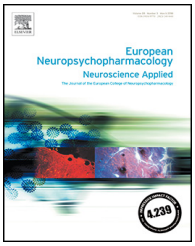
European Neuropsychopharmacology (2019) 29, 66–75

www.elsevier.com/locate/euroneuro



Predicting treatment response to antidepressant medication using early changes in emotional processing

Michael Browning[a,b,c,*], **Jonathan Kingslake**[c], **Colin T. Dourish**[c], **Guy M. Goodwin**[a,b], **Catherine J Harmer**[a,b], **Gerard R. Dawson**[c]

[a] *Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, United Kingdom*
[b] *Oxford Health NHS Trust, Warneford Hospital, Oxford, United Kingdom*
[c] *P1vital Ltd, Manor House, Howbery Park, Wallingford, Oxfordshire, United Kingdom*

Received 7 April 2018; received in revised form 2 October 2018; accepted 9 November 2018

KEYWORDS
Emotional bias;
Machine learning;
Antidepressant;
Treatment;
Prediction;
Depression

Abstract
Antidepressants must be taken for weeks before response can be assessed with many patients not responding to the first medication prescribed. This often results in long delays before effective treatment is started. Antidepressants induce changes in the processing of emotional stimuli early in the course of treatment. In the current study we assessed whether changes in emotional processing and subjective symptoms over the first week of antidepressant treatment predicted clinical response after 4–8 weeks of treatment. Such a predictive test may shorten the time taken to initiate effective treatment in depressed patients. Seventy-four depressed primary care patients completed measures of emotional bias and subjective symptoms before starting antidepressant treatment and then again 1 week later. Response to treatment was assessed after 4–6 weeks. The performance of classifiers based on these measures was assessed using a leave-one-out validation procedure with the best classifier then tested in an independent sample from a second study of 239 patients. The combination of a facial emotion recognition task and subjective symptoms predicted response with 77% accuracy in the training sample and 60% accuracy in the independent study, significantly better than possible using baseline response rates. The face based measure of emotional bias provided good quality data

* Corresponding author at: Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, United Kingdom.
E-mail address: Michael.browning@psych.ox.ac.uk (M. Browning).

https://doi.org/10.1016/j.euroneuro.2018.11.1102
0924-977X/© 2018 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

with high acceptability ratings. Changes in emotional processing can provide a sensitive early measure of antidepressant efficacy for individual patients. Early treatment induced changes in emotional processing may be used to guide antidepressant therapy and reduce the time taken for depressed patients to return to good mental health.
© 2018 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

A number of effective treatments for major depressive disorder have been developed, including a range of antidepressant medications (Cipriani et al., 2018; NICE, 2009). However, many patients do not respond to the first antidepressant prescribed (Rush et al., 2006) and clinical guidelines suggest waiting at least four to six weeks (NICE, 2009) before subjective symptom response to treatment can be confidently assessed. In practice the delays may be longer than that and assessment may not be systematic. Unsurprisingly, this can lead to patients trying a series of different antidepressant drugs, one after the other, which often results in long delays before patients' symptoms resolve. These delays are likely to increase the risk of incomplete resolution of symptoms and reduce the chances of full functional recovery. One way in which delays may be reduced is by using predictive tests which are able to detect, early in the course of treatment, whether individual patients will go on to respond to the treatment or not (Cattaneo et al., 2016; Chekroud et al., 2016; de Vries et al., 2018; Dinteren et al., 2015; Etkin et al., 2015; Leuchter et al., 2009). There has been particular recent interest in machine learning approaches to develop classifiers which combine a range of different patient level data (in the prediction literature, called *features*) in order to provide a patient level prediction of this kind. The most common classifiers are based on patient characteristics or biomarkers, which are assessed *before* treatment is started (Cattaneo et al., 2016; Chekroud et al., 2016; Dinteren et al., 2015; Etkin et al., 2015). It is argued that this information may be used in clinical settings to select a specific treatment if it is predicted to have a higher chance of success (Chekroud et al., 2016) or to suggest that a patient may be generally treatment resistant, which would justify earlier use of second line therapies (Cattaneoet al., 2016).

An alternative approach to baseline classification uses change in biomarkers or symptoms measured during initial antidepressant therapy, for example across the first week of treatment, to predict subsequent response to a specific drug (de Vries et al., 2018; Leuchter et al., 2009). This approach may provide a tailored prediction of whether a particular patient will respond to a specific antidepressant, rather than to a broad class of drugs and guide individual treatment by, for example, changing antidepressant treatment after one week if the predictive test suggests a likely non-response, rather than waiting the standard four to six weeks.

