Clinical Neurophysiology 121 (2010) 1998–2006

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy

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

[a] Electrical and Computer Eng. Dept., McMaster University, Hamilton, ON, Canada L8S 4K1
[b] Dept. of Psychiatry and Behavioral Neurosciences, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada L8S 4L8
[c] Mood Disorders Program, Centre for Mountain Health Services, St. Joseph Hospital, Hamilton, ON, Canada L8N 3K7
[d] School of Biomedical Engineering, McMaster University, Hamilton, ON, Canada L8S 4K1

ARTICLE INFO

Article history:
Accepted 11 May 2010
Available online 17 June 2010

Keyword:
Schizophrenia
Clozapine
EEG
Treatment-efficacy prediction
Machine learning
Psychiatry

ABSTRACT

Objective: To investigate whether applying advanced machine learning (ML) methodologies to pre-treatment electroencephalography (EEG) data can predict the response to clozapine therapy in adult subjects suffering from chronic schizophrenia.
Methods: Pre-treatment EEG data are collected in 23 + 14 schizophrenic adults. Treatment outcome, after at least one year follow-up, is determined using clinical ratings by a trained clinician blind to EEG results. First, a feature selection scheme is employed to select a reduced subset of features extracted from the subjects' EEG that is most statistically relevant to our treatment-response prediction. These features are then entered into a classifier, which is realized in the form of a kernel partial least squares regression method that performs response prediction. Various scales, including the positive and negative syndrome scale (PANSS) are used as treatment-response indicators.
Results: We determined that a set of discriminating EEG features do exist. A low-dimensional representation of the feature space showed significant clustering into clozapine responder and non-responder groups. The minimum level of performance of the proposed prediction methodology, tested over a range of conditions using the leave-one-out cross-validation method using the original 23 subjects, with further testing in an independent sample of 14 subjects, was 85%.
Conclusions: These findings indicate that analysis of pre-treatment EEG data can predict the clinical response to clozapine in treatment resistant schizophrenia.
Significance: If replicated in a larger population, this novel approach to EEG analysis may assist the clinician in determining treatment-efficacy.

1. Introduction

Compared with other antipsychotic medications the atypical antipsychotic medication clozapine is recognized to have superior therapeutic effectiveness in the treatment of chronic medication-resistant schizophrenia (e.g., Essali et al., 2009). However, clozapine may produce serious side effects such as seizures, cardiac arrhythmias or bone marrow suppression with neutropenia (Young et al., 1998). According to a recent Cochrane review, about 34% of treatment-resistant patients respond to clozapine while 3.2% develop blood problems (Essali et al., 2009). As the hematological side effects can be life threatening, blood samples to monitor the white blood cell count must be collected as long as the drug is used, at weekly to monthly intervals. The logistic difficulties for the patient and the treatment team are substantial. A method that could reliably determine, before the onset of therapy, whether a given patient will or will not respond to clozapine would greatly assist the clinician in determining whether the risks and logistic complexity of clozapine are outweighed by the potential benefits.

Quantitative electroencephalography (QEEG or EEG) may offer some promise in this regard. EEG abnormalities in schizophrenic subjects and EEG changes due to clozapine therapy have been the focus of a number of clinical studies (see e.g., Gunther et al., 1993; Malow et al., 1994; Freudenreich et al., 1997; Hughes and

* Corresponding author. Tel.: +1 905 525 9140x22895; fax: +1 905 521 2922.
E-mail address: reillyj@mcmaster.ca (J.P. Reilly).

1388-2457/$36.00
doi:10.1016/j.clinph.2010.05.009

John, 1999; Knott et al., 2001; Adler et al., 2002; Knott et al., 2002; Birca et al., 2006; Coburn et al., 2006; Dunki and Dressel, 2006; Oikonomou et al., 2006; Sakkalis et al., 2006; Boutros et al., 2008).

Based on findings in 17 schizophrenic subjects, Knott et al. (2000) found that the clozapine-induced improvement of psychopathology symptom ratings using the Positive and Negative Syndrome Scale (PANSS) was correlated with pre-treatment QEEG inter and intra-hemispheric spectral power asymmetry. Greater pre-treatment anterior to posterior asymmetry in the delta frequency range was associated with greater improvement in negative symptoms while greater pre-treatment anterior to posterior theta asymmetry predicted improvement of positive symptoms and global improvement. Larger inter-hemispheric asymmetry in the theta and beta frequencies in the central and anterior temporal regions were, respectively, predictive of greater improvement in positive and negative symptoms. Gross et al. (2004) also found that changes in the theta frequency in QEEG with clozapine treatment, particularly in the midline electrodes over the fronto-central scalp area, were a more sensitive indicator for the evaluation of clozapine treatment efficacy than the serum clozapine level. Though these methods reveal important relationships between QEEG variables and clinical outcome, a series of simple correlational analyses do not readily yield a "responder" or "non-responder" dichotomous categorization for an individual patient.

The above analyses employed standard statistical methods. On the other hand, a more mathematically sophisticated analysis including pattern recognition and dimensionality reduction methods (which together may be categorized as *machine learning* techniques) can perform a more comprehensive data analysis. Machine learning techniques are finding increasing application in psychiatry, particularly when multi-dimensional, noisy, highly complex data or multi-modal data sets are analyzed together, (see e.g., Gallinat and Heinz, 2006). For example, support vector machine (SVM) techniques that select spectro-temporal patterns from multichannel magnetoencephalogram (MEG) data collected during a verbal working memory task have been used to distinguish schizophrenic from control subjects (Ince et al., 2008). Machine learning algorithms using structural brain magnetic resonance (MRI) images (Fan et al., 2007), functional MRI (fMRI) data (Guo et al., 2008; Kim et al., 2008) and combined genomic and clinical data (Struyf et al., 2008) have been employed to separate schizophrenic, bipolar and healthy control subjects.

Machine learning approaches have also been applied to prediction of clozapine treatment-efficacy. Lin et al. (2008) describes a study in which a feed-forward multilayer perceptron network (with a back-propagation error training technique) is employed using clinical and pharmacogenetic data to predict clozapine response in schizophrenic subjects. Five pharmacogenetic variables and five clinical variables (including gender, age, height, baseline body weight, and baseline body mass index) were collated from 93 schizophrenic subjects taking clozapine, including 26 responders. Using this method, they obtained an overall prediction accuracy rate of 83.3%.

Guo et al. (2008) describes a Bayesian hierarchical model using pre-treatment fMRI and positron emission tomography (PET) information coupled with patient characteristics (e.g. medical or family history and genotype) as training data to predict changes in brain activity in 16 schizophrenic subjects following treatment with two atypical antipsychotics (risperidone or olanzapine). The authors postulated that predicting drug-induced changes in brain activity would assist the clinician in determining optimal drug choice.

