Neuropsychopharmacology (2015) 40, 1332–1342

www.neuropsychopharmacology.org

A Cognitive–Emotional Biomarker for Predicting Remission with Antidepressant Medications: A Report from the iSPOT-D Trial

Amit Etkin*[,1,2], Brian Patenaude[1,2], Yun Ju C Song[3], Timothy Usherwood[4], William Rekshan[5,6], Alan F Schatzberg[1], A John Rush[7] and Leanne M Williams*[,1,2,3]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA; [2]Sierra-Pacific Mental Illness Research, Education, and Clinical Center (MIRECC) Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA; [3]Brain Dynamics Center, University of Sydney Medical School and Westmead Millennium Institute for Medical Research at Westmead Hospital, Sydney, NSW, Australia; [4]Department of General Practice, Sydney Medical School, Westmead, University of Sydney, Sydney, NSW, Australia; [5]Brain Resource, Sydney, NSW, Australia; [6]Brain Resource, San Francisco, CA, USA; [7]Duke-National University of Singapore, Singapore, Singapore

Depression involves impairments in a range of cognitive and emotional capacities. It is unknown whether these functions can inform medication choice when considered as a composite predictive biomarker. We tested whether behavioral tests, grounded in the neurobiology of cognitive and emotional functions, predict outcome with common antidepressants. Medication-free outpatients with nonpsychotic major depressive disorder ($N = 1008$; 665 completers) were assessed before treatment using 13 computerized tests of psychomotor, executive, memory–attention, processing speed, inhibitory, and emotional functions. Matched healthy controls ($N = 336$) provided a normative reference sample for test performance. Depressed participants were then randomized to escitalopram, sertraline, or venlafaxine–extended release, and were assessed using the 16-item Quick Inventory of Depressive Symptomatology (QIDS-SR_{16}) and the 17-item Hamilton Rating Scale for Depression. Given the heterogeneity of depression, analyses were furthermore stratified by pretreatment performance. We then used pattern classification with cross-validation to determine individual patient-level composite predictive biomarkers of antidepressant outcome based on test performance. A subgroup of depressed participants (approximately one-quarter of patients) were found to be impaired across most cognitive tests relative to the healthy norm, from which they could be discriminated with 91% accuracy. These patients with generally impaired cognitive task performance had poorer treatment outcomes. For this impaired subgroup, task performance furthermore predicted remission on the QIDS-SR_{16} at 72% accuracy specifically following treatment with escitalopram but not the other medications. Therefore, tests of cognitive and emotional functions can form a clinically meaningful composite biomarker that may help drive general treatment outcome prediction for optimal treatment selection in depression, particularly for escitalopram.
Neuropsychopharmacology (2015) **40,** 1332–1342; doi:10.1038/npp.2014.333; published online 21 January 2015

INTRODUCTION

Major depressive disorder (MDD) is a common and disabling condition (WHO, 2011). There is a range of treatment options, but only approximately one-third of patients reach remission with any single antidepressant (Rush *et al*, 2006; Trivedi *et al*, 2006a, b). Unfortunately, there are no widely accepted, clinically applicable predictors of outcomes to guide treatment choice.

One approach for improving prediction of outcome is to use tests that quantify specific neurobiological impairments that are inherent to MDD and are targeted by antidepressants. These impairments include the loss of cognitive and emotional capacities, and has been extensively described in more than three decades of work using behavioral tests (Gotlib and Joormann, 2010; Snyder, 2013) (Supplementary Table 1). This work supports the formulation that depression is characterized by perturbations in psychomotor response speed, processing speed, executive functions (eg, attention and working memory), memory encoding, and recall and emotion processing. Within this work, there are also suggestions that performance on some of these tests may predict antidepressant medication outcomes (Supplementary Table 1) that formed in part the basis for the International Study to Predict Optimized Treatment in Depression (iSPOT-D) (Williams *et al*, 2011). Specifically, prior smaller-scale studies suggest that poor cognitive

*Correspondence: Dr A Etkin, Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA, Tel: +1 650 725 5736, Fax: +1 650 724 9900,
E-mail: amitetkin@stanford.edu or Dr LM Williams, Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA, Tel: +1 650 723 3579, E-mail: leawilliams@stanford.edu
Received 19 July 2014; revised 12 November 2014; accepted 13 November 2014; accepted article preview online 30 December 2014

performance, such as working memory on the N-back task (Gorlyn *et al*, 2008), information processing speed on the digit symbol task (Leuchter *et al*, 2004), executive functioning and flexibility on the Wisconsin Card sort task (Dunkin *et al*, 2000), and color naming on the color and word Stroop task (which may reflect psychomotor slowing) (Taylor *et al*, 2006), all predicted worse outcome with an acute course of antidepressant treatment (Supplementary Table 1). Importantly, these findings regarding treatment outcome suggest that there may be a relationship between the pathophysiology of depression, which includes a broad-based dysfunction in cognition, and the capacity of these individuals to respond to treatment.

This report investigates whether performance on a standardized computerized battery of 13 tests of cognitive and emotional capacities (Table 1), given to antidepressant-medication-free, depressed outpatients before treatment, predicts remission or response of depressive symptoms after 8 weeks of acute treatment. These tests include not only similar tests previously shown to predict treatment outcome, but also significantly extend upon these in both scope and breadth. As such, the combination of these tests can be taken as composite elements of a single predictive test (ie, a 'biomarker' of treatment outcome), particularly because no single prior behavioral test has provided sufficient predictive utility in isolation.

Using performance on this behavioral battery, we aimed to make three determinations. First, we examined the formulation of depression as a disorder in which cognitive and emotional capacities are perturbed by determining whether performance in the behavioral battery would differentiate depressed participants from matched healthy controls, and whether heterogeneity in behavioral performance within the depression group contributed to the differentiation from controls. Second, we assessed whether behavioral performance could predict treatment outcome to each of the three antidepressant medications, predicting based on prior work ((Gorlyn *et al*, 2008; Leuchter *et al*, 2004; Dunkin *et al*, 2000; Taylor *et al*, 2006) and Supplementary Table 1) that nonresponders would be characterized by impaired cognitive functioning relative to responders. Third, we tested whether the prediction generated for one antidepressant medication supported the ability to select between medications in the study, by virtue of differentially predicting outcome to this medication *vs* the others. Importantly, prediction analyses were done with cross-validation to evaluate their ability to generalize to new individuals.

