Clinical Neurophysiology 124 (2013) 1975–1985

Contents lists available at SciVerse ScienceDirect

Clinical Neurophysiology

journal homepage: www.elsevier.com/locate/clinph

A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder

Ahmad Khodayari-Rostamabad [a,c,*], James P. Reilly [a], Gary M. Hasey [b,c], Hubert de Bruin [a], Duncan J. MacCrimmon [b]

[a] Dept. of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada L8S 4K1
[b] Dept. of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada L8S 4L8
[c] Mood Disorders Program, Mountain Health Services, St. Joseph's Healthcare, Hamilton, ON, Canada L8N 3K7

ARTICLE INFO

Article history:
Accepted 5 April 2013
Available online 15 May 2013

Keywords:
Prediction
Machine learning
Mood disorders
Major depressive disorder
EEG
Antidepressants
Personalized medicine
Biomarkers

HIGHLIGHTS

- Salient quantitative biomarkers or features extracted from the pre-treatment EEG data are automatically identified that discriminate responders from non-responders to an antidepressant treatment for major depressive disorder (MDD).
- The proposed machine learning methodology produces improved prediction performance compared to previous methods.
- This methodology introduces a novel, personalized approach to the treatment of MDD, that could improve treatment efficacy and reduce health care costs.

ABSTRACT

Objective: The problem of identifying, in advance, the most effective treatment agent for various psychiatric conditions remains an elusive goal. To address this challenge, we investigate the performance of the proposed machine learning (ML) methodology (based on the pre-treatment electroencephalogram (EEG)) for prediction of response to treatment with a selective serotonin reuptake inhibitor (SSRI) medication in subjects suffering from major depressive disorder (MDD).
Methods: A relatively small number of most discriminating features are selected from a large group of candidate features extracted from the subject's pre-treatment EEG, using a machine learning procedure for feature selection. The selected features are fed into a classifier, which was realized as a mixture of factor analysis (MFA) model, whose output is the predicted response in the form of a likelihood value. This likelihood indicates the extent to which the subject belongs to the responder vs. non-responder classes. The overall method was evaluated using a "leave-*n*-out" randomized permutation cross-validation procedure.
Results: A list of discriminating EEG biomarkers (features) was found. The specificity of the proposed method is 80.9% while sensitivity is 94.9%, for an overall prediction accuracy of 87.9%. There is a 98.76% confidence that the estimated prediction rate is within the interval [75%, 100%].
Conclusions: These results indicate that the proposed ML method holds considerable promise in predicting the efficacy of SSRI antidepressant therapy for MDD, based on a simple and cost-effective pre-treatment EEG.
Significance: The proposed approach offers the potential to improve the treatment of major depression and to reduce health care costs.

* Corresponding author at: Dept. of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada L8S 4K1. Tel.: +1 905 730 2474; fax: +1 905 521 2922.
E-mail addresses: khodaya@mcmaster.ca (A. Khodayari-Rostamabad), reillyj@mcmaster.ca (J.P. Reilly), ghasey@sympatico.ca (G.M. Hasey), debruin@mcmaster.ca (H. de Bruin), maccrim@mcmaster.ca (D.J. MacCrimmon).

1388-2457/$36.00
http://dx.doi.org/10.1016/j.clinph.2013.04.010

1. Introduction

Major depressive disorder (MDD) is a serious and common mental disorder and is a major cause of workplace disability with costs very similar to those of diabetes and heart disease (Druss et al., 2000). By the year 2020, depression is expected to account for about 15% of total global disease burden, second only to ischemic heart disease (Lopez and Mathers, 2006). In industrialized countries mental illnesses may account for about 16% of total health care costs (Jäger et al., 2008) and for about 30% of disability claims (Dewa et al., 2004).

There are more than 20 available anti-depressant medications divided into several classes for the treatment of MDD. Despite the high prevalence and disabling nature of this illness, objective procedures for selecting which of these medications is optimal for a specific individual are lacking. Therefore the determination of an effective treatment often devolves into a trial-and-error procedure. Typically about two out of every three patients treated do not remit after the first antidepressant medication trial (Trivedi et al., 2006). In the STAR*D trial, one of the world's largest naturalistic studies of the treatment of major depression, 67% of those treated eventually reached remission; however, for many subjects, up to 4 different antidepressant treatment trials were required, each taking 6 weeks or longer (Rush et al., 2006; Malone, 2007).

The personal and economic cost of delayed or ineffective antidepressant therapy is substantial (Simon et al., 2006; Malone, 2007). Untreated MDD is associated with a 2- to 7-fold increase in non-psychiatric medical costs (Revicki et al., 1998; Druss et al., 2000). Prolonged, depression-related disability results in unnecessary personal suffering, delayed return to work, loss of income and productivity, and increased risk of suicide.

In this paper, we show that machine learning (ML) methods, based on an analysis of the patient's pre-treatment electroencephalogram (EEG), can be used to predict the response of a particular subject to a commonly-prescribed class of anti-depressant medications, the selective serotonin reuptake inhibitors (SSRIs). Such a procedure has the potential to significantly improve treatment efficiency, since considerable time and resources can be saved by avoiding SSRI treatment in the significant proportion of subjects who are likely to be non-responsive. Furthermore, if the proposed method can be extended to predict response to a wider range of anti-depressant therapies,[1] then the physician will be able to determine, in advance, which treatment modality is most effective for a particular individual. This will improve the probability of a response to the first trial, thereby diminishing the need for the extensive trial-and-error process that is current practice.

Over the years, strategies have been developed to employ resting EEG, or quantitative EEG (QEEG) data, as a method for understanding the biological heterogeneity of psychiatric syndromes and predicting treatment outcome in depressed subjects (Kemp et al., 2008; Leuchter et al., 2009a; Iosifescu et al., 2009, 2011; DeBattista et al., 2011; Spronk et al., 2011; Moore, 2011; Alhaj et al., 2011; Tenke et al., 2011).

Pre-treatment right-hemisphere delta and theta absolute powers are reportedly increased in drug-free depressed subjects compared to controls (Kwon et al., 1996), suggesting that inter-hemispheric factors may play an important role in the pathophysiology of MDD. This is further supported by the observations of Bruder et al. (2001) who, employing EEG together with dichotic listening, found that greater alpha power in the right compared with the left hemisphere was associated with favorable response to the SSRI drug fluoxetine. The opposite hemispheric asymmetry predicted poor response.

With respect to the application of more sophisticated mathematical approaches to treatment response prediction in MDD, a quantitative approach using artificial neural networks (ANN) employing clinical and demographic (but not EEG) features has been developed (Serretti et al., 2007). However this approach correctly predicted response at a level of 62%, only slightly better than chance (50%). Others using LORETA (Low Resolution Electromagnetic Tomography Analysis) (Pascual-Marqui et al., 1999) have found significantly increased pre-treatment resting theta activity in the rostral anterior cingulate cortex but not in the posterior cingulate cortex in responders to the drug reboxetine, and to a less significant level to citalopram, an SSRI (Mulert et al., 2007a). Using LORETA, Korb et al. (2009) studied the rostral anterior cingulate and the medial orbitofrontal cortex in subjects receiving fluoxetine, venlafaxine, or placebo, and found that pretreatment theta current density in these regions is higher in responders to medication compared to non-responders.

Spronk et al. (2011) investigated EEG, genetic and neurophysiological biomarkers in 25 subjects with MDD and found that auditory evoked responses (the odd-ball N1 potential) and elevated levels of pre-treatment frontal theta power in the eyes-closed condition was predictive of response to a variety of anti-depressant medications. Tenke et al. (2011) studied the predictive ability of posterior alpha to predict response to a set of antidepressant medications in a sample of 41 MDD and 41 healthy subjects and found that non-responders had significantly less posterior alpha, compared to responders or healthy controls. Using medial alpha of healthy control subjects as the discrimination threshold, a specificity of 92.3% and a sensitivity of 50% were obtained.

Several different research groups have noted that the change in prefrontal theta band cordance (a measure including absolute and relative power) between pretreatment values and those after 1 week of treatment are predictive of final response to a variety of antidepressants (Bares et al., 2007, 2008; Hunter et al., 2007; Cook et al., 2002). In these studies at least two EEGs are required as pretreatment cordance by itself is not predictive of treatment response. Iosifescu et al. (2009) and Leuchter et al. (2009a) investigated the use of the ATR (*Antidepressant Treatment Response*) index, a non-linear combination of prefrontal QEEG theta and alpha powers calculated by measuring the difference between pretreatment and post 1 week of medication treatment values. ATR, determined after the first week of treatment, is reported to predict response to escitalopram (with 74% accuracy), and venlafaxine (with 70% accuracy), (Iosifescu et al., 2009) measured several weeks later. While potentially saving time and money, this method has the disadvantage of requiring two separate EEGs measured at least 1 week apart thereby committing the patient to at least 1 week of potentially ineffective drug therapy.

Reference guided EEG (rEEG) is a rules-based method that compares the QEEG of a given subject with those in a database of others who have demonstrated response to a particular pharmacotherapy (Suffin et al., 2007). Using rEEG (DeBattista et al., 2011) reported that *reference-EEG* guided antidepressant therapy is more effective (at a 65% response rate) than the response rate (39%) of a control group, who were prescribed the treatment based on the STAR-D study algorithms (Rush et al., 2006).

