Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach

V. Rocha-Rego1,2†, J. Jogia†, A. F. Marquand1, J. Mourao-Miranda1,2, A. Simmons1,2,4 and S. Frangou5

1 Department of Neuroimaging, Institute of Psychiatry, King’s College London, UK
2 NIHR Biomedical Research Centre for Mental Health at South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, King’s College London, UK
3 Computer Science Department, Centre for Computational Statistics and Machine Learning, University College London, UK
4 MRC Centre for Neurodegeneration Research, Institute of Psychiatry, King’s College London, UK
5 Psychosis Research Program, Icahn School of Medicine at Mount Sinai, Icahn Medical Institute, New York, NY, USA

Background. Bipolar disorder (BD) is one of the leading causes of disability worldwide. Patients are further disadvantaged by delays in accurate diagnosis ranging between 5 and 10 years. We applied Gaussian process classifiers (GPCs) to structural magnetic resonance imaging (sMRI) data to evaluate the feasibility of using pattern recognition techniques for the diagnostic classification of patients with BD.

Method. GPCs were applied to gray (GM) and white matter (WM) sMRI data derived from two independent samples of patients with BD (cohort 1: n=26; cohort 2: n=14). Within each cohort patients were matched on age, sex and IQ to an equal number of healthy controls.

Results. The diagnostic accuracy of the GPC for GM was 73% in cohort 1 and 72% in cohort 2; the sensitivity and specificity of the GM classification were respectively 69% and 77% in cohort 1 and 64% and 99% in cohort 2. The diagnostic accuracy of the GPC for WM was 69% in cohort 1 and 78% in cohort 2; the sensitivity and specificity of the WM classification were both 69% in cohort 1 and 71% and 86% respectively in cohort 2. In both samples, GM and WM clusters discriminating between patients and controls were localized within cortical and subcortical structures implicated in BD.

Conclusions. Our results demonstrate the predictive value of neuroanatomical data in discriminating patients with BD from healthy individuals. The overlap between discriminative networks and regions implicated in the pathophysiology of BD supports the biological plausibility of the classifiers.

Received 24 December 2011; Revised 3 October 2012; Accepted 9 April 2013; First published online 5 June 2013

Key words: Bipolar disorder, diagnosis, Gaussian process classifiers, imaging, pattern recognition.

Introduction

Bipolar disorder (BD) is a complex psychiatric disorder characterized by severe mood dysregulation (APA, 1994). BD is associated with significant psychosocial morbidity and mortality (Angst et al. 2002; Baldessarini & Tondo, 2003) and is among the leading causes of disability worldwide (WHO, 2004). Despite illness severity, the delay between onset and accurate diagnosis is typically between 5 and 10 years (Lish et al. 1994; Hirschfeld et al. 2003; Berk et al. 2007). Surveys of BD patients conducted over the past 20 years show no evidence of improvement in timely illness recognition (Hirschfeld et al. 2003; WHO, 2004; Berk et al. 2007). Delayed diagnosis in BD has adverse consequences in terms of increased periods in episode, greater psychosocial morbidity (Stensland et al. 2008) and emerging treatment resistance (Post et al. 2003; Ketter et al. 2006). Delayed diagnosis is also associated with increased treatment costs (Stensland et al. 2008) whereas early and accurate identification of BD leads to significant cost savings (Menzin et al. 2009). The importance of early recognition in BD is further underscored by emerging yet compelling evidence that the illness is associated with evolving neurobiological changes that may drive subsequent clinical
deterioration (Berk et al. 2009; Kapczinski et al. 2009). Therefore, the timely diagnosis of BD is currently the single most important unmet need in enhancing clinical and functional outcomes.

Neuroimaging studies to date have established that structural magnetic resonance imaging (sMRI) can be used to identify brain morphological differences between BD patients and controls. Meta-analytic studies that have synthesized the extensive available literature have confirmed that BD is reliably associated with structural abnormalities in the ventral prefrontal cortex, the cingulate gyrus, amygdala/parahippocampal complex and the basal ganglia (Kempton et al. 2008; Arnone et al. 2009; Vita et al. 2009; Bora et al. 2010; Ellison-Wright & Bullmore, 2010; Kempton et al. 2011; Selvaraj et al. 2012). Despite the contribution of these findings to our understanding of the pathophysiology of BD, their clinical usefulness has been negligible. This is primarily because conventional sMRI data analyses compute mean group differences in spatial localized anatomical regions and do not make use of information about the distributed pattern of relationships among regions or voxels. Information about these spatial patterns is of particular relevance as it can be used for the diagnostic classification of individual patients, thus bridging the gap between neuroscience and clinical practice. In this respect, recent advances in multivariate pattern recognition techniques represent a major development.

