Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study

Christian A. Webb1, Madhukar H. Trivedi2, Zachary D. Cohen3, Daniel G. Dillon1, Jay C. Fournier4, Franziska Goer5, Maurizio Fava5, Patrick J. McGrath6, Myrna Weissman6, Ramin Parsey7, Phil Adams6, Joseph M. Trombello2, Crystal Cooper2, Patricia Deldin6, Maria A. Oquendo6, Melvin G. McInnis6, Quentin Huys3, Gerard Bruder6, Benji T. Kurian2, Manish Jha2, Robert J. DeRubeis3 and Diego A. Pizzagalli1

Abstract

Background. Major depressive disorder (MDD) is a highly heterogeneous condition in terms of symptom presentation and, likely, underlying pathophysiology. Accordingly, it is possible that only certain individuals with MDD are well-suited to antidepressants. A potentially fruitful approach to parsing this heterogeneity is to focus on promising endophenotypes of depression, such as neuroticism, anhedonia, and cognitive control deficits.

Methods. Within an 8-week multisite trial of sertraline v. placebo for depressed adults (n = 216), we examined whether the combination of machine learning with a Personalized Advantage Index (PAI) can generate individualized treatment recommendations on the basis of endophenotype profiles coupled with clinical and demographic characteristics.

Results. Five pre-treatment variables moderated treatment response. Higher depression severity and neuroticism, older age, less impairment in cognitive control, and being employed were each associated with better outcomes to sertraline than placebo. Across 1000 iterations of a 10-fold cross-validation, the PAI model predicted that 31% of the sample would exhibit a clinically meaningful advantage [post-treatment Hamilton Rating Scale for Depression (HRSD) difference ⩾3] with sertraline relative to placebo. Although there were no overall outcome differences between treatment groups (d = 0.15), those identified as optimally suited to sertraline at pre-treatment had better week 8 HRSD scores if randomized to sertraline (10.7) than placebo (14.7) (d = 0.58).

Conclusions. A subset of MDD patients optimally suited to sertraline can be identified on the basis of pre-treatment characteristics. This model must be tested prospectively before it can be used to inform treatment selection. However, findings demonstrate the potential to improve individual outcomes through algorithm-guided treatment recommendations.

Introduction

Meta-analyses reveal that average differences in depressive symptom improvement between antidepressant medications [ADMs; most commonly, selective serotonin reuptake inhibitors (SSRIs)] and placebo are often small (i.e. between-group differences in symptom change of <3 points on the Hamilton Depression Rating Scale (Hamilton, 1960; Moncrieff et al., 2004; Kirsch et al., 2008; Fournier et al., 2010; Kirsch, 2015; Jakobsen et al., 2017; Cipriani et al., 2018)). A potential reason for this modest differentiation is that major depressive disorder (MDD) is a highly heterogeneous condition in terms of symptom presentation and, likely, underlying pathophysiology (Wakefield and Schmitz, 2013; Fried and Nesse, 2015a; 2015b; Baldessarini et al., 2017). Accordingly, it is possible that subsets of depressed individuals are better suited to SSRIs, whereas others may derive limited benefit. For example, for certain depressed individuals, the mere passage of time – possibly coupled with the expectation of improvement – may result in symptom remission (e.g. ‘spontaneous remitters’). Such individuals may not require SSRIs. Instead a less costly, low-intensity alternative intervention with minimal or no side effects may be sufficient for symptom remission (e.g. internet-based cognitive behavioral therapy (CBT), which is included in the National Institute for Health and
Care Excellence Guidelines (NICE, 2018) as an efficacious intervention. Currently, treatment selection is largely based on trial-and-error. Approximately 55–75% of depressed individuals in primary care fail to achieve remission to first-line antidepressants, and 8–40% will switch to at least one other medication (Rush et al., 2006; Marcus et al., 2009; Schultz and Joish, 2009; Vuoriilehto et al., 2009; Milea et al., 2010; Saragossi et al., 2012; Thomas et al., 2013; Ball et al., 2014; Mars et al., 2017). Identifying predictors of antidepressant response may ultimately inform the development of algorithms generating personalized predictions of optimal treatment assignment for clinicians and patients to consider in their decision-making regarding which intervention to select.