Machine learning (ML) (*aka* data mining or pattern recognition) methods have been previously used in several EEG applications, including the analysis of EEG signals for epilepsy (Ghosh-Dastidar et al., 2008; Güler and Übeyli, 2007), in evaluating residual functional deficits following concussion (Cao et al., 2008), to classify sleep stage in animals (Crisler et al., 2008), and for distinguishing age of infants (Ravan et al., 2011).

The machine learning methodology proposed in this paper employs mathematically-structured, optimization-based machine learning techniques using only the pre-treatment EEG. The

[1] This is work currently in progress by the authors.

Table 1
Available clinical characteristics of SSRI responders (R) and non-responders (NR) who participated in the study. Mean (± standard deviation) are shown for relevant variables.

Information	R	NR	Total	*p* Values
Age [years]	46.4 (± 8.3)	46.9 (± 11.2)	46.8 (± 10.1)	$p = 0.904$
Gender (female/male)	6/1	5/10	11/11	$p = 0.015$
Handedness[a]	0.54	0.73	0.67	$p = 0.542$
Pre-treatment HamD-17	27.1 (± 7)	21.5 (± 3.5)	23.3 (± 5.4)	$p = 0.082$
Post-treatment HamD-17	14.4 (± 6.4)	20.9 (± 3.8)	18.8 (± 5.6)	$p = 0.039$
Pre-treatment BDI[b]	38.1 (± 10.6)	37.2 (± 8.9)	37.5 (± 9.2)	$p = 0.842$
SSRI treatment[c]	Z:3,L:2,C:1,P:1	Z:12,L:1,C:1,P:1	Z:15,L:3,C:2,P:2	
Co-morbidities:[d]				
Dysthymia	5 (=71.4% of R)	10 (=66.7% of NR)		
OCD	1 (=14.3% of R)	1 (=6.7% of NR)		
PTSD	1 (=14.3% of R)	0		
Social Phobia	1 (=14.3% of R)	1 (=6.7% of NR)		
Panic/Agora	1 (=14.3% of R)	2 (=13.3% of NR)		
GAD	1 (=14.3% of R)	0		
Borderline PD	1 (=14.3% of R)	0		

[a] Handedness: 1:Right-handed (13 subjects, including 3 R and 10 NR subjects), 0.75: Mostly right, but some left (5 subjects, including two R subjects and three NR subjects), 0.25: some right, some left, but right is used more than left (1 R subject), −0.25: some left, some right, but left is used more than right (1 NR subject), −1: Left-handed (2 subjects, including 1 R and 1 NR subject) (based on Annett handedness rating).
[b] BDI stands for Beck Depression Inventory, Beck et al. (1961).
[c] SSRI medication administered: Z: Zoloft (Sertraline hydrochloride), C: Celexa (Citalopram), L: Luvox (Fluvoxamine), P: Paxil (Paroxetine).
[d] OCD stands for Obsessive Compulsive Disorder, PTSD stands for Post-Traumatic Stress Disorder, GAD stands for Generalized Anxiety Disorder, and PD stands for Personality Disorder.

proposed method automatically selects salient or discriminative features from a large list of candidate features extracted from the EEG. These selected features are used to predict response. The prediction is generated using a classifier, which outputs a response based on an optimal multi-dimensional decision rule. This mathematically structured approach produces a prediction performance that is improved over that of previous methods.

2. Methods

2.1. Participants

Our study was approved by the Research Ethics Board of St. Joseph's Health Care, Hamilton, ON, Canada. Twenty-two subjects (11 males, 11 females, age 20.6–62.6 years) diagnosed with MDD using the internationally recognized Diagnostic and Statistical Manual – IV diagnostic criteria (American Psychiatric Association, 2000) were treated with a 6 week course of an SSRI (mainly, the drug Sertraline hydrochloride). Twenty one subjects had co-morbidities (see Table 1). Our subjects were recruited from a tertiary Mood Disorders Clinic and all were considered treatment resistant, defined as failure to respond to at least two previous adequate trials of different antidepressant medications, (Burrows et al., 1994). Depression severity was measured by trained raters at baseline and after 6 weeks of treatment using the 17 item Hamilton depression rating scale (HamD-17) (Hamilton, 1960). Participants had a pre-treatment score of $\geqslant 18$ on the HamD-17. A score greater than 18 is an indication of at least moderately severe depression. Table 1 summarizes the available clinical and demographic information of the participants. Table 2 lists the details of previous treatments. Responders did not differ significantly (with a 5% significance level) from non-responders in age, handedness or co-morbidity. However, responders were more likely to be females and to have higher pre-treatment HamD-17 scores. In column 5 of Table 1, the result of a two-tailed Welch's t test comparing R and NR groups is shown. The gender imbalance in the responder group is very likely a consequence of our small sample size.[2]

[2] Despite this fact, care must be exercised with regard to interpretation of the results, since a bias may be introduced if the machine learning process is anomalously biased towards predicting gender. This point is discussed further in Section 4.

Table 2
Previous medication treatment classes in SSRI responder (R) versus non-responder (NR) subjects who participated in the study. Note that the EEG signals were recorded after approximately 10 days of medication withdrawal and before administering a 6-week trial of SSRI treatment. Columns 2 and 3 show the average number of medications taken from each medication class by each subject in the R and NR groups, respectively.

Medication Class	R	NR	*p* Values
SSRI[a]	2.29 (± 1.25)	2.33 (± 1.05)	0.932
SNRI[b]	0.86 (± 0.38)	0.8 (± 0.41)	0.754
Aminoketone	0.43 (± 0.53)	0.87 (± 0.35)	0.081
Phenylpiperazine	0.43 (± 0.53)	0.13 (± 0.35)	0.217
Tetracyclic	0 (± 0)	0.07 (± 0.26)	0.334
Tricyclic	0.86 (± 1.22)	1 (± 0.93)	0.789
MAOI[c]	0.71 (± 0.76)	0.4 (± 0.51)	0.345
Psycho stimulant	0.14 (± 0.38)	0.27 (± 0.46)	0.515
Antipsychotic	0.29 (± 0.49)	0.27 (± 0.46)	0.932
Lithium augmentation	0.71 (± 0.49)	0.47 (± 0.52)	0.297
Buspirone augmentation	0 (± 0)	0.13 (± 0.35)	0.164
Anticonvulsant	0.57 (± 1.51)	0.33 (± 0.62)	0.7
Thyroid hormone	0.29 (± 0.49)	0.4 (± 0.51)	0.622
Tryptophan augmentation	0.14 (± 0.38)	0.13 (± 0.35)	0.956
Benzodiazepine	0.43 (± 0.79)	0.27 (± 0.46)	0.627
SARI[d]	0.14 (± 0.38)	0.07 (± 0.26)	0.641

[a] Previous SSRI medications received by subjects include: Sertraline, Paroxetine, Fluvoxamine, Fluoxetine, Citalopram, and Escitalopram.
[b] SNRI stands for serotoninnorepinephrine reuptake inhibitors class.
[c] MAOI stands for monoamine oxidase inhibitors class.
[d] SARI stands for serotonin antagonist and reuptake inhibitor.

EEG recordings were also taken 2 weeks after onset of treatment for 21 out of the 22 subjects who participated in the study, using a protocol identical to that of the pre-treatment EEG collection. The subject for whom the post-2-week data is not available is a 20.6-year-old right-handed male who is NR. In addition to the primary data explained above, the EEG data of 91 healthy adult subjects were also collected. These are used for normalization of the features extracted from the pre-treatment EEG.

2.2. EEG Measurements and the pre-processing procedure

The standard 10–20 EEG recording procedure with a sampling frequency of 205 Hz, referenced to linked ears (LE), was used in this study. Recordings were taken after approximately 10 days of medication withdrawal and before a 6-week trial of antidepressant

treatment was administered. Subsequent analysis determined that the inter- and intra-hemispheric characteristics of the EEG data were dominant; therefore, in the interests of saving computational and hardware resources, the data from the centreline electrodes were discarded. Thus only the 16 electrodes Fp1, Fp2, F3, F4, F7, F8, T3, T4, C3, C4, T5, T6, P3, P4, O1 and O2 were used in this study. A QSI-9500 EEG system is used, which filters the signals between [0.5–80 Hz] and applies a notch filter at 60 Hz. The patient was in a semi-recumbent position in a sound attenuated, electrically shielded room. The process was administered by an experienced technician who prompted patients on signs of drowsiness. Sessions were arranged in the mornings and patients were requested to avoid coffee, drugs, alcohol and smoking immediately prior to the recording. Only resting EEGs were recorded.

For de-artifacting, an expert blind to treatment response visually inspected the entire EEG waveforms and discarded the parts which had eye-movements, eye blinks, saturation/clipping, movement artifacts or excessive muscle artifacts. The signals were then digitally bandpass filtered after recording between 2 Hz and 35 Hz. A disadvantage of the band-pass filtering used in this process is the loss of potentially important data in lower part of the delta and all gamma frequency bands. An alternative option for de-artifacting is to use an automated method such as the one discussed in Mourad et al. (2010), or a method based on machine learning which is trained using a visually de-artifacted data set.

