**­Supplementary Online Content**

**Personalized prediction of antidepressant versus placebo response:**

**Evidence from the EMBARC study**

**Supplemental Methods**

**Exclusion Criteria**

In addition to the exclusion criteria noted in the text, participants were excluded when any of the following criteria were met: 1) current pregnancy, breastfeeding, no use of contraception; 2) unstable psychiatric or general medical conditions requiring hospitalization; 3) study medication contraindication; 4) clinically significant laboratory abnormalities; 5) history of epilepsy or condition requiring an anticonvulsant; 6) electroconvulsive therapy (ECT), vagal nerve stimulation (VNS), transcranial magnetic stimulation (TMS) or other somatic treatments in the current episode; 7) medications (including but not limited to antipsychotics and mood stabilizers); 8) current psychotherapy; 9) significant suicide risk; or 10) failure to respond to any antidepressant at adequate dose and duration in the current episode. To minimize clinical heterogeneity, only patients reporting early onset (before age 30) MDD that was chronic (episode duration > 2 years) or recurrent (≥ 2 recurrences including the current episode) were enrolled. See supplement of Pizzagalli et al. (2018) for the detailed study protocol. Data from a second phase of the study – in which placebo non-responders are switched to SSRI and SSRI non-responders are switched to Bupropion (and with a smaller sample size) – are not reported in this publication.

**Probabilistic Reward Task (PRT**)

During the PRT, participants were asked to determine, via button press, whether a short (11.5 mm) or a long (13 mm) mouth (superimposed on a previously mouthless cartoon face) was presented on a computer screen. In this study, two blocks consisting of 100 trials were presented. Within each block, an equal number of short and long mouths were presented. Each trial began with a fixation cross (jittered 750-900 ms) followed by a mouthless face (500 ms), after which either the short or a long mouth appeared on the face (100 ms). Critically, to induce a response bias towards a more frequently rewarded stimulus, an asymmetric reinforcer ratio was employed. Namely, correct identification of either the long or short mouth was rewarded (“Correct!! You won 5 Cents”) 3x more frequently (“rich” stimulus) than the other mouth (“lean” stimulus). Participants were told at the start of the task that the purpose of the game was to win as much money as possible, but that not every correct response would be rewarded. Keys and conditions (long or short mouth as “rich” stimulus) were counterbalanced across participants. Participants who met any of the following quality control checks were excluded: (1) less than 80 valid trials in each block (i.e., less than 20% outlier responses, as defined by RT shorter than 150 ms or greater than 2,500 ms and the log-transformed RT exceeding the participant’s mean±3SD; for more detail see (Pizzagalli *et al.* 2008a)); (2) less than 24 rich rewards or less than 7 lean reward in each block; (3) rich-to-lean reward ratio < 2.5 in any block; and (4) rich or lean accuracy < 0.40 in any block.

 The primary variable of interest was change in *response bias* (RB) (Pizzagalli *et al.* 2008b) scores from the first to the second block (RBBlock2 – RBBlock1), which captures reward learning*.* RB scores capture a participant’s preference for the most frequently rewarded (“rich”) stimulus, and were calculated as:

log b = 0.5\*log{[(RichCorrect+0.5) \* (LeanIncorrect+0.5)] / [(RichIncorrect+0.5) \* (LeanCorrect+0.5)]}.

**Eriksen Flanker Task**

Participants began by completing a practice session consisting of 15 congruent and 15 incongruent trials. The flanking arrows were first presented alone (100 ms) and then a central arrow appeared (50 ms), for a total stimulus duration of 150 ms. Participants indicated, via button press, whether the center arrow pointed left or right. Both accuracy and reaction time (RT) were recorded.After the practice session, participants completed five blocks consisting of 70 trials each (46 congruent, 24 congruent), for a total of 350 trials. To ensure sufficient task difficulty, a response deadline was established for each block that corresponded to the 85th percentile of the RT distribution from incongruent trials in the preceding block (in the first block, the practice RT distribution was used). Stimulus presentation was followed by a fixation cross (1400 ms). If the participant failed to respond by the response deadline, a screen reading “TOO SLOW!” was presented (300 ms). Participants were instructed to speed up if they saw this screen. Each trial ended with presentation of the fixation cross for an additional 200-400 ms. In sum, total trial time varied between 2050-2250 ms. The sequence of congruent and incongruent trials was established with optseq2 (http://surfer.nmr.mgh.harvard.edu/optseq/) and was identical across participants.