However, the clinical utility of these previous approaches is negatively impacted by the expense and unavailability of complex methods such as fMRI, PET, genetic screening and MEG. In contrast, electroencephalography (EEG) is an inexpensive, non-invasive technique widely available in smaller hospitals and in community laboratories. Therefore, predictive algorithms dependent on EEG measurements are more practical. Furthermore, since the required EEG data is acquired during the resting state, only minimal cooperation is required from the patient. Thus, an EEG based method of predicting treatment response would have many advantages over imaging methods such as MRI, PET or MEG.

The goal of the present pilot study is to examine the utility of machine learning (ML) methods for processing EEG signals to predict the response of schizophrenic subjects to clozapine.

2. Methods

2.1. Quantitative EEG recordings

We collected pre-clozapine resting EEG data from chronically ill, treatment-resistant schizophrenic subjects prior to beginning clozapine therapy. The data were collected without change to the patient's current medication regimen. EEG was recorded with the patient in a semi-recumbent position in a sound attenuated, electrically shielded room 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. A maximum of ten and a half minutes of eyes-closed (EC) and of eyes-open (EO) data respectively were collected in up to three separate 3.5 min runs using a QSI-9500 system, giving a total of 3 EO and 3 EC files. Electrodes were placed in the 10/20 configuration referenced to linked ears with impedances below 5 kΩ. The signals were band pass filtered between [0.5 and 80 Hz] and notch filtered at 60 Hz by the QSI system during the recording. Data were digitized at a rate of 204.8 Hz. Since our selected features were either intra- or inter-hemispherical, we discarded the data from the midline electrodes (FZ, CZ, PZ, and OZ) in the interests of saving computational resources. The 16 remaining EEG electrodes used in our study were Fp1, Fp2, F3, F4, F7, F8, T3, T4, C3, C4, T5, T6, P3, P4, O1 and O2.

For de-artifacting, the data were partitioned into segments of 1 s duration. If the input signal on any electrode saturated the acquisition hardware at approximately plus or minus 160 μv, the entire segment was rejected. The signals were then digitally bandpass filtered after recording between 4 and 42 Hz to partially mitigate the effects of eye movement and muscle artifacts. For each EEG file, the first 60 segments of the de-artifacted part of the 3.5 min of data were used, since several segments were heavily artifacted, leaving only this number of segments that were uncorrupted on all electrodes. The selected data in each of the three files for both the EO and EC cases were divided into 2 epochs of 40 s duration with 50% overlap, to give a nominal 12 epochs per subject. These epochs were used to extract statistical quantities (such as absolute powers, power spectral densities, coherences, etc.) that became the candidate features as described below. When estimating these statistical quantities, each epoch was divided into overlapping 1 s windows with 60% overlap between adjacent windows. The respective statistical quantity was then calculated over each window and the desired result obtained by averaging over all windows. In the experimental results which follow, all EO and EC epochs were combined, to make maximum use of the available data.

2.2. Description of subjects and the clinical assessment procedures

Subjects, comprising both in-patients and out-patients, were recruited from the schizophrenia program at St. Joseph's Hospital, Centre for Mountain Health Services, Hamilton, Ontario. All subjects met both DSM-IV criteria for schizophrenia and the Kane

Table 1
Demographic information of the 23 subjects (denoted Group A) who participated in the study. The lower 4 items in the table are scales related to the PANSS clinical rating score.

Information	Range
Age at start of treatment [years]	Average = 41.2, std = 8.4, min = 28.8, max = 57
Gender:	
Male	12 (52%)
Female	11 (48%)
Educational Level[1]	Average = 3.1, std = 1.4, min = 2, max = 7
Age at symptom onset [years]	Average = 21.2, std = 5, min = 14, max = 32
Total # of hospitalizations (Pre-clozapine)	Average = 9.7, std = 13, min = 0, max = 63
Duration total of hospitalization (Pre-clozapine) [days]	Average = 615.7, std = 928, min = 0, max = 3789
Chlorpromazine equivalents (Pre-clozapine) [mg/day]	Average = 726.6, std = 636, min = 40, max = 2485
Clozapine dose [mg/day]	Average = 344.6, std = 157, min = 50, max = 600
Post-treatment Positive Symptoms Scale	Average = 17.8, std = 3.4, min = 11, max = 24
Post-treatment Negative Symptoms Scale	Average = 23, std = 3.9, min = 12, max = 32
Post-treatment General Symptoms Rank (GR)	Average = 46.3, std = 5.7, min = 32, max = 56
Post-treatment Total Rank (PSS + NSS + GR)	Average = 87.2, std = 10.9, min = 58, max = 101

[1] Education level rating: 1: grade 6 or less, 2: grade 7 to 12 without graduating, 3: graduated high school, 4: part college, 5: graduate 2 years college, 6: graduate 4 years college, 7: part graduated/professional school, 8: completed graduated professional school.

et al. (1988) criteria for treatment resistance. Patients meeting these criteria may be considered to be "severely symptomatic", i.e., as suffering acutely from schizophrenia. All subjects gave informed consent.

Data from two groups of schizophrenic subjects were used in this retrospective study. The first group (Group A) consists of 23 subjects. Group B is an independent sample of 14 subjects. Available socio-demographic and clinical information for Groups A and B are shown in Tables 1 and 2. Symptom severity after clozapine treatment is measured in Group A using the positive and negative syndrome scale (PANSS) score (Kay et al., 1987). PANSS evaluations are not available for Group B subjects. As PANSS scores were not available for Group A subjects prior to clozapine treatment, pre-treatment symptom severity was assessed through a quantitative clinical assessment (QCA) conducted by review of the clinical record guided by the structure of the PANSS. The QCA procedure is outlined in Appendix A. As all QCA ratings were completed before initiation of this study raters were blind to the machine learning outcome predictions. QCA was used to assess psychopathology both pre and post clozapine treatment in Group B.

We now discuss how we determine whether a patient is a responder (R) or non-responder (NR). In this retrospective pilot study quantifying clinical response is complicated by the absence of pre-treatment PANSS scores. We were therefore obliged to define response on the basis of a single post-treatment PANSS score. To do this we created post-treatment PANSS score[1] thresholds δ_1 to assess response: first we rank-ordered all subjects by post-treatment PANSS score then chose a value of δ_1 (88.5) such that our 23 subjects were divided into responder (R) and non-responder (NR) classes with roughly equal number of subjects (R = 12, NR = 11).

