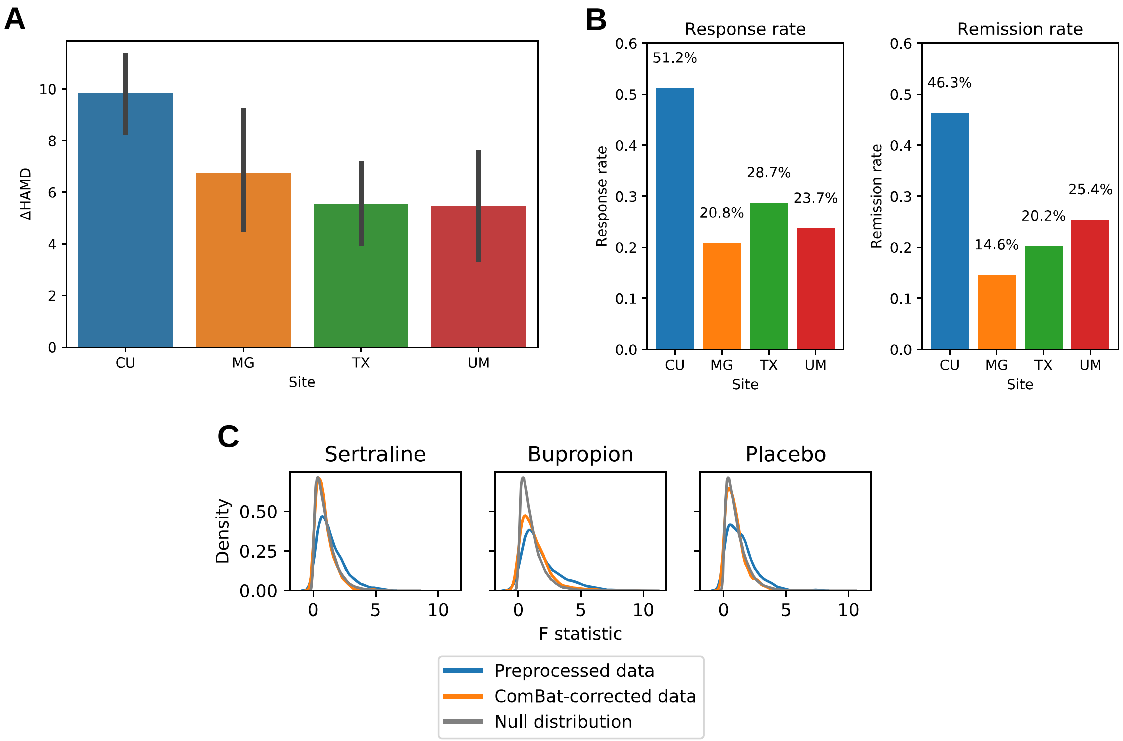

Figure S3. Site differences in treatment outcome and imaging features. Study sites included Columbia University (CU), Massachusetts General Hospital (MG), University of Texas Southwestern Medical Center (TX), and University of Michigan (UM). a) Mean ΔHAMD with 95% confidence intervals for each site. b) Response and remission rates for each site. c) For each treatment group, F-statistics were computed for all imaging features in the original preprocessed data (blue) and after ComBat correction (orange), and kernel density estimates of the distributions are shown. For comparison, a null distribution (grey) was generated by randomly permuting the site labels across participants 100 times.