A range of pre-treatment variables (e.g. baseline clinical, demographic, and neurobiological characteristics) have been examined as predictors of SSRI response. Perhaps the most well-supported clinical moderator of SSRI v. placebo response is baseline depressive symptom severity (Khan et al., 2002; Kirsch et al., 2008; Fournier et al., 2010). Meta-analyses indicate that in patients with MDD, lower levels of depressive symptom severity predicts minimal to no advantage of ADM over placebo, but that as depression severity increases, so does the magnitude of the advantage of ADM over placebo (Khan et al., 2002; Kirsch et al., 2008; Fournier et al., 2010). Other relevant predictors of greater ADM response include younger age (Fournier et al., 2009), being female (Trivedi et al., 2006; Jakubovski and Bloch, 2014), higher education (Trivedi et al., 2006), being employed (Fournier et al., 2009; Jakubovski and Bloch, 2014), lower anhedonia (McMakin et al., 2012; Uher et al., 2012a), non-chronic depression (Souery et al., 2007), and lower anxiety (Fava et al., 2008). Although each of these variables has limited predictive power when considered individually, recent advances in multi-variable machine learning approaches allow for the combination of large sets of variables to predict treatment response (Gillan and Whelan, 2017). Critically, to be clinically useful for treatment selection, predictors of treatment response must be applicable to individual patients. Consistent with the goals of precision medicine, such work aims to translate treatment outcome moderation findings to actionable, algorithm-guided treatment recommendations (Cohen and DeRubeis, 2018).

We sought to use machine learning coupled with a recently published Personalized Advantage Index (PAI) (DeRubeis et al., 2014; Huibers et al., 2015) to predict treatment outcome at the individual level on the basis of pre-treatment patient data. Our aim was to use the above approach to identify the subset of patients who may be optimally suited to SSRI. With regards to machine-learning approaches, we used four complementary variable selection procedures in an effort to identify a reliable and stable set of predictors from the initial, larger set of baseline variables. These procedures rely on different algorithms, such as decision tree-based ensemble learning methods [e.g. Random Forests (RF)] and regression-based methods [e.g. Elastic Net Regularization (ENR)]. This approach encouraged the selection of a set of predictors that emerged consistently across differing variable selection algorithms (see ‘Variable selection’ section below). Data were derived from the multi-site EMBARC (Establishing Modulators and Biosignatures of Antidepressant Response for Clinical Care) clinical trial comparing SSRI (sertraline) v. placebo (Trivedi et al., 2016). Of relevance, in a recent study based on EEG and cluster analyses, we reported that the substantial heterogeneity of MDD could be parsed by considering three putative endophenotypes of depression: neuroticism, blunted reward learning, and cognitive control deficits (Webb et al., 2016). Endophenotypes are hypothesized to lie on the pathway between genes and downstream symptoms, and are traditionally defined as meeting the following criteria (Gottesman and Gould, 2003): (1) associated with the disease, (2) heritable, (3) primarily state-independent, (4) cosegregate within families, (5) familial association, and (6) measured reliably (Goldstein and Klein, 2014). We posited that depressed patients with certain endophenotype profiles may be differentially responsive to certain interventions (e.g. the cluster of depressed patients defined by relatively high levels of neuroticism may be more responsive to SSRIs). Indeed, there is an evidence that depressed individuals characterized by elevated neuroticism may derive relatively greater therapeutic benefit from SSRIs relative to CBT (Bagby et al., 2008) or placebo (Tang et al., 2009). Thus, we examined whether the combination of putative endophenotypes (neuroticism, reward learning, cognitive control deficits, anhedonia) with both baseline clinical (depressive symptom severity, depression chronicity, anxiety severity) and demographic (gender, age, marital status, employment status, years of education) variables previously linked with antidepressant response could be used to identify individual depressed patients optimally suited to SSRIs. Plausible neuroimaging predictor variables ( McGrath et al., 2013; Pizzagalli et al., 2018) were excluded from this particular study given that they are substantially more costly and time-consuming than the above set of clinical, demographic, and behavioral variables, the latter of which could be reasonably integrated into a current psychiatric clinic for the purpose of treatment selection.

### Methods and materials

After providing informed consent, participants completed several behavioral and self-report assessments prior to enrolling in an 8-week, double-blind, placebo-controlled clinical trial of sertraline v. placebo. The clinical trial design has been described in detail in a previous publication (Trivedi et al., 2016).