Having R and NR groups of similar size has advantages with respect to the machine learning process; however, this assumes that clinically significant improvement is seen in about 50% of those treated with clozapine. Others have reported that, on average, only 34% of treatment-resistant schizophrenic patients will respond to clozapine. For this reason we also reanalyzed our data using a value $\delta_1 = 83.5$ which yields a 30% response rate (i.e. with 7 R and 16 NR subjects in group A).

We must confirm that the pre-treatment QCA means of the R and NR subgroups of group A subjects are not significantly different, so that the post-treatment PANSS rating alone accurately indicates the effect of the treatment on the subject. To this end, we conducted a hypothesis test on the means, assuming the QCA data points are independent and normally distributed, and that the variances of the R and NR groups are identical. It is straightforward to show that the respective likelihood ratio is F-distributed. In this case, df = 10, 11 for the numerator and denominator, respectively, with F = 1.1056 and p = 0.43. Thus, there is no evidence to suggest the pre-treatment QCA means of the two groups are significantly different.

Group B subjects are defined as *responders* to clozapine therapy if there is an improvement of at least 25% between the pre- and post-QCA scores. This level of relative change represents a clinically significant improvement in symptom severity considering the fact that all the subjects in our study were in the treatment-resistant population (Leucht et al., 2005). See e.g., Kane et al. (1988) who used a 20% relative change as response indicator.

2.3. Overview of the machine learning process

We now present a brief overview of the machine learning process used for prediction of clozapine response. A necessary component of this process is the collection of a training set. In our case, the training set consists of M_p EEG epochs from each of M subjects, for a total of M_t epochs altogether. In our experiments[2], $M = 23$, $M_p = 12$ and $M_t = 270$. The training set also includes the set of response outcomes y_i ,$i = 1,\ldots,M_t$ corresponding to each epoch; i.e., if the subject corresponding to the ith EEG epoch is a responder (non-responder), then the value of y_i is R (NR), determined by the response criterion discussed previously.

There are three phases in a machine learning procedure. These are the *design*, *operational* and *evaluation* phases, as outlined in Fig. 1. The design phase, which consists of the feature extraction, feature selection and classification components, is now described.

The Design Process is depicted in Fig. 1(a). The first step is to extract candidate *features* from each epoch of pre-treatment EEG data. In our study, these features are statistical quantities including coherence[3] between all electrode pairs at various frequencies, correlation and cross-correlation coefficients, mutual information between all sensor pairs (Cover and Thomas, 2006), absolute and relative power levels at various frequencies[4], the left-to-right hemi-

[1] Using the PANSS data, the 'total rank' (TR) score is used as the clinical assessment in our experiments. TR is the sum of three scales in PANSS: 1. general rank, (GR), 2. positive (or productive) symptoms scale, (PSS), 3. negative (or deficit) symptoms scale, (NSS). This means that TR = GR + PSS + NSS.

[2] The total number of epochs is nominally 12 x 23 = 276. However, there are only 8 and 10 available epochs for 2 of the subjects, leaving only 270 net epochs.

[3] We calculated the magnitude squared coherence estimate using the averaged periodogram method of Welch by the MathWorks MATLAB software, ver. 7.1. See www.mathworks.com.

[4] Using power spectral density (PSD) estimate via Welchs' averaged modified periodogram method in MATLAB.

Table 2
Available demographic information of the 14 subjects denoted by Group B.

Information	Range
Age at start of treatment [years]	Average = 35.7, std = 10, min = 22, max = 55.5
Gender:	
Male	8 (57%)
Female	6 (43%)
Educational level[1]	Average = 3.3, std = 1.64, min = 2, max = 7
Age at symptom onset [years]	Average = 21.3, std = 5.28, min = 15, max = 31
Total # of hospitalizations (Pre-clozapine)	Average = 6.43, std = 6.9, min = 0, max = 18
Duration total of hospitalization (Pre-clozapine) [days]	Average = 470.8, std = 627, min = 0, max = 1879
Chlorpromazine equivalents (Pre-clozapine) [mg/day]	Average = 628, std = 404, min = 40, max = 1169
Clozapine dose [mg/day]	Average = 396.4, std = 101, min = 200, max = 500

[1] See Table 1 for definition.

sphere power ratio[5], the anterior/posterior power gradient across many frequencies and between electrodes (calculated using logarithm difference of power spectral density values). These quantities can all be readily calculated from the measured EEG signal. The number N_c of such candidate features can be quite large. In our experiments, using 1 Hz frequency resolution and considering all possible electrode pairs, in addition to various electrode combinations used in the power ratio group of features, we have $N_c = 8468$. The feature extraction process is applied over all epochs from all subjects. The result of the feature extraction process is a set of M_t vectors $\tilde{\mathbf{x}}_i,\ i = 1, \ldots, M_t$, each of dimension N_c.

Notice that the majority of these candidate features are statistical characterizations of the measured EEG process and as such at least partially describe the underlying statistical behaviour of the EEG signal. Many of these quantities have been used as features in previous related work; e.g., mutual information was used by Kwak and Choi (2002), and coherences were used by Knott et al. (2002).

After extracting candidate features, the second step in the design phase is *feature reduction,* or 'feature selection' which is critical to the performance of the resulting classifier or predictor. Feature selection is an ongoing topic of research in the machine learning community. Typically, only a relatively small number of the above candidate features bear any significant statistical relationship with the post-treatment response. We therefore identify those features which share the strongest statistical dependencies with the post-treatment-response variable. The result of the feature selection process is to reduce the number N_c of candidate features to a much smaller number N_r of most-relevant features. Our proposed prediction procedure uses the "regularized feature selection" method[6] of Peng et al. (2005). This procedure proceeds in a sequence of N_r steps, where one feature is selected in each step. At each step, the feature which is selected from the list of (remaining) candidate features is the one which has the best combination of maximum statistical dependence with the treatment-response variable, and minimal statistical dependence with respect to the set of features already chosen in previous steps. In Peng's method, statistical dependence is quantified using mutual information. Further details are provided in the reference. The output of the feature selection process is a set of indices that identify which of the N_c candidate features are to be included in the set of N_r most relevant features. In this study, the useful range for N_r is between 8 and 14.