Some authors e.g., (Qin et al., 2010; Nunez, 2010) recommend the use of the average reference (AR) as well as the EEG infinity referencing system (IR), otherwise known as the *reference electrode standardization technique* (REST) (Qin et al., 2010). It is a relatively straightforward mathematical procedure to first transform the recorded linked ear (LE) referenced data to the AR referencing system, and from this into the IR reference system. In this process, we used the same head model and design parameters as in Qin et al. (2010). For the purposes of comparison, in this study we analyzed the data using each of the LE, AR and IR referencing methods; however, following the recommendation by Qin et al. (2010), our results are reported using only the IR referencing system.

For each patient, a nominal 6 EEG data files, each of 3.5 min duration, were collected. Of these, 3 were under the eyes open (EO) condition, and 3 were eyes closed (EC). Often in studies using the EEG, it is necessary to distinguish between the EO and EC cases. Many studies involving EEG use only EC data, in part because posterior EEG alpha power, which is analyzed as a feature, is more prominent for the EC condition. However, at least one study found greater pre-treatment alpha power asymmetry in responders to an SSRI vs. non-responders for EO EEG (Bruder et al., 2001). In the present study, it was also found that even though particular features corresponding to the EO, EC, and EO- and EC-combined cases were different, the overall final performance did not vary significantly. We therefore used combined EO and EC EEG measurements in the following experiments to make maximal use of the available data.

The first 60 s[3] of de-artifacted data from each of the six files of each subject were compiled into a single epoch. Statistical quantities such as power spectral densities and magnitude coherences (which become candidate features to be described later) are calculated using a Welch modified periodogram method (Proakis et al., 1992) over each epoch. The individual windows required for this process are obtained by dividing each epoch into windows of 2 s. duration with 50% overlap.[4] These settings result in a nominal six epochs per subject. A 2 s. window length enables analysis of frequencies above approximately 2.5 Hz.

2.3. Definition of response

The definition of a responder to the SSRI medication in this case was taken to be at least a 30% improvement between the pre- and post-treatment HamD-17 scores. Although "response" is often defined as at least 50% improvement of depression rating scales, a recent review of the threshold for clinically significant improvement (Bandelow et al., 2006) concluded that this value is overly conservative. In this review it was noted that many subjects considered by their clinicians to show clinically meaningful improvement showed only 23–42% improvement in scores using standardized symptom rating scales, such as the Hamilton Depression rating scale. Furthermore it has been demonstrated that a 25% improvement in depression rating scale scores in the first few weeks of treatment may be predictive of more extensive improvement several weeks later (Henkel et al., 2009). Indeed, only 50% of those patients who go on to reach full remission do so in the first 6 weeks (Rush et al., 2009). For these reasons we believe that an improvement of 30% in Hamilton depression rating scale scores is clinically significant. Further, in our study sample of treatment-resistant patients, the number of subjects showing 50% improvement or more was insufficient to generate a reliable training set.

2.4. Overview of the machine learning methodology

Pre-treatment resting or spontaneous EEG signals were collected from 22 subjects with MDD who participated in the study. All subjects were then treated with standard doses of an SSRI medication for 6 weeks under the supervision of an experienced psychiatrist. The treatment response was determined by a trained interviewer using the 17 item Hamilton Depression rating scale after 6 weeks of antidepressant medication therapy. The response is denoted by y_i, $i = 1, \ldots, M_t$, where M_t is the number of training epochs available. The possible values for the y_i are either "R" (responder), or "NR" (non-responder). In our study, there are a nominal six epochs per subject. However, for one subject, only two instead of the three EO EEG files were available, and one of the EC files for another subject had less than 60 s of recorded data after visual de-artifacting. Therefore, the total number of available epochs for our experiments is $M_t = (6 \text{ epochs/subject} \times 22 \text{ subjects} - 2) = 130$.

Each epoch of the measured EEG signals was processed to obtain a large number N_c of candidate features that might be relevant for prediction. These features were assembled into a set of candidate feature vectors $\tilde{\boldsymbol{x}}_i, \quad i = 1, \ldots, M_t$ of dimension N_c. Only those features from the $\tilde{\boldsymbol{x}}_i$ that were most statistically relevant were then selected, using a feature selection procedure to be described. The set of selected reduced-dimensionality vectors are denoted as $\boldsymbol{x}_i$, of dimension N_r. The training set for this study, which consists of the available set of selected features and the corresponding responses, was then denoted by $\mathscr{D} = \{(\boldsymbol{x}_i, y_i), \quad i = 1, \ldots, M_t\}$.

In this study, there are two options considered. The first is the 1-D case where individual features on their own were used to predict response. The second is the M-D case, where the multiple features contained in $\boldsymbol{x}_i$ were fed into a classifier, which outputs the predicted response to the treatment using an optimal multi-dimensional decision rule.

As is made evident later, the 1-D option is considerably simpler than the M-D implementation. The question as to why the M-D case should be considered at all naturally arises. The answer is associated with the need for higher prediction performance.

[3] We found that increasing the epoch length beyond a length of 60 s. did not result in improved performance.

[4] The 50% overlap figure and 60 s. epoch length were chosen somewhat arbitrarily and can be altered with reasonable limits without significant impact on performance. A 50% overlap results in a total of 59 windows in each epoch, which is a high enough number to obtain a relatively low variance power spectral density estimate.

Consider the brain process, or model, that maps the underlying response characteristic of the brain into the observed EEG patterns (In essence, the proposed ML procedure attempts to "learn" this model). Since this model is likely to be complex in structure, it can be described more accurately if multiple variables (as opposed to a single variable) are used to describe it. Thus in principle, a multi-dimensional description of the model can be more accurate. However, in the finite data case, it could be that the simpler 1-D description performs better. It is due to this uncertainty that we investigate both options in this paper.

A similar ML methodology is described in Khodayari-Rostamabad et al. (2010a), which investigates the use of machine learning methods for the prediction of response to clozapine therapy for schizophrenia. The methodology of the present study for prediction of response to SSRI therapy is also briefly described in Khodayari-Rostamabad et al. (2010b).

The components of ML methodology are feature extraction, feature selection, classification and performance evaluation. In the following sub-sections, we briefly explain the role of each of these components as they apply to the problem at hand.

2.5. Feature extraction and feature selection

The first step in the ML prediction method is to extract the candidate feature or biomarker vector $\tilde{\boldsymbol{x}}_i$ from each epoch of the measured EEG waveforms. The set of candidate features extracted from each data epoch consist of the following statistical parameters: the log power spectral density (PSD) levels at each individual electrode at all frequencies of interest, magnitude squared spectral coherence between all electrode pairs at each frequency of interest, the mutual information between all electrode pairs, the log ratio of left-to-right hemisphere powers and anterior/posterior log power ratios, between several electrodes and again at all frequencies of interest. Basically, the candidate feature set consists of as broad a range of statistical parameters extracted from the EEG as possible. The extraction of candidate features is done without any *a priori* concern as to whether they are predictive or not.

In this study, we obtain better performance if a frequency resolution finer than that of the classical EEG frequency band demarcations is used. We therefore calculated the frequency dependent parameters in the band 3–30 Hz in 1 Hz increments. A 1 Hz resolution is significant in this study, as subtle predictive relationships that might occur at only one specific frequency may be obscured when the data are collapsed across the traditional frequency bands. In this analysis, there are $N_c = 5552$ candidate features. This number is a function of the selected frequency resolution and the number of EEG electrodes used.

The candidate feature values were normalized using a *z-score* normalization procedure. That is, the ith candidate feature value f_i used in subsequent processing was replaced with the value $\frac{f_i-\mu_i}{\sigma_i}$, where μ_i and σ_i respectively are the means and standard deviations of the corresponding feature, evaluated over the EEG data gathered from the healthy subjects.

The use of 2nd-order statistics as candidate features can be justified by considering that a complete set of statistical moments fully describes a multi-variate random process. In our case, the statistical quantities we select as candidate features are a subset of the complete set of moments, and therefore offer a partial description of the process generating the EEG observations. Such quantities have been used in previous related studies; e.g., PSD values are used in (Cook et al., 2002; Hunter et al., 2007; Hinrikus et al., 2009; Knott et al., 2001; Kwon et al., 1996; Bruder et al., 2001); left-to-right hemisphere powers are used in Hinrikus et al. (2009), Knott et al. (2001); and anterior/posterior power ratios are used in Knott et al. (2001). The work of Knott et al. (2001), Hinrikus et al. (2009), Knott et al. (2002) used coherence between electrode pairs to assess the effect of the anti-psychotic drug *clozapine* and characterize depression, respectively. Also Kwon et al. (1996), Ulrich et al. (1994), Bruder et al. (2001, 2008) have used inter and intra-hemispheric power ratios as numerical indicators or predictors of treatment response. Mutual information was used in (Na et al., 2002) to analyze EEG data abnormalities in 10 schizophrenic subjects as compared to 10 normals.