MATERIALS AND METHODS

Details of the iSPOT-D study design and protocols have been reported elsewhere (Williams *et al*, 2011). In brief, 1008 adults (18–65 years old) with first-onset or recurrent, nonpsychotic MDD (age: 37.8 years (SD 12.6), education: 14.5 years (SD 2.8), and 57% female) were enrolled at 17 study sites using broad inclusion and minimal exclusion criteria to recruit a sample consistent with outpatient clinical practice (Supplementary Figure 1) (see Saveanu *et al*, in press, for detailed recruitment information and sociodemographic features, with relevant information excerpted in the Supplementary Methods and Supplementary Figure 2 of this report). The study also recruited 336 healthy controls matched in age, gender, and years of education (age: 37.0 years (SD 13.1), education: 14.4 years (SD 3.6), and 57% female). The study received approval by local institutional review boards. After providing a complete description of the study to the participants, written informed consent was obtained.

Protocol Treatment

Before randomization, all psychotropic medications—except sleep aids and anxiolytics—were discontinued for at least 1 week. Participants were randomized with equal probability to receive escitalopram, sertraline, or venlafaxine–extended release (venlafaxine-XR). Doses were adjusted by the participant's usual treating clinician according to their routine clinical practice. Given the practical trial design, participants and treating clinicians were not blind to treatment assignment. Medications were allowed for associated symptoms (eg, insomnia), adverse drug reactions (eg, nausea), and concurrent general medical conditions.

Assessments and Outcome Measures

DSM-IV diagnoses were made based on the structured diagnostic Mini-International Neuropsychiatric Interview and were confirmed by licensed and trained MD or PhD clinicians (American Psychiatric Association, 1994; Sheehan *et al*, 1998). Study visits occurred at week 0 (pretreatment or baseline) and week 8. At both visits, blinded clinician raters completed the 17-item Hamilton Rating Scale for Depression ($HRSD_{17}$) (Hamilton, 1960), and participants completed the 16-item Quick Inventory of Depressive Symptomatology–Self-Report (QIDS-SR_{16}, ratings that cannot be blinded) (Rush *et al*, 2003). Study site personnel made telephone calls to participants at day 4 and weeks 2, 4, and 6 to monitor antidepressant dosage, compliance, concomitant medications, and adverse events (Williams *et al*, 2011). For this report, we considered as primary outcomes remission as defined by either an $HRSD_{17}$ score ≤ 7 or a QIDS-SR_{16} score ≤ 5 and response ($\geq 50\%$ decrease from baseline in either the $HRSD_{17}$ or QIDS-SR_{16} score), adjusting family-wise error for multiple comparisons across this full family of outcomes. Response is commonly used to define a clinically meaningful benefit, but it is an arbitrary end point that is dependent on the length of the trial. However, remission, perhaps a more definitive end point as well as the ultimate goal of treatment, has the disadvantage that many patients experience a large decrease in symptoms (ie, responders) but do not fully remit. As such, we considered response and remission to each be meaningful targets for classification. We also note that recent large-scale trials, such as STAR*D, use both clinician ratings of symptoms (typically the $HRSD_{17}$) and self-reported symptoms (on the QIDS-SR_{16}). It is unknown whether prediction using biological assays (eg, behavioral tests) aligns with outcome on either or both of the $HRSD_{17}$ or the QIDS-SR_{16}. Each scale captures a different mix of depression-related symptoms. Thus, we chose to use both as targets for classification with appropriate control for multiple comparisons. iSPOT-D was designed *a priori* to include both $HRSD_{17}$ and QIDS-SR_{16} as treatment outcome end points (Williams *et al*, 2011).

Table 1 Cognitive and Emotional Tests Taken by Participants as Part of the Standard Computerized Test Battery

Summary measure name	Test	Construct	Outcome measures	Test description	Tests assessing equivalent construct
Psychomotor function	Motor Tapping	Psychomotor function	Number and variability of taps	Tapping index finger as fast as possible for 30 s; assessing sensorimotor response speed	Finger Tapping
Decision speed	Choice reaction time	Simple decision RT	Average RT, variability of RT	Respond to one of four circles as they light up; assesses decision-related reaction time. Assessing sensorimotor coordination and speed	Corsi Blocks
Verbal memory	Memory recall	Declarative verbal memory	Accuracy (recall, intrusion errors), learning rate	Learn and then recall lists of 12 words; assesses learning, memory recall.	Rey Auditory Verbal Learning Test California Verbal Learning Test
Working memory	Digit span	Working memory	Accuracy (total recall, maximum recall span)	Repeat a series of digits in forward and backward order; assessing working memory	Digit span
Cognitive flexibility	Verbal interference (color-word Stroop)	Cognitive control	Accuracy (errors), RT	Respond to the name of color word (ignore color) and then color word presented (ignore name); assessing suppression of automatic responses	Stroop
Attention	Continuous performance test	Sustained attention–working memory	Accuracy (total, false positive, false negative errors), RT, variability of RT	Sustained attention to series of letters (D, C, G, or T). Identify when same letter is 1-back. Requires working memory updating	Conners CPT, TOVA
Response inhibition	Go/No-Go	Response inhibition	Accuracy (total, false positive, false negative errors), RT, variability of RT	Press response pad as quickly as possible to 'Go' (green) trials, and withhold to 'No-Go' (red) trials. Assessing impulsivity *vs* inhibition	
Information processed speed	Switching of attention	Information processing speed–executive function	Accuracy (switching errors), completion time, connection time	Connect a sequence of alternating numbers and letters; assesses information processing efficiency	Trails A and B (paper and pencil)
Executive function– maze navigation	Executive maze	Executive function	Accuracy (total, overrun errors), completion time	Discover (by trial and error) a maze path; reflecting planning, monitoring feedback, and error correction	Austin maze
Emotion identification accuracy and RT (summary measures for emotion ID accuracy, RT, and relative RT (emotion minus neutral))	Explicit emotion identification	Explicit emotion processing	Accuracy, RT	Identify emotion shown on a facial expressions (anger, disgust, fear, sadness, happiness)	Penn Emotion Test
Emotion bias RT (summary measure is relative RT (emotion minus neutral))	Emotion attention bias	Implicit emotion processing	RT	Implicit influence of prior exposure to emotion on subsequent 'old/new' recognition of a face	

Abbreviations: CPT, continuous performance test; RT, reaction time; TOVA, test of variables of attention.