### Participants

Eligible participants (ages 18–65) met Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria for a current MDD episode (SCID-I/P), scored ≥14 on the 16-item Quick Inventory of Depression Symptomatology (QIDS-SR16) (Rush et al., 2003), and were medication-free for ≥3 weeks prior to completing any study measures. Exclusion criteria included: history of bipolar disorder or psychosis; substance dependence (excluding nicotine) in the past 6 months or substance abuse in the past 2 months; active suicidality; or unstable medical conditions (see online Supplementary Methods). Data from 216 MDD subjects who passed quality control criteria for both Flanker and Probabilistic Reward Task and completed at least 4 weeks of treatment (American Psychiatric Association, 2010; Fournier et al., 2013) were included (online Supplementary Methods).
Endophenotype measures

NEO Five-Factor Inventory-3 (NEO-FFI-3) (McCrae and Costa, 2010). The 12-item neuroticism subscale from the NEO-FFI was used.

Probabilistic Reward Task (PRT). The PRT uses a differential reinforcement schedule to assess reward learning (i.e. the ability to adapt behavior as a function of rewards), and has been described in detail in previous publications (Pizzagalli et al., 2005, 2008a) (see online Supplementary Methods).

Snauth-Hamilton Pleasure Scale (SHAPS) (Snauth et al., 1995). The SHAPS is a 14-item self-report scale, with items asking about hedonic experience in the ‘last few days’ for a variety of pleasurable activities. Items consist of four response categories, with ‘strongly agree’ (=1), ‘agree’ (=2), ‘disagree’ (=3), ‘strongly disagree’ (=4). Higher scores indicate higher anhedonia.

Flanker Task (Eriksen and Eriksen, 1974). An adapted version of the Eriksen Flanker Task that included an individually titrated response window was used to assess cognitive control (see online Supplementary Methods) (Holmes et al., 2010).

Clinical measures

Hamilton Rating Scale for Depression (HRSD) (Hamilton, 1960). The 17-item HRSD, a clinician-administered measure of depressive symptom severity, was administered by trained clinical evaluators.

Mood and Anxiety Symptoms Questionnaire (MASQ) (Watson et al., 1995). The anxious arousal subscale from a 30-item adaptation of the MASQ (MASQ-AA) assessed anxiety.

Data acquisition and reduction

PRT. The primary variable of interest was reward learning, which has been found to predict response to antidepressant treatment among inpatients with MDD (Vrieze et al., 2013). As in prior studies (Pizzagalli et al., 2008b; Vrieze et al., 2013), reward learning was defined as change in response bias (RB) scores throughout the task [here, from the first to the second block (RB_{block2}−RB_{block1})].

Flanker Task. The primary variable of interest was the interference effect on accuracy, defined as lower accuracy on incongruent relative to congruent trials, computed as (Accuracy_{compatible trials} − Accuracy_{incompatible trials}). Higher scores reflect greater interference (i.e. reduced cognitive control).

Data Pre-Processing. Missing data were imputed using a RF-based imputation strategy [missForest (Stekhoven and Bühlmann, 2012) package in R (R Core Team, 2013)] (see online Supplementary Methods) (Waalje et al., 2013). This approach can handle both categorical and continuous variables, and generates a single imputed dataset via averaging across multiple regression trees. Consistent with the recommendation of Kraemer and Blasey (Kraemer and Blasey, 2004), continuous variables were mean-centered and categorical variables were transformed into binary variables with the values of −0.5 and 0.5. Of the 216 individuals in this sample, 10.19% were missing data for the outcome variable (week 8 HRSD) and thus had their data imputed. 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]. For additional analyses on dropout rates and medication/placebo adherence, see online Supplementary Methods.

Statistical analyses

Variable selection

Prior to implementing the PAI algorithm, pre-treatment variables that interact with treatment group (SSRI or placebo) in predicting HRSD outcome (week 8 scores) must be selected. We implemented (1) RF modeling [using the mobForest (Garge et al., 2013) package in R (R Core Team, 2013)], (2) ENR [glmnet package (Friedman et al., 2010)], and (3) Bayesian Additive Regression Trees [BART; bartMachine package (Kapelner and Bleich, 2016)]. For each of these three models, we entered all of our selected pre-treatment variables simultaneously: four endophenotype variables [neuroticism (NEO-FFI-3), cognitive control (Flanker interference effect on accuracy), reward learning (PRT), and anhedonia (SHAPS)], three clinical variables [baseline severity of depressive symptoms (HRSD), baseline severity of anxiety (MASQ-AA), and chronic MDD (yes/no)] and five demographic variables (age, gender, marital status, employment status, and years of education). Variables showing treatment group × predictor variable interactions in two of the three models were entered into a final stepwise AIC-penalized bootstrapped variable selection [using the bootStepAIC package (Austin and Tu, 2004)]. For details on each of these approaches and how variables are selected from each model, see online Supplementary Methods.