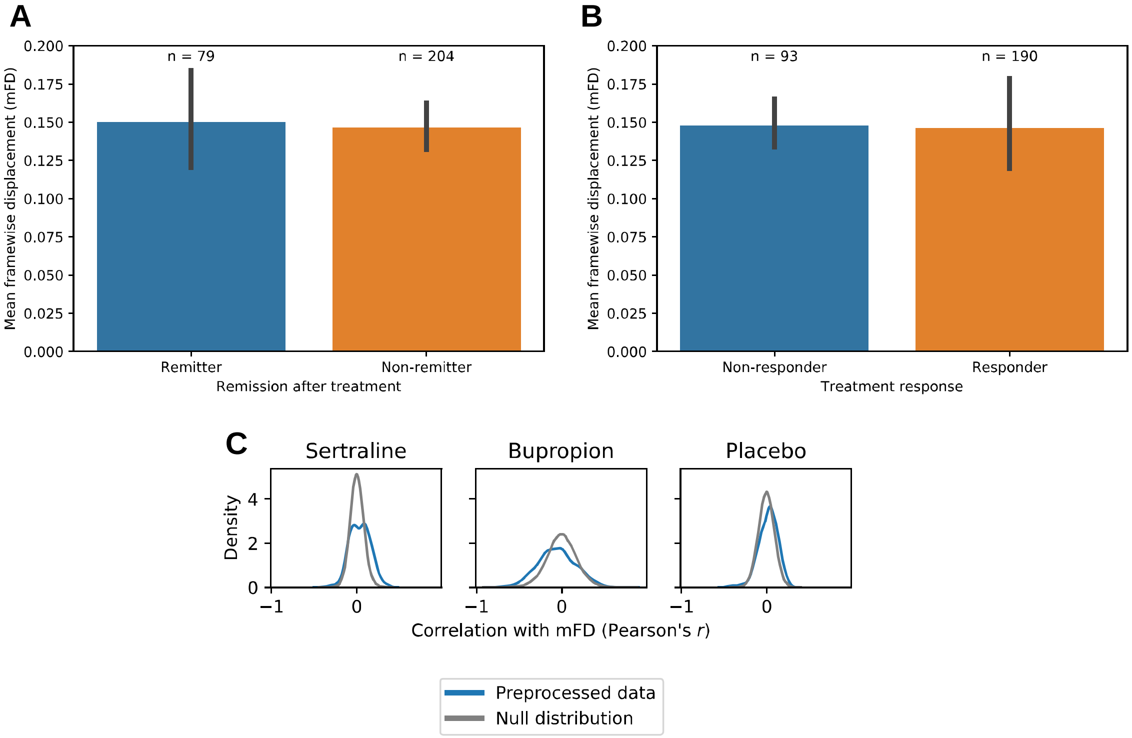

Figure S4. Effect of motion in treatment outcome and imaging features. Magnitude of head motion during fMRI acquisition was computed using the mean framewise displacement (mFD), as defined by Power et al. 2012. Mean mFD (with 95% confidence intervals) did not differ between a) remitters and non-remitters or b) responders and non-responders. c) For each treatment group, correlations across participants were computed between mFD and each imaging feature in the preprocessed data (blue). For comparison, a null distribution (grey) was generated by permuting mFD values across participants 100 times and recomputing the correlations.