The feature selection process yields a set of reduced N_r dimensional vectors, $\mathbf{x}_i,\ i = 1, \ldots, M_t$. Each of these vectors correspond to a point in an N_r - dimensional feature space. Ideally, these points should cluster into two distinct non-overlapping regions in the feature space, corresponding to the R and NR groups, respectively. A feature vector $\mathbf{x}_i$ maps into the R region if the subject corresponding to the i-th epoch is a responder, and into the NR region otherwise. In practice however, the clusters overlap somewhat, so that feature vectors from a few epochs of the R subjects map into the NR region, and vice versa. As we demonstrate in Section 3, these miss-located points result in a prediction error for that subject. An example of such clustering behaviour (shown in only two dimensions) for the current prediction problem is shown in Fig. 2, where it is seen that the feature vectors corresponding to the R and NR subjects indeed lie in distinct (although slightly overlapping) regions of the feature space. The selection of "better" features; i.e., features with greater statistical dependence on the outcome variable, leads to the formation of tighter clusters with smaller variances and with greater separation between the means of the clusters of different classes, resulting in improved performance.

We normalized feature values to improve performance. Certain feature values, such as coherence and correlation, are inherently limited to an interval $[-1, 1]$ and so normalization is not required in these cases. However, for other feature values, such as e.g., spectral power levels, etc., normalization is desirable. In this study the "z-score" normalization method was used. The EEG data of 91 normal (or healthy) adult subjects were measured and the means μ_l and standard deviations σ_l, $l = 1,\ldots, N_c$ for each feature are calculated over the healthy subject sample. Then for schizophrenic subjects, the corresponding l-th feature value x_l is replaced with its normalized z-score value $z_l = \frac{x_l - \mu_l}{\sigma_l}$ before being fed to the feature selection and classifier processes.

Because many of the candidate features are highly correlated, there are many possible subsets of features that may be selected by our proposed feature selection algorithm, resulting in approximately equivalent prediction performance. The set of selected features is dependent on the normalization method used, the feature selection process, the response criterion and the definition of the target values y in the training data.

The next step in the design phase of the prediction process is the specification of the classifier. The job of the classifier is to input a reduced feature vector $\mathbf{x}$ and output the corresponding predicted response value y, which has a discrete value corresponding to either R or NR. In this way, the classifier output gives us the predicted response of the subject to the clozapine therapy.In this study, the classification process was implemented using a kernelized *partial least squares regression* (KPLSR) procedure (Rosipal and Kramer, 2006). The kernel matrix required by the KPLSR method was chosen to have a Gaussian structure. The KPLSR method determines a regression function using the available training data that approximates the value 1 over the region of the feature space corresponding to non-responders (i.e., the non-responder cluster), and the value 2 over the responder cluster. (The numerical values 1 and 2 are chosen arbitrarily). In the proposed method, all available M_p reduced feature vectors corresponding to the epochs available

[5] The power ratio is calculated via the difference of natural logarithm of PSD values.
[6] This method is also referred to as the "minimum-redundancy maximal relevance" criterion.

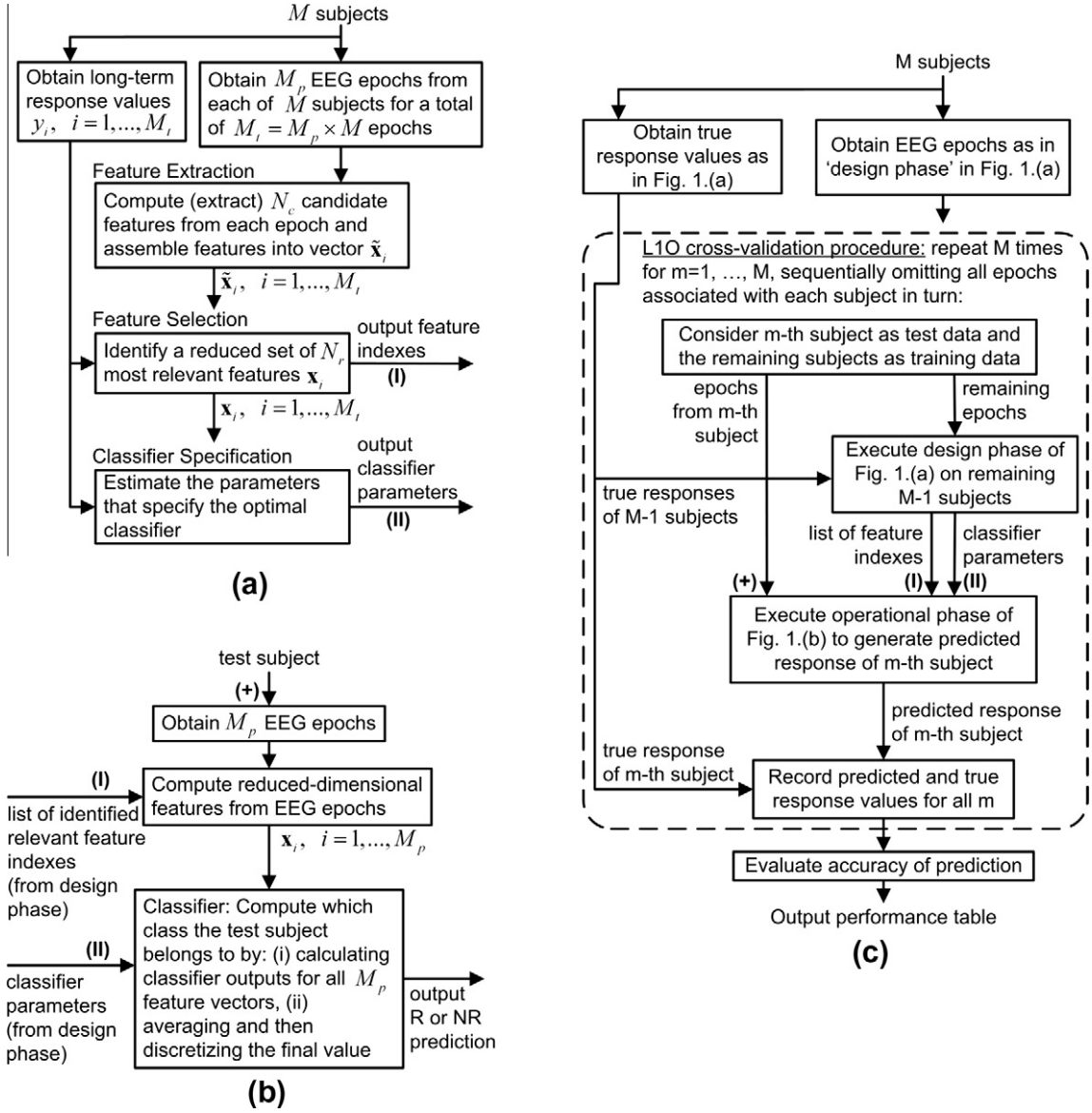

Fig. 1. A simplified schematic drawing of the data analysis steps: (a). The design phase, (b). The operational phase, (c). The L1O cross-validation procedure.

for a given subject are fed into the regression function, which outputs values $\hat{y}_j, j = 1, \ldots, M_p$. Ideally, these quantities are exactly 1 or 2, but in practice, they only approximate these values. The mean of these $\hat{y}$- values is evaluated and then quantized to the closest integer 1 or 2, to yield the corresponding NR or R prediction value.