The number N_c of *candidate* features is large; many of the extracted features are highly correlated amongst themselves, and most are statistically independent of the y_i. We therefore select only a subset of N_r most relevant features, where $N_r \ll N_c$. It is desirable to choose N_r as small as possible to minimize the redundancy in the feature set to avoid over-fitting of the inherent model, yet simultaneously choose N_r large enough so that under-fitting is avoided. The topic of feature selection is an on-going field of research in machine learning, and many methods are available for this purpose. In this study, the features were selected based on the Fisher discriminant ratio (FDR) (Hastie et al., 2009), which is defined as follows

$$\mathrm{FDR}(j) = \frac{(\mu_N(j) - \mu_R(j))^2}{\sigma_N^2(j) + \sigma_R^2(j)} \tag{1}$$

where $(\mu_R(j),\ \sigma_R(j))$ are the mean and standard deviation, respectively, of the jth feature values for the responder (R) group. Corresponding values for the non-responder (NR) group are denoted by subscript N. The FDR (j) value is calculated for all $j = 1, \ldots, N_c$ candidate features. The N_r features with the largest FDR values form the selected feature set. Roughly speaking, the FDR selects features for which the squared difference of the means between the R and NR groups is large, relative to the sum of the variances for that feature. In an approximate sense, features with higher FDR value are more discriminative. There are many possible alternative algorithms for feature selection; for example, the method of Peng et al. (2005) was also found to perform well for the purposes of this study.

To increase robustness of the selected features, we apply the following procedure. We divide the training set into 10 contiguous subsets of 2 subjects each. Then feature selection is repeated over 10 iterations, where in each iteration, the epochs from each subset in turn are omitted, whereupon the feature selection process is performed using only the remaining subsets. In each round, a list of $k\ N_r$, $k > 1$ most relevant feature indexes are selected using the FDR feature selection method. ($k = 4$ is used in this study). Then, the reduced feature set is selected as the set of N_r most commonly occurring feature indexes amongst the 10 available lists. This is referred to as a *feature polling procedure*, in which the feature indices with the maximum number of votes among the available subsets are chosen. The selected features are assembled into a vector $\boldsymbol{x}_i \in \mathbb{R}^{N_r}, i = 1, \ldots, M_t$. In this study, the value for N_r was selected to be five or less on the basis that using a fewer input variables to the predictor results in simpler prediction model with less chance to over-fitting in a small sample size problem.

2.6. Using features to predict response

The subject response must be predicted from the selected feature set. This is determined using a classifier or predictor designed using the training data. In this case, all selected features are fed into the classifier, which then outputs the predicted response. In this case, we used the *mixture of factor analysis* (MFA) classifier (Ghahramani et al., 1996), which is based on a *maximum likelihood* classification rule. This method is used previously in Khodayari-Rostamabad et al. (2010c) to build statistical diagnosis models to discriminate psychiatric disorders. The MFA method outputs a continuously-varying likelihood value $\ell(\boldsymbol{x})$, where $0 \leqslant \ell(\boldsymbol{x}) \leqslant 1$, which indicates the degree to which the subject belongs to the responder

category. If $\ell(\boldsymbol{x})$ is above a specified threshold value, the subject is declared a responder, and a non-responder otherwise. In contrast, most other classifier structures output only a discrete "R" or "NR" value. The MFA method is desirable in this application since it is capable of adapting to any inherent nonlinearities that may exist in the feature space, thus improving performance over methods which implicitly assume a linear feature space. Further, since the MFA likelihood indicates degree of responsiveness, it is useful if predicting response to multiple forms of treatment. In the case where the subject may respond to more than one of these treatments, the MFA method allows rank ordering of treatment options in terms of response likelihood. Note that the MFA method has two associated parameters; i.e., the number of mixtures and the number of factors, that together have a considerable impact on performance. A procedure for optimizing these parameters is discussed in the next sub-section.

We process multiple epochs for each subject. The final prediction result for each subject is obtained by averaging the MFA likelihood values over all available epochs for that subject. The resulting averaged value is then quantized, and then the predicted class (i.e., R or NR) of the subject is the one with the highest average likelihood.

2.7. Performance evaluation

A fair evaluation requires the assessment of performance over the range of selected features and classifier designs that correspond to a wide range of subjects. To address this consideration, we evaluate performance using a leave-*n*-subjects-out (L*n*O) cross-validation procedure. The 22 subjects available in our study are divided into 11 contiguous numbered subsets each of size $n = 2$ subjects, so that each subject is contained in only one subset. The process iterates through each subset. In the pth iteration (i.e., fold), the epochs contained within the pth subset, along with the corresponding outcome variable y_i, are sequentially omitted from the training set.[5] The epochs in the pth omitted subset are referred to as the test subset, denoted as $\mathscr{D}_p^{\text{test}}$. The remaining epochs (belonging to the remaining 20 subjects) constitute the set denoted as $\mathscr{D}_p^{\text{train}}$, which is used for selecting features, parameter optimization and the training of the classifier. The resulting structure is then tested using $\mathscr{D}_p^{\text{test}}$.

The training process implicitly includes the selection of features and the optimization of the parameters. It is tempting to perform both these operations once, outside the main cross-validation loop in the interests of computational cost. However, in so doing, the test set is inherently involved in these processes, with the result that the estimated performance is optimistically biased. To remedy this situation, we perform a separate parameter optimization and feature selection process within each fold of the cross-validation procedure, as suggested by Varma and Simon (2006), Hastie et al. (2009).

The results of the evaluation procedure, can be compiled into a contingency matrix (or table) $\boldsymbol{T}$, (with elements t_{ij}), that conveniently expresses the level of performance. In the case of two classes, the *specificity* of the method is evaluated as $\frac{t_{11}}{t_{11}+t_{12}}$, whereas *sensitivity* is evaluated as $\frac{t_{22}}{t_{21}+t_{22}}$. An overall performance index is given as the average of the specificity and sensitivity performance indexes.

2.8. Estimation of a confidence interval

The performance figures obtained using the L*n*O procedure, for the M-D case, over our limited data set are only *estimates* of the true performance level of the method. We would therefore like to determine a confidence interval associated with our estimate. To do so, we calculate the probability that the true performance is above a minimum acceptable prediction level, which in this study we take to be 75%.

To generate such statistics of the performance estimate, we use a randomized permutation cross-validation method similar to the Monte–Carlo cross-validation procedure, e.g., (Molinaro et al., 2005), but where the entire L*n*O cross-validation process is repeated over H iterations. In each iteration, the ordering of the subjects in the data set $\mathscr{D}$ is randomly permuted beforehand. This has the desired effect of the test subsets being selected randomly in each iteration. It also guarantees the independence of the training and test sets in every fold and is an efficient method for confidence estimation for small sample sizes. The final performance measure is calculated from the final contingency matrix $\boldsymbol{T}$, which is the average of the contingency tables obtained over all H cross-validation iterations. In our study, $H = 100$.

2.9. Low-dimensional representation

Kernelized principal component analysis (KPCA) (Müller et al., 2001) is used to visualize the clustering behaviour. Here, the N_r-dimensional feature space is reduced into a two-dimensional subset of the feature space, so it can be represented on a two-dimensional plane. In essence, this procedure rotates and nonlinearly transforms the coordinate axes of the feature space to render the view giving the most compact representation. This gives us insight into the clustering and discriminating performance of the feature set, and aids in the identification of outliers.

3. Results

A list of the most relevant discriminating features selected by the proposed FDR procedure is shown in Table 3. The following notation is used: $coh(s_1,s_2,f)$ denotes the coherence between two EEG electrodes s_1 and s_2 at frequency f Hz, $ftb(s_1,s_2,f)$ denotes

Table 3
A list of the most discriminating features, sorted based on FDR value, showing the mean and standard deviation of each feature over the non-responder (μ_N, σ_N) and responder groups (μ_R, σ_R). Notation is described in the text.

#	*Selected feature*	μ_N, ($\pm\ \sigma_N$)	μ_R, ($\pm\ \sigma_R$)	FDR
1	*coh*(F4,T5,16)*	0.348 (± 0.066)	0.189 (± 0.081)	2.34
2	*coh*(Fp2,T5,16)	0.334 (± 0.05)	0.225 (± 0.051)	2.34
3	*coh*(F4,T3,16)	0.159 (± 0.057)	0.066 (± 0.025)	2.24 ◇
4	*coh*(F4,T3,13)*	0.177 (± 0.073)	0.064 (± 0.025)	2.17 ◇
5	*coh*(F4,T3,20)*	0.148 (± 0.06)	0.055 (± 0.026)	2.01
6	*coh*(F4,T5,17)	0.329 (± 0.068)	0.187 (± 0.076)	1.94 ◇
7	*coh*(F7,C4,26)	0.225 (± 0.074)	0.111 (± 0.038)	1.9 ◇
8	*coh*(F4,T3,12)	0.173 (± 0.07)	0.064 (± 0.04)	1.82
9	*coh*(Fp2,T5,17)	0.33 (± 0.066)	0.205 (± 0.065)	1.81 ◇
10	*coh*(F8,C3,14)	0.19 (± 0.072)	0.08 (± 0.039)	1.79
11	*coh*(F4,T3,11)	0.175 (± 0.066)	0.078 (± 0.033)	1.75
12	*coh*(F4,T3,14)*	0.165 (± 0.069)	0.066 (± 0.03)	1.73 ◇
13	*coh*(F4,T5,18)	0.337 (± 0.066)	0.194 (± 0.096)	1.52
14	*coh*(F4,T3,15)	0.157 (± 0.073)	0.061 (± 0.029)	1.51 ◇
15	*coh*(F7,C4,15)	0.203 (± 0.053)	0.117 (± 0.047)	1.47
16	*coh*(F4,T3,21)	0.145 (± 0.057)	0.064 (± 0.036)	1.47
17	*coh*(F4,T5,20)	0.316 (± 0.069)	0.183 (± 0.086)	1.44
18	*coh*(F4,T5,13)	0.403 (± 0.124)	0.232 (± 0.073)	1.42
19	*coh*(P3,O1,3)*	0.338 (± 0.145)	0.517 (± 0.042)	1.4
20	*coh*(F4,T3,17)	0.141 (± 0.068)	0.054 (± 0.028)	1.4 ◇
21	*coh*(F8,T5,28)	0.232 (± 0.057)	0.151 (± 0.038)	1.39
22	*coh*(F4,T5,12)	0.412 (± 0.119)	0.244 (± 0.081)	1.37
23	*psd*(F4,29)	−0.419(± 0.327)	0.142 (± 0.376)	1.27
24	*psd*(Fp2,29)	−0.397(± 0.296)	0.117 (± 0.347)	1.27
25	*coh*(Fp2,T3,13)	0.117 (± 0.053)	0.051 (± 0.025)	1.27
26	*psd*(F4,30)	−0.483(± 0.35)	0.067 (± 0.349)	1.24
27	*coh*(F7,C4,24)	0.205 (± 0.066)	0.115 (± 0.047)	1.24 ◇