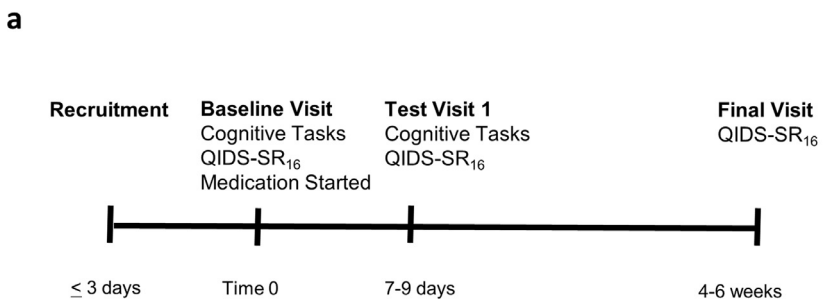
A promising biomarker of antidepressant response is change in the automatic processing of emotional information early on in antidepressant therapy (Harmer et al., 2011). For example, brief treatment with antidepressant medication increases the tendency for patients to categorise ambiguous facial expressions as positive (e.g. happy) relative to negative (e.g. fearful) (Harmer et al., 2009). Moreover, using prospective designs, early change in measures of emotional bias in depressed patients treated with antidepressants has been positively correlated with the improvement in patients' symptoms of depression across a full 6–8 weeks of treatment (Shiroma et al., 2014; Tranter et al., 2009) suggesting that such early changes may be predictive of treatment outcome. However, a number of outstanding questions remain. Most obviously, no previous study has assessed whether a classifier incorporating change in emotional bias over the first week of antidepressant treatment is able to predict whether an individual patient will go on to respond to that treatment. A related question is whether collecting data over this initial week offers any advantage over classifiers based on pre-treatment baseline data. Finally, the ultimate aim of developing predictive classifiers is not simply to demonstrate their accuracy—rather it is to develop a test that, when used in clinical practice, results in a better patient outcome. Thus, administration of the test must be practicable and acceptable in the clinical context in which it is to be used. We are aware of no previous work which has assessed this question.

In the present study, we have assessed the feasibility of deploying a computer based assessment of emotional bias and depressive symptoms in primary care patients and shown that a classifier based on measures before and one week after initiation of the serotonin reuptake inhibitor (SRI) citalopram was able to predict response after 4–6 weeks. We also compared the predictive performance of classifiers based on baseline data (i.e. collected before treatment administration) with those based on data collected over the first week of treatment to assess if collecting data over this week conferred an advantage in predicting response. Finally, we assessed the performance of the winning classifier in a fully held out sample from a second study.

2. Experimental procedures

2.1. Overview

A single group of depressed primary care patients, whose treating clinician had made the decision to prescribe citalopram, were enrolled and completed 3 visits in the classifier development study (Fig. 1(a)). At visit 1 patients completed a questionnaire measuring depressive symptoms, the Quick Inventory of Depressive Symptoms, 16 item self-report version (QIDS-SR_{16}) (Rush et al., 2003) and the emotional bias



b

Female : Male N(%)	**34:23 (60%:40%)**
Age, yrs Mean (SD)	41.1 (13.1)
Years education Mean (SD)	13.6 (2.7)
Previous depressive episodes Mean (SD)	1.6 (1)
Responders: Non-Responders N(%)	22:35 (39%:61%)

Fig. 1 Timeline of the classifier development study (panel a). Demographic and treatment response details of participants (panel b).
QIDS-SR_{16} Quick Inventory of Depressive Symptoms, 16 item self-report version.

tasks (see below for description). Following this patients started citalopram treatment. The second visit occurred 1 week later. During this visit patients repeated the QIDS-SR_{16} and the bias tasks. The final visit occurred 4–6 weeks (min 28 days, max 48 days, and mean 35.9 days) after the baseline visit during which patients completed the QIDS-SR_{16} for a final time as well as completing acceptability questionnaires designed to capture challenges to the administration of the tasks in primary care settings. Change in bias and questionnaire scores between baseline and week 1, or just the baseline scores, were used to train support vector machine (SVM) classifiers (Cortes and Vapnik, 1995) to predict response status at week 6. Response was defined as a 50% or greater reduction from baseline QIDS-SR_{16} score (Rush et al., 2006) and classifier performance was estimated using a leave-one-out (LOO) validation procedure. The performance of the winning classifier was then tested in a fully held out sample from a second study. The details below describe the main classifier development study, details of the study used for validation are described in the section "validation sample study". Both studies received ethical approval from the National Research Ethics Service (reference numbers 14/NW/0250, 16/NE/0095) and all patients provided written informed consent on enrolment.

2.2. Population

Patients aged between 18 and 65 years who attended one of twelve general practices (GP) in the UK and who, in the opinion of their general practitioner required treatment with citalopram, were recruited to the study. Older patients were excluded as they are often initiated on a reduced dose of citalopram (10 mg rather than 20 mg) which is increased after 3–4 days. Potential participants were excluded if they were currently taking antidepressant, antipsychotic or regular hypnotic medication, or if they required immediate referral to secondary care mental health services.

2.3. Treatment during the study

The decision to initiate treatment with citalopram and the dosing regimen to use was made by the treating clinician before trial entry.

2.4. Measures of emotional bias

Patients completed standard emotional word-based memory encoding (ECAT) and recall (EREC) tasks and a face-based emotional recognition task (FERT) (Browning et al., 2007). The tasks, which are described in detail in the supplementary methods, were administered at baseline and week 1 using dedicated personal computers located within the primary care surgeries. Patients were supported in completion of the tasks by research nurses located within the GP practices. Task instructions were presented on screen for patients to read before task completion. The psychometric properties of the cognitive tasks used have previously been assessed by Thomas et al. (2016), with intraclass correlation coefficients ranging from 0.4 to 0.8.