Generating PIs

Briefly, to generate treatment recommendations with the PAI approach, a regression model is built and used to predict treatment outcome (week 8 HRSD) for each patient in SSRI and placebo separately. 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 reverse. Moreover, the magnitude of the absolute value of the PAI reflects the strength of the differential prediction, such that patients with larger PIs, in either direction, are those who are most likely to evidence a substantially better outcome in their PAI indicated, relative to their PAI non-indicated treatment. To limit bias that could occur when evaluating model performance on individuals whose data were used to set model weights, PIs were generated using 10-fold cross-validation. This procedure ensures that each model is estimated absent any data from the patient whose outcome will be predicted (see PAI generation and PAI evaluation in the online Supplementary Methods for details; see also Alternative PAI models section below).

Evaluating PIs

To assess whether PAI scores moderate treatment group differences in depression outcomes, we tested a treatment group × PAI score interaction with week 8 HRSD scores as the dependent variables. Next, and similar to previous PAI publications (DeRubeis et al., 2014; Huibers et al., 2015), to evaluate the utility of the PIs, we compared mean week 8 HRSD scores for SSRI-indicated individuals who were randomized to SSRI in comparison to SSRI-indicated participants who received placebo. We performed the analogous comparison for those identified as ‘placebo-indicated’. We then evaluated the above comparisons with only those patients for whom the absolute value of the PAI was 3 or greater (i.e. predicted to have a ‘clinically significant’ advantage in one treatment condition) (DeRubeis et al., 2014).
Finally, the entire 10-fold cross-validation procedure and evaluation was repeated 1000 times to generate stable estimates.

Results

Variable selection

See Table 1 for variable selection results, including which variables were selected during each stage. The following variables survived the four-step procedure and were included in the final model (see Fig. 1 and Table 2):  

\[ Y = \text{treatment} \times (\text{depression severity} \times [\text{HRSD} + \text{neuroticism}(\text{NEO – FFI – 3}) + \text{cognitive control} [\text{Flanker Interference(Accuracy)]} + \text{age + employment status}). \]

Predicted outcomes and PAIs

The average absolute value of PAI scores was 3.4 (s.d. = 2.6), indicating that our model predicted an average 3.4-point difference in week 8 HRSD scores between indicated and non-indicated treatment assignment. The absolute value of the PAI was 3 or greater in approximately half (48.6%) of the sample (see Fig. 2 for distribution of PAI scores). Specifically, 31.5% of the sample was predicted to have a ‘clinically significant’ advantage (DeRubeis et al., 2014) in the SSRI condition (PAI \( \leq -3 \)), whereas this value was 17.1% for placebo (PAI \( \geq 3 \)). In contrast, the model indicates that 51.4% of the sample was predicted to exhibit relatively minimal differences in outcome between treatment conditions.

**Observed outcomes in indicated v. non-indicated treatment condition**

**Full sample**

First, it is important to highlight that, in the full sample, patients randomized to SSRI (\( M = 10.86; \text{s.d.} = 6.27 \)) and placebo (\( M = 11.88; \text{s.d.} = 7.37 \)) did not significantly differ in mean week 8 HRSD outcomes (adjusting for baseline HRSD scores) \( [F(1,213) = 0.92; p = 0.339; \text{Cohen’s } d = 0.15; \text{ Fig. 3, left panel}] \). Critically, a significant \( \text{treatment group} \times \text{PAI} \) interaction emerged in predicting week 8 HRSD scores, indicating that PAI scores moderated treatment group differences in outcome \( [F(1,212) = 6.68; p = 0.010] \). For the full sample, patients randomized to their PAI-indicated treatment condition (\( M = 10.39; \text{s.d.} = 6.97 \)) were observed to have lower week 8 HRSD scores relative to those randomized to their contraindicated condition (\( M = 12.38; \text{s.d.} = 6.70 \)) \( [d = 0.29, t(214) = 2.16; p = 0.032] \). For patients predicted to have better outcomes to SSRI than placebo (PAI < 0), those randomized to SSRI (\( M = 10.57; \text{s.d.} = 6.48 \)) were observed to have lower week 8 HRSD scores than those randomized to placebo (\( M = 13.12; \text{s.d.} = 7.03 \)) \( [d = 0.38, t(121) = 2.08; p = 0.040; \text{ see Fig. 3, right panel}] \). However, for patients predicted to have better outcomes to placebo (PAI > 0), those who received placebo (\( M = 10.18; \text{s.d.} = 7.54 \)) did not differ significantly in outcome relative to those who received SSRI (\( M = 11.23; \text{s.d.} = 6.04 \)) \( [d = 0.16, t(91) = 0.74; p = 0.460; \text{ see Fig. 3, right panel}] \).