[5] The value of $n = 2$ in this study was chosen on the basis that most statistically reliable results are obtained when the number of folds is in the range 5–20 (Kohavi, 1995).

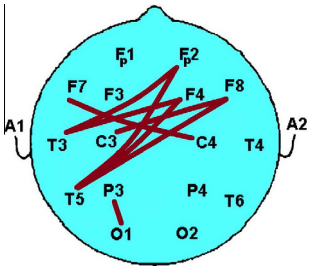

Fig. 1. A rough schematic drawing showing the most relevant features by connections between corresponding EEG electrodes, as reflected in Table 3. Electrodes A1 and A2 denote the linked ear reference used for recording.

front-to-back PSD ratio s_1/s_2 at f Hz, and $psd(s_1 f)$ denotes the PSD of electrode s_1 at f Hz. For each feature, columns 3 and 4 reflect the means and standard deviations of non-responder (μ_N, σ_N) and responder groups (μ_R, σ_R) before *z-score* normalization. These values however depend on the pre-processing, feature extraction and normalization procedures. To calculate standard deviation, we first determined the intra-subject average of each discriminating feature over all subject epochs, and then calculated the inter-subject standard deviation of the averaged feature values. A feature is listed in this table if it is used at least once throughout the entire randomized permutation L*n*O procedure. The set of N_r features is different in each fold; however, the items marked with * in Table 3 indicate an example of such a set of 5 features. It should be noted that because some of the features have strong statistical dependencies, (especially those closely spaced in frequency) the set of selected features in Table 3 is not unique. Some of the features may be replaced with others, with small penalty in performance.

An additional interpretation of the features is given in Fig. 1, which presents a graphical depiction of the most-relevant features listed in Table 3. A connection between two electrode sites in the figure corresponds to a selected feature which involves those two locations. Connections are shown by solid thick lines. This roughly indicates relations between EEG sensors that convey relevant information for our response-prediction task.

We first present prediction results for the 1-D case. Fig. 2 shows results for two individual features, the first and last feature (#27) corresponding to highest and lowest FDR values in Table 3. The horizontal dashed line through each panel is the threshold, which

Table 4
The contingency table for the 1-D case, using either Feature #1 or #27 from Table 3. This table is generated using the randomized permutation cross-validation procedure described in Sections 2.7 and 2.8 but using only Feature 1 (part a) or Feature 27 (part b). The figures shown are the averages over the $H = 100$ LnO runs.

	Predicted NR	Predicted R	% Correct
Part a. Using Feature #1.			
Actual NR	9.03	5.97	60.2% (specificity)
Actual R	1	6	85.7% (sensitivity)
Part b. Using Feature #27.			
Actual NR	9.23	5.77	61.5% (specificity)
Actual R	1	6	85.7% (sensitivity)

in these figures, is the average value of the feature over all subjects, both R and NR. For each feature shown, the responder subjects (those on the left in dark blue) should have values lower than the threshold, while the value of the NR subjects should be higher. Note that this decision threshold is obtained using all available data samples, a process which implicitly includes the test data in the calculation. This results in overfitting which in turn generates an optimistic bias in the results.

The contingency tables for the above two 1-D examples are shown in Table 4, parts a and b. In order to avoid the bias indicated above, the randomized permutation cross-validation procedure described in Sections 2.7 and 2.8 is used, with only the single respective feature. The entries in each cell represent the respective number of subjects, averaged over the 100 random permutation iterations. The corresponding total prediction accuracies are 73.0% and 73.6%, respectively.

The performance in the M-D case obtained by the randomized permutation cross-validation procedure is documented in Table 5 for the value $N_r = 5$. The parameters associated with the MFA classifier (i.e., the number of mixtures and the number of factors) were optimized for best performance, independently in each fold, over the range of values [1, 2,..., 6] and [1,...,4] respectively. The cross-validation procedure yielded an average prediction rate of 87.9% (specificity = 80.93%, sensitivity = 94.86%), with an unbiased standard deviation estimate of $\tilde{\sigma} = 5.35$. Using these figures with a one-tailed t-distribution with 22 degrees of freedom, it is determined that the true prediction rate is in the interval [75–100]% with 98.76% confidence.

An additional set of features (for the M-D case) was identified from the pre-treatment data that was also predictive of the final

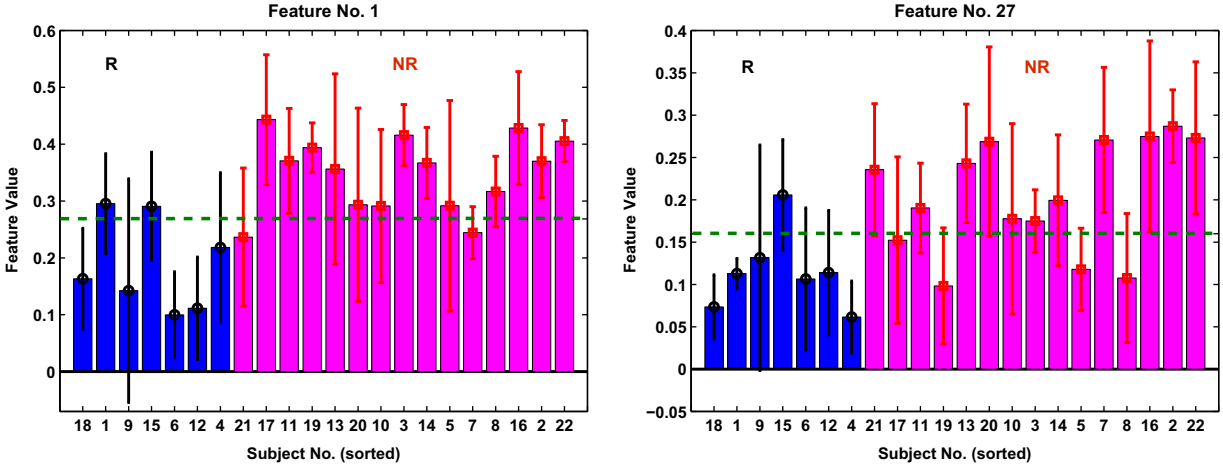

Fig. 2. Plots of feature value vs. subject index, for two features in Table 3. The subjects have been sorted so that the R and NR groups appear in dark blue on the left, and pink on the right, respectively for each feature. The dashed horizontal line represents the average value of the respective feature over both groups, that could be used as a threshold value. Note that this threshold is found using all subjects (including the test subset), and therefore this is an over-fitting case. The R and NR subjects should be below, above the threshold, respectively, for a correct decision. The error bars show the standard deviation of the feature over the six available epochs.

Table 5
The contingency table for the M-D case, generated by the randomized permutation cross-validation procedure, for predicting response to SSRI therapy for subjects with MDD. $N_r = 5$. The figures shown are the averages over the $H = 100$ LnO runs.

	Predicted NR	Predicted R	% Correct
Actual NR	12.14	2.86	80.93% (specificity)
Actual R	0.36	6.64	94.86% (sensitivity)

Table 6
Contingency table for the M-D case obtained by testing the treatment-response predictor on the post-2-week data, when the predictor is trained using the available pre-treatment data. $N_r = 5$.

	Predicted NR	Predicted R	% Correct
Actual NR	11	3	78.57%
Actual R	0	7	100%

response when the post-2-week data was used instead. These features are marked in Table 3 with a diamond symbol. As an additional test of performance, this *post-2-week data*, which in this case consists of 125 epochs, was tested using this set of features, for the multi-dimensional case. We observed that the average response prediction performance in this case is 89.29% (specificity = 0.786, sensitivity = 1), as reflected in Table 6. The three misclassified NR subjects include two males and one female. Since we are not interested in a confidence interval in this example, we used a single *stratified* cross-validation procedure (Kohavi, 1995) instead of the randomized permutation procedure as before. With this former procedure, the subjects are ordered beforehand so that the R and NR subjects are balanced throughout the list. Note that these performance figures are within our confidence interval, so the difference is statistically insignificant.