Behavioral Tests of Cognitive and Emotional Functions

At baseline, participants completed a computerized battery of tests designed to evaluate a range of cognitive and emotional capacities including attention, working memory, psychomotor response speed, cognitive flexibility of task shifting, response inhibition, verbal memory, processing speed, decision speed, emotion identification, and emotional biasing of memory for faces (Paul *et al*, 2005; Mathersul *et al*, 2009; Williams *et al*, 2009) (Table 1). The commercially available battery (Brain Resource), presented at a grade 5 reading level, was run locally at each study site on a computer equipped with dedicated software and a touch screen. The software precluded access to other programs or the internet. Behavioral performance on the tests was measured by reaction times and accuracies. To create summary performance measures of each of the 13 tests, we normalized each measure to the benchmark from

the 336 healthy controls (ie, as standardized *z*-scores relative to a control mean of 0) and averaged normalized measures (eg, accuracy and reaction time) within each test. Normalization was therefore required to be able to combine across different measures such as accuracy and reaction times. Values on each measure were aligned such that positive meant better performance and negative meant worse performance. By doing so, we could more readily interpret the weights on our summary measures relative to healthy performance.

Characterization of Heterogeneity in Cognitive and Emotional Test Performance

As MDD is well known to be a heterogeneous condition with a spectrum of impairments in cognitive and emotional capacities, we used data-driven assumption-free methods to characterize baseline heterogeneity in task performance. To do so, we performed a clustering analysis on participants' cognitive and emotional test scores in the full sample of 1008 MDD participants. This approach draws on procedures in other established areas of psychiatry, such as the quantification of cognitive behavioral heterogeneity in schizophrenia (Horan and Goldstein, 2003; Dawes *et al*, 2011). To identify the number of cohesive clusters in the sample, we used *K*-means method to cluster the summary scores. We used the 'elbow' method to compare 1 with 10 cluster solutions to identify the optimal solution. This method looks at the percentage of variance explained as a function of the number of clusters to identify the point beyond which there was only a marginal gain in variance explained. We included all behavioral tests in our cluster analyses to ensure we identified clusters based on profile of relative impairment across the range of cognitive and emotional capacities previously implicated as abnormal in depression, but in most cases have not been assessed in the same patients nor understood in relation to treatment outcome prediction.

Ensemble Pattern Classification for Patient-Level Prediction

We chose to analyze data using a cross-validated multivariate pattern classification approach rather than conventional regression methods (Hastie *et al*, 2009). The pattern classification approach (1) incorporates complex interactions between variables, (2) potentially maximizes our sensitivity for the detection of predictive effects, and (3) provides a reliable single patient-level prediction. The cross-validation is important as it (1) reduces the bias of the generalizability estimate of the predictive models and (2) establishes an estimate of generalization accuracy, sensitivity, and specificity (which is critical for evaluating the clinical relevance of our findings). To implement the classifier, we focused our analyses on participants who completed the study per protocol, as determined by having baseline and week 8 clinical scores and taking the randomized medication ($N = 655$). This was done because imputation or mixed model approaches that are typically used in clinical outcome studies to account for study attrition in an intent-to-treat framework are difficult to implement within the context of a classifier, and imputing missing data before running the classifier would violate assumptions of the cross-validation. As an additional control analysis, we examined whether outcome prediction held in an intent-to-treat framework in which we conservatively considered all dropouts to be nonremitters/nonresponders in the model. Our classifier comprised three major components: a data transformation component, a discriminant function (using linear discriminant analysis), and an ensemble classifier.

Cross-validation overview. The cross-validation procedure used to evaluate the classifier consisted of repeatedly dividing the data into a training set that was used to establish a predictive model (on 80% of a bootstrap subsample) and a test set that consisted of the remaining left-out data and upon which the predictive model was applied (Supplementary Figure 3). Each bootstrap subsample was also limited by the size of the smaller outcome group, with an equal number of participants included from each outcome group (eg, if a sample contained 80 remitters and 100 nonremitters, a bootstrap subsample would contain a random set of 64 remitters and 64 nonremitters). This was repeated 1000 times for each classifier to form an ensemble classifier using a majority vote rule.

Data transformation. Within each bootstrap subsample, the ranges of each predictor (emotion or cognition capacity summary score) were normalized to the range $(-1, 1)$ before estimating the linear discriminant analysis (LDA) model parameters. The data transformation was all done within the cross-validation loop (ie, parameters were estimated from the training set and then applied to the test set).

Discriminant function (LDA). This describes the method used for modeling the relationship between cognitive and emotional test summary scores and outcome measure (eg, remitter *vs* nonremitter) in a given bootstrap training sample. We used LDA for this purpose as this method has been well described for decades (Hastie *et al*, 2009), its interpretation is straightforward (and related to commonly understood principle components analysis methods), and it has been shown to yield robust classification with relatively minimal skew in sensitivity and specificity (Maroco *et al*, 2011).

Ensemble classifier. LDA, as used above on a training sample, provides a model for each bootstrap subsample and a classification for each left out subject (ie, the remaining 20%). This resulted in multiple predictions for each participant that were summarized into a single classification using a majority vote procedure across the family of models. Confidence intervals (95%) were determined by using 1000 bootstrap samples without replacement.

Statistical Analyses

The statistical significance of each classifier was determined by permuting responder/nonresponder labels for each participant 2500 times, resulting in a one-sided *p*-value for the observed classification accuracy. To correct for multiple comparisons made across the 24 behavioral

performance treatment prediction models that were run (see Table 1), we used a Bonferroni correction controlling family-wise error at $p<0.05$ (Dunn, 1961), yielding an uncorrected critical value of $p=0.0021$. We also reported multivariate effect sizes for the linear discriminant analyses using the Mahalanobis distance (Mahalanobis, 1936; Kline, 2013). Subsequent analyses in the tables and text used logistic regression (eg, to predict remission based on predictor outcome), independent sample *t*-tests (eg, to assess differences in a measure between participants who were predicted to remit *vs* not to remit), or χ^2 tests (eg, gender distributions).