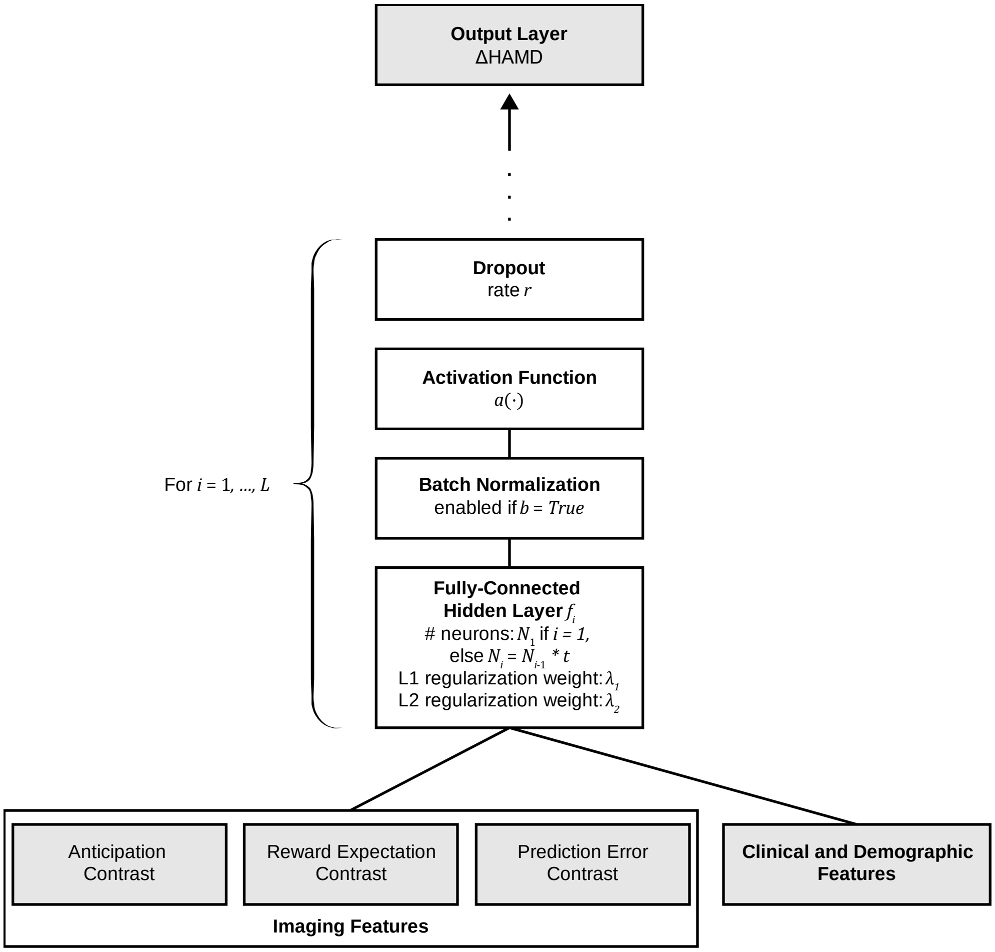

Figure S5. Schematic for the feed-forward neural networks developed in this work. Hyperparameters are indicated for each layer and were optimized using a random search for each treatment. Ranges of hyperparameters that were searched are listed in **Table S3**. Inputs to these models included imaging features, extracted from the contrast maps for each of the three task conditions, and clinical and demographic features for the sertraline and placebo models. This data is fed through a series of fully-connected hidden layers f_i for $i = 1, \dots, L$, and the number of layers L was optimized during the random search. Regularization parameters λ_1 and λ_2, use of batch normalization b, dropout rate r, and activation function $a(\cdot)$ were included as optimized hyperparameters. The final output layer returns the ΔHAMD prediction.