The most commonly used pattern recognition algorithm has been the support vector machine (SVM) classifier. The SVM classifier has been used for the classification of patients with Alzheimer’s disease (Klöppel et al. 2008; Vemuri et al. 2008), autism (Ecker et al. 2010), aphasia (Wilson et al. 2009) and psychosis (Koutsouleris et al. 2009; Mourao-Miranda et al. 2012) and to predict clinical variables based on patterns of brain activation in functional MRI (Fu et al. 2008; Marquand et al. 2008). However, the SVM classifier yields binary (case or control) and not probabilistic outcomes. For many applications, probabilistic predictions are desirable as they have two key advantages: they provide accurate quantification of predictive uncertainty, reflecting variability within subject groups (e.g. in quantifying the probability that a subject has a psychiatric disorder within a population where illness severity can be expected to vary between individuals), and they allow adjustment of predictions to compensate for different frequencies of diagnostic classes within the general population (Bishop, 2006). Gaussian process classifiers (GPCs) represent a significant advance over SVM as they are fully probabilistic pattern recognition models based on Bayesian probability theory. For neuroimaging, GPCs combine equivalent predictive performance to SVM with the additional benefit of probabilistic classification (Marquand et al. 2010).

Therefore, we used GPCs to examine the predictive value of whole-brain gray (GM) and white matter (WM) anatomy in discriminating patients with BD from healthy individuals. We embedded the classifier in a recursive feature elimination (RFE) framework (Guyon et al. 2002; Marquand et al. 2011) to identify and localize the subset of brain voxels that provide optimal discrimination accuracy. We focused on sMRI, rather than other neuroimaging techniques, as it is widely available, safe and has an established role in the diagnosis and management of brain disorders. Thus, a diagnostic aid based on sMRI data could be easily incorporated into routine clinical practice and is likely to have high patient acceptability. We enrolled patients with bipolar disorder, type 1 (BP-I; APA, 1994), whose diagnosis was further confirmed following detailed clinical assessment. Patients were in remission, free of any other lifetime psychiatric co-morbidity and matched to healthy controls on age, sex and general intellectual ability (IQ). This careful sample selection was designed to maximize the probability that discrimination between patients and controls would be attributable to brain structural changes relating to BD rather than other factors such as co-morbidity or general cognitive ability. Furthermore, we included two independent cohorts of patients and controls to determine the reliability of our findings.

Method

Samples

Cohort 1 comprised 26 patients fulfilling criteria for BP-I according to DSM-IV criteria (APA, 1994) and 26 healthy controls derived from participants in the Maudsley Bipolar Disorder Project (Frangou, 2005; Frangou et al. 2005). Demographic and clinical information on the sample is shown in Table 1. Nineteen BD patients were prescribed psychotropic medications, often in combination (typical antipsychotics=4, atypical antipsychotics=4, lithium=10, carbamazepine=7, sodium valproate=2). None of the patients were prescribed benzodiazepines, anticholinergics or any other medication.