To examine whether results from a given classifier developed on one medication could predict differential outcome between medications, we first determined the outcome prediction of each significant classifier for every participant, independent of which medication they had actually received. This was possible as classification was determined through cross-validation. Hence, data for participants who did not receive a medication were handled in a similar manner to data for participants who did receive a medication but were left out of a bootstrap subsampling. We then conducted logistic regression analyses in SPSS 20 (SPSS, Chicago, IL) with factors of classifier prediction (eg, remit/not remit) and medication received in contrasts comparing the medication that was the target of the classifier with the other medications while controlling for age, education, gender, study site, and depressive severity ($HRSD_{17}$, QIDS-SR_{16}). Inclusion of these covariates ensured that easily ascertained clinical and demographic variables did not account for prediction effects that were possible using our behavioral battery. Effect sizes were quantified with odds ratios, and where appropriate a number needed to treat (NNT). In this case, NNT refers to the minimal number of participants who would need to be assessed with the classifier to capture one additional remission/response event.

RESULTS

Identifying Depression Subgroups Based on Pretreatment Cognitive and Emotional Test Performance

Cluster analysis showed that the MDD participants fell into two subgroups in terms of cognitive and emotional test performance. The first was an 'intact' subgroup, composed of approximately 3/4 of the MDD participants who performed on average within the healthy range (Table 2). The second 'impaired' subgroup was composed of MDD participants with a test performance well below the healthy norm for 11 of the 13 aspects of function. Table 2 shows that the impaired subgroup was older, less educated, and had a modestly greater depressive severity than the intact performance cluster. The intact subgroup had a better overall response to treatment. The distribution of patients by medication arm did not differ across the two subgroups ($\chi^2=1.575$, $p=0.455$). Thus, clustering by intact *vs* impaired performance, we could describe heterogeneity in test performance in a way that also mapped onto the general likelihood of achieving successful treatment outcomes.

When MDD participants were considered as a single group, our pattern classification analysis significantly differentiated them from healthy controls, but only at a modest level (56% accuracy, $p=0.002$). In contrast, when we considered intact and impaired subgroups separately, the separation of both of these subgroups from controls improved: intact 57% accuracy ($p<0.001$), and impaired 91% accuracy ($p<0.001$). These data illustrate the impact of heterogeneity in cognitive and emotional function in MDD and support the use of the clustering results for stratifying treatment prediction analyses.

Cognitive–Emotional Composite Biomarker Prediction of Patient-Level Medication Outcomes

After correction for multiple comparisons, we observed prediction of remission outcomes by pretreatment tests of cognitive–emotional function for the escitalopram arm of the impaired subgroup (Table 3). Specifically, we classified remission to escitalopram on the QIDS-SR_{16} scale with 72% accuracy ($p<0.001$, corrected $p=0.048$). As shown in Figure 1, QIDS-SR_{16} remission rates with escitalopram treatment were higher for individuals predicted to remit with escitalopram (58%) than for those predicted to not remit (16%; logistic regression odds ratio (OR) 7.5, $p=0.001$). For comparison, the remission rate was 37% if classifier prediction was not taken into consideration. Patients predicted to not remit with escitalopram were characterized by generally impaired cognitive functioning. When each capacity was considered separately (after a Bonferroni correction for multiple comparisons), impairments were greatest in patients predicted to be nonremitters to escitalopram for attention, decision speed, working memory, and speed of emotion identification (p's<0.003). This subset of tests overlaps with those previously demonstrated to predict treatment outcome in other studies, as noted in Supplementary Table 1. However, on its own, this focal subset of tests failed to yield significant classification of remission status (see Supplementary Table 2), suggesting that the full profile of behavioral performance needs to be taken into account for robust prediction that survives cross-validation and correction for multiple comparisons.

We also examined whether the escitalopram prediction held in a conservative intent-to-treat analysis in which we assumed that all noncompleters were nonremitters. We still found significant classification (67% accuracy, $p=0.0012$). Moreover, in an analysis of completers *vs* noncompleters, we were unable to differentiate the groups (54% accuracy, $p=0.064$), suggesting that baseline task performance differences associated with attrition are minimal.

Next, we tested whether the patient-level predictions generated by the QIDS-SR_{16} remission escitalopram classifier differentially predicted outcome in a comparison of escitalopram with the other two medications among participants in the impaired performance subgroup (see Materials and Methods). Covariates included age, education, gender, depressive severity on the $HRSD_{17}$ and QIDS-SR_{16}, and study site. Logistic regression analyses revealed a significant interaction between medication arm (as a three-level factor) and prediction from the escitalopram classifier (Figure 1; OR 0.3, $p=0.019$). This interaction was driven by prediction results for escitalopram, as illustrated by a significant interaction between a medication arm factor coding for escitalopram *vs* the two other medications, and

Table 2 Depression Subgroup Clustering Defined by Profiles of 'Intact' and 'Impaired' Cognitive–Emotional Function, with Comparison with Each Other and with the Healthy Controls (to Whom Behavioral Scores Were Standardized; $z = 0$)

Demographics	Intact ($N = 735$)		Impaired ($N = 273$)		*P*-value intact *vs* impaired	*P*-value healthy *vs* intact	*P*-value healthy *vs* impaired
	Mean	SD	Mean	SD			
Age (years)	35.0	11.6	45.6	11.9	<0.001	0.014	<0.001
Education (years)	14.8	2.6	13.8	3.1	<0.001	0.513	<0.001
Pretreatment symptoms							
$HRSD_{17}$	21.7	3.9	22.4	4.5	0.025	<0.001	<0.001
$QIDS\text{-}SR_{16}$	14.4	3.8	14.5	3.9	0.721	<0.001	<0.001
Pretreatment cognitive–emotional test performance (z-score)							
Attention	−0.16	0.58	−0.73	0.96	<0.001	<0.001	<0.001
Cognitive flexibility	0.08	0.54	−0.75	0.93	<0.001	0.278	<0.001
Decision speed	0.04	0.65	−0.76	1.16	<0.001	0.925	<0.001
Executive function	0.14	0.45	−1.50	1.34	<0.001	0.029	<0.001
Information processing speed	0.09	0.59	−0.72	0.57	<0.001	0.116	<0.001
Psychomotor response speed	−0.04	0.70	−0.60	1.00	<0.001	0.162	<0.001
Response inhibition	−0.03	0.49	−0.97	1.30	<0.001	0.056	<0.001
Verbal memory	−0.05	0.80	−1.10	0.85	<0.001	0.406	<0.001
Working memory	0.16	0.89	−0.74	0.88	<0.001	0.015	<0.001
Explicit emotion identification accuracy	0.04	0.48	−0.71	0.82	<0.001	0.331	<0.001
Explicit emotion identification speed	0.02	0.64	−0.50	0.83	<0.001	0.960	<0.001
Explicit emotion identification bias	0.01	0.68	0.03	0.94	0.590	0.977	0.645
Implicit emotion priming of recognition speed	−0.03	0.73	0.08	1.12	0.139	0.625	0.276
	%	*N*	%	*N*		0.791	0.681
Female gender (% female)	56	191	59	160	0.444		
Treatment outcome (completers per protocol)							
$HRSD_{17}$ remission	48	232	39	71	0.043		
$HRSD_{17}$ response	65	316	56	100	0.030		
$QIDS\text{-}SR_{16}$ remission	40	188	32	54	0.057		
$QIDS\text{-}SR_{16}$ response	55	249	49	80	0.242		