**Largest PAIs (PAI \( \geq 3 \))**

Among this subset, patients randomized to their indicated treatment condition (\( M = 9.53; \text{s.d.} = 6.68 \)) were observed to have lower week 8 HRSD scores relative to those randomized to their contraindicated condition (\( M = 14.09; \text{s.d.} = 6.42 \)) \( [d = 0.70, t(103) = 3.59; p < 0.001] \). SSRI-indicated patients randomized to SSRI (\( M = 10.68; \text{s.d.} = 7.04 \)) were observed to have lower week 8 HRSD scores than those randomized to placebo (\( M = 14.66; \text{s.d.} = 6.83 \)) \( [d = 0.58, t(66) = 2.34; p = 0.023; \text{ see Fig. 3, right panel}] \). Conversely, placebo-indicated patients randomized to placebo (\( M = 7.65; \text{s.d.} = 5.64 \)) had better outcomes than those randomized to SSRI (\( M = 13.06; \text{s.d.} = 5.57 \)) \( [d = 1.01; t(35) = 3.07; p = 0.004; \text{ see Fig. 3, right panel}] \).

**Alternative PAI models**

See online Supplementary Material for results from two alternative PAI models. First, a PAI model was run including all 12 \( a \text{ priori} \) baseline variables, rather than the reduced set of five moderators emerging from our variable selection procedure. In other words, in the former model including all \( a \text{ priori} \) variables, our variable selection procedure was not performed. The fact that a similar pattern of findings emerged in this control PAI analysis suggests that our findings are likely not attributable to overfitting due to running our PAI analysis on a reduced set of variables emerging from our variable selection steps. Second, to evaluate the utility of treatment recommendations based solely on

\[
\text{Table 1. Variable selection results}
\]

<table>
<thead>
<tr>
<th>Pre-treatment variable</th>
<th>Random Forest</th>
<th>Elastic Net</th>
<th>BART</th>
<th>Included in bootstrap AIC?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression severity (HDRS)*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Anxiety severity (MASQ-AA)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Chronic MDD (yes/no)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Neuroticism (NEO-FFI-3)*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Anhedonia (SHAPS)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Reward learning (PRT)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Cognitive control (Flanker ACC)*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gender</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Age*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Marital status</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Employment status*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Years of education</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

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 and Costa, 2010); SHAPS, Snaith-Hamilton Pleasure Scale (Snaith et al., 1995); PRT, Probabilistic Reward Task (Pizzagalli et al., 2000); Flanker ACC, Flanker Interference Accuracy score \( = \text{Accuracy}(\text{compatible trials} – \text{Accuracy}(\text{incompatible trials})) \); higher scores indicate more interference (i.e. reduced cognitive control); BART, Bayesian Additive Regression Tress. *=Variables selected by Stepwise AIC to be included in the final model.
**Table 2. Final model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>s.e.</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>11.51</td>
<td>0.43</td>
<td>0.00**</td>
</tr>
<tr>
<td>Treatment</td>
<td>−0.65</td>
<td>0.85</td>
<td>0.44</td>
</tr>
<tr>
<td>Depression severity (HDRS)</td>
<td>2.17</td>
<td>0.44</td>
<td>0.00**</td>
</tr>
<tr>
<td>Neuroticism (NEO-FFI-3)</td>
<td>0.42</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>Cognitive control (Flanker ACC)</td>
<td>−0.31</td>
<td>0.45</td>
<td>0.49</td>
</tr>
<tr>
<td>Age</td>
<td>0.85</td>
<td>0.45</td>
<td>0.06</td>
</tr>
<tr>
<td>Employment status</td>
<td>−2.61</td>
<td>0.87</td>
<td>0.00**</td>
</tr>
<tr>
<td>Treatment × depression severity (HDRS)</td>
<td>−1.29</td>
<td>0.88</td>
<td>0.14</td>
</tr>
<tr>
<td>Treatment × neuroticism (NEO-FFI-3)</td>
<td>−2.56</td>
<td>0.90</td>
<td>0.01**</td>
</tr>
<tr>
<td>Treatment × cognitive control (flanker ACC)</td>
<td>1.86</td>
<td>0.89</td>
<td>0.04*</td>
</tr>
<tr>
<td>Treatment × age</td>
<td>−2.25</td>
<td>0.91</td>
<td>0.01*</td>
</tr>
<tr>
<td>Treatment × employment status</td>
<td>−3.21</td>
<td>1.74</td>
<td>0.07</td>
</tr>
</tbody>
</table>

HDRS, Hamilton Depression Rating Scale (17-item) (Hamilton, 1960); NEO-FFI-3, NEO Five-Factor Inventory – 3 (McCrae and Costa, 2010); Flanker ACC, Flanker Interference Accuracy score (AccuracyCompatible trials − AccuracyIncompatible trials).