Cohort 2 comprised 14 patients fulfilling DSM-IV criteria for BP-I and 14 healthy controls derived from participants in the Vulnerability Indicators to Bipolar Disorder Study (VIBES; Frangou, 2009). Demographic and clinical information on the sample is shown in Table 1. All patients in this sample were on treatment with anticonvulsant (sodium valproate=10, carbamazepine=4) monotherapy and did not receive any other type of medication.
For both samples, patients had an established diagnosis of BD and were receiving outpatient treatment within secondary care services. They were individually matched on sex, age and IQ to an equal number of healthy controls without a personal or family history of any DSM-IV Axis I disorders. All participants were screened to exclude past, current and hereditary neurological disorders, current medical conditions, DSM-IV current or lifetime drug or alcohol dependence or abuse, and other DSM-IV Axis I current or lifetime co-morbidity and contraindications to MR imaging. All participants were assessed by qualified psychiatrists using the Structured Clinical Interview for DSM-IV for Axis I Disorders, Patient or Non-Patient Version (SCID-I/P and SCID-I/NP; First et al. 2002a,b) and the Family Interview for Genetic Studies (FIGS; Maxwell, 1992), with additional information supplemented by medical notes as appropriate. Psychopathology was assessed using the Hamilton Depression Rating Scale (HAMD; Hamilton, 1960) and the Mania Rating Scale (MRS; Spitzer & Endicott, 1978), and psychosocial functioning was assessed with the Global Assessment of Functioning (GAF) scale (APA, 1994). Patients were scanned when in remission operationalized as (a) the absence of syndromal episode for ≥3 months, (b) being prescribed the same type and dose of medication for ≥3 months, and (c) having HAMD and MRS total scores of <10 on the day of scanning. An estimate of general intellectual ability was obtained using the National Adult Reading Test (NART; Nelson & Wilson, 1992). Patients in cohort 2 were younger, had an earlier age of onset and higher IQ than those in cohort 1 (p<0.01).

This study was approved by the Joint Ethics Committee of the Institute of Psychiatry and the South London and Maudsley National Health Service (NHS) Foundation Trust. Written informed consent was obtained from all participants after a detailed description of the study.

**MRI data acquisition**

Participants were scanned using a 1.5-T GE NV/i Signa MR system (GE Medical Systems, USA) at the Maudsley Hospital, London. The whole brain was imaged with a three-dimensional (3D) inversion recovery prepared fast spoiled gradient-recalled acquisition in the steady state (SPGR) T1-weighted dataset. These T1-weighted images were obtained in the axial plane with 1.5-mm contiguous sections (echo time=5.1 ms, repetition time=18 ms, flip angle=20°, slice thickness=1.5 mm, in-plane resolution=0.9375 × 0.9375 mm, number of excitations=1). Image contrast for all datasets was chosen with the aid of optimizing software (Simmons et al. 1996).

**Data preprocessing**

For both samples, all images were first visually inspected for artifacts or gross structural abnormalities using criteria described previously (Simmons et al. 1996, 2011). Subsequently, images were preprocessed using SPM5 (www.fil.ion.ucl.ac.uk/spm/software/spm5). Using the unified segmentation step included in SPM5, images were normalized and segmented (Ashburner & Friston, 2005). Normalized and modulated GM and WM segmented images were then smoothed with 8-mm isotropic Gaussian kernels and used as input into the classification algorithms.

**Pattern classification analysis**

The probability of group membership was determined separately in each cohort using GPCs to the MRI data.

---

**Table 1. Demographic and clinical characteristics of the study samples**

<table>
<thead>
<tr>
<th></th>
<th>Cohort 1</th>
<th>Cohort 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients with BD (n=26)</td>
<td>Healthy controls (n=26)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>41.5 (11.27)</td>
<td>41.3 (11.65)</td>
</tr>
<tr>
<td>Male:Female</td>
<td>12:14</td>
<td>12:14</td>
</tr>
<tr>
<td>IQ</td>
<td>102.3 (15.93)</td>
<td>103.3 (17.37)</td>
</tr>
<tr>
<td>Age of onset (years)</td>
<td>25.7 (9.22)</td>
<td>–</td>
</tr>
<tr>
<td>HAMD</td>
<td>6.6 (3.39)</td>
<td>–</td>
</tr>
<tr>
<td>MRS</td>
<td>1.8 (4.11)</td>
<td>–</td>
</tr>
<tr>
<td>GAF</td>
<td>71.1 (11.31)</td>
<td>–</td>
</tr>
</tbody>
</table>

BD, Bipolar disorder; HAMD, Hamilton Depression Rating Scale; MRS, Mania Rating Scale; GAF, Global Assessment of Functioning.

Continuous data expressed as mean (standard deviation).
Technical descriptions of GPC inference have been presented elsewhere (Bishop, 2006; Rasmussen & Williams, 2006; Marquand et al. 2008, 2010) and are summarized in the online Supplementary Material. In brief, the classifier is first trained to determine a predictive distribution that best distinguishes cases from controls; any parameters controlling the behavior of this distribution are computed by maximizing the logarithm of the marginal likelihood on the training data only. Then, in the test phase, the classifier predicts the group membership of a previously unseen example. This is achieved by integrating over the predictive distribution for the test case and passing the output through a sigmoidal function, resulting in predictive probabilities scaled between 0 and 1 that precisely quantify the predictive uncertainty of the classifier for the test case.