2.5. Questionnaire measures

The QIDS-SR_{16} is a well validated 16 item, self-report scale of depressive symptoms which is increasingly being used as the primary outcome in clinical trials of treatments for depression (Rush et al., 2006, 2003). Treatment response using this scale is defined as a 50% or greater reduction of the patient's baseline score (Rush et al., 2006). The QIDS-SR_{16} questionnaire was completed following 1 week of treatment as well as at baseline and week 6. This allowed us to compare the ability of the emotional bias tasks to predict treatment response with the ability of changes in QIDS-SR_{16} over the same time period (NB treating each of the 16 items as separate features). It also allowed us to assess predictive algorithms which combined the questionnaire items with task features.

An acceptability questionnaire was developed specifically for this study and was composed of three questions (see Supplementary Table 5) which participants could answer by circling one of three responses (yes/partly/no). Participants were also encouraged to provide free text assessments of the positive and negative aspects of the tasks.

2.6. Development of the predictive algorithm

Assessment of the specific tasks to include in the predictive algorithm was guided by the data quality and predictive

performance of the algorithms based on the separate tasks, in the classifier development study, in a primary care setting. The primary measure of data quality was the proportion of patients completing the task who met minimal performance parameters, which are detailed in the supplementary methods.

A linear support vector machine (SVM) was used to combine task and questionnaire features into binary predictions (i.e. responder/non-responder). SVMs are a widely used and robust method of deriving binary classifications, particularly when the ratio of data points to features is relatively low, as is the case in this study. Analysis was performed using Matlab (version R2015b, Mathworks). Performance of the algorithm within the classifier development study was assessed using a leave-one-out validation procedure during which a training set consisting of all but one participant was used. The training set was used for feature selection, estimation of the C-parameter and model training, with the left out sample being used solely for validation (Hastie et al., 2009). Note that this approach results in variability in the features selected, the C-parameter used and the model weights for each iteration of the leave-one-out procedure. The value of the C-parameter used was selected based on the achieved accuracy within the training set using 50 values of the parameter ranging from 0.01 to 100. Feature selection was achieved by selecting the features with the highest area under the curve for predicting response in the training set. Missing values of a given feature in either the training or testing set (e.g. reaction times for choices, which were not made by a particular participant could not be calculated) were entered as the mean value for that feature, calculated from the training set. The unbalanced nature of the data set (i.e. unequal numbers of responders and non-responders) was dealt with by setting the weight of each observation to 1/(number of observations of a given class) in the training set (Huang and Du, 2005).

Separate analyses were completed to test the predictive ability of the emotional bias tasks and QIDS-SR_{16} as well as using different proportions of task features (10%, 50% or 100% of available features). The rationale for assessing this range of proportions of task features is that, if most information about treatment response is contained in only a few task features then the classifier which uses just these features will perform better, whereas if information about treatment response is distributed throughout many task features then the more inclusive classifiers will perform better. In addition, separate classifiers were created using features calculated as the change from baseline, that is the difference in the feature between visit 1 when the participants had taken 1 week of citalopram and baseline, or using just the baseline values. This allowed us to determine whether classifiers based solely on baseline data performed differently to those based on the early effects of treatment. The degree to which the accuracy scores of the best performing analyses were robust to overfitting to outlying data points in the classifier development study was further tested using sensitivity analyses (reported in the supplementary results) in which 10% of patients were randomly removed from the sample and the analyses then rerun 1000 times. We used classifier accuracy as the primary metric of response but also report positive and negative predictive value and sensitivity and specificity of the derived classifiers.

Analyses were performed on all patients who had completed the relevant study measures (i.e. completed the QIDS-SR_{16} at baseline, visit 1 and on the final visit and completed the cognitive tasks at baseline and visit 1) and who continued to take citalopram for the duration of the study.

Statistical assessment of the achieved prediction accuracies was performed using one tailed *z*-tests, which assess whether the achieved prediction is significantly better than a comparator performance. Two comparator conditions in the classifier development process were used: first predictions were compared against an accuracy of 50%, which represents no prior knowledge of the response rate in the current cohort, and would be a reasonable estimate of the response rate in open label trials (Rush et al., 2006). Secondly, the prediction was compared against the accuracy which could be achieved from knowledge of the baseline response rate in the current cohort (i.e. in the classifier development study the response rate was 39%, therefore, theoretically, an accuracy of 61% could be achieved by labelling all participants as non-responders).

2.7. Validation sample study

Data from 239 participants taken from a separate study (The PReDicT Study; Kingslake et al., 2017) were used to assess the performance of the winning classifier identified from the above process in a fully held out sample. Importantly, these data were kept completely separate from the classification development process described above and only one classifier was tested on it. The ongoing PReDicT study is recruiting a sample of patients with depression, who are being started on antidepressant medication, and who complete the FERT task and QIDS-SR_{16} at baseline and after 1 week as in the classifier development study. A complete description of the PReDicT study protocol is available (Kingslake et al., 2017), the pertinent differences between the validation and classifier development studies are that in the validation sample study: (a) patients were started on any SRI (other than Fluoxetine) rather than only citalopram, (b) treatment response was assessed following 8–10 weeks of treatment rather than 4–6, (c) participants completed only the FERT task online rather than 3 bias measure tasks on a dedicated PC in the GP surgery, (d) the study was performed across 5 European countries (UK, the Netherlands, France, Germany and Spain) rather than only in the UK, (e) the overall response rate in the validation sample was 50%, as compared to 39%. Note that the PReDicT study randomises participants into a group receiving treatment as usual and a group in which antidepressant treatment is guided by the classifier. Here we test classifier performance solely in the treatment as usual group. Performance of the classifier in this sample was assessed using a one tailed *z*-test with a comparison performance of 50% (i.e. which is also the baseline response rate in this sample).