Abbreviations: $HRSD_{17}$, 17-item Hamilton Rating Scale for Depression; $QIDS\text{-}SR_{16}$, 16-item Quick Inventory of Depressive Symptomatology—Self-Rated.
Note that *P*-values reflect independent sample *t*-tests for all except gender and treatment outcome, for which they reflect logistic regression tests.

prediction from the escitalopram classifier (OR 6.5; $p = 0.020$). Specifically, individuals predicted to remit with escitalopram remitted at a higher rate if they received escitalopram compared with the other two medications (58% *vs* 32%; NNT = 3.8, $p = 0.016$), whereas those predicted to not remit with escitalopram remitted at a lower rate if they received escitalopram compared with the other two medications (16% *vs* 26%; NNT = 9.7, $p = 0.300$).

When restricting the interaction analysis to pairwise comparisons between medications, we found a significant interaction between medication arm in a comparison of escitalopram and venlafaxine-XR and prediction from the escitalopram classifier (OR 8.7; $p = 0.029$), and a trend when comparing escitalopram with sertraline (OR 5.4; $p = 0.064$). For the intact cognition subgroup, we found no interaction between medication arm (escitalopram *vs* the other two medications) and predictions from the escitalopram classifier ($p = 0.223$).

Validation of Classifier Specificity

We next compared using independent sample *t*-tests and χ^2 tests (gender only) to compare clinical/demographic characteristics between participants who received escitalopram and were predicted to remit compared with those predicted to not remit. We found no significant differences in age, gender, or $HRSD_{17}$ score (Table 4), but found that participants predicted to remit had slightly lower depressive severity on the $QIDS\text{-}SR_{16}$, were more educated, and received a lower dose of escitalopram at week 8. Nonetheless, controlling for these clinical/demographic variables, as well as final dose and study site, yielded similar results (OR 9.9, $p = 0.005$).

Identification of the Key Tests that Contribute to Classifier Performance

To visualize the relative contribution of each of the behavioral tests to the classifier, we plotted the average of the absolute value of the LDA weights on each test score across the prediction models that comprised the final ensemble classifier (Figure 2a). These weights together yield the classification 'equation.' Using this metric, all tests have broadly similar weights, with emotion identification differential reaction time (ie, emotion minus neutral), speed, and psychomotor function having the greatest weights, and

Table 3 Classification Results for the Different Outcomes, Divided by Medication Arm

Outcome criterion	Performance cluster	Medication	Number	Accuracy	± 95% confidence interval	Sensitivity	Specificity	Uncorrected *p*-value	Corrected *p*-value	Mahalanobis distance
$HRSD_{17}$ remission	Intact	ESC	152	58	0.046	61	55	0.082	n.s.	1.84
		SER	175	51	0.060	51	46	0.668	n.s.	1.66
		VEN	152	38	0.057	36	37	0.500	n.s.	1.69
	Impaired	ESC	65	58	0.099	53	59	0.235	n.s.	1.68
		SER	59	59	0.110	57	61	0.115	n.s.	1.60
		VEN	52	54	0.064	60	53	0.151	n.s.	1.63
$HRSD_{17}$ response	Intact	ESC	152	53	0.055	59	46	0.245	n.s.	1.62
		SER	175	55	0.057	60	39	0.309	n.s.	1.62
		VEN	152	38	0.080	35	40	0.500	n.s.	1.61
	Impaired	ESC	65	56	0.097	57	48	0.307	n.s.	1.60
		SER	59	49	0.085	51	48	0.501	n.s.	1.62
		VEN	52	54	0.110	57	45	0.550	n.s.	1.60
$QIDS\text{-}SR_{16}$ remission	Intact	ESC	152	36	0.068	29	40	0.500	n.s.	1.59
		SER	175	51	0.042	57	51	0.232	n.s.	1.61
		VEN	152	43	0.071	45	40	0.992	n.s.	1.60
	Impaired	ESC	65	72	0.061	79	69	0.002	0.048	1.58
		SER	59	64	0.057	66	63	0.021	n.s.	1.59
		VEN	52	58	0.120	58	56	0.126	n.s.	1.59
$QIDS\text{-}SR_{16}$ response	Intact	ESC	152	50	0.047	50	46	0.901	n.s.	1.64
		SER	175	62	0.029	65	64	0.004	n.s.	1.62
		VEN	152	46	0.057	42	51	0.891	n.s.	1.62
	Impaired	ESC	65	67	0.100	72	60	0.016	n.s.	1.63
		SER	59	45	0.130	46	45	0.810	n.s.	1.61
		VEN	52	61	0.095	57	64	0.084	n.s.	1.64

Abbreviations: CI, confidence interval; ESC, escitalopram; $HRSD_{17}$, 17-item Hamilton Rating Scale for Depression; $QIDS\text{-}SR_{16}$, 16-item Quick Inventory of Depressive Symptomatology–Self-Report; SER, sertraline; VEN, venlafaxine.
P-values reflect permutation tests on classifier accuracy.

information processing speed and verbal memory having the lowest weights.