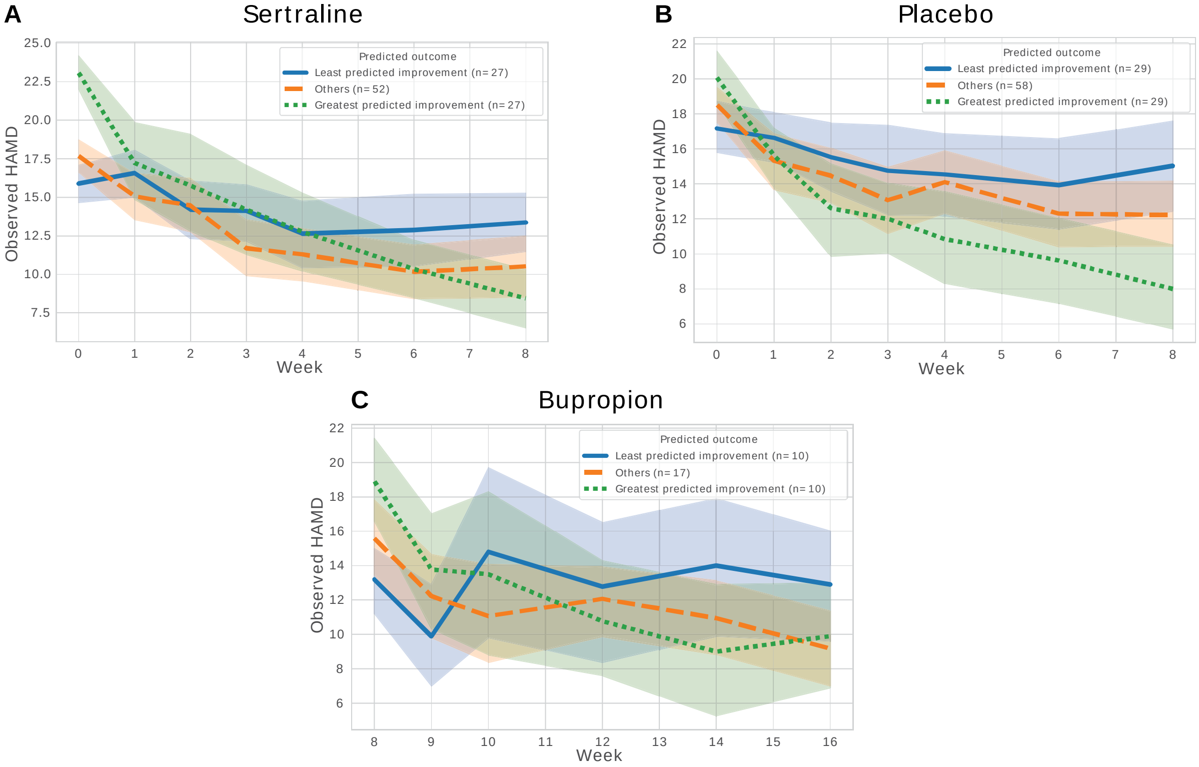

Figure S6. Association between predicted treatment outcomes and patterns of clinical improvement (unnormalized). For each treatment group—a) sertraline, b) placebo, and c) bupropion—participants were ranked by predicted treatment outcome. Participants were then grouped into the 25% with the greatest predicted improvement, the 25% with least predicted improvement, and the remaining 50% ("others"). Mean *true* HAMD scores over the 8-week treatment period are shown for each group. The 95% confidence interval is illustrated by the shaded area around each line.