3. Results

3.1. Population

Demographic information on participants from the classifier development study is reported in Fig. 1(b). A total of 74

Table 1 Accuracies, positive predictive values (percentage of patients predicted to respond who actually respond) and negative predictive values (percentage of patients predicted not to respond who do not respond) in held out patients using changes in the different cognitive tasks and the QIDS-SR_{16} questionnaire over the first week of treatment. Three different levels of feature selection were employed. Sensitivity and specificity are reported in supplementary Table 1.

	Percentage of total features selected		*10%*	*50%*	*100%*
Tasks used in algorithm	Total number of features	*Accuracy (PPV, NPV)*	*Accuracy (PPV, NPV)*	*Accuracy (PPV, NPV)*
QIDS-SR$_{16}$	16	46% (37% 59%)	61% (50%, 65%)	56% (36%, 61%)
FERT	36	51% (41%, 64%)	70% (78%, 69%)	49% (27%, 57%)
ECAT	6	56% (46%, 81%)	51% (41%, 65%)	58% (46%, 68%)
EREC	6	63% (53%, 68%)	54% (38%, 61%)	54% (38%, 61%)
FERT + QIDS-SR$_{16}$	52	54% (43%, 66%)	77% (70%, 82%)	68% (59%, 74%)
FERT + ECAT + EREC + QIDS-SR$_{16}$	64	56% (44%, 66%)	79% (73%, 83%)	51% (25%, 58%)

QIDS-SR_{16}; Quick Inventory of Depressive Symptoms, Self-Rate 16 item version. FERT; Facial Emotion Recognition Task. ECAT; Emotional Categorisation Task. EREC; Emotional Recall Task. PPV; Positive Predictive Value. NPV; Negative Predictive Value.

patients were enrolled in the study. Of these 57 patients completed all necessary study assessments. 6 patients did not attend the study visits, 8 discontinued citalopram treatment during the study, 1 was withdrawn due to the discovery of an exclusion criterion shortly after enrolment and 2 did not fully complete the QIDS-SR_{16} questionnaire at the final visit. The overall response rate to citalopram was relatively low for an open label study (39%) meaning that the majority of patients (61%) did not respond to the citalopram they were prescribed.

3.2. Task data quality

Of the 57 patients who completed the study, all 57 had useable data for the face-based FERT task (minimum mean accuracy 36%), 54 had useable data for the ECAT memory encoding task (3 patients scored less than chance) and 53 had useable data for the EREC memory recall task (3 lost due to poor ECAT, EREC results from 1 patient lost). In summary successful data collection was more consistently achieved from the FERT task than the ECAT or EREC tasks.

3.3. Prediction accuracy

The accuracies of the classifiers based on the change in cognitive and questionnaire measures over the first week of treatment when predicting response vs. non-response status in held out patients is summarised in Table 1 (see Supplementary Table 1 for test sensitivity and specificity). As can be seen using self-report symptoms only, as measured using early change in the QIDS-SR_{16}, accuracies of around 60% were achieved. Using only features from the FERT task achieved accuracies of 70%, with the ECAT and EREC managing only 55–60% accuracies. A clear improvement in accuracy was seen when moving from 10% to 50% of the most informative features of the FERT and QIDS-SR_{16}, suggesting that information is spread throughout these task features rather than being concentrated in a small number. However accuracies were generally lower when all features (100%) were included in the algorithm indicating that removal of the more noisy features is beneficial to prediction.

Next, we assessed whether combining the most informative features across tasks and questionnaires improved overall prediction. As can be seen combining the features from the two best performing classifiers (QIDS-SR_{16} and FERT) led to a higher overall accuracy (Fig. 2; 77%) than either separately. Adding the EREC and ECAT did not improve the prediction of the classifier (79%). As can be seen from Fig. 2, the algorithm based on the QIDS-SR_{16} and FERT achieved a specificity of approximately 80% and a sensitivity of 73%, with a reasonably consistent selection of features from across both the FERT and QIDS-SR_{16} measures over the iterations of the LOO procedure. Performance of this algorithm was statistically better than both an uniformed, 50% comparator ($z = 2.9$, $p = 0.002$) and an informed comparator, based on the baseline non-response rate of 61% ($z = 1.83$, $p = 0.03$). An additional sensitivity analysis of these results, suggested that the classifiers based on 10% of features were more susceptible to misestimating the classifier performance and is described in the supplementary results file.