The operational phase is depicted in Fig. 1(b). Once the machine learning prediction process is designed, it may be applied e.g., in an operational mode in a clinical setting, or, in this context, on Group B subjects. Here, EEG recordings are taken from the patient, and the set of reduced features identified in the design phase are computed from the EEG data, to give a sequence of feature vectors $\mathbf{x}_j, j = 1, \ldots, M_p$. These feature vectors are fed into the classifier or regression function which is specified from the classifier parameters determined in the design phase. The classifier outputs the predicted response of the subject to the proposed clozapine treatment, in the manner described above.

In the current situation however, we are interested in evaluating the performance of the machine learning prediction procedure resulting from the design phase, using the available training data. This is the *evaluation phase*, depicted in Fig. 1(c). In this respect, a *leave-one-out* (L1O) cross-validation procedure is used, where the data from one subject at a time is sequentially removed from the training set. The feature selection and classifier design processes are then executed using all remaining data. The resulting machine learning structure is then tested using the omitted subject. The classifier output is then compared to the known response of the subject, and a performance tally is recorded. The process repeats, each time omitting a different subject, until all subjects have been omitted once. The overall performance figure for the prediction process is then the aggregate performance over all iterations (or folds) of the L1O cross-validation process. With this method, we test over all available data and in each trial we use the largest possible training set. Further, the method is "fair", since the tested data is not part of the training set used in the design phase. The number of latent variables in the KPLSR approach and the variance parameter associated with the Gaussian kernel are determined using a simple multi-dimensional grid search optimization within the cross-validation loop, in a manner consistent with the methodology of (Varma et al., 2006).

Since in effect a different training set is used in each L1O iteration, the set of selected reduced features may vary from one

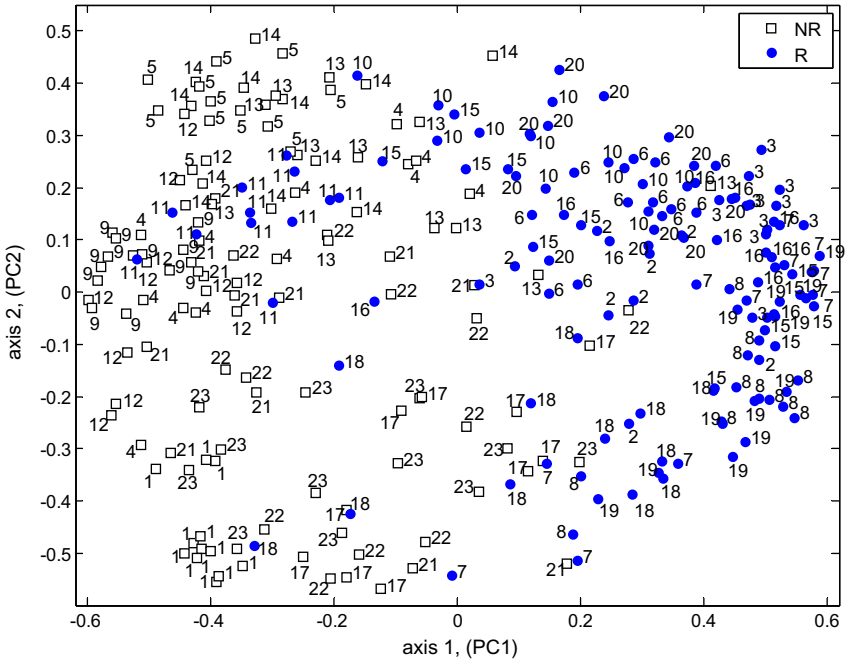

Fig. 2. A demonstration of the clustering behaviour of the proposed ML procedure. The $N_r = 8$ dimensional feature space compressed into 2 dimensions using the KPCA method. There are nominally 12 data points corresponding to multiple EEG epochs from each subject. The subject index corresponding to each point is indicated on the plot.

iteration to another. In the operational phase discussed above, we need a single set of N_r features that best represents the entire training set. We could identify a single set of reduced features simply by applying the feature selection process once on the entire training set. The difficulty with this approach however, is that it is possible that the data from a subset of subjects can dominate the feature selection process. A convenient method of avoiding this possibility is, at each L1O iteration, to select a list of k N_r features, where k is a constant greater than unity, typically greater than 3 in our experiments. Then the desired single set of N_r features is chosen as those which occur most frequently amongst the lists generated over all L1O iterations. In this way, the features are selected on an equitable basis from different combinations of the data. To find a proper value for k, this procedure is repeated with increasing values of k, until at least N_r common features (out of the available k N_r features) can be found among all iterations of the L1O test.

For optimal performance of the proposed scheme, the classifier must operate in an N_r - dimensional feature space, where in our experiments the value of N_r is 8. However, if we wish to visualize the feature space on a plane, it is necessary to compress the feature space. It is readily verified that an optimal *linear* basis for dimensionality compression is the set of principal components of the feature space, obtained by principal component analysis (PCA). Better visualization performance can sometimes be obtained through a *nonlinear* principal component method, in which case kernelization techniques (Muller et al., 2001) are applied to PCA. We refer to the nonlinearized version of PCA as kernel PCA (KPCA). In our study, the KPCA method is used only for the purposes of displaying the clustering results, as in Figs 2 and 3, and is not used in the prediction process.

3. Results

3.1. Treatment-efficacy prediction performance

The first set of results uses data from Group A which consists of 23 subjects. The set of candidate features were extracted from the pre-treatment EEG data and then reduced into a set of N_r = 8 most-relevant features using the available training set data, as discussed in Section 2.3. The prediction performance was then evaluated using the leave-one-out cross-validation procedure discussed previously. The performance evaluation results using the combined EO and EC EEG data sets together for the 23 subjects, for a response threshold value $\delta_1 = 88.5$ and $N_r = 8$ are summarized in Table 3(i), where it is seen that the overall prediction performance is 87.12%. When δ_1 is reduced to 83.5 corresponding to a 30% responder rate, the overall performance becomes 89.7%. Two major latent variables are used for the kernel PLSR method. These results indicate that it is indeed possible to predict the response to clozapine therapy using the proposed methods. Further experiments were performed using a range of δ_1 from 83.5 to 92.5; prediction performance was above 85% in all cases.