*p < 0.10. **p < 0.05. ***p < 0.01.

**Fig. 1.** Plots of baseline predictor by treatment group interactions from the final model.

**Fig. 2.** Frequency histogram displaying distribution of Personalized Advantage Index (PAI) scores, computed as the predicted difference in week 8 HRSD scores for SSRI minus placebo. Accordingly, a PAI score <0 signifies that SSRI was indicated, whereas a PAI score >0 indicates that placebo was expected to yield a better outcome. The kernel density estimate illustrates the expected distribution of PAI scores in the population.
depression severity (rather than our five moderator variables), we re-ran the above analysis using only baseline depressive symptom (HRSD) severity to inform the PAI, which did not yield significant findings.

Discussion

This study used the variable selection approach proposed by Cohen et al. (Cohen et al., 2017) combining machine learning with a previously published PAI algorithm (DeRubeis et al., 2014; Huibers et al., 2015) to generate individualized treatment recommendations on the basis of (i) putative behavioral endophenotypes of depression (Goldstein and Klein, 2014; Webb et al., 2016) and (ii) clinical and demographic characteristics previously linked with antidepressant response. Ultimately, the goal is to translate research on predictors of antidepressant response to actionable treatment recommendations for individuals. First, it is important to highlight that the baseline moderators emerging from our machine learning variable selection steps are largely consistent with prior research. In particular, depressed individuals with higher baseline severity of depressive symptoms (Khan et al., 2002; Kirsch et al., 2008; Fournier et al., 2010), higher neuroticism (Tang et al., 2009), and who were employed (Fournier et al., 2009; Jakubovski and Bloch, 2014) had better outcomes to SSRI than placebo. In addition, relatively older patients and those with lower deficits in cognitive control (i.e. smaller Flanker accuracy interference effect) also exhibited better outcomes to SSRI. Of note, owing to their minimal cost and relatively low time burden, these baseline measurements could be more easily integrated into a treatment clinic than baseline assessments involving neuroimaging.

Perhaps the most well-supported clinical moderator of SSRI v. placebo response is baseline depressive symptom severity (Khan et al., 2002; Kirsch et al., 2008; Fournier et al., 2010). It should be noted that total depression score at baseline is not the only meaningful marker of depression severity. Other relevant variables such as episode chronicity and anhedonia were included in our initial models but did not survive the variable selection steps. Chronicity is known to be linked with poor response to placebo (Khan et al., 1991; Dunner, 2001), yet did not emerge as a moderator of SSRI v. placebo response. Consistent with prior work, higher neuroticism was associated with greater response to SSRI relative to placebo, which may in part be due to the role of SSRIs in blunting negative affect (Quilty et al., 2008; Tang et al., 2009; Soskin et al., 2012). It is important to highlight that elevated neuroticism moderated SSRI v. placebo response above and beyond the contribution of baseline depression (i.e. while the baseline HRSD × treatment group interaction was included in the model).

The interpretation of the cognitive control finding is less clear. Namely, those with more intact cognitive control exhibited better outcomes in SSRI v. placebo; whereas those with greater impairments showed little between-group differences in outcome. Continued cognitive impairments – even following symptom remission – are among the most common residual symptoms of depression (Herrera-Guzmán et al., 2009; Lam et al., 2014). Moderation may be more likely to be observed when comparing a treatment that more successfully targets cognitive control deficits [e.g. vortioxetine, (Mahableshwarkar et al., 2015)] v. one with limited pro-cognitive effects (also see Etkin et al., 2015).