While the accuracy of a classifier gives an estimate of its overall performance, the clinical application of the specific classifiers described in this paper – patients who are classified as non-responders would have their treatment

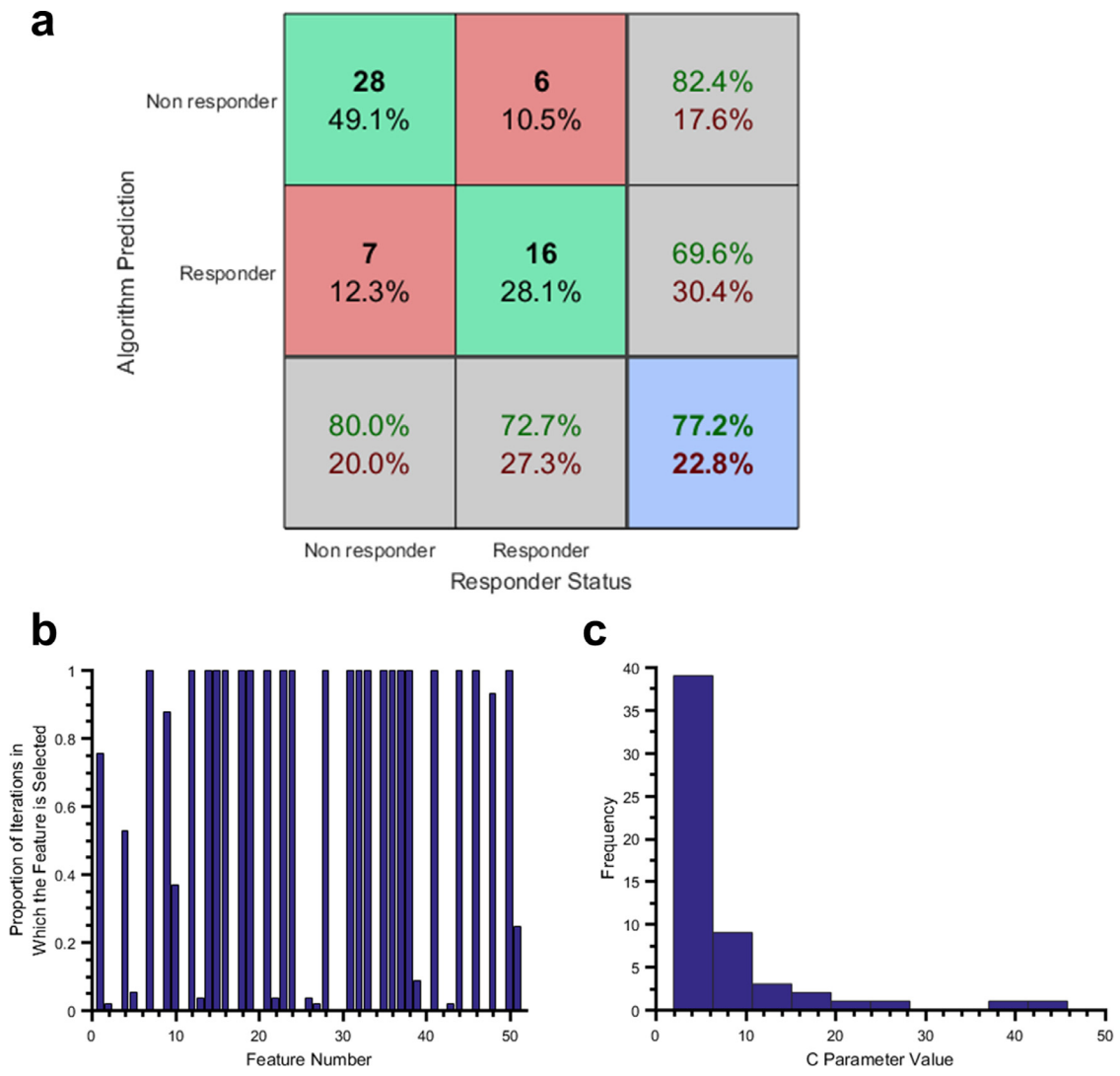


Fig. 2 Performance of the classifier combining the QIDS-SR_{16} and FERT and Selecting the 50% most informative features. (a) A confusion matrix illustrating the number of participants correctly classified by the algorithm (green squares) and those misclassified (red squares). Grey squares along the bottom report algorithm specificity (80%) and sensitivity (72%), grey squares on the right report the negative predictive value (82%) and positive predictive value (70%). (b) Feature selection consistency across iterations of the leave-one-out (LOO) validation process. Feature numbers 1–36 are derived from the FERT task, numbers 37–52 are the items of the QIDS-SR_{16}. The *y* axis reports the proportion of LOO iterations in which the feature was selected. As can be seen most features which are selected, are selected on every iteration, although there is some variability. (c) Optimal value of the C-parameter used for each LOO iteration. Histogram illustrating the frequency with which optimal *C* values were estimated across iterations suggests a relatively consistent selection of low values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

altered after one week (Kingslake et al., 2017) – suggests other relevant, clinically intuitive, performance metrics. In particular, the proportion of all patients who would have their treatment changed appropriately (i.e. they would not have responded to the original medication) by the classifier is described by the negative predictive value (NPV) and estimates how likely a classifier guided active change in treatment is to be appropriate. As can be seen from Table 1 the NPV based on the QIDS-SR_{16} alone is approximately 65% meaning that 35% of patients would have their treatment changed even though they were going to respond to the original treatment. Addition of the FERT task to the classifier increases the NPV to 82%, reducing the rate of inappropriate treatment changes to 18% of patients.

3.4. Classification using baseline data

The classification accuracies achieved using just the baseline data are summarised in Table 2. The scores achieved using baseline data are generally lower than those achieved using change scores, with only one classifier performing better than chance (based on QIDS-SR_{16} only and 10% of features) and no classifier predicting response significantly better than the baseline non-response rate.