During the task, block-by-block feedback was added to maintain performance at desired levels. Namely, if participants made fewer than three incongruent errors in a block, they were shown a screen stating, “Remember to respond as QUICKLY as possible while still being accurate”. If six or more incongruent errors were committed, the screen read, “Remember to respond as ACCURATELY as possible while still being fast”. Otherwise, the screen read, “Please respond as quickly and accurately as possible”.

Similar to the PRT, quality control checks were implemented to exclude datasets characterized by unusually poor performance. First, outlier trials were defined as those in which the raw RT < 150 ms or the log-transformed RT exceeded the participant’s mean±3SD (computed separately for congruent and incongruent stimuli). Second, we excluded datasets with: 35 or more RT outliers (i.e., greater than 10% of trials), fewer than 200 outlier-free congruent trials, fewer then 90 outlier-free incongruent trials, or lower than 50% correct for congruent or incongruent trials. Trials characterized by RT outliers were excluded from all analyses. In our previous EMBARC publication (Webb *et al.* 2016), significant whole-brain correlations emerged with Flanker interference effect on accuracy scores but not RT. As well, a cluster analysis in this paper, which differentiated subgroups of depressed patients, was informed by the Flanker interference effect on accuracy - as well as reward learning [PRT] and neuroticism [NEO-FFI-3] - but not the interference effect on RT. Accordingly, for our Flanker measure we focus on the interference effect on accuracy.

**Imputation**

There was little missing data in the baseline predictors. Only anxiety severity (MASQ-AA, 0.93%), anhedonia (SHAPS, 0.46%), race (0.46%), marital status (0.63%), employment status (0.93%), and years of education (1.39%) contained missing observations. Of the 216 individuals in this sample, 10.19% were missing data for the outcome variable, week 8 HRSD. Importantly, there were no significant differences in week 8 completion rates between the SSRI (88.0%) or placebo (91.5%) conditions (*χ2* (1)= 0.41, *p* = 0.52), nor between patients for whom SSRI was predicted to be superior to placebo (PAI < 0 or “SSRI-indicated*”*) (91.0%), and those for whom placebo was expected to be superior to SSRI (PAI > 0 or “placebo-indicated”) (88.3%) (*χ2* (1)= 0.42, *p* = 0.52). Similarly, adherence to medication/placebo pills was assessed at each visit, allowing us to estimate mean adherence over the 8-week study. There were no significant differences in adherence between the SSRI (94.5%) or placebo (95.0%) conditions (*t*(213) = ‑0.17, *p* = 0.86), nor between SSRI-indicated (95.3%) and placebo-indicated patients (93.3%) (*t*(213) = -0.52, *p* = 0.61). Missing data for were imputed using the baseline variables as well as longitudinal HRSD outcome measurements from weeks 0, 1, 2, 3, 4, and 6. Imputation was performed using a Random Forest (RF) approach (missForest (Stekhoven & Buhlmann 2012) package in R (R Core Team 2013)) that generates a single imputed dataset by averaging over multiple regression trees, thus giving the benefit of multiple imputation without needing to run the primary analyses across multiple imputed datasets (as is necessary when using multivariate imputation by chained equations (Buuren & Groothuis-Oudshoorn 2011)). This RF approach has been found to outperform other methods of imputation under many conditions (e.g., when complex and non-linear interactions are present) and has the additional benefit of being able to handle both continuous and categorical variables (Stekhoven & Buhlmann 2012; Shah *et al.* 2014).