The fact that we can use ML analysis of EEG signals even after 2 weeks of antidepressant treatment is important, since depressed patients may not need to come off antidepressant medication to be tested using the proposed ML methodology. The study by Bruder et al. (2008) reports that features such as alpha power and hemispheric asymmetry also do not change significantly after 12 weeks of treatment with the anti-depressant fluoxetine.

We now show how low-dimensional representations based on kernelized nonlinear principal components, as discussed in Section 2.9, can be used to visualize the clustering behaviour of the multi-dimensional feature space. Fig. 3 shows a scatter plot of the $M_t = 130$ available pre-treatment training samples projected onto only the first two major nonlinear principal components. This figure was generated using the KPCA method with a Gaussian kernel using the $N_r = 5$ selected features. This figure shows one point for each epoch from each subject; i.e., a nominal 6 points per subject. Each subject is arbitrarily assigned an exclusive index within the range $[1, \ldots, 22]$. For clarity of presentation, in Fig. 3 we label only the points corresponding to the two (female) subjects 18 and 22 (R and NR respectively), to show how the projected data vary between epochs for a given subject. These two subjects were arbitrarily selected, with one subject in each class.

Fig. 4 shows the case where these 6 points from each subject are averaged together into a single point. The shape of the clusters in Figs. 3 and 4 depend on many factors, and in particular which specific features are chosen by the feature selection procedure. Each point in Fig. 4 is labelled with its corresponding subject index.

This example lends credibility to the idea that it is possible to select a set of features from the background EEG which are indicative of response. An advantage of the low-dimensional representation of Fig. 4 is that it visually confirms that the classes do indeed cluster into distinct separable regions in the feature space, indicating that prediction is feasible.

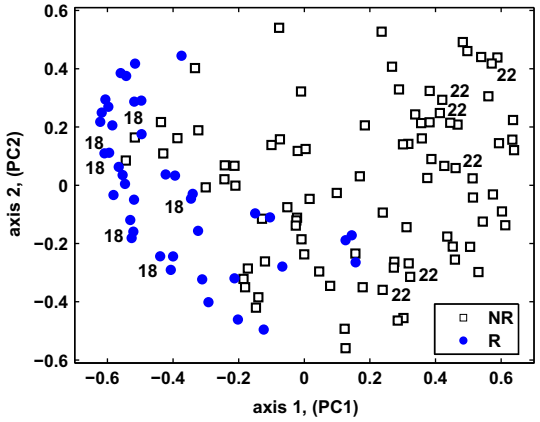

Fig. 3. The M-D case: scatter plot of the projection of the $N_r = 5$-dimensional feature vectors from all $M_t = 130$ available training epochs (nominally 6 epochs per subject) onto the first 2 major principal components, which are obtained using the kernelized principal component analysis (KPCA) method with a Gaussian kernel. The numbers identify the epochs belonging to two, arbitrarily-chosen subjects, one in each class.

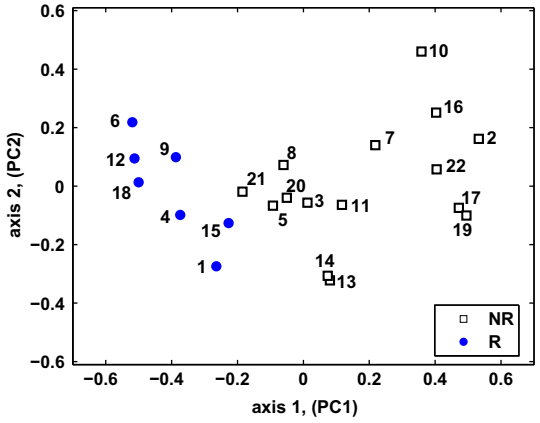

Fig. 4. The same as Fig. 3, except that all epochs corresponding to each subject have been averaged together. The clustering behaviour between the R and NR groups is clearly evident.

4. Discussion and conclusion

Better methods of determining the best antidepressant medication are desperately needed. The STAR*D trial (Rush et al., 2006; Malone, 2007), as well clinical experience tell us that the current best practice methods of determining optimal treatment are highly inefficient, resulting in delayed treatment response, unnecessary personal suffering, extended disability, higher risk of suicide and avoidable medical expense. In our pilot study we are bringing forward for possible further investigation by other teams, a new biomarker-based method for the prediction of response to an SSRI treatment for MDD.

The proposed method introduces statistical machine learning principles to the psychiatric treatment planning process. The method consists of first, extracting a large set of candidate features, primarily in the form of second-order statistical parameters, from the subject's pre-treatment EEG. Then, only the most discriminative features are selected from this large set using the Fisher discriminant ratio (the FDR). The features selected by this procedure are suboptimal, in that they offer close to the best possible prediction performance. The method is capable of identifying highly predictive features that have been missed in the previous literature. The performance of the proposed method suggests that

we may be able to extract features from a single pretreatment EEG that will predict a patient's subsequent response to a particular antidepressant medication, in this case an SSRI.

Two options are investigated for predicting response from these features. The first is the 1-D case, where individual features on their own are used for prediction. The second is the M-D case, where the features are fed jointly into a classifier. Our results suggest that the performance in the 1-D case is diminished over that of the M-D case.

We must be careful that our machine learning process is actually predicting response, rather than responding to some form of underlying variable. For example, as seen in Table 1, the proportion of females is significantly higher in the responder group, and therefore it is possible that our algorithms are actually predicting gender, rather than response itself. However, we found that the misclassified subjects are shared between males and females, in a way that indicates gender does not play a significant role amongst misclassified cases. Nonetheless, we cannot completely rule out the effect of gender in our study. A larger, gender-balanced data sample is required to obtain higher reliability.

There are also differences in the pre-treatment depression rating scores, with responders having higher HamD-17 scores, i.e., more severe depression, than non-responders. This association is well known, e.g., (Kilts et al., 2009), but in fact of very little predictive value for an individual patient. Furthermore, pre-treatment Beck depression inventory (BDI) scores from this study indicate that pre-treatment subjective depressive severity was no different ($p = 0.842$) in responders and non-responders, a fact which contradicts the conclusion that the responders are more depressed that non-responders at the onset of treatment.

Our proposed method uses a straightforward mathematical process to select a (sub) optimal set of discriminative features from a very large set of candidate features. The features selected by our algorithm have been shown to be highly predictive through the use of a classifier. Previous methods on the other hand hypothesize, either from theoretical insight into the neurophysiology of depression, or from trial-and-error, that a particular feature may be predictive. This hypothesis is then verified by experiment. As such, the scope for the selection of features with these methods is limited. In contrast, the feature selection process used in the proposed method automatically yields features that can render the best (or close to the best) possible prediction performance from a very broad range or scope of candidate features. It is for this reason that the performance of the proposed method exceeds that of previous methods.

The candidate features are extracted from the subject's pretreatment EEG on an atheoretical basis using little *a priori* knowledge concerning their discriminative ability. This procedure is in fact a latent advantage with regard to the proposed method, in that if the set of candidate features consisted only of those which had been identified in previous studies, some of the most discriminating features identified in this study would have been missed. For example, as we discuss in subsequent paragraphs, coherences between the T3 or T5 and the F4 electrodes at various frequencies are identified in this study as being highly predictive. To our knowledge, these quantities have never been identified as such in previous studies. A further encouraging fact is that many of the features selected in the proposed manner are in a similar frequency range and at similar brain sites to those previously identified by other researchers as relevant to depression and its treatment (Duffy et al., 2011; Leuchter et al., 2012; Anand et al., 2009).

When comparing the three EEG referencing methods in our study (LE, AR and IR), our analysis indicates that the AR and IR referencing methods result in very similar selected features, but which are considerably different from those produced by the LE method. For a low-resolution EEG system like ours, this result is in confirmation with the claims in Qin et al. (2010), who report similarity between the AR and IR methods in reflecting the correct EEG source locations, compared to results given by the LE reference. In this study, the IR reference was used exclusively. In practice however, we can build our prediction model based on any of these references, given that the same referencing method is used consistently in both the training and test data sets.

Our findings are consistent with the results of Cook et al. (2002). Cook et al. found that pre-treatment absolute power levels, relative power levels and cordance values in all four EEG bands in the regions implicated in mood disorders (i.e., prefrontal Fp1-Fp2-Fpz, FC1-FC2-Cz, left temporal T3-T5; and right temporal T4-T6) are not significant predictors of response.

It is evident from Table 3 that coherence plays prominently in the list of selected features. Coherence, which is a measure of brain connectivity, has been used in previous studies, but mostly for distinguishing subjects with various mental disorders from normals, rather than for prediction. For example, coherence has been used to discriminate subjects with MDD from normals (Hinrikus et al., 2009; Knott et al., 2001), and in the study of the effect of clozapine therapy in schizophrenia (Knott et al., 2002). Others have shown that connectivity is generally lower across several brain regions in subjects with severe mood disorder (such as Bipolar Disorder and MDD) compared with healthy volunteers (Anand et al., 2009; Leuchter et al., 2012). However, this finding is not universally seen and particularly, Leuchter et al. (2012) found that compared to healthy controls, depressed subjects had higher coherence values over the entire EEG frequency band spanning [0.5–20 Hz].