Table 2 Accuracies, positive predictive values (percentage of patients predicted to respond who actually respond) and negative predictive values (percentage of patients predicted not to respond who do not respond) in held out patients using baseline scores from the different cognitive tasks and the QIDS-SR_{16} questionnaire. Three different levels of feature selection were employed. Sensitivity and specificity are reported in supplementary Table 2.

	Percentage of total features selected		*10%*	*50%*	*100%*
Tasks used in algorithm		Total number of features	*Accuracy (PPV, NPV)*	*Accuracy (PPV, NPV)*	*Accuracy (PPV, NPV)*
QIDS-SR$_{16}$		16	70% (60%, 78%)	61% (50%, 74%)	58% (46%, 68%)
FERT		36	53% (40%, 63%)	53% (41%, 63%)	39% (28%, 50%)
ECAT		6	60% (49%, 77%)	51% (42%, 67%)	49% (40%, 64%)
EREC		6	51% (36%, 60%)	53% (39%, 62%)	42% (31%, 54%)
FERT + QIDS-SR$_{16}$		52	60% (47%, 66%)	60% (48%, 70%)	58% (46%, 66%)
FERT + ECAT + EREC + QIDS-SR$_{16}$		64	54% (42%, 64%)	65% (54%, 74%)	65% (54%, 73%)

QIDS-SR_{16}; Quick Inventory of Depressive Symptoms, Self-Rate 16 item version. FERT; Facial Emotion Recognition Task. ECAT; Emotional Categorisation Task. EREC; Emotional Recall Task. PPV; Positive Predictive Value. NPV; Negative Predictive Value.

3.5. Acceptability questionnaires

Quantitative responses from the acceptability questionnaire are summarised in Supplementary Table 5 (note responses to the acceptability questionnaires are included for participants with missing task data who were not included in the classifier performance analysis). The large majority of patients felt that they would be able to complete the task without researcher assistance. Approximately 10% of respondents reported that they would not be able to repeat the task at home (due to lack of home internet). The main difficulties with the tasks reported in the free text responses were the overall duration of the test (i.e. time to complete all tasks) which was thought to be too long and the perceived difficulty of the tasks (i.e. the short duration the faces were presented and the level of difficulty of the memory task).

3.6. Performance in a held out sample

The above analysis indicates that the classifier based on change scores from the $QIDS_{16}$ and FERT task was the most robust in terms of both prediction accuracy and data quality. The performance of this classifier was therefore tested in the sample of 251 patients recruited to the treatment as usual arm of the PReDicT study (Kingslake et al., 2017). The classifier accurately predicted response to treatment in 60.3% of patients which was significantly better than classification based on the baseline response of 50% ($z = 2.13$, $p = 0.02$). More details about the sample characteristics and classifier performance is provided in the supplementary materials.

4. Discussion

Induced changes in emotional bias and subjective symptoms, measured after one week of treatment, may be used to predict antidepressant treatment response at the level of the individual patient. These measures can feasibly be collected in primary care settings. This suggests that it may be possible to use cognitive and symptomatic measures to guide antidepressant treatment in depressed patients.

4.1. Data quality, predictive performance and acceptability of the emotional bias tasks

The current study tested three different behavioural measures of emotional bias, all of which have previously been reported to be influenced by treatment with antidepressant medication in experimental settings (Harmer and Cowen, 2013). Of these three measures, the facial expression recognition task (FERT) provided a higher level of data quality, with all patients completing it adequately, unlike the word based encoding and recall tasks (ECAT and EREC tasks) in which approximately 5% of data were lost due to difficulties with task comprehension despite the support of research nurses during task completion. The FERT task also provided a superior prediction of participant response, with changes in task scores over the first week of treatment correctly predicting the response status of 70% of patients using a leave-one-out validation procedure, as opposed to prediction accuracies of between 55% and 60% for the memory tasks. Together, these results indicate that the FERT task provides the most reliable and useful measure of emotional bias tested in this study.

The predictive performance of the algorithm based on change in task scores over the first week of treatment was improved when change in subjective symptoms, measured with the QIDS-SR_{16} questionnaire, was added to the emotional bias task based features. While the FERT task alone predicted response rates to a greater extent than symptom scores (70 % vs. 61%) combining both sets of features led to an improved performance (77%) suggesting that the different measures carry different information about response likelihood and that their combination is likely to extend the range of patients whose response may be predicted.

Assessment of the acceptability of the tasks used in the current study indicate that computer based e-assessment is generally acceptable to patients in primary care and that they may be administered with relatively little support. Patients in the current study did feedback that the duration of testing was too long. Therefore acceptability could be improved further by reducing the duration of the assessment process, this may best be achieved by focussing only on the FERT task and QIDS-SR_{16} questionnaires which would reduce the total testing time from 30–40 min to approximately 15–20 min. The FERT task has the added advantage that it is relatively culturally unbiased as the specific emotional expressions used in the task appear to be recognised across cultures (Ekman et al., 1987) and the task does not depend on linguistic ability.