**Variables selection**

There are many different approaches to variable selection (Steyerberg 2008). In the original publication of the PAI (DeRubeis *et al.* 2014) (see also Huibers *et al.* 2015), a domain-based backwards stepwise-regression approach was used (Fournier *et al.* 2009). More recent PAI efforts have used a RF machine learning approach for variable selection (Garge *et al.* 2013a) as well as a two-step approach that combines RF followed by stepwise AIC-penalized bootstrapped variable selection (Keefe *et al.* under review). Other commonly employed variable selection procedures include Support Vector Machines (SVM)(Koutsouleris *et al.* 2016) and Elastic Net Regularization (ENR)(Iniesta *et al.* 2016). Of relevance, one study found that Bayesian Additive Regression Trees (BART) outperformed several alternative approaches (e.g., RF and Lasso) in variable selection (Bleich *et al.* 2014). These variable selection approaches rely on different algorithms, such as decision tree-based ensemble learning methods (e.g., RF) and regression-based approaches (e.g., ENR). With the development of a variety of different machine learning approaches it is unclear how best to determine the optimal method to employ. In response, Cohen et al. (2017) recently combined four complementary variable selection approaches (i.e., RF, ENR, BART, and a stepwise AIC-penalized bootstrapped variable selection procedure [bootStepAIC (Austin & Tu 2004)]) to identify reliable and stable predictors across variable selection methods. We used this new approach in our study. Specifically, all potential baseline predictors were entered simultaneously into each of the first three approaches (RF, ENR, BART). The variables that were consistently identified as having important interactions with treatment by these approaches (selected by at least 2 of the 3 methods) were entered together into the final variable selection step (bootStepAIC). The variables that emerged from bootStepAIC comprised those that were used in the final model (see Supplemental Table 1 for correlations between each continuous predictor in the final model). Additional details on each method are provided below.

**Random Forest (RF).** RF is a recursive partitioning approach to modeling that can accommodate large numbers of predictor variables as well as complex relationships (e.g., non-linear associations and higher-order interactions) (Garge *et al.* 2013b). RF can be made to focus on predictor by treatment group interactions during the model building process by forcing the approach to select “splits” that maximize the difference in the treatment condition coefficient between subgroups (by specifying the model function of RF as “y ~ treatment”). Variables that surpass the permutation-based importance threshold are selected (see **Supplemental Figure 1**). The variables identified by RF were: Depression Severity (HDRS), Neuroticism (NEO-FFI-3), Cognitive Control (Flanker Accuracy interference effect), Age, and Employment Status.

**Elastic Net Regularization (ENR).** ENR combines the L1 and L2 penalizations (represented by the Lasso and Ridge regression approaches) to allow for the selection of a parsimonious set of variables that predict outcome (Hastie *et al.* 2016). ENR can accommodate large numbers of variables and is robust to high predictor covariance. Unlike RF and BART, ENR does not have the capability to consider unspecified non-linear relationships, or to focus on unspecified higher-order interactions between variables (Genuer *et al.* 2010). Because current statistical packages for ENR do not accommodate variable selection for models in which moderators are of primary interest, we split the training sample into the two separate treatment groups and identified variables that were prognostic within each group. Variables that were identified in only one condition, or that were indicated in both conditions but whose coefficients differed, were selected as potential moderators of treatment effects. The variables selected by ENR were: Depression Severity (HDRS), Anxiety Severity (MASQ-AA), Chronic MDD, Neuroticism (NEO-FFI-3), Cognitive Control (Flanker Accuracy effect), Age, Employment Status, and Years of Education.

**Bayesian Additive Regression Trees (BART).** BART (Chipman *et al.* 2010) builds on ensemble-of-tree methods (e.g., RF) by incorporating an underlying Bayesian probability model. BART and RF are similar in their ability to handle high numbers of predictors and to model non-linear and higher-order interactions. Bleich and Kapelner (Kapelner & Bleich 2013; Bleich *et al.* 2014) adapted BART to extract information about variable importance, and provide an interaction plot feature that can be used to identify potential 3-way interactions, and with Goldstein and colleagues (Goldstein *et al.* 2015) added the ICEbox package in R that allows for the visualization of non-linear relationships from BART models. Kapelner and Bleich adapted the bartMachine R package for the purposes of variable selection for the PAI to help focus model building on variables that predict differential treatment response. To facilitate this goal, they introduced a parameter that forces the variable splitting search to focus more on treatment group than other variables, thus introducing more interactions between treatment and other variables. This can be thought of similar to what researchers do when they only consider interactions between treatment group and baseline variables (and not interactions between baseline variables themselves), or to how (as described above) RF can specify the splitting criteria to evaluate the difference in the treatment coefficient for the model y ~ treatment. The N most important interactions identified by BART are retained, where N was decided based on the number of variables selected by RF (which uses a permutation test to determine an importance threshold cutoff). To account for variability in model structure, BART was run five times with the treatment importance parameter set to 5 and five times with it set to 10. Variables ranked within the top five strongest moderators, for both importance parameters, were selected. The variables selected by BART were the same as those selected by RF: Depression Severity (HDRS), Neuroticism (NEO-FFI-3), Cognitive Control (Flanker Accuracy interference effect), Age, and Employment Status.