To get a different summary, we conducted another analysis in which we removed each predictor variable one at a time, and quantified the reduction in model accuracy as a consequence (see Figure 2b), thus evaluating the criticalness of the presence of a variable. This analysis suggested that the variables that reflect cognitive control capacities (ie, working memory, attention, response inhibition, and information processing speed) were most essential for the classification results.

DISCUSSION

This study provides robust evidence that performance on a battery of standardized cognitive and emotional tests before treatment can predict a depressed patient's likelihood of responding to a medication. This study and analytic approach are notable by the large sample size and our ability to generate and test treatment prediction within this group using cross-validation, hence demonstrating the ability of the predictive test to generalize to new individuals. These findings also have clinical significance because the cognitive and emotional tests are easily administered and the effect sizes are considerable. Performance on these tests revealed a meaningful subgrouping within depression. One subgroup—approximately a quarter of the total depressed sample—was notable in its poor performance across tests relative to the other depressed subgroup and healthy controls. Overall, this impaired performance subgroup had worse treatment response, consistent with prior findings of greater impairments in related tasks (working memory, cognitive flexibility, information processing speed) in treatment of nonresponders (see Supplementary Table 1). Furthermore, only in this subgroup were we able to predict treatment response based on individual differences in performance across these cognitive and emotional tests, independent of potential clinical and demographic confounders.

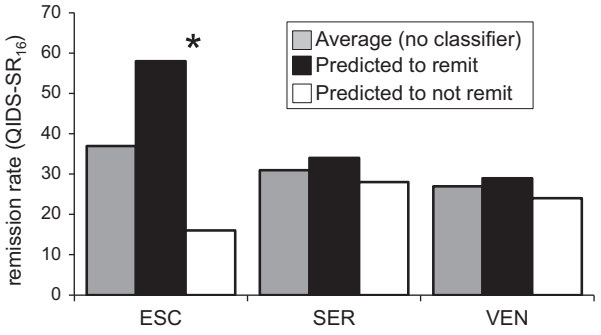

Figure 1 QIDS-SR_{16} remission rates based on predictions generated by the escitalopram QIDS-SR_{16} remission classifier, separated by medication arm. Plotted are remission rates when classifier outcome is not considered (gray bars, ie, current clinical practice), and response or remission rates when the QIDS-SR_{16} remission escitalopram classifier predicts that a participant will remit (black bars) or will not remit (white bars). *Significant difference between participants predicted to remit *vs* predicted not to remit on escitalopram (logistic regression odds ratio: 7.5; $p = 0.001$). ESC, escitalopram; QIDS-SR_{16}, 16-item Quick Inventory of Depressive Symptomatology; SER, sertraline; VEN, venlafaxine–extended release.

Table 4 Pretreatment Characteristics of Participants Predicted to Remit and to Not Remit within the 'Impaired' Subgroup Divided by Predictions from the Escitalopram QIDS-SR_{16} Remission Classifier

Characteristics	Predicted to not remit ($N = 34$)		Predicted to remit ($N = 31$)		*p*-value
	N	**%**	***N***	**%**	
Female gender (% female)	22	65	18	58	0.583
	Mean	**SD**	**Mean**	**SD**	
Age (years)	45.3	11.8	48.0	10.1	0.327
Education (years)	12.6	2.4	14.2	3.4	0.030
Symptoms					
$HRSD_{17}$	23.3	4.4	21.5	2.9	0.059
QIDS-SR_{16}	15.9	3.8	13.3	3.5	0.006
Medication final dose					
Escitalopram (mg)	14.6	9.0	11.3	3.6	0.046

Abbreviations: $HRSD_{17}$, 17-item Hamilton Rating Scale for Depression; QIDS-SR_{16}, 16-item Quick Inventory of Depressive Symptomatology–Self-Report.

Classification was verified to be robust through cross-validation, and confidence intervals were very tight around the classifier's mean accuracy. Individual patient-level predictions from our escitalopram classifier were specific to this medication and thus could be used to drive selection of optimal treatment for patients based on this information. In other words, for patients in the impaired subgroup, prediction of remission by the escitalopram classifier suggests choice of escitalopram with a very clinically significant effect size (NNT 3.8), whereas prediction of nonremission suggests choice of another medication other than escitalopram with a more modest effect size (NNT 9.7).

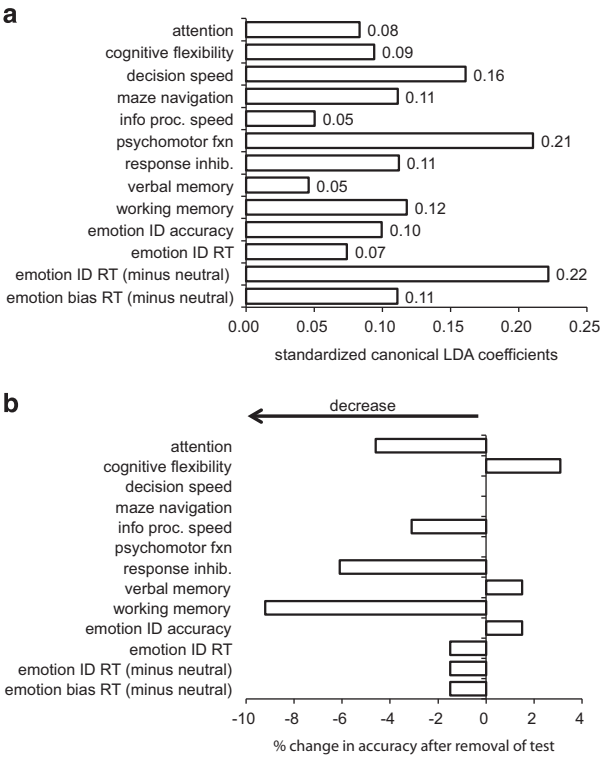

Figure 2 Relationships between individual tests and classifier performance. (a) Weights on each variable from the linear discriminant analysis. Shown are the means across all rounds of the classifier (not considering the majority vote ensemble step in the classifier). Together, these comprise the average classification 'equation' that can be applied to future data. (b) Impact of individual behavioral tests on classification accuracy for the QIDS-SR_{16} escitalopram remission classifier. Plotted is the change in accuracy after removal of each test. ID, identification; LDA, linear discriminant analysis; QIDS-SR_{16}, 16-item Quick Inventory of Depressive Symptomatology; RT, reaction time.