4.2. Comparison of classifiers based on baseline data versus induced change

A number of previous studies have used baseline measures to predict response to antidepressant treatment (Cattaneo et al., 2016; Chekroud et al., 2016; Dinteren et al., 2015; Etkin et al., 2015). The results have been somewhat mixed although any test which predicts the best treatment *before* it is prescribed is potentially superior to our post hoc approach. In the current study, classifiers based on baseline measures of the cognitive tasks and QIDS-SR_{16} did not perform as well as those which utilised change scores suggesting that, using these emotional bias measures at least, there is a benefit to basing classification on data collected over the first week of treatment.

The performance of the classifier developed in this paper in a fully held out sample, was approximately 60%, similar to that reported by other classifiers (Chekroud et al., 2016; Etkin et al., 2015). While this result provides evidence that the classifier is able to predict response better than chance, in a relatively naturalistic sample of patients receiving a range of different medications, it does not answer the key clinical question—does use of the algorithm in clinical practice meaningfully improve patient outcome? Ultimately, this question of clinical efficacy will not be settled using the study designs employed in this and previously published studies which measure classification accuracy. Rather clinical efficacy needs to be tested using a randomised controlled design in which patients are randomised to have either treatment as usual, or treatment guided by a specific algorithm. A superior outcome in patients who use the algorithm will provide strong evidence for its adoption into routine care. We are currently completing such a study, in which antidepressant treatment is guided by the algorithm described in this paper (Kingslake et al., 2017). In this ongoing trial patients will complete the FERT and QIDS-SR_{16} after one week of treatment and, should they be predicted to be not responding, will have their medication changed at this stage, with the prediction process repeated again after a further week.

4.3. Study limitations

The limitations of the current study should be acknowledged. Firstly, the sample size of 57 patients in the classifier development study is relatively small. While this does not influence the estimated accuracy of the classifier in the held out sample it does raise the possibility that a more accurate classifier could be developed if trained on a larger data set. Against this limitation, the current study has collected targeted data from a reasonably representative clinical sample, whereas many previous studies have used samples of convenience from previous clinical trials.

A second limitation is that patients were only recruited to the classifier development study if they had been prescribed citalopram. This medication was chosen as it is commonly prescribed in the UK and we wanted to reduce between patient variance in this initial test of the algorithm. While this raises the potential risk that the classifier may only predict response to citalopram, the demonstration of reasonable classification accuracy in the held out sample (with superior performance in those who received sertraline) suggests that classifier performance is relatively constant, at least across SRI medications. However, it remains possible that overall classifier performance could have been improved if the training dataset had included participants on a broader range of medications.

Lastly, the primary outcome measure in the current study, which we sought to predict, was treatment response rather than remission. Response is often used as the outcome of interest in clinical trials, however remission is clearly a more desirable target for the individual patient. Predicting remission presents additional methodological challenges as it is less common and should ideally be assessed after a longer follow up. This may be addressed in future studies which incorporate a longer course of treatment.

In conclusion, induced changes in emotional bias may be utilised together with symptomatic change over the first week of citalopram treatment to predict treatment response after 4–6 weeks. Our classifier is based on a clear mechanistic hypothesis about the mode of action of antidepressants. An algorithm based on these measures could be used to direct antidepressant prescription with the goal of reducing the time taken to initiate patients on effective treatment and the time to remission.

Funding

This study was funed by a grant from the NHS SBRI Healthcare programme (SBRI-COLAB-1719) to P1vital Products Ltd. The funder had no role in te design, conduct or reporting of the study

Conflict of interest

MB, CD and GD are employees of P1vital Ltd. CD, GD, JK, CH and GG own shares in P1vital Ltd. JK is an employee of P1vital Products Ltd. CD, MB, GD, GG and JK own shares in P1vital Products Ltd, which sponsored the study and which has developed the tests of emotional bias described in the study. MB has received fees for travel expenses from Lundbeck for attending conferences and consultancy fees from Johnson and Johnson. GG is a NIHR Senior Investigator, holds a grant from the Wellcome Trust and has served as a consultant, advisor or CME speaker for Allergan, Angelini, Compass pathways, MSD, Lundbeck, Minerva, Otsuka, Takeda, Medscape, P1vital, Pfizer, Servier, Shire and Sun Pharma. CJH has received consultancy fees from Lundbeck, P1vital, Servier and Johnson and Johnson. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

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.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroneuro.2018.11.1102.

References

Browning, M., Reid, C., Cowen, P.J., Goodwin, G.M., Harmer, C.J., 2007. A single dose of citalopram increases fear recognition in healthy subjects. J. Psychopharmacol. Oxf. Engl. 21, 684–690. doi:10.1177/0269881106074062.

Cattaneo, A., Ferrari, C., Uher, R., Bocchio-Chiavetto, L., Riva, M.A., Pariante, C.M.MRC ImmunoPsychiatry Consortium, 2016. Absolute measurements of macrophage migration inhibitory factor and interleukin-1-β mRNA levels accurately predict treatment response in depressed patients. Int. J. Neuropsychopharmacol. Off. Sci. J. Coll. Int. Neuropsychopharmacol (CINP) doi:10.1093/ijnp/pyw045.

Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R., 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 3, 243–250. doi:10.1016/S2215-0366(15)00471-X.