It is also evident from Table 3 that the selected feature set includes a significant representation of coherences specifically involving the T3, T5 and to a lesser extent, other proximal electrodes, over a range of frequencies. This region has been prominent in previous studies, but these again have focussed on discriminating MDD subjects from normals. For example, coherences involving this region distinguish subjects with MDD from those with other conditions (Duffy et al., 2011; Leuchter et al., 2012). Furthermore, glucose metabolism in the left temporal and left parietal cortex are reported to correlate with depression severity (Milak et al., 2010) and grey matter volume in the parietal cortex in patients with MDD is less than that of healthy controls (Inkster et al., 2011).

With regard to the present study, in Table 3, 19 out of the 27 selected features are coherences between a T3 or T5 electrode and a right-sided lead. Of these, 15 out of the 27 are coherences specifically between F4 and either T3 or T5 at frequencies in the 12–20 Hz range (low β-band). Thus it appears that these coherences in this region are strongly predictive of response. As may be seen, these coherence values are all lower for responders than non-responders. It has been further shown that SSRI medications may increase connectivity in temporal, parietal and frontal regions (Anand et al., 2007; Diaconescu et al., 2011). Thus SSRI antidepressants may work well only in those subjects whose connectivity in these regions is abnormally suppressed. Pre-treatment connectivity levels in non-responders on the other hand are already at more normal levels, thus rendering SSRI treatment ineffective in these subjects.

To compare with the findings of Lee et al. (2011), who studied eyes-closed data, we also looked at a list of discriminating features when only EC data are used for prediction, and found that *ftb*(F8,T6,23), *ftb*(F8,T6,25) as well as *ftb*(F8,T6,12) are predictive of response with an FDR value of approximately 0.9. However, these features are not among the first 30 most predictive features in this case.

Other QEEG based techniques, such as the referenced EEG methods or ATR have also been shown to be useful in predicting response to antidepressant medications. The referenced EEG method (Suffin et al., 2007; DeBattista et al., 2011) operates on the assumption that psychiatric diagnosis is not relevant. This

approach may be difficult to reconcile with current established practice as it may lead to recommendations for administration of drugs not currently approved for use with the psychiatric condition being treated (Iosifescu, 2011).

The ATR method (Iosifescu et al., 2009; Leuchter et al., 2009a, b) which delivers predictive accuracy of about 74%, suffers from the disadvantage of requiring two EEG analyses, one prior to treatment and another after 1 week of daily drug administration. In contrast, the method we propose delivers response prediction with over 80% accuracy after only a single pretreatment EEG. Not only is our method more efficient, but it also eliminates the need to commit to at least 1 week of treatment with a potentially ineffective medication.

In a manner similar to Tenke et al. (2011), we also investigated an alternative approach in which the *difference* between the averaged EO and EC selected EEG features are used as predictors of response. The average of each selected feature value over the (nominally three) EO epochs was subtracted from the corresponding average of the EC epochs for each subject, to give one point per subject. These differences were used as input to the prediction analysis. The resulting level of performance is very similar to the previous EO, EC combined case. It was found that, in agreement with Tenke et al. (2011), the posterior power spectral density at electrode P4 at 11 Hz does have discriminative power in the EO minus EC case, corresponding to an FDR = 0.7. This feature appears as the 20th most relevant feature in the respective list of selected features for this case. For comparison, the top-ranking feature in this case was *coh*(P3,O2,24) with an FDR = 2.3. Note also that this posterior alpha PSD feature was not among the selected features in the combined EC and EO case in Table 3, and therefore on the basis of the FDR as a feature selection criterion, features derived from posterior alpha PSD are not as predictive as those selected in the present study.

The data analysis procedure and methodology described in this paper are currently being extended to construct models that predict the response to various other treatments available for patients with MDD or other psychiatric disorders. Furthermore, it may be possible also to incorporate information from other sources (such as symptom rating scales, scores from personality inventories and other psychiatric evaluations, the levels of various hormones etc. in the blood, demographic and socioeconomic information, medical imaging data, etc.) as candidate features.

Our pilot data results are encouraging and suggest that the objective of personalizing the choice of antidepressant medication may be within the realm of possibility. The methodology we propose would require only the relatively inexpensive and accessible EEG. Furthermore, our observation that EEG data collected 2 weeks after drug has been administered would suggest that the patient need not be entirely off psychotropic medication for the test to still be useful. In a study using ML analysis of pretreatment EEG to predict response to clozapine we also found that concurrent treatment with psychotropic medications did not appear to interfere with predictive accuracy (Khodayari-Rostamabad et al., 2010a). Nonetheless, caution is warranted as our findings are based on a small sample of treatment-resistant depressed patients. Our findings are currently being replicated in larger studies to assess the potential clinical usefulness of this methodology.

Acknowledgements

The authors wish to acknowledge the contribution of the reviewers, whose comments have resulted in a much improved version of the paper. The authors also gratefully acknowledge the support of Magstim Company Ltd., Carmarthenshire, Wales, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Psychiatric Research Foundation, the Ontario Mental Health Foundation and the Stanley Foundation. This work is partly supported by an Etherden Fellowship at McMaster University and St Joseph's Healthcare Foundation, Hamilton, ON. The authors would also like to thank Cathy Ivanski, Rose Marie Mueller, Jackie Heaslip, Sandra Chalmers and Joy Fournier for their help with the clinical experiments.

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR, 4th ed. USA: American Psychiatric Publishing; 2000.

Alhaj H, Wisniewski G, McAllister-Williams RH. The use of the EEG in measuring therapeutic drug action: focus on depression and antidepressants. J Psychopharmacol 2011;25:1175–91.

Anand A, Li Y, Wang Y, Gardner K, Lowe MJ. Reciprocal effects of antidepressant treatment on activity and connectivity of the mood regulating circuit: an fMRI study. J Neuropsychiatry Clin Neurosci 2007;19:274–82.

Anand A, Li Y, Wang Y, Lowe MJ, Dzemidzic M. Resting state corticolimbic connectivity abnormalities in unmedicated bipolar disorder and unipolar depression. Psychiatry Res: Neuroimaging 2009;171:189–98.

Bandelow B, Baldwin DS, Dolberg OT, Andersen HF, Stein DJ. What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder? J Clin Psychiatry 2006;67:1428–34.

Bares M, Brunovsky M, Kopecek M, Stopkova P, Novak T, Kozeny J, et al. Changes in QEEG prefrontal cordance as a predictor of response to antidepressants in patients with treatment resistant depressive disorder: a pilot study. J Psychiatric Res 2007;41:319–25.

Bares M, Brunovsky M, Kopecek M, Novak T, Stopkova P, Kozeny J, et al. Early reduction in prefrontal theta QEEG cordance value predicts response to venlafaxine treatment in patients with resistant depressive disorder. Eur Psychiatry 2008;23:350–5.

Beck AT, Ward CH, Mendelson M, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.

Bruder GE, Stewart JW, Tenke CE, McGrath PJ, Leite P, Bhattacharya N, et al. Electroencephalographic and perceptual asymmetry differences between responders and nonresponders to an SSRI antidepressant. Biol Psychiatry 2001;49:416–25.

Bruder GE, Sedoruk JP, Stewart JW, McGrath PJ, Quitkin FM, Tenke CE. Electroencephalographic alpha measures predict therapeutic response to selective serotonin reuptake inhibitor antidepressant: pre- and post-treatment findings. Biol Psychiatry 2008;63:171–1177.

Burrows GD, Norman TR, Judd FK. Definition and differential diagnosis of treatment-resistant depression. Int Clin Psychopharmacol 1994;9(Suppl. 2): 5–10.

Cao C, Tutwiler RL, Slobounov S. Automatic classification of athletes with residual functional deficits following concussion by means of EEG signal using support vector machine. IEEE Trans Neural Syst Rehabil Eng 2008;16:327–35.

Cook IA, Leuchter AF, Morgan M, Witte E, Stubbeman WF, Abrams M, et al. Early changes in prefrontal activity characterize clinical responders to antidepressants. Neuropsychopharmacology 2002;27:120–31.

Crisler S, Morrissey MJ, Anch AM, Barnett DW. Sleep-stage scoring in the rat using a support vector machine. J Neurosci Methods 2008;168:524–34.

DeBattista C, Kinrys G, Hoffman D, Goldstein C, Zajecka J, Kocsis J, et al. The use of referenced-EEG (rEEG) in assisting medication selection for the treatment of depression. J Psychiatric Res 2011;45:64–75.

Dewa CS, Lesage A, Goering P, Craveen M. Nature and prevalence of mental illness in the workplace. Healthcare Papers 2004;5:12–25.

Diaconescu AO, Kramer E, Hermann C, Ma Y, Dhawan V, Chaly T, et al. Distinct functional networks associated with improvement of affective symptoms and cognitive function during citalopram treatment in geriatric depression. Hum Brain Mapping 2011;32:1677–91.

Druss BG, Rosenheck RA, Sledge WH. Health and disability costs of depressive illness in a major U.S. corporation. Am J Psychiatry 2000;157:1274–8.