Furthermore, a lower dose of escitalopram may be required for those predicted to remit, further indicative of the clinical relevance of this classifier result.

Several points are important to note regarding the specifics of our treatment prediction classifier. First, classification was only found for patients who were impaired relative to healthy controls at baseline. Thus, cognitive–emotional predictors may be most relevant for those patients in whom they reflect domains of limited capacity (ie, not at 'ceiling'). That we only observed robust prediction of treatment outcome in participants with impaired task performance at baseline may limit the generalizability of the findings to the overall population of depressed patients (most of whom are not impaired on these tests). However, demonstration that prediction relates to the magnitude of the patient-related abnormality also helps strengthen the link between an understanding of the pathophysiology of depression (and potential heterogeneity in it), and the capacity of patients to recover with treatment. This also suggests that the impaired subgroup may reflect one potentially more homogenous 'type' of depressive pathology characterized by impaired cognitive task performance relative to the general population of depressed patients. In doing so, the findings may help move us closer to identifying illness subtypes that are neurobiologically

informed and also help guide treatment selection. In other words, membership of a patient in the impaired subgroup indicates generally poorer response to treatment, but even among this subgroup a sensitive classifier can nonetheless identify those patients whom one might expect to do less well on escitalopram. Treatment-response patients can thus be identified by a two-stage model: (1) normal-range task performance predicts better response, and (2) among patients performing below normal, the specific behavioral profile identified by the classifier predicts better response (to escitalopram specifically).

Second, by removing one variable at a time we discovered which functions were most critical for successful classification (ie, the 'glue capacities'). These critical functions could all be broadly characterized as related to cognitive control, consistent with prior work that has implicated improved cognitive control at baseline in predicting better treatment outcome (see Supplementary Table 1). Absence of classification for the intact performance subgroup may be related to performance being closer to ceiling, and thus with less variance with which to predict outcome (although the distribution was not itself truncated by ceiling performance). Thus, future testing with more adaptive behavioral designs that could identify more subtle deficits may be warranted.

Our findings also support the notion that a more direct assessment of neurobiologically relevant measures, such as performance on well-characterized behavioral tests that tap into the functioning of specific brain circuits, may allow greater insight into a patient's likelihood of response than a simple assessment of symptoms and demographics, as is commonly done now (and which does not provide information regarding medication-specific response). Furthermore, cognitive and emotional functions are not secondary reflections of these other easily obtained clinical variables (eg, demographics or clinical factors), but rather appear to be core predictive features in a clinically important subgroup of depressed patients. These features likely reflect the underlying neurobiology of the disorder, thus encouraging a linkage between brain-relevant measures and the definition of the disease and its treatments, without the need to rely on symptom reports. This view is furthermore consistent with recent reformulations of psychopathology, such as those in the National Institute of Mental Health's Research Domain Criteria project (Insel *et al*, 2010).

This study has several important translational limitations. First, because the focus was on pretreatment prediction of acute response to a first-line treatment, it is not known how the test battery would work in patients who are already taking psychotropic medications and for whom the prediction of options for switching or augmenting may be needed. Second, average doses of sertraline and venlafaxine-XR reflected usual practice in primary care but were at the lower end of the range that is typically used in psychiatric practice, even though response and remission rates were substantial. It is unknown whether the predictors identified would change if only higher doses of the medications were used, or if a specific dose regimen (rather than usual practice) was used. Similarly, it is unknown whether differential prediction of escitalopram is related to the relatively low doses of the medications that may serve to either increase or decrease their specificity for particular neural substrates. In addition, although all of the medications have serotonergic activity, failure to identify predictors of sertraline and venlafaxine might reflect the fact that these medications have additional significant dopaminergic and noradrenergic effects, respectively, especially at higher doses, whereas escitalopram is a more purely serotonergic medication (Tatsumi *et al*, 1997). Thus, classification may not be as specific for venlafaxine and sertraline at the broad, and generally low, set of doses represented in this study, whereas escitalopram prediction may simply be less heterogeneous.

We also note that a placebo group was not included in the study because it was designed as a practical study with translation in mind. The goal of the practical design was to identify predictors of treatment outcome for antidepressants in common use and with previously established efficacy. Moreover, by developing treatment prediction algorithms within the context of usual practice, we also advance the goal of translating our findings to clinical care. Related to these design issues, neither participants nor prescribers were blind to treatment arm. Nonetheless, the observed outcome predictions in this study provide information directly relevant to the typical context under which antidepressant medication is prescribed (ie, in a nonblinded manner). It is important to note that our aim was to predict individual differences in treatment outcome in the context of real-world clinical practice, and not to evaluate mean differences in overall outcome between treatment arms (which is the aim of traditional clinical trials).

In conclusion, we found that response with antidepressant medication can be reliably predicted for outpatients with MDD by their pretreatment performance on a standardized test battery of cognitive and emotional function. Moreover, prediction for one medication (escitalopram) was specific for this medication, and thus may be used to support medication selection based on test performance. Importantly, this prediction was only evident in a subgroup of participants who had impaired performance across these tests relative to other depressed participants and healthy controls. These findings have the potential to inform personalized care and enhance our understanding of the cognitive and emotional neurocircuitry of depression. Though our results were derived from a rigorous classification procedure and all information presented regarding the classifiers reflected cross-validated results in independent subsamples within our larger sample, ultimate verification of our findings will require replication in a new patient sample. In light of the readily deployable nature of our standardized behavioral assessment battery, which can be performed on home or office computers, these findings also open a practical avenue for testing the replicability of these findings as well as testing other antidepressants and contexts in which patients cannot be assessed medication free.

Trial Registry

Registry Name: ClinicalTrials.gov; Registration Number: NCT00693849; URL: http://www.clinicaltrials.gov/ct2/show/NCT00693849?term=ispot-D&rank=1.