We now present results using data from both subject groups A and B. For this second experiment, we train the classifiers using

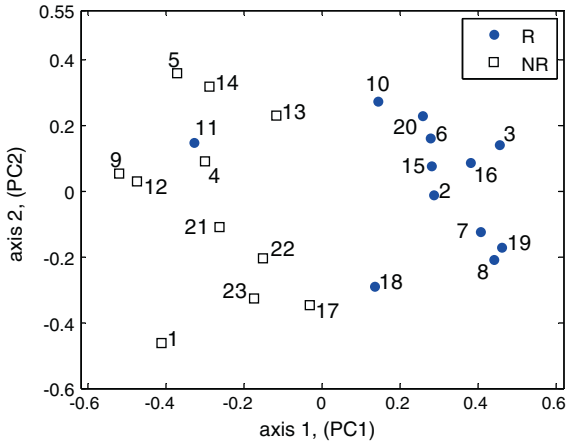

Fig. 3. Same as Fig. 2, except that all data points belonging to each subject in Fig. 2 are averaged to provide one point per subject. The clustering behaviour between the R and NR groups is clearly evident.

Table 3
Performance results predicting the response to clozapine therapy in Group A subjects using $N_r = 8$. Subjects with a post-treatment PANSS score of less than or more than δ_1 are considered responders (R) and non-responders (NR), respectively.

(i). $\delta_1 = 88.5$ (corresponds to 52% response rate)	Predicted R	Predicted NR	% Correct
Actual R	10	2	83.33% = Sensitivity
Actual NR	1	10	90.91% = Specificity
			Average = 87.12%
(ii). $\delta_1 = 83.5$ (corresponds to 30% response rate)	Predicted R	Predicted NR	% Correct
Actual R	6	1	85.7% = Sensitivity
Actual NR	1	15	93.75% = Specificity
			Average = 89.7%

only Group A as training data, and then test the prediction performance over Group B. A group B responder in this case is defined as a subject having an improvement of at least 25% between the pre- and post-QCA scores. The average treatment-efficacy prediction performance for this experiment was 85.7% as reflected in Table 4. This shows a satisfactory prediction performance under different conditions when the classifier is trained on one set, and then tested on another independent set.

We now show an example illustrating the clustering behaviour for the proposed scheme, using Group A data. Fig. 2 shows a scatter plot containing 270 points corresponding to the $M_t = 270$ available epochs of EEG data from the Group A subjects. This figure was generated using the kernel PCA method with a Gaussian kernel. Filled circles correspond to responders and squares to non-responders. In this figure, there are nominally 12 points associated with each subject; however, there are 2 subjects that have only 10 or 8 points. The number written beside each point is the corresponding subject index, which is assigned arbitrarily. Averaging the location of all points corresponding to each subject results in Fig. 3, in which each subject is shown with one point. The clustering between the R and NR groups is clearly evident in this figure. The clustering performance shown in this figure is indicative that the proposed machine learning procedure will perform well, as the results of Table 3 suggest.

3.2. A list of discriminating features

We show a list of 20 most relevant EEG features of interest in Table 5. These are the features that are most strongly discriminative of response to clozapine. Each of the features listed in the table is selected at least once over all L1O iterations. Fig. 4 is a depiction of the most-relevant features selected in Table 5. A connection between two electrode sites in the figure corresponds to a selected feature which involves those two locations. It roughly indicates any relations between EEG sensors that convey relevant information for our prediction problem. This figure depicts how the selected features could give clues about the locality and interconnection of neurological mechanisms associated with a positive response to clozapine. Further investigation of this matter remains a promising topic for future work.

Table 4
Independent test performance using subjects in group A as training data (with $\delta_1 = 88.5$ and $N_r = 8$), and group B as test subjects. Response to clozapine therapy is defined as more than a 25% improvement in the QCA score. Subjects with a post-treatment QCA score of less than or more than δ_1 are considered responders (R) and non-responders (NR), respectively.

	Predicted R	Predicted NR	% Correct
Actual R	6	1	85.7% = Sensitivity
Actual NR	1	6	85.7% = Specificity
			Average = 85.7%

4. Discussion

Our findings support the potential utility of machine learning methods in clinical psychiatry. In the current example we have been able to predict, in advance of the first dose, whether a treatment–resistant patient will or will not respond to a powerful but potentially toxic medication. In various experiments, we evaluated the performance of advanced prediction models in conjunction with kernelization methods to analyze pre-treatment EEG to predict the responsiveness to clozapine. These results support the idea that resting EEG data contains embedded salient information related to clozapine treatment-outcome that can be extracted using machine learning techniques.

We can provide some further evidence of the validity of the proposed prediction method, as follows. First, the clustering behaviour shown in Fig. 3 shows clean separation of the clusters, which is a strong indication that the reduced features can indeed discriminate long-term response. Also, with the L1O cross-validation procedure, different test and training samples are used in each iteration, and yet overall, a reasonable performance level is attained. This suggests the proposed machine learning procedure is consistent across variations of the input data. A final argument to suggest validity of the proposed method is with regard to the results of Table 4. Here, the prediction procedure is trained on Group A data and tested on a completely independent set of Group B data. Even though performance degrades somewhat, the resulting performance of 85.7% is still quite satisfactory.

We can further examine the integrity of the proposed prediction procedure by evaluating the probability that our demonstrated prediction performance would have been due to chance alone.

Table 5
A list of discriminating features for treatment-efficacy prediction using pre-treatment EEG information. Note that the discriminative feature subset is not unique and there is statistical dependence among them. $\delta_1 = 88.5$.

#	Selected EEG-driven Feature
1	Mutual Information between T3 & P3
2	Mutual Information between T3 & O1
3	Mutual Information between C3 & P3
4	Correlation between F8 & T4
5	Coherence at $f = 6$ Hz between T3 & O1
6	Coherence at $f = 6$ Hz between T3 & P3
7	Coherence at $f = 6$ Hz between C3 & O1
8	Coherence at $f = 7$ Hz between F3 & P3
9	Coherence at $f = 8$ Hz between T6 & P3
10	Coherence at $f = 9$ Hz between T3 & O1
11	Coherence at $f = 10$ Hz between T3 & T5
12	Coherence at $f = 10$ Hz between T3 & P3
13	Left to right PSD-ratio at $f = 10$ Hz, T5/T6
14	Left to right PSD-ratio at $f = 11$ Hz, T5/T6
15	Coherence at $f = 11$ Hz between C3 & P3
16	Coherence at $f = 11$ Hz between T3 & P3
17	Left to right PSD-ratio at $f = 12$ Hz, T5/T6
18	Coherence at $f = 12$ Hz between T3 & T5
19	Coherence at $f = 13$ Hz between F7 & F3
20	Left to right PSD-ratio at $f = 16$ Hz, T5/T6

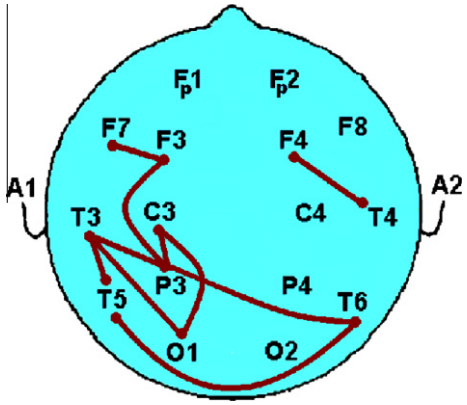

Fig. 4. A rough schematic drawing which shows a list of some relevant features by connections, as reflected in Table 5. Connections are shown by solid thick lines. Electrodes A1 and A2 represent the linked ears reference.