Duffy FH, McAnulty GB, McCreary MC, Cuchural GJ, Komaroff AL. EEG spectral coherence data distinguish chronic fatigue syndrome patients from healthy controls and depressed patients – a case control study. BMC Neurol 2011;11:82.

Ghahramani Z, Hinton GE. The EM algorithm for mixtures of factor analyzers. Department of Computer Science Technical Report, CRG-TR-96-1. Toronto, Canada: University of Toronto; 1996.

Ghosh-Dastidar S, Adeli H, Dadmehr N. Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection. IEEE Trans Biomed Eng 2008;55:512–8.

Güler I, Übeyli ED. Multiclass support vector machines for EEG-signals classification. IEEE Trans Inf Technol Biomed 2007;11:117–26.

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

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. USA: Springer; 2009.

Henkel V, Seemüller F, Obermeier M, Adli M, Bauer M, Mundt C, et al. Does early improvement triggered by antidepressants predict response/remission? – Analysis of data from a naturalistic study on a large sample of inpatients with major depression. J Affect Disord 2009;115:439–49.

Hinrikus H, Suhhova A, Bachmann M, Aadamsoo K, Võhma Ü, Lass J, et al. Electroencephalographic spectral asymmetry index for detection of depression. Med Biol Eng Comput 2009;47:1291–9.
Hunter AM, Cook IA, Leuchter AF. The promise of the quantitative electroencephalogram as a predictor of antidepressant treatment outcomes in major depressive disorder. Psychiatric Clin North Am 2007;30:105–24.
Inkster B, Rao AW, Ridler K, Nichols TE, Saemann PG, Auer DP, et al. Structural brain changes in patients with recurrent major depressive disorder presenting with anxiety symptoms. J Neuroimaging 2011;21:375–82.
Iosifescu DV, Greenwald S, Devlin P, Mischoulon D, Denninger JW, Alperta JE, et al. Frontal EEG predictors of treatment outcome in major depressive disorder. Eur Neuropsychopharmacology 2009;19:772–7.
Iosifescu DV. Electroencephalography-derived biomarkers of antidepressant response. Harvard Rev Psychiatry 2011;19:144–54.
Jäger M, Sobocki P, Rössler W. Cost of disorders of the brain in Switzerland – with a focus on mental disorders. Swiss Medical Weekly 2008;138:4–11.
Kemp AH, Gordon E, Rush AJ, Williams LM. Improving the prediction of treatment response in depression: integration of clinical, cognitive, psychophysiological, neuroimaging, and genetic measures. CNS Spectr 2008;13:1066–86.
Khodayari-Rostamabad A, Hasey GM, MacCrimmon DJ, Reilly JP, DeBruin H. A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. Clin Neurophysiol 2010a;121:1998–2006.
Khodayari-Rostamabad A, Reilly JP, Hasey GM, DeBruin H, MacCrimmon DJ. Using pre-treatment EEG data to predict response to SSRI treatment for MDD. Conf Proc IEEE Eng Med Biol Soc 2010b;2010:6103–6.
Khodayari-Rostamabad A, Reilly JP, Hasey GM, DeBruin H, MacCrimmon DJ. Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model. Conf Proc IEEE Eng Med Biol Soc 2010c;2010:4006–9.
Kilts CD, Wade AG, Andersen HF, Schlaepfer TE. Baseline severity of depression predicts antidepressant drug response relative to escitalopram. Expert Opin Pharmacother 2009;10:927–36.
Knott V, Mahoney C, Kennedy S, Evans K. EEG power, frequency, asymmetry and coherence in male depression. Psychiatry Res: Neuroimaging Sect 2001;106: 123–40.
Knott VJ, LaBelle A, Jones B, Mahoney C. EEG coherence following acute and chronic clozapine in treatment-resistant schizophrenics. Exp Clin Psychopharmacol 2002;10:435–44.
Kohavi R. A study of cross validation and bootstrap for accuracy estimation and model selection. Proc Int Joint Conf Artif Intell 1995:1137–43.
Korb AS, Hunter AM, Cook IA, Leuchter AF. Rostral anterior cingulate cortex theta current density and response to antidepressants and placebo in major depression. Clin Neurophysiol 2009;120:1313–9.
Kwon JS, Youn T, Jung HY. Right hemisphere abnormalities in major depression: quantitative electroencephalographic findings before and after treatment. J Affect Disord 1996;40:169–73.
Lee TW, Wu YT, Yu YW, Chen MC, Chen TJ. The implication of functional connectivity strength in predicting treatment response of major depressive disorder: a resting EEG study. Psychiatry Res 2011;194:372–7.
Leuchter AF, Cook IA, Marangell LB, Gilmer WS, Burgoyne KS, Howland RH, et al. 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 2009a;169:124–31.
Leuchter AF, Cook IA, Gilmer WS, Marangell LB, Burgoyne KS, Howland RH, et al. Effectiveness of a quantitative electroencephalographic biomarker for predicting differential response or remission with escitalopram and bupropion in major depressive disorder. Psychiatry Res 2009b;169:132–8.
Leuchter AF, Cook IA, Hunter AM, Cai C, Horvath S. Resting-state quantitative electroencephalography reveals increased neurophysiologic connectivity in depression. PLoS One 2012;7:e32508.
Lopez AD, Mathers CD. Measuring the global burden of disease and epidemiological transitions: 2002–2030. Ann Trop Med Parasitol 2006;100:481–99.
Malone DC. A budget-impact and cost-effectiveness model for second-line treatment of major depression. J Managed Care Pharm 2007;13:S8–S18.
Milak MS, Keilp J, Parsey RV, Oquendo MA, Malone KM, Mann JJ. Regional brain metabolic correlates of self-reported depression severity contrasted with clinician ratings. J Affect Disord 2010;126:113–24.
Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21:3301–7.
Moore SK. The psychiatrist in the machine. IEEE Spectr 2011;48:11–2.
Mourad N, Reilly JP, Hasey G, MacCrimmon D. Temporally constrained SCA with applications to EEG data. IEEE ICASSP 2010;2010:4138–41.
Mulert C, Juckel G, Brunnmeier M, Karch S, Leicht G, Mergl R, et al. Prediction of treatment response in major depression: integration of concepts. J Affect Disord 2007a;98:215–25.
Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B. An introduction to kernel-based learning algorithms. IEEE Trans Neural Networks 2001;12:181–201.
Na SH, Jin S-H, Kim SY, Ham BJ. EEG in schizophrenic patients: mutual information analysis. Clin Neurophysiol 2002;113:1954–60.
Nunez PL. REST: a good idea but not the gold standard. Clin Neurophysiol 2010;121:2177–80.
Pascual-Marqui RD, Lehmann D, Koenig T, Kochi K, Merlo MC, Hell D, et al. Low resolution brain electromagnetic tomography (LORETA) functional imaging in acute, neuroleptic-naive, first episode, productive schizophrenia. Psychiatry Res 1999;90:169–79.
Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226–38.
Proakis JG, Rader CM, Ling FY, Nikias CL. Advanced digital signal processing. New York: MacMillan; 1992.
Qin Y, Xu P, Yao D. A comparative study of different references for EEG default mode network: the use of the infinity reference. Clin Neurophysiol 2010;121: 1981–91.
Ravan M, Reilly JP, Trainor LJ, Khodayari-Rostamabad A. A machine learning approach for distinguishing age of infants using auditory evoked potentials. Clin Neurophysiol 2011;122:2139–50.
Revicki DA, Simon GE, Chan K, Katon W, Heiligenstein J. Depression, health-related quality of life, and medical cost outcomes of receiving recommended levels of antidepressant treatment. J Fam Pract 1998;47:446–52.
Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry 2006;163:1905–17.
Rush AJ, Warden D, Wisniewski SR, Fava M, Trivedi MH, Gaynes BN, et al. STAR*D: revising conventional wisdom. CNS Drugs 2009;23:627–47.
Serretti A, Olgiati P, Liebman MN, Hu H, Zhang Y, Zanardi R, et al. Clinical prediction of antidepressant response in mood disorders: linear multivariate vs. neural network models. Psychiatry Res 2007;152:223–31.
Simon GE, Khandker RK, Ichikawa L, Operskalski BH. Recovery from depression predicts lower health services costs. J Clin Psychiatry 2006;67:1226–31.
Spronk D, Arns M, Barnett KJ, Cooper NJ, Gordon E. An investigation of EEG, genetic and cognitive markers of treatment response to antidepressant medication in patients with major depressive disorder: a pilot study. J Affect Disord 2011;128: 41–8.
Suffin SC, Emory WH, Gutierrez G, Arora G, Schiller MJ, Kling A. A QEEG database method for predicting pharmacotherapeutic outcome in refractory major depressive disorders. J Am Phys Surg 2007;12:104–8.
Tenke CE, Kayser J, Manna CG, Fekri S, Kroppmann CJ, Schaller JD, et al. Current source density measures of electroencephalographic alpha predict antidepressant treatment response. Biol Psychiatry 2011;70:388–94.
Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry 2006;163:28–40.
Ulrich G, Haug HJ, Fähndrich E. Acute vs. chronic EEG effects in maprotiline- and clomipramine-treated depressive patients in the prediction of therapeutic outcome. J Affect Disord 1994;32:213–7.
Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinf 2006;7:91.