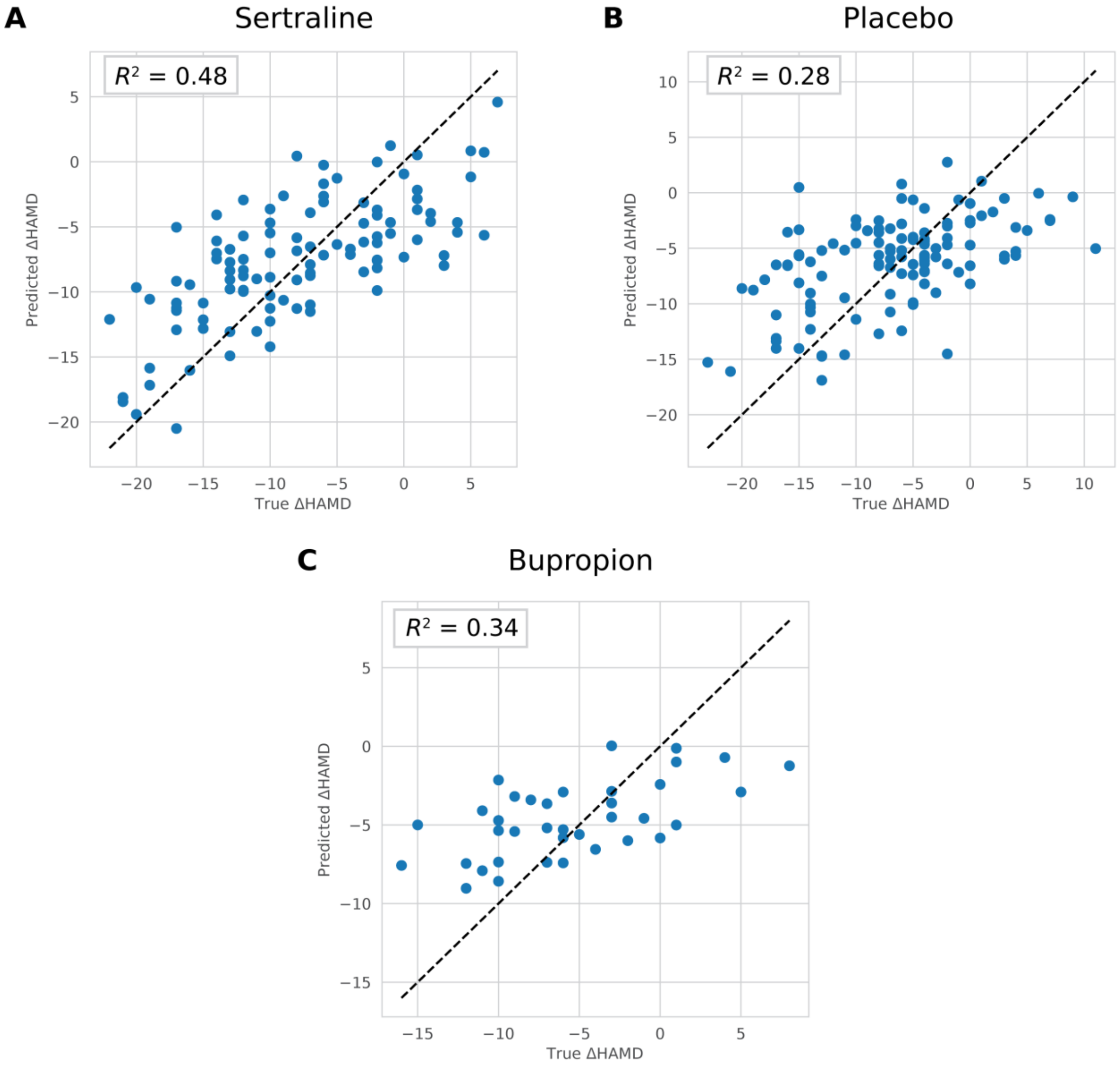

Figure S7. Predicted vs. true ΔHAMD for the sertraline, placebo, and bupropion models.

Table S1. Scanner and pulse sequence information for each study site.

	Columbia University	**Massachusetts General Hospital**	**University of Michigan**	**UT Southwestern Medical Center**
Scanner	General Electric Signa HDx 3T	Siemens TrioTim 3T	Philips Achieva 3T	Philips Ingenia 3T
Structural MRI				
Sequence	FSPGR	MPRAGE	TFE	MPRAGE
TR/TI/TE	6.0ms/900ms/2.4ms	2300ms/900ms/2.54ms	8.2ms/1100ms/3.7 ms	2100ms/1100ms/3.7ms
Flip angle	9°	9°	12°	12°
Dimensions	256 x 256 x 174	256 x 256 x 176	256 x 256 x 178	256 x 256 x 178
Voxel size	1 x 1 x 1 mm	1 x 1 x 1 mm	1 x 1 x 1 mm	1 x 1 x 1 mm
Functional MRI				
Sequence	GE-EPI	GE-EPI	GE-EPI	GE-EPI
TR/TE	2000ms/28ms	2000ms/28ms	2000ms/28ms	2000ms/28ms
Flip angle	90°	90°	90°	90°
Dimensions	64 x 64 x 39	64 x 64 x 39	64 x 64 x 39	64 x 64 x 39
Voxel size	3.2 x 3.2 x 3.1 mm	3.2 x 3.2 x 3.1 mm	3.2 x 3.2 x 3.1 mm	3.2 x 3.2 x 3.1 mm
Dummy scans	5	5	5	5
Number of volumes, reward task	240	240	240	240
Total acquisition time, resting state	480 s	480 s	480 s	480 s
Number of volumes, resting state	180	180	180	180
Total acquisition time, resting state	360 s	360 s	360 s	360 s

Table S2. Clinical features used as inputs for sertraline and placebo predictive models.