With reference to Table 3(i), there are 12 responders and 11 non-responders, so the probability p of a responder may be taken as 12/23 = 0.5212. Assuming all subjects are independent, the probability of a prediction error is governed by a binomial distribution, which is parameterized by N, the number of samples, and p, in this case the probability of a responder. Therefore, the probability of this level of performance (10 classifications as R and 2 as NR out of N = 12 true responders) occurring due to chance alone is evaluated from the binomial distribution as 0.0226. Similarly, the value of p for the non-responder case is 0.4783, so the probability of estimating 10 NR and 1 R out of 11 non-responders due to chance alone is 0.0036. Similarly, for the case of Table 3(ii), the corresponding figures are 0.0039 and 0.0211 for the R and NR groups, respectively. Thus we see that these figures are negligibly small and we can conclude the prediction results are almost certainly a consequence of the distinguishing characteristics of the EEG measurements obtained from the two groups.

By employing more advanced analytical models, the present study was designed to extend and improve upon the utility of the EEG in predicting the responsiveness to clozapine as investigated in other studies. Although Gross et al. (2004) found that changes in EEG features correlated with outcome, post-treatment EEG data was required. Our methodology is more potentially useful to the clinician as prediction is possible using EEG data collected before this potentially toxic treatment is initiated. Further, even though Knott et al. (2000) were successful at identifying features which were indicative of response, they did not incorporate their findings into a quantitative prediction algorithm. We have therefore been able to extend their work by accomplishing this purpose.

Our proposed feature selection method is novel in the respect that a small number of maximally discriminative features are *automatically* identified from a very large list of candidate features. This is in contrast to the previous approaches, which inherently require a trial-and-error procedure. The previous approach consists of hypothesizing that a single feature may be discriminative, and then verifying or rejecting the hypothesis by experiment. Thus our method can identify salient features that could easily be missed using previous methods.

It is gratifying to note that our proposed feature selection procedure did select some features that were identified from previous studies. This serves as a verification of our method and provides a useful connection with the previous research. Nevertheless, the mathematical structure produced by our ML methods was created from the training data alone without an *a priori* model or previous research findings (e.g. regarding QEEG differences between responders and non-responders). As such it has the advantage of not being constrained by the theoretical constructs derived from previous studies. Without devaluing previous work, or discounting the importance of replication, limiting feature selection to only a group made up of those reported to be useful in previous studies decreases the probability that new and highly salient features will be discovered. Also we have not employed traditional EEG frequency bands and instead used frequency components individually within a 1 Hz resolution window. This maximizes the possibility of detecting potentially important EEG features that might otherwise be obscured when power is integrated over a broad range of frequencies in a given band, e.g. a 10 Hz signal might be lost in the 8–12 Hz alpha band.

The goal of this paper is to propose a new clinical data analysis method and derive an empirical set of EEG features predictive of response to clozapine, not to derive neurological information regarding the pathophysiology of schizophrenia. Nevertheless the clustering of relevant EEG features in the temporo-parietal area of the dominant hemisphere, as seen in Table 5 and in Fig. 4, may be of some interest to those studying regional brain activity patterns in patients with schizophrenia. Others have described bilateral reduced grey matter volume in the temporal lobes (e.g., Okugawa et al., 2002) and electrophysiological abnormalities in the left temporo-parietal region on EEG (e.g., Faux et al., 1987) in schizophrenic patients.

This retrospective study suffers from some weaknesses. Most notably our QCA clinical rating is based on chart review and therefore likely to be less accurate than a standardized PANSS. However, our raters were clinicians expert in the treatment of schizophrenia and familiar with the subjects being evaluated. The QCA would therefore have reasonable clinical validity. The high predictive accuracy of our algorithm in both Group A and B subjects even in the face of this source of outcome variance may speak to the robustness of this methodology. As QCA and PANSS ratings were completed years before this project they could not have been influenced by the machine learning assignment into responder and non-responder groups.

It must be noted the results of this pilot study are derived using a relatively small quantity of data. Our findings must be replicated in a much larger sample of training and test subjects before they can be accepted with confidence. Notwithstanding these issues, our data suggest that machine learning methods of analyzing EEG signal may be employed to create a useful psychiatric management tool. Furthermore, the methodology described in this paper could be extended to construct models that predict the response to various other treatments available for patients with schizophrenia or with other psychiatric conditions. Finally, it may be possible to incorporate a range of other clinical and laboratory data beyond EEG measurements, such as personality inventory scores, personal and demographic information and treatment history to improve clustering behaviour and prediction performance.

An additional topic for future consideration is to investigate the minimum number of channels needed to yield adequate prediction performance. It may be that a reduced configuration of electrodes concentrated over the left side (as suggested by Fig. 4), will still yield an acceptable level of performance, but at a reduced cost.

5. Appendix A. The QCA clinical rating procedure

The QCA clinical rating procedure was devised in the context of an un-related earlier naturalistic retrospective un-published clinical study of treatment-resistant schizophrenic patients being considered for clozapine treatment. The subjects in the present study were included in this previous study. An experienced clinician reviewed all the available clinical descriptive information of the patient's symptomatology prior to beginning a course of clozapine.

Reported symptoms, corresponding to those described in the PANSS, were rated as: present, moderate or severe on a one to six point scale. Only explicitly described symptoms were scored and the clinical rater was instructed not to infer the presence of potential symptoms. The same rating was repeated, based on case records describing current symptoms at the time (usually after approximately six months) when the decision was made to either discontinue or continue with on-going maintenance clozapine therapy.