In each cohort, the GPCs for GM and WM were implemented separately in the PROBID software package (http://www.kcl.ac.uk/iop/depts/Neuroimaging/research/imaginganalysis/Software/PROBID.aspx). We embedded each classifier in a recursive feature elimination (RFE) framework (Guyon et al. 2002; Marquand et al. 2011), which enabled us to identify the subset of

**Fig. 1.** A hypothetical example based on a simplified version of the Gaussian process classifier (GPC) decision function considering two gray matter (GM) voxels per image. To provide an intuitive interpretation of the maps, we consider a simplified version of the GPC decision function. The class probability \( p \) is given by \( p(y=\text{class} \, 1 \mid x, w) = \sigma(w^T x) \), where \( y \) is the class label of a test subject (if \( y>0 \) corresponds to class 1, otherwise it corresponds to class 2), \( x \) is the feature vector containing gray matter voxels for the test subject, \( w \) is referred to as the maximum a posteriori estimate of the GPC weight vector, and is the best point estimate of the GPC decision function (i.e. the mode of the Gaussian approximation in voxel space) and \( \sigma \) is a sigmoid function that maps the values to the interval \([0,1]\). (A) Training Phase: Through training the GPC classifier assigns for each voxel a weight value \( w=(+5, -5) \); \(+5\) represents a voxel containing predictive contribution for class 1 displayed in red and \(-5\) represents a voxel containing predictive contribution for class 2 (displayed in blue). (S) Test Phase: The feature vector (vector containing gray matter probability assigned for each voxel) is shown within each of the two voxels in the two-voxel image example. During the test phase, for classifying a new example we first multiplied each voxel by its corresponding coefficient in the weight vector. After that we add all multiplied values and pass the sum through a sigmoid function in order to obtain an output (i.e. predictive probabilities) between 0 and 1. In the illustrative example above: Subject 1: The feature vector for this subject is \((0.5, 0.2)\). The predictive value for this subject is \( 0((+5*0.5)+(-5*0.2))=0(0.5) \) which corresponds to a predictive probability above 0.5. Therefore subject 1 will be classified as class 1. For the subject 1, a low value in the voxel 2, which has negative coefficient in the weight vector, contributed to the classification of this subject as class 1. Subject 2: The feature vector for this subject is \((0.5, 0.8)\). The predictive value for this subject is \( 0((+5*0.5)+(-5*0.8))=0(-1.5) \) which corresponds to a predictive probability below 0.5. Therefore subject 2 will be classified as class 2. For subject 2, a high value in the voxel 2, which has a positive coefficient in the weight vector, contributed to the classification of this subject as class 2.
brain voxels that provided the optimal discrimination accuracy and to accurately localize the most discriminative brain voxels. To achieve this, we used nested (three-way) cross-validation where we first excluded a matched pair of subjects (one from each group) to comprise the test set, and then performed a second split where we repeatedly repartitioned the remaining subject pairs into a validation and a training set. We then repeatedly trained the classifier on the training set, removing a subset of the least informative features at each iteration, until no features remained. We used a common ranking criterion based on the GPC predictive weights to quantify the information content of each voxel at each iteration (Marquand et al. 2011) and used a small step size (~1% of voxels) to provide fine-grained control over the number of features retained. In each case we selected the number of features that produced maximal accuracy on the validation set before applying it to the test set. We thresholded the probabilistic predictions at 0.5 to convert the probabilistic predictions to class labels and computed the proportion of subjects having the correct label across all test splits to estimate the classification accuracy. The statistical significance of each classifier was determined by permutation testing. This test was used to derive a p value to determine whether the classification accuracy exceeded chance levels (50%). To achieve this, we permuted the class labels from the training set 1000 times (i.e. each time randomly assigning class labels to each structural MRI pattern) and repeated the entire RFE procedure. We then counted the number of times the permuted test accuracy was equal to or greater than the one obtained for the true labels. Dividing this number by 1000, we derived a p value of the classification accuracy.