Clinical assessment name	Items used
Body mass index	
Clinical history	Number of suicide attempts, lifetime suicide rating
17-item Hamilton Rating Scale for Depression ($HAMD_{17}$)	Total
24-item Hamilton Rating Scale for Depression ($HAMD_{24}$)	Total
Altman Self-Rating Mania Scale (ASRM)	Total
Anger Attack Questionaire (AAQ)	Total
Childhood Trauma Questionnaire (CTQ)	Emotional Abuse, Emotional Neglect, Physical Abuse, Physical Neglect, Sexual Abuse, and Validity subscores
Columbia Suicide Severity Rating (CSSRS)	Baseline intensity score
Concise Health Risk Tracking (CHRTP)	Propensity score, risk score
Edinburgh Handedness Inventory (EHI)	Total
Fagerstrom Test of Nicotine Dependence (FTND)	Current cigarette-smoking status
Family History Screen (FHS)	All items
Mood and Anxiety Symptoms Questionnaire (MASQ)	Anxious Arousal, Anhedonic Depression, and General Distress subscores
Mood Disorders Questionnaire (MDQ)	Total
NEO-Five Factor Inventory	Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness subscores
16-item Quick Inventory of Depressive Symptomatology (QIDS SR16)	Total
Structured Clinical Interview for DSM-5 (SCID)	Current episode duration Current episode specifier (melancholic, atypical, or catatonic) Number of episodes Presence of anxious distress, mixed features, insomnia, hypersomnia, psychomotor agitation, psychomotor retardation History of alcoholism, generalized anxiety, bipolar disorder, panic disorder, or psychotic symptoms
Self-Administered Comorbidity Questionnaire (SCQ)	Total
Snaith-Hamliton Pleasure Scale (SHAPS)	Ordinal and dichotomous total
Social Adjustment Scale (SAS) short form	Total and mean
Speilberger State Anxiety Inventory (STAI)	Pre-fMRI and post-fMRI score
Standard Assessment of Personality Abbreviated Scale (SAPAS)	Total
Visual Analog Mood Scales (VAMS)	Happy-sad, quick witted, relaxed-tense scores

Table S3. Hyperparameter ranges used during hyperparameter optimization. Hyperparameter values were selected uniform randomly from these ranges to create 500 model configurations, which were tested with nested cross-validation to identify an optimal model configuration for the predictive task for each treatment. A model schematic is illustrated in **Figure S5** and is labelled accordingly with these hyperparameters.

Hyperparameter	**Search range**
Number of fully-connected hidden layers, L	1, 2, 3
Number of neurons in the first hidden layer, N_1	[8 … 128]
L1 regularization weight, λ_1	[10^{-3} … 3 x 10^{-1}]
L2 regularization weight, λ_2	[10^{-3} … 3 x 10^{-1}]
% decrease in hidden layer size from previous layer, t	10%, 75%
Batch normalization, b	True, False
Dropout rate, r	[0.1 … 0.8]
Activation function, $a(\cdot)$	ReLU, LeakyReLU, ELU, PReLU
Learning rate	[10^{-5} … 10^{-3}]

Table S4. Treatment outcome prediction performance for each treatment group, with and without clinical and demographic features and with or without data augmentation. Rows with **bold** text are the results presented in the main text. Performance metrics are coefficient of determination (*R^2*, 95% confidence interval in parentheses) and root mean squared error (RMSE) for predicting the numerical target of ΔHAMD. For predicting the binary targets of remission and response, performance metrics include positive predictive value (PPV), and area under the receiver operating characteristic curve (AUROC).