**Stepwise AIC-penalized bootstrapped variable selection.** The variables selected by at least two of the three procedures described above were then reduced using a fourth approach. Our choice to use linear regression for our final model made it important that all included variables function in a linear regression context. The variables selected by RF or BART could lead to a model with poor fit, if, for example, these variables rely on unspecified non-linear relationships or higher-order interactions. The bootStepAIC package performs variable selection using a stepwise AIC-penalized bootstrapped approach (Austin & Tu 2004). Ten thousand (N=10,000) bootstrapped training samples were drawn, and backwards elimination was performed within each of the samples to select variables that independently contribute to predicting outcome. We examined the consistency in the sign of the moderator coefficients (which indicates consistency of the direction of the predictive relationships) to determine which variables to include in the final model; this allowed us to include variables with smaller effects that are consistent in the direction with which they predict differential response across treatments, and thus may be useful for treatment selection. All five variables met our criteria of at least 95% consistency in sign of the moderator coefficient.

The variance inflation factors (VIFs) in our final model were low (all VIFs < 1.1), indicating minimal collinearity among variables. In addition, an anonymous reviewer requested that we test whether each significant *Treatment Group* x *Predictor* interaction in the final model (see Table 2 in the main text) remained significant when included in a model without the other predictors (only including baseline HRSD). When these models were run all significant *Treatment Group* x *Predictor* interactions from the final model remained significant (all *ps <* .042).

**PAI Generation**

Once the variables that predict differential response have been identified, a statistical model can be constructed to generate treatment recommendations. This model is a linear regression predicting week 8 HRSD and includes the main effect of treatment and the five moderators (i.e., their main effects and terms representing their interactions with treatment). For each patient, a predicted outcome in SSRI is generated by inputting the patient’s values on the predictors and the value for SSRI (0.5), and then a prediction for placebo outcome is generated by changing the value for treatment to placebo (-0.5). A patient’s PAI is the signed difference between the two predictions (i.e., week 8 HRSD predicted in SSRI minus week 8 HRSD predicted in placebo), where a negative value reflects a predicted better outcome in SSRI, and a positive value reflects the opposite. To limit bias that could occur when evaluating model performance on individuals whose data were used to set model weights, PAIs were generated using 10-fold cross validation, ensuring each model is estimated absent any data from the patient whose week 8 HRSD score will be predicted. The folds were balanced on treatment condition so that the ratio from the full sample was maintained.

**PAI Evaluation**

Once a PAI value identifying the indicated optimal treatment has been generated for each patient, the expected utility of treatment selection was evaluated. To do this, the sample was split into two groups: those for whom SSRI was predicted to be superior to placebo (PAI < 0 or “SSRI-indicated*”*), and those for whom placebo was expected to be superior to SSRI (PAI > 0 or “placebo-indicated”). The distribution of week 8 HRSD scores for SSRI-indicated patients who were randomized to SSRI vs. placebo was compared using a paired student’s t-test and a Cohen’s d-type effect size. The analogous comparison was performed for those who were placebo-indicated. The strength of the treatment recommendation could vary from a large predicted advantage of one treatment over another (i.e., large absolute PAI value) to a small or potentially meaningless predicted advantage (i.e., PAI close to 0). We would not expect large, clinically meaningful differences in outcome between treatments for individuals who have a negligibly small PAI. In order to investigate the importance of the strength of these recommendations, we also evaluated the above comparisons within the subset of individuals for whom clinically meaningful advantage (DeRubeis *et al.* 2014) of one treatment over the other was predicted (i.e., a predicted advantage of ≥ 3 HRSD points). Finally, the entire 10-fold cross-validation procedure and evaluation was repeated 1000 times to generate stable estimates (for a distribution of these estimates see **Supplemental Figure 2**; see Cohen et al., 2017 for additional details on the modeling approach).