FUNDING AND DISCLOSURE

Drs Patenaude, Song and Usherwood declare no conflict of interest. William Rekshan is employed as a biostatistician at Brain Resource, in which he has stock options. A John Rush has received consulting fees from Brain Resource, Duke-National University of Singapore, Eli Lilly, Takeda, Medavante, the University of Colorado, and the National Institute of Drug Abuse; speaker fees from University of California at San Diego, the Hershey Medical Center, the American Society of Clinical Pharmacology, the New York State Psychiatric Institute, and Otsuka Pharmaceutical; royalties from Guilford Publications and the University of Texas Southwestern Medical Center; a travel grant from CINP; and research support from Duke-NUS. Alan Schatzberg has served as a consultant to BrainCells, CeNeRx, CNS Response, Depomed, Eli Lilly, Forest Labs, GSK, Jazz, Lundbeck, Merck, Neuronetics, Novadel, Novartis, Pathway Diagnostics, Pfizer, PharmaNeuroBoost, Quintiles, Sanofi-Aventis, Sunovion, Synosia, Takeda, Xytis, and Wyeth. Dr Schatzberg has equity in Amnestix, BrainCells, CeNeRx, Corcept (co-founder), Delpor, Forest, Merck, Neurocrine, Novadel, Pfizer, PharaNeuroBoost, Somaxon, and Synosis, and was named an inventor on pharmacogenetic use patents on prediction of antidepressant response. Dr Schatzberg has also received speaking fees from GlaxoSmithKline and Roche. Drs Etkin and Williams have received research funding from Brain Resource for iSPOT-D. Leanne Williams has served as a consultant to Brain Resource and was a stockholder in Brain Resource.

ACKNOWLEDGEMENTS

The iSPOT-D is sponsored by Brain Resource Company Operations. We gratefully acknowledge the contributions of principal and co-investigators at each site, the editorial support of Jon Kilner (Pittsburgh, PA, USA), and the Scoring Server management by Donna Palmer (Brain Resource). We also acknowledge the statistical advice of Eugene Laska (New York University).

REFERENCES

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. 4th edn American Psychiatric Association: Washington, DC.

Dawes SE, Jeste DV, Palmer BW (2011). Cognitive profiles in persons with chronic schizophrenia. *J Clin Exp Neuropsychol* **33**: 929–936.

Dunkin JJ, Leuchter AF, Cook IA, Kasl-Godley JE, Abrams M, Rosenberg-Thompson S (2000). Executive dysfunction predicts nonresponse to fluoxetine in major depression. *J Affect Disord* **60**: 13–23.

Dunn OJ (1961). Multiple comparisons among means. *J Am Stat Assoc* **56**: 52–64.

Gorlyn M, Keilp JG, Grunebaum MF, Taylor BP, Oquendo MA, Bruder GE *et al* (2008). Neuropsychological characteristics as predictors of SSRI treatment response in depressed subjects. *J Neural Transm* **115**: 1213–1219.

Gotlib IH, Joormann J (2010). Cognition and depression: current status and future directions. *Annu Rev Clin Psychol* **6**: 285–312.

Hamilton M (1960). A rating scale for depression. *J Neurol Neurosurg Psychiatry* **23**: 56–62.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. 2nd edn Springer-Verlag: New York.

Horan WP, Goldstein G (2003). A retrospective study of premorbid ability and aging differences in cognitive clusters of schizophrenia. *Psychiatry Res* **118**: 209–221.

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K *et al* (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* **167**: 748–751.

Kline RB (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*. American Psychological Association.

Leuchter AF, Morgan M, Cook IA, Dunkin J, Abrams M, Witte E (2004). Pretreatment neurophysiological and clinical characteristics of placebo responders in treatment trials for major depression. *Psychopharmacology (Berl)* **177**: 15–22.

Mahalanobis PC (1936). On the generalized distance in statistics. *Proc Natl Inst Sci India* **2**: 49–55.

Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A (2011). Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes* **4**: 299.

Mathersul D, Palmer DM, Gur RC, Gur RE, Cooper N, Gordon E *et al* (2009). Explicit identification and implicit recognition of facial emotions: II. Core domains and relationships with general cognition. *J Clin Exp Neuropsychol* **31**: 278–291.

Paul RH, Lawrence J, Williams LM, Richard CC, Cooper N, Gordon E (2005). Preliminary validity of "integneuro": a new computerized battery of neurocognitive tests. *Int J Neurosci* **115**: 1549–1567.

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN *et al* (2003). The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry* **54**: 573–583.

Rush AJ, Trivedi MH, Wisniewski SR, Stewart JW, Nierenberg AA, Thase ME *et al* Team SDS (2006). Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. *N Engl J Med* **354**: 1231–1242.

Saveanu R, Etkin A, Duchemin A-M, Gyurak A, DeBattista C, Schatzberg AF *et al*. The International Study to Predict Optimized Treatment for Depression (iSPOT-D): Outcomes from the acute phase of antidepressant treatment. *J Psyc Res* in press.

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E *et al* (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* **59**(Suppl 20): 22–33 quiz 34-57.

Snyder HR (2013). Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. *Psychol Bull* **139**: 81–132.

Tatsumi M, Groshan K, Blakely RD, Richelson E (1997). Pharmacological profile of antidepressants and related compounds at human monoamine transporters. *Eur J Pharmacol* **340**: 249–258.

Taylor BP, Bruder GE, Stewart JW, McGrath PJ, Halperin J, Ehrlichman H *et al* (2006). Psychomotor slowing as a predictor of fluoxetine nonresponse in depressed outpatients. *Am J Psychiatry* **163**: 73–78.

Trivedi MH, Fava M, Wisniewski SR, Thase ME, Quitkin F, Warden D *et al* Team SDS (2006a). Medication augmentation after the failure of SSRIs for depression. *N Engl J Med* **354**: 1243–1252.

Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L *et al* Team SDS (2006b). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry* **163**: 28–40.

World Health Organization, *Mental health: mental health atlas 2011*. World Health Organization; Switzerland, 2012 (cited June 28, 2011) http://www.who.int/mental_health/publications/mental_health_atlas_2011/en/index.html.

Williams LM, Mathersul D, Palmer DM, Gur RC, Gur RE, Gordon E (2009). Explicit identification and implicit recognition of facial emotions: I. Age effects in males and females across 10 decades. *J Clin Exp Neuropsychol* **31**: 257–277.

Williams LM, Rush AJ, Koslow SH, Wisniewski SR, Cooper NJ, Nemeroff CB *et al* (2011). International Study to Predict Optimized Treatment for Depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials* **12**: 4.

Supplementary Information accompanies the paper on the Neuropsychopharmacology website (http://www.nature.com/npp)