Of the 12 a priori variables we initially included, seven did not survive our four-step variable selection procedure. It may be that some of these variables are prognostic predictors of outcome, but were not selected as they do not moderate SSRI v. placebo response. For example, higher anhedonia (McMakin et al., 2012; Uher et al., 2012a) and blunted reward learning (Vrieze et al., 2013) have each been shown to predict worse antidepressant outcome. Although anhedonia did not moderate of SSRI v. placebo response, it did emerge as a prognostic predictor of worse outcome across groups (t = 3.51, p < 0.001; reward learning ns; see online Supplementary Results). With regards to the specific variable selection approaches used, both RF and BART identified the same five variables; whereas ENR selected a larger set of eight variables. Differences in results between these approaches are not unexpected, and may be due to the fact that both RF and BART rely on a similar decision tree-based ensemble learning algorithm, whereas ENR is a variant of classic regression. As well, unlike ENR, both RF and BART consider both unspecified non-linear relationships and higher order interactions between variables.

Fig. 3. Comparison of mean week 8 HRSD for patients randomized to SSRI or placebo (left panel) (n = 216). Comparison of mean week 8 HRSD scores for patients randomly assigned to their PAI-indicated treatment v. those assigned to their PAI-contraindicated treatment for the full sample (n = 216) v. including only patients for whom the algorithm predicted a clinically significant advantage in one treatment condition (PAI ≥ 3); n = 105 (right panel). Error bars represent standard error.
Importantly, there were no overall differences in depression outcomes between patients randomized to SSRI and placebo in the overall sample ($d = 0.15$). These findings are in line with meta-analyses of SSRI v. placebo indicating small overall differences in outcome (Moncrieff et al., 2004; Kirsch et al., 2008; Fournier et al., 2010; Kirsch, 2015; Jakobsen et al., 2017; Cipriani et al., 2018). However, overall between-group comparisons obscure any meaningful between-patient characteristics that may moderate SSRI v. placebo differences in outcome. Indeed, we identified five patient characteristics that moderated group differences in depression outcome. These variables were subsequently entered into a PAI algorithm (DeRubeis et al., 2014; Huibers et al., 2015) to generate patient-specific predictions of SSRI v. placebo outcome. Results using our PAI model indicated that approximately one-third of the sample would have a clinically significant advantage (DeRubeis et al., 2014) with SSRI relative to placebo (PAI ≤ 3). Intriguingly, and unexpectedly, the model also predicted that a subset (17%) of depressed individuals would exhibit a clinically significant advantage in placebo.

As the treatment recommendations for some individuals indicated almost no advantage of one treatment over the other (e.g. see distribution of PAI scores near 0 in Fig. 2), one might reason that baseline depressive symptom severity alone is not as informative as our model incorporating baseline data on five variables. Second, a similar pattern of findings emerged in a control PAI analysis (in which all 12 a priori variables were included) relative to our primary analysis, suggesting that our findings are likely not attributable to overfitting due to running our PAI analysis on a reduced set of variables emerging from our variable selection steps.

**Limitations**

Several limitations should be noted. First, and importantly, prospective tests are needed in which a PAI model is built in one sample, and then tested in a separate sample. The $k$-fold cross-validation approach we used approximates such a test by leaving each patient’s data out of the model used to generate their predicted outcomes. However, although we implemented cross-validation during the weight-setting stage, we used the full sample for variable selection which can lead to overfitting and inflated associations (Hastie et al., 2009; Fiedler, 2011). Until such models are tested and replicated in separate samples, it will be difficult to determine the extent to which overfitting contributes to findings and whether models generalize to new sets of treatment-seeking depressed individuals. Second, we focused on clinical, demographic, and putative behavioral endophenotypes that could be collected at low cost and with relatively minimal clinic staff and patient burden. The extent to which neural assessments provide incremental predictive validity above and beyond such variables is an important direction for research, particularly with regards to relatively less costly and non-invasive imaging approaches (e.g. EEG). Third, it is unclear whether findings would generalize to depressed individuals who do not meet the inclusion/exclusion criteria of this trial. In addition, as others have highlighted (Uher et al., 2012b), measures of outcome (HRSD) and predictors include a certain amount of error, which may significantly attenuate the magnitude of observed predictor–outcome associations. Fourth, sample size was relatively small. Finally, the current PAI model relies on randomized designs (i.e. to examine outcomes for those randomly assigned to their indicated v. non-indicated treatment). An important future direction for research is to adapt these statistical models for the investigation of optimal treatment assignment in current clinical practice settings in which patients are not randomly assigned to interventions. These limitations notwithstanding, our findings demonstrate the potential for precision medicine to improve individual outcomes through model-guided treatment recommendations rather than the current practice of trial-and-error. Findings from replicated prescriptive algorithms could ultimately be used to inform the development of web-based ‘treatment selection calculators’ available to clinicians and patients to facilitate decision-making.