**Supplemental Results**

**Alternative PAI Models**

We tested two alternative PAI models. First, although we implemented k-fold cross-validation in the PAI analyses to minimize overfitting, the variable selection steps and PAI algorithm were ultimately implemented in the same sample. Thus, the extent to which model overfitting contributes to the observed pattern of findings is unclear. Namely, to what extent are results biased due to the fact that the PAI analyses were implemented using the 5 moderator variables emerging from our machine learning variable selection procedure? There are several approaches to evaluating model overfitting, one of which would be to test our model in an independent validation sample. However, we are not aware of any clinical trial of depression with the same set of variables (e.g., Flanker task, PRT, NEO, etc.) and with sufficiently similar inclusion criteria and study design. An alternative approach is to skip the variable selection procedure and implement the cross-validation PAI procedure including *all* 12 a priori moderator variables (rather than only selecting those variables that demonstrate moderation from our machine learning variable selection procedures). Evidence of overfitting would be reflected in the predictive model developed in the cross-validation training sets exhibiting poor fit in the test set (i.e., the model would fit too well to the peculiarities of the training set and not generalize to the test set, yielding non-significant findings). Second, the most well-supported clinical moderator of SSRI vs. placebo response is baseline depressive symptom severity (Khan *et al.* 2002; Kirsch *et al.* 2008; Fournier *et al.* 2010). To evaluate the utility of treatment recommendations based solely on depression severity (rather than our 5 moderator variables), we re-ran the above analysis using only baseline depressive symptom (HRSD) severity to inform the PAI.

***PAI Based on All 12 a Priori Baseline Variables.*** In this analysis, patients randomized to their PAI-indicated treatment condition (M = 10.39; SD = 6.78) were observed to have significantly lower week 8 HRSD scores relative to those randomized to their contraindicated condition (M = 12.41; SD = 6.88) (*d* = .29, *t*(214)= 2.16; *p* = .032). SSRI-indicated patients randomized to SSRI (M = 10.43; SD = 6.45) were observed to have significantly lower week 8 HRSD scores than those randomized to placebo (M = 12.96; SD = 7.33) (*d* = .37, *t*(124)= 2.04; *p* = .043). However, for patients predicted to have better outcomes to placebo, those who received placebo (M = 10.36; SD = 7.20) did not differ significantly in outcome relative to those who received SSRI (M = 11.48; SD = 6.02) (*d* = .17; *t*(88)= 0.79; *p* = .43).

***PAI Based Solely on Baseline Depression Severity.*** In this analysis, patients assigned to their PAI-indicated condition (M = 11.21; SD = 6.67) did not exhibit significantly lower HRSD scores than those randomized to the contraindicated group (M = 11.60; SD = 7.10) (*d* = .06; t(214) = 0.45; p = .62). For patients predicted to have better outcomes to SSRI than placebo (PAI < 0), those randomized to SSRI (M = 11.65; SD = 6.43) did not exhibit significantly lower HRSD scores than those randomized to placebo (M = 12.49; SD = 7.37) (*d* = .12; t(156) = 0.76; p = .45). Likewise, for patients predicted to have better outcomes to placebo (PAI > 0), those who received placebo (M = 10.17; SD = 7.17) did not have significantly better depression outcomes than those randomized to SSRI (M = 8.85; SD = 5.39) (*d* = .20; t(56) = -0.78; p = .44).

**Defining Outcome as Percent Change in Symptoms and Response Status**

In response to a request from an anonymous reviewer, primary analyses were re-run with the following two dependent variables: (1) percent change in HRSD score and (2) response status (i.e., response defined as ≥ 50% improvement). To do so, average PAI scores were computed for each individual across the 1000 10-fold cross-validations. For the first re-analysis, we ran t-tests comparing mean percent change in HRSD scores from intake to week 8 for SSRI-indicated (PAI > 0) patients who were randomized to SSRI vs. those randomized to placebo. The corresponding analysis was also run for placebo-indicated (PAI < 0) patients. The same pattern of findings emerged as in the primary analyses. For SSRI-indicated patients, those randomized to SSRI (M = 44.9%) were observed to have greater percent improvement in depressive symptoms relative to those randomized to placebo (M = 29.6%) (*t*(121)= -2.44; *p* = .016). For patients predicted to have better outcomes to placebo (PAI > 0), those who received placebo (M = 45.3%) did not differ significantly in this outcome relative to those who received SSRI (M = 31.6%) (*t*(91)= 1.60; *p* = .114). Next, we ran a corresponding analysis comparing group differences in response rates. Among SSRI-indicated patients, those who were randomized to SSRI had significantly higher response rates (50%) than those who received placebo (29%) (*χ2* (1)= 5.39, *p* = 0.02). Among placebo-indicated patients, the difference in response rates between those who received SSRI (38%) and placebo (53%) was not significantly different (*χ2* (1)= 2.21, *p* = 0.14). It is important to note that when a continuous depression outcome variable is dichotomized the predictive power of a pretreatment predictor is approximately halved (Uher et al., 2012). This is consistent with prior work demonstrating that dichotomization of continuous variables results in significant loss of information and statistical power (Kunz, 2011; Altman & Royston, 2006; Devi et al., 1998; Streiner, 2002). For these reasons, and consistent with prior PAI work, we excluded categorical analyses from the main text.