Cipriani, A., Furukawa, T.A., Salanti, G., Chaimani, A., Atkinson, L.Z., Ogawa, Y., Leucht, S., Ruhe, H.G., Turner, E.H., Higgins, J.P.T., Egger, M., Takeshima, N., Hayasaka, Y., Imai, H., Shinohara, K., Tajika, A., Ioannidis, J.P.A., Geddes, J.R., 2018. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet Lond. Engl. doi:10.1016/S0140-6736(17)32802-7.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. doi:10.1007/BF00994018.

de Vries, Y.A., Roest, A.M., Bos, E.H., Burgerhof, J.G.M., van Loo, H.M., de Jonge, P., 2018. Predicting antidepressant response by monitoring early improvement of individual symptoms of depression: individual patient data meta-analysis. Br. J. Psychiatry J. Ment. Sci. 1–7. doi:10.1192/bjp.2018.122.

Dinteren, R.van, Arns, M., Kenemans, L., Jongsma, M.L.A., Kessels, R.P.C., Fitzgerald, P., Fallahpour, K., Debattista, C., Gordon, E., Williams, L.M., 2015. Utility of event-related potentials in predicting antidepressant treatment response: an iSPOT-D report. Eur. Neuropsychopharmacol. J. Eur. Coll. Neuropsychopharmacol. 25, 1981–1990. doi:10.1016/j.euroneuro.2015.07.022.

Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., 1987. Universals and cultural differences in the judgments of facial expressions of emotion. J. Pers. Soc. Psychol. 53, 712–717.

Etkin, A., Patenaude, B., Song, Y.J.C., Usherwood, T., Rekshan, W., Schatzberg, A.F., Rush, A.J., Williams, L.M., 2015. A cognitive-emotional biomarker for predicting remission with antidepressant medications: a report from the iSPOT-D trial. Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol. 40, 1332–1342. doi:10.1038/npp.2014.333.

Harmer, C.J., Cowen, P.J., 2013. 'It's the way that you look at it' – a cognitive neuropsychological account of SSRI action in depression. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368, 20120407. doi:10.1098/rstb.2012.0407.

Harmer, C.J., Cowen, P.J., Goodwin, G.M., 2011. Efficacy markers in depression. J. Psychopharmacol. doi:10.1177/0269881110367722.

Harmer, C.J., O'Sullivan, U., Favaron, E., Massey-Chase, R., Ayres, R., Reinecke, A., Goodwin, G.M., Cowen, P.J., 2009. Effect of acute antidepressant administration on negative affective bias in depressed patients. Am. J. Psychiatry 166, 1178–1184. doi:10.1176/appi.ajp.2009.09020149.

Hastie, Tibshirani, Friedman, 2009. Model assessment and selection. The Elements of Statistical Learning. Spinger-Verlag.

Huang, Y.-M., Du, S.-X., 2005. Weighted support vector machine for classification with uneven training class sizes. In: Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, 7, pp. 4365–4369. doi:10.1109/ICMLC.2005.1527706.

Kingslake, J., Dias, R., Dawson, G.R., Simon, J., Goodwin, G.M., Harmer, C.J., Morriss, R., Brown, S., Guo, B., Dourish, C.T., Ruhé, H.G., Lever, A.G., Veltman, D.J., van Schaik, A., Deckert, J., Reif, A., Stäblein, M., Menke, A., Gorwood, P., Voegeli, G., Pérez, V., Browning, M., 2017. The effects of using the PReDicT test to guide the antidepressant treatment of depressed patients: study protocol for a randomised controlled trial. Trials 18, 558. doi:10.1186/s13063-017-2247-2.

Leuchter, A.F., Cook, I.A., Marangell, L.B., Gilmer, W.S., Burgoyne, K.S., Howland, R.H., Trivedi, M.H., Zisook, S., Jain, R., McCracken, J.T., Fava, M., Iosifescu, D., Greenwald, S., 2009. Comparative effectiveness of biomarkers and clinical indicators for predicting outcomes of SSRI treatment in major depressive disorder: results of the BRITE-MD study. Psychiatry Res. 169, 124–131. doi:10.1016/j.psychres.2009.06.004.

NICE, 2009. Treatment and Management of Depression in Adults, Including Adults with a Chronic Physical Health Problem. NICE, London.

Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 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, A.J., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., Niederehe, G., Thase, M.E., Lavori, P.W., Lebowitz, B.D., McGrath, P.J., Rosenbaum, J.F., Sackeim, H.A., Kupfer, D.J., Luther, J., Fava, M., 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am. J. Psychiatry 163, 1905–1917. doi:10.1176/appi.ajp.163.11.1905.

Shiroma, P.R., Thuras, P., Johns, B., Lim, K.O., 2014. Emotion recognition processing as early predictor of response to 8-week citalopram treatment in late-life depression. Int. J. Geriatr. Psychiatry 29, 1132–1139. doi:10.1002/gps.4104.

Thomas, J.M., Higgs, S., Dourish, C.T., 2016. Test–retest reliability and effects of repeated testing and satiety on performance of an emotional test battery. J. Clin. Exp. Neuropsychol. 38, 416–433. doi:10.1080/13803395.2015.1121969.

Tranter, R., Bell, D., Gutting, P., Harmer, C., Healy, D., Anderson, I.M., 2009. The effect of serotonergic and noradrenergic antidepressants on face emotion processing in depressed patients. J. Affect. Disord. 118, 87–93. doi:10.1016/j.jad.2009.01.028.