Acknowledgements

The authors would like to thank Margarita Criollo, Joy Fournier, and Eleanor Bard for their help in clinical experiments. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Adler G, Grieshaber S, Faude V, Thebaldi B, Dressing H. Clozapine in patients with chronic schizophrenia: serum level, EEG and memory performance. Pharmacopsychiatry 2002;35:190–4.

Birca A, Carmant L, Lortie A, Lassonde M. Interaction between the flash evoked SSVEPs and the spontaneous EEG activity in children and adults. Clin. Neurophysiol. 2006;117:279–88.

Boutros NN, Arfken C, Galderisi S, Warrick J, Pratt G, Iacono W. The status of spectral EEG abnormality as a diagnostic test for schizophrenia. Schizophr. Res. 2008;99:225–37.

Coburn KL, Lauterbach EC, Boutros NN, Black KJ, Arciniegas DB, Coffey CE. The Value of Quantitative Electroencephalography in Clinical Psychiatry: A Report by the Committee on Research of the American Neuropsychiatric Association. J. Neuropsychiatry Clin. Neurosci. 2006;18:460–500.

Cover TM, Thomas, JA. Elements of Information Theory, 2nd Ed. John Wiley & Sons, 2006.

Dunki RM, Dressel M. Statistics of biophysical signal characteristics and state specificity of the human EEG. Physica A 2006;370:632–50.

Essali A, Haj-Hasan NA, Li C, Rathbone J. Clozapine versus typical neuroleptic medication for schizophrenia. Cochrane Database of Systematic Reviews 2009; John Wiley and Sons Ltd, Art No.CD000059.

Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Medi. Imaging 2007;26:93–105.

Faux SF, Shenton ME, McCarley RW, Torello MW, Duffy FH. P200 topographic alterations in schizophrenia: evidence for left temporal-centroparietal region deficits. Electroencephalogr. Clin. Neurophysiol. Suppl. 1987;40:681–7.

Freudenreich O, Weiner RD, McEvoy JP. Clozapine-induced electroencephalogram changes as a function of clozapine serum levels. Biol. Psychiatry 1997;42:132–7.

Gallinat J, Heinz A. Combination of multimodal imaging and molecular genetic information to investigate complex psychiatric disorders. Pharmacopsychiatry 2006;39:S76–9.

Gross A, Joutsiniemi SL, Rimon R, Appelberg B. Clozapine-induced QEEG changes correlate with clinical response in schizophrenic patients: a prospective, longitudinal study. Pharmacopsychiatry 2004;37:119–22.

Gunther W, Baghai T, Naber D, Spatz R, Hippius H. EEG alterations and seizures during treatment with clozapine: a retrospective study of 283 patients. Pharmacopsychiatry 1993;26:69–74.

Guo Y, Bowman FD, Kilts C. Predicting the brain response to treatment using a Bayesian hierarchical model with application to a study of schizophrenia. Hum. Brain Mapp. 2008;29:1092–109.

Hughes JR, John ER. Conventional and Quantitative Electroencephalography in Psychiatry. J. Neuropsychiatry Clin. Neurosc. 1999;11:190–208.

Ince N F, Goksu F, Pellizzer G, Tewfik A, Stephane M. Selection of spectro-temporal patterns in multichannel MEG with support vector machines for schizophrenia classification. Proc. Annual Int. Conf. IEEE Eng. in Medicine and Biology Society 2008; 3554–3557.

Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 1987;13:261–76.

Kane J, Honigfeld G, Singer J. Meltzer H and the Clozaril Collaborative Study Group. Clozapine for the treatment-resistant schizophrenic: A double-blind comparison with chlorpromazine. Archives of General Psychiatry 1988;45:789–96.

Kim D, Burge J, Lane T, Pearlson GD, Kiehl KA, Calhoun VD. Hybrid ICA-Bayesian network approach reveals distinct effective connectivity differences in schizophrenia. Neuroimage 2008;42:1560–8.

Knott V, Labelle A, Jones B, Mahoney C. EEG hemispheric asymmetry as a predictor and correlate of short-term response to clozapine treatment in schizophrenia. Clin. Electroencephalogr. 2000;31:145–52.

Knott V, Labelle A, Jones B, Mahoney C. Quantitative EEG in schizophrenia and in response to acute and chronic clozapine treatment. Schizophr. Res. 2001;50:41–53.

Knott VJ, LaBelle A, Jones B, Mahoney C. EEG coherence following acute and chronic clozapine in treatment-resistant schizophrenics. Experiment. Clin. Psychopharmacolo. 2002;10:435–44.

Kwak N, Choi C-H. Input feature selection by mutual information based on Parzen window. IEEE Trans Pattern Analysis and Machine Intelligence 2002;24:1667–71.

Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel RR. What does the PANSS mean? Schizophr. Res. 2005;79:231–8.

Lin C, Wang Y, Chen J, Liou Y, Bai Y, Lai I, Chen T, Chiu H, Li Y. Artificial neural network prediction of clozapine response with combined pharmacogenetic and clinical data. Comput. Methods Programs Biomed. 2008;91:91–9.

Malow BA, Reese KB, Sato S, Bogard PJ, Malhotra AK, Tung-Ping S, Pickar D. Spectrum of EEG abnormalities during clozapine treatment. Electroencephalogr. Clin. Neurophysiol. 1994;91:205–11.

Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. IEEE Trans Neural Networks 2001;12:181–201.

Oikonomou T, Sakkalis V, Tollis IG, Micheloyannis S. Searching and visualizing brain networks in Schizophrenia. Springer Lecture Notes in Computer Science. Biological and Medical Data Analysis 2006;4345:172–82.

Okugawa G, Sedvall GC, Agartz I. Reduced grey and white matter volumes in the temporal lobe of male patients with chronic schizophrenia. Eur. Arch. Psychiatry Clin. Neurosci. 2002;252:120–3.

Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Analysis and Machine Intelligence 2005;27:1226–38.

Rosipal R, Kramer N. Overview and recent advances in partial least squares. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J, editors. Subspace, latent structure and feature selection techniques. Lecture Notes in Computer Science: Springer; 2006. p. 34–51.

Sakkalis V, Oikonomou T, Pachou E, Tollis I, Micheloyannis S, Zervakis M. Time-significant wavelet coherence for the evaluation of Schizophrenic brain activity using a graph theory approach. Proceedings Int Conference of the IEEE Engineering in Medicine and Biology 2006:4265–8.

Struyf J, Dobrin S, Page D. Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. BMC Genomics 2008; 9:(531).

Varma, S, R. Simon, Bias in error estimation when using cross-validation for model selection, BMC Bioinformatics, 2006; 7:(91).

Young CR. Bowers Jr. MB, Mazure CM. Management of the adverse effects of clozapine. Schizophrenia Bulletin 1998;24:381–90.