**Cross-validation**

The performance of the GM and WM classifiers for each cohort separately was estimated in four ways. First, for each classifier within each sample we computed the proportion of images correctly classified as BD patients or controls (i.e. classification accuracy). Second, we quantified the sensitivity and specificity of each classifier defined as: sensitivity = TP/(TP+FN) and specificity = TN/(TN+FP), where TP is the number of true positives (number of images of patients correctly classified), TN is the number of true negatives (number of images of controls correctly classified), FP is the number of false positives (number of images of controls classified as patients) and FN is the number of false negatives (number of images of patients classified as controls). Third, we compared the results obtained through the GPCs to those derived from conventional univariate voxel-based morphometry (VBM)
implemented in SPM5. In VBM analyses we used a statistical threshold of \( p < 0.0001 \) uncorrected for multiple comparisons. Thus we preserve a reasonable degree of specificity in favor of increased sensitivity as the aim of this analysis was to assist in making inferences about the contribution of different regions to the spatially distributed pattern associated with a diagnosis of BD. This is in contrast to the more stringent inferential methods used for controlling type I error to find highly localized, spatially segregated focal group differences. Fourth, regression analyses, thresholded at \( p < 0.001 \) uncorrected, were implemented in SPM5 using cumulative exposure to lithium or antipsychotics (based on doses transformed to chlorpromazine equivalents) to identify potential effects of medication on the GM and WM volumes of patients.

**GPC discrimination maps**

The discrimination map (Rasmussen & Williams, 2006) is a spatial representation of the vector of GPC predictive weights and describes the relative contribution of each brain voxel to the classifier decision. Technical details of GPC discrimination mapping have been

<table>
<thead>
<tr>
<th>Region</th>
<th>Laterality</th>
<th>BA</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>Cluster size</th>
<th>( W_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle frontal gyrus</td>
<td>R</td>
<td>10</td>
<td>33.3</td>
<td>49.9</td>
<td>11.7</td>
<td>466</td>
<td>−16.7</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>10</td>
<td>−31.6</td>
<td>55.3</td>
<td>7.7</td>
<td>397</td>
<td>−16.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>−39</td>
<td>23.9</td>
<td>27.7</td>
<td>93</td>
<td>−21.8</td>
</tr>
<tr>
<td>Inferior frontal gyrus</td>
<td>L</td>
<td>11</td>
<td>−30.6</td>
<td>35.8</td>
<td>−16.9</td>
<td>71</td>
<td>−14.0</td>
</tr>
<tr>
<td>Parietal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferior parietal lobule</td>
<td>R</td>
<td>40</td>
<td>39.6</td>
<td>−54</td>
<td>41</td>
<td>221</td>
<td>−22.4</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>40</td>
<td>−43.1</td>
<td>−35.8</td>
<td>44.5</td>
<td>569</td>
<td>−16.8</td>
</tr>
<tr>
<td>Temporal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superior temporal gyrus</td>
<td>R</td>
<td>41</td>
<td>48.1</td>
<td>−32.1</td>
<td>15.2</td>
<td>76</td>
<td>−14.5</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>39</td>
<td>−52.8</td>
<td>−55.4</td>
<td>23.5</td>
<td>832</td>
<td>−18.4</td>
</tr>
<tr>
<td>Fusiform gyrus</td>
<td>L</td>
<td>37</td>
<td>−45.7</td>
<td>−59.7</td>
<td>−14.7</td>
<td>210</td>
<td>−16.1</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>R</td>
<td>21</td>
<td>37.3</td>
<td>−37.6</td>
<td>−2.4</td>
<td>155</td>
<td>−13.6</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>37</td>
<td>−56.6</td>
<td>−41.1</td>
<td>−8.9</td>
<td>204</td>
<td>−16.3</td>
</tr>
<tr>
<td>Occipital lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lingual gyrus</td>
<td>R</td>
<td>19</td>
<td>33.3</td>
<td>−63.1</td>
<td>2.7</td>
<td>1759</td>
<td>−18.2</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>17</td>
<td>7.2</td>
<td>−94.1</td>
<td>2.9</td>
<td>111</td>
<td>−18.0</td>
</tr>
<tr>
<td>Cuneus</td>
<td>L</td>
<td>18</td>
<td>−4.3</td>
<td>−76.4</td>
<td>6.3</td>
<td>109</td>
<td>−10.0</td>
</tr>
<tr>
<td>Superior occipital gyrus</td>
<td>L</td>
<td>19</td>
<td>−31.6</td>
<td>−72.9</td>
<td>26.4</td>
<td>84</td>
<td>−15.6</td>
</tr>
<tr>
<td>Middle occipital gyrus</td>
<td>R</td>
<td>19</td>
<td>29</td>
<td>−76.3</td>
<td>24.2</td>
<td>53</td>
<td>−18.0</td>
</tr>
<tr>
<td>Limbic lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parahippocampal gyrus</td>
<td>R</td>
<td>28</td>
<td>20.8</td>
<td>−14.6</td>
<td>−19.6</td>
<td>55</td>
<td>−12.8</td>
</tr>
<tr>
<td>Anterior cingulate gyrus</td>
<td>L</td>
<td>32</td>
<td>−0.4</td>
<td>48.8</td>
<td>5.55</td>
<td>179</td>
<td>−11.7</td>
</tr>
<tr>
<td>Posterior cingulate gyrus</td>
<td>R</td>
<td>31</td>
<td>0.4</td>
<td>−37.9</td>
<td>40.4</td>
<td>759</td>
<td>−12.2</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>30</td>
<td>−9.5</td>
<td>−60.4</td>
<td>12.2</td>
<td>71</td>
<td>−9.2</td>
</tr>
<tr>
<td>Insula</td>
<td>R</td>
<td>13</td>
<td>34.7</td>
<td>9.6</td>
<td>2.2</td>
<td>707</td>
<td>−14.1</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>13</td>
<td>−38.7</td>
<td>16.2</td>
<td>−1.4</td>
<td>747</td>
<td>−15.5</td>
</tr>
<tr>
<td>Thalamus</td>
<td>L</td>
<td>−</td>
<td>0</td>
<td>−18.1</td>
<td>6</td>
<td>184</td>
<td>−14.6</td>
</tr>
<tr>
<td>Striatum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caudate</td>
<td>R</td>
<td>−</td>
<td>11</td>
<td>15.7</td>
<td>5.7</td>
<td>50</td>
<td>−12.2</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>−</td>
<td>7</td>
<td>16.9</td>
<td>6.7</td>
<td>77</td>
<td>−16.1</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>L</td>
<td>−</td>
<td>20.8</td>
<td>−61.4</td>
<td>−53.4</td>
<td>582</td>
<td>−16.2</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>−</td>
<td>−28.9</td>
<td>−66.3</td>
<td>−52.6</td>
<td>257</td>
<td>−12.5</td>
</tr>
</tbody>
</table>