			ΔHAMD		**Remission**		**Response**	
Treatment	**Features used**	**Augmentation used**	***R^2***	**RMSE**	**PPV**	**AUROC**	**PPV**	**AUROC**
Sertraline	**Imaging, clinical**	**Yes**	**48% (33%–61%)**	**5.15**	**0.69**	**0.60**	**0.68**	**0.62**
	Imaging	Yes	12% (3%–26%)	6.68	0.52	0.56	0.57	0.55
	Imaging, clinical	No	< 0%	7.51	0.70	0.56	0.75	0.54
	Clinical	N/A	< 0%	7.81	0.63	0.54	0.50	0.51
	Imaging	No	< 0%	8.35	0.20	0.48	0.46	0.50
Placebo	**Imaging, clinical**	**Yes**	**28% (15%–42%)**	**5.87**	**0.81**	**0.65**	**0.69**	**0.67**
	Imaging	Yes	18% (7%–26%)	6.25	0.64	0.59	0.65	0.65
	Imaging, clinical	No	< 0%	7.17	0.86	0.57	0.75	0.59
	Clinical	N/A	< 0%	7.04	0.86	0.57	0.70	0.57
	Imaging	No	< 0%	7.60	0.75	0.53	0.67	0.52
Bupropion	**Imaging, clinical**	**Yes**	**34% (10%–59%)**	**4.46**	**0.75**	**0.71**	**1.00**	**0.57**
	Imaging	Yes	38% (13%–62%)	4.32	0.86	0.73	0.60	0.56
	Imaging, clinical	No	< 0%	6.78	1.00	0.58	0.00	0.50
	Clinical	N/A	< 0%	6.27	0.60	0.59	0.67	0.54
	Imaging	No	< 0%	6.53	1.00	0.63	0.50	0.51

Table S5. Treatment outcome prediction performance measured using 20x20 Monte Carlo cross-validation (20 repetitions of 20-fold cross-validation). The mean performance and 95% confidence interval over 20 repetitions is reported. Performance metrics for ΔHAMD 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).

	Prediction target							
	ΔHAMD		**Remission**			**Response**		
Treatment	R^2	RMSE	NNT	PPV	AUROC	NNT	PPV	AUROC
Sertraline	33% (31-36%)	5.81 (5.69-5.92)	5.91 (4.51-7.31)	0.59 (0.56-0.63)	0.57 (0.56-0.58)	5.08 (4.34-5.81)	0.69 (0.66-0.72)	0.62 (0.61-0.64)
Placebo	21% (19-23%)	6.12 (6.05-6.19)	2.98 (2.69-3.28)	0.68 (0.65-0.71)	0.61 (0.60-0.62)	3.61 (3.24-3.98)	0.65 (0.65-0.71)	0.61 (0.61-0.62)
Bupropion	43% (41-46%)	4.13 (4.04-4.22)	2.40 (2.15-2.66)	0.76 (0.73-0.79)	0.68 (0.67-0.70)	3.83 (2.79-4.86)	0.73 (0.68-0.79)	0.59 (0.57-0.62)