**Financial support.** The EMBARC study was supported by the National Institute of Mental Health of the National Institutes of Health under award numbers U01MH092221 (Trivedi M.J.) and U01MH092250 (McGrath P.J., Parsey R.V., Weissman M.M.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported by the EMBARC National Coordinating Center at UT Southwestern Medical Center (Madhukar H. Trivedi, M.D., Coordinating PI) and the Data Center at Columbia and Stony Brook Universities. Christian A. Webb and Diego A. Pizzagalli were partially supported by SK23MH108752 and 2R37MH068376, respectively. Zachary D. Cohen and Robert J. DeRubeis are supported in part by a grant from MQ: Transforming mental health MQ144PM_27. The opinions and assertions contained in this article should not be construed as reflecting the views of the sponsors.
Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291718001708

Acknowledgements. This study was supported by NIMH grants (M.H.K., grant number U01MH092221; P.J.M., R.V.P., M.M.W., grant number U01MH092250).

Conflict of interest. In the last three years, the authors report the following financial disclosures, for activities unrelated to the current research: Dr Trivedi reports the following lifetime disclosures: research support from the Agency for Healthcare Research and Quality, Cyberonics Inc., National Alliance for Research on Schizophrenia and Depression, National Institute on Mental Health, National Institute on Drug Abuse, National Institute of Diabetes and Digestive and Kidney Diseases, Johnson & Johnson, and consulting and speaker fees from Abbott Laboratories Inc., Akzo (Organon Pharmaceuticals Inc.), Allergan Sales LLC, Alkermes, AstraZeneca, Axon Advisors, Brintellix, Bristol-Myers Squibb Company, Cephalon Inc., Cerecor, Eli Lilly & Company, Evotec, Fabeŕ Kramer Pharmaceuticals Inc., Forest Pharmaceuticals, GlaxoSmithKline, Health Research Associates, Johnson & Johnson, Lundbeck, MedAvante Medscape, Medtronic, Merck, Mitsubishi Tanabe Pharma Development America Inc., MSI Methylation Sciences Inc., Nestle Health Science-PamLab Inc., Naurex, Neuroretinics, One Carbon Therapeutics Ltd., Otsuka Pharmaceuticals, Pamlab, Parke-Davis Pharmaceuticals Inc., Pfizer Inc., PgxHealth, Phoenix Marketing Solutions, Rexahn Pharmaceuticals, Ridge Diagnostics, Roche Products Ltd., Sepracor, SHIRE Development, Sierra, SK Life and Science, Sunovion, Takeda, Tal Medical/Puretech Venture, Targacept, Transcept, VantagePoint, Vivus, and Wyeth-Ayerst Laboratories. Dr Dillon reports funding from NIMH, consulting fees from Pfizer Inc. Dr Fava reports the following lifetime disclosures: http://mglicme.org/faculty/faculty-detail/maurizio_fava. Dr Weissman reports funding from NIMH, the National Alliance for Research on Schizophrenia and Depression (NARSAD), the Sackler Foundation, and the Templeton Foundation; royalties from the Oxford University Press, Perseus Press, the American Psychiatric Association Press, and MultiHealth Systems. Dr Oquendo reports funding from NIMH; royalties for the commercial use of the Columbia-Suicide Severity Rating Scale. Her family owns stock in Bristol Myers Squibb. Dr McInnis reports funding from NIMH; consulting fees from Janssen and Otsuka Pharmaceuticals. Dr McGrath has received research grant support from Naurex Pharmaceuticals (now Allergan), Sunovion, and the State of New York. Dr Pizzagalli reports funding from NIMH and the Dana Foundation; consulting fees from Akili Interactive Labs, BlackThorn Therapeutics, Boehringer Ingelheim, Takeda Pharmaceuticals USA and Posit Science. Dr Trombello currently owns stock in Merck and Gilead Sciences and within the past 36 months previously owned stock in Johnson & Johnson. Drs Adams, Cohen, Bruder, Cooper, Deldin, DeRubeis, Fournier, Huys, Jha, Kurian, McGrath, Parsey, Webb, Ms Goer report no financial conflicts.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References


Fiedler K (2011) Voodoo correlations are everywhere – not only in neurocience. Perspectives on Psychological Science 6, 163–171.


Fried EL and Nesse RM (2015b) Depression sum-scores don’t add up: why analyzing specific depression symptoms is essential. BMC Medicine 13, 72.