BA, Brodmann area; R, right; L, left.
Coordinates are shown in Montreal Neurological Institute (MNI) standard space; x=sagittal, y=coronal, z=axial.
\(^a\) Number of voxels.
\(^b\) Highest weights within individual clusters.
published elsewhere (Bishop, 2006; Marquand et al. 2010) and are described in the online Supplementary Material. The process is illustrated in Fig. 1 based on a simplified hypothetical example of a two-voxel image.

**Results**

**Prediction accuracy**

Classification accuracy reflects the predictive power of the algorithm and is therefore of direct diagnostic relevance. For cohort 1, classification accuracy using GPC analysis of GM images was 73% with a sensitivity and specificity of 77% and 69% respectively. In other words, based on a GM anatomical scan, if a participant had a clinical diagnosis of BD, the probability of correct classification was 0.77. Conversely, if a participant did not have BD, the probability of being correctly classified as a control was 0.69. The GPC analysis using WM images for cohort 1 yielded an accuracy of 69% with a sensitivity of 69% and specificity of 69%.

For both cohorts, the models were significant at \(p<0.001\).

**Discrimination maps**

Discrimination maps showing the global spatial pattern by which the groups differ are illustrated in Fig. 2 and detailed in Tables 2 (GM) and 3 (WM) for cohort 1 and in Fig. 3 and Tables 4 (GM) and 5 (WM) for cohort 2. For both cohorts, the maps highlight those regions that, according to our GPC-RFE classification approach, contain the most discriminative voxels between BD patients and controls. This optimal discriminative pattern was obtained following removal of \(\sim 98\%\) of all voxels. Because of the multivariate character of the GPC, the discrimination maps should not be interpreted as describing focal effects within individual brain regions. Instead they represent a spatially distributed pattern of coefficients that quantify the contribution of each voxel to the GPC decision function (i.e. the value of a voxel in the discrimination map reflects its contribution or predictive value towards one class or the other