**Anhedonia and Reward Learning**

Given the study’s focus, models focused on the identification of prescriptive predictors (moderators) of SSRI vs. placebo response, rather than on prognostic predictors (i.e., main effects of predictors on outcome). Although they have received substantial attention in the depression literature, neither anhedonia nor reward learning emerged as moderators in our variable selection steps. In response to an anonymous reviewer, we tested whether they served as prognostic predictors. Specifically, we ran the final model (with week 8 HRSD as the dependent variable), removing all predictors with the exception of *Anhedonia* and *Anhedonia* x *Treatment* terms (and adjusting for baseline HRSD scores). Higher anhedonia predicted worse outcome (t=3.51, *p* < .001), but the *Anhedonia* x *Treatment* term was not significant (t=0.078, p= 0.93). A corresponding model with reward learning yielded non-significant findings for both terms (*p*s > .92).

**Effect of Site**

Recruiting sites were Columbia University (CU), Massachusetts General Hospital (MGH), the University of Texas Southwestern Medical Center (UT), and the University of Michigan (UM). There was no effect of study site on average PAI scores (*F*(3,212)= 1.21; *p* = .306). In addition, analyses controlling for site are similar to those reported in the main text. Specifically, when controlling for site – and when considering the full sample – SSRI-indicated patients (PAI < 0) randomized to SSRI had lower week 8 HRSD scores than those randomized to placebo (*F*(1,117)= 4.08; *p* = .046). However, for patients predicted to have better outcomes to placebo (PAI > 0), those who received placebo did not differ significantly in outcome relative to those who received SSRI (*F*(1,89)= 0.55; *p* = .462).

**Supplementary References**

**Austin PC, Tu JV** (2004). Bootstrap Methods for Developing Predictive Models. *The American Statistician* **58**, 131–137.

**Bleich J, Kapelner A, George EI, Jensen ST** (2014). Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics* **8**, 1750–1781.

**Buuren S, Groothuis-Oudshoorn K** (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software* **45**

**Chipman HA, George EI, McCulloch RE** (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**, 266–298.

**Cohen ZD, Kim T, Van H, Dekker J, Driessen E** (2017). Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *PsyArXiv,* <https://osf.io/6qxve/>.

**DeRubeis RJ, Cohen ZD, Forand NR, Fournier JC, Gelfand LA, Lorenzo-Luaces L** (2014). The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *PLOS ONE* **9**, e83875.

**Fournier JC, DeRubeis RJ, Shelton RC, Hollon SD, Amsterdam JD, Gallop R** (2009). Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. *Journal of Consulting and Clinical Psychology* **77**, 775–787.

**Garge NR, Bobashev G, Eggleston B** (2013a). Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC Bioinformatics* **14**, 125.

**Garge NR, Bobashev G, Eggleston B** (2013b). Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC bioinformatics* **14**, 125.

**Genuer R, Poggi J-M, Tuleau-Malot C** (2010). Variable selection using random forests. *Pattern Recognition Letters* **31**, 2225–2236.

**Goldstein A, Kapelner A, Bleich J, Pitkin E** (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* **24**, 44–65.

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

**Hastie T, Tibshirani R, Friedman J** (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2nd edition. Springer: New York, NY.

**Huibers MJH, Cohen ZD, Lemmens LHJM, Arntz A, Peeters FPML, Cuijpers P, DeRubeis RJ** (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLOS ONE* **10**, e0140771.

**Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, Henigsberg N, Dernovsek MZ, Souery D, Stahl D, Dobson R, Aitchison KJ, Farmer A, Lewis CM, McGuffin P, Uher R** (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research* **78**, 94–102.

**Kapelner A, Bleich J** (2013). Bartmachine: A powerful tool for machine learning. *stat* **1050**, 8.

**Keefe J, Wiltsey-Stirman S, Cohen ZD, DeRubeis RJ, Smith BN, Resick P** (under review). In Rape-Trauma PTSD, Patient Characteristics Indicate Which Trauma-Focused Treatment They are Most Likely to Complete.

**Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, Derks EM, Fleischhacker WW, Hasan A** (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry* **3**, 935–946.

**McCrae RR, Costa PT** (2010). NEO Inventories professional manual. *Lutz, FL: Psychological Assessment Resources*

**Pizzagalli DA, Evins AE, Schetter EC, Frank MJ, Pajtas PE, Santesso DL, Culhane M** (2008a). Single dose of a dopamine agonist impairs reinforcement learning in humans: Behavioral evidence from a laboratory-based measure of reward responsiveness. *Psychopharmacology* **196**, 221–232.

**Pizzagalli DA, Iosifescu D, Hallett LA, Ratner KG, Fava M** (2008b). Reduced Hedonic Capacity in Major Depressive Disorder: Evidence from a Probabilistic Reward Task. *Journal of psychiatric research* **43**, 76–87.

**Pizzagalli DA, Jahn AL, O’Shea JP** (2005). Toward an objective characterization of an anhedonic phenotype: A signal-detection approach. *Biological Psychiatry* **57**, 319–327.

**R Core Team** (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing: Vienna, Austria.

**Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H** (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology* **179**, 764–774.

**Snaith RP, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P** (1995). A scale for the assessment of hedonic tone: The Snaith-Hamilton Pleasure Scale. *British Journal of Psychiatry* **167**, 99–103.

**Stekhoven DJ, Buhlmann P** (2012). MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118.

**Steyerberg EW** (2008). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2009 edition. Springer: New York, NY.

**Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA** (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology* **104**, 15–25.

**Webb CA, Dillon DG, Pechtel P, Goer FK, Murray L, Huys QJ, Fava M, McGrath PJ, Weissman M, Parsey R, Kurian BT, Adams P, Weyandt S, Trombello JM, Grannemann B, Cooper CM, Deldin P, Tenke C, Trivedi M, Bruder G, Pizzagalli DA** (2016). Neural Correlates of Three Promising Endophenotypes of Depression: Evidence from the EMBARC Study. *Neuropsychopharmacology* **41**, 454–463.

*Means, Standard Deviations, and Correlations for Continuous Variables*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| 1. Depression Severity (HDRS) | .23\*\* | .07 | .41\*\* | .12 | -.02 | .05 | -.06 |
| 2. Anxiety Severity (MASQ-AA) | - | .25\*\* | .14\* | .03 | -.11 | -.04 | -.01 |
| 3. Neuroticism (NEO-FFI-3) |  | - | .19\*\* | .08 | .06 | -.23\*\* | .08 |
| 4. Anhedonia (SHAPS) |  |  | - | .05 | -.09\* | -.12 | -.09 |
| 5. Reward Learning (PRT) |  |  |  | - | -.07 | -.03 | .00 |
| 6. Cognitive Control (Flanker ACC) |  |  |  |  | - | -21\*\* | -.20 |
| 7. Age |  |  |  |  |  | - | .17\* |
| 8. Years of Education |  |  |  |  |  |  | - |

Note. HDRS: Hamilton Depression Rating Scale (17-item)(Hamilton 1960); MASQ-AA: Mood and Anxiety Symptoms Questionnaire, Anxious Arousal subscore (Watson *et al.* 1995), MDD: Major Depressive Disorder; NEO-FFI-3: NEO Five-Factor Inventory – 3 (McCrae & Costa 2010); SHAPS: Snaith-Hamilton Pleasure Scale (Snaith *et al.* 1995); PRT: Probabilistic Reward Task (Pizzagalli *et al.* 2005); Flanker ACC: Flanker Interference Accuracy score (= AccuracyCompatible trials – AccuracyIncompatible trials).

 \* p < .05, \*\* p < .01.

**Figure Captions**

**Supplemental Figure S1.** Plot displaying baseline moderators emerging from Random Forest (RF) analysis. Variables to the right of red dotted line (permutation-based importance threshold) are deemed significant.

**Supplemental Figure S2.** Distribution of mean observed week 8 Hamilton Depression Rating Scale (HRSD) scores for contraindicated minus indicated treatment across each 1000 10-fold cross-validations. Distributions are shown separately for SSRI-indicated (PAI < 0) and placebo-indicated (PAI>0).



****