REFERENCES

1. Shorten C, Khoshgoftaar TM (2019): A survey on Image Data Augmentation for Deep Learning. *J Big Data* 6.
2. Nguyen KP, Chin Fatt C, Treacher A, Mellema C, Trivedi MH, Montillo A (2020): Anatomically-informed data augmentation for functional MRI with applications to deep learning. In: Landman BA, Išgum I. *Medical Imaging: Image Processing*: SPIE, pp 28–33.
3. Nguyen KP, Raval V, Minhajuddin A, Carmody T, Trivedi MH, Dewey RB*, et al.* (2021): The BLENDS Method for Data Augmentation of 4-Dimensional Brain Images. *BiorXiv:* 10.1101/2021.06.02.446748.
4. Iglesias JE, Liu C-Y, Thompson PM, Tu Z (2011): Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging* 30: 1617–1634.
5. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2010): A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration. *NeuroImage* 54: 2033–2044.
6. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang M-C*, et al.* (2009): Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage* 46: 786–802.
7. Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL*, et al.* (2014): Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Human Brain Mapping* 35: 1981–1996.
8. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A*, et al.* (2019): fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods* 16: 111–116.
9. Calhoun VD, Wager TD, Krishnan A, Rosch KS, Seymour KE, Nebel MB*, et al.* (2017): The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. *Human Brain Mapping* 38: 5331–5342.
10. Dohmatob E, Varoquaux G, Thirion B (2018): Inter-subject Registration of Functional Images: Do We Need Anatomical Images? *Frontiers in Neuroscience* 12: 64.
11. Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF (2015): ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage* 112: 267–277.
12. Greenberg T, Chase HW, Almeida JR, Stiffler R, Zevallos CR, Aslam HA*, et al.* (2015): 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. *American Journal of Psychiatry* 172: 881–891.
13. Greenberg T, Fournier JC, Stiffler R, Chase HW, Almeida JR, Aslam H*, et al.* (2019): Reward related ventral striatal activity and differential response to sertraline versus placebo in depressed individuals. *Molecular Psychiatry* 25: 1526–1536.
14. Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ*, et al.* (2018): Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex* 28: 3095–3114.
15. Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS (2012): A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping* 33: 1914–1928.
16. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N*, et al.* (2002): Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15: 273–289.

17. Rao A, Monteiro JM, Mourao-Miranda J (2017): Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage* 150: 23–49.

18. Yu M, Linn KA, Cook PA, Phillips ML, McInnis M, Fava M*, et al.* (2018): Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping* 39: 4213–4227.

19. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012): Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* 59: 2142–2154.

20. Bergstra J, Bardenet R, Bengio Y, Kégl B (2011): Algorithms for Hyper-Parameter Optimization. *Proceedings of the 24th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc, pp 2546–2554.

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

22. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019): Machine learning algorithm validation with a limited sample size. *PloS one* 14: e0224365.

23. Wu W, Zhang Y, Jiang J, Lucas MV, Fonzo GA, Rolle CE*, et al.* (2020): Antidepressant-Responsive Brain Signature in Major Depression Defined by Electroencephalography. *Nature Biotechnology* 38: 439–447.

24. Roose SP, Rutherford BR, Wall MM, Thase ME (2016): 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.

25. Ojala M, Garriga GC (2010): Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research* 11: 1833–1863.

26. Dimopoulos Y, Bourret P, Lek S (1995): Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Process Lett* 2: 1–4.

27. Olden JD, Joy MK, Death RG (2004): An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* 178: 389–397.

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

29. Bondareff W, Alpert M, Friedhoff AJ, Richter EM, Clary CM, Batzar E (2000): Comparison of sertraline and nortriptyline in the treatment of major depressive disorder in late life. *American Journal of Psychiatry* 157: 729–736.

30. Mitchell AJ, Subramaniam H (2005): Prognosis of depression in old age compared to middle age: a systematic review of comparative studies. *American Journal of Psychiatry* 162: 1588–1601.

31. Jain FA, Hunter AM, Brooks JO, Leuchter AF (2013): Predictive socioeconomic and clinical profiles of antidepressant response and remission. *Depression and anxiety* 30: 624–630.

32. Novick D, Montgomery W, Vorstenbosch E, Moneta MV, Dueñas H, Haro JM (2017): Recovery in patients with major depressive disorder (MDD): results of a 6-month, multinational, observational study. *Patient preference and adherence* 11: 1859–1868.

33. Fava M, Rush AJ, Alpert JE, Balasubramani GK, Wisniewski SR, Carmin CN*, et al.* (2008): Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR*D report. *American Journal of Psychiatry* 165: 342–351.

34. Zisook S, Johnson GR, Tal I, Hicks P, Chen P, Davis L*, et al.* (2019): General Predictors and Moderators of Depression Remission: A VAST-D Report. *American Journal of Psychiatry* 176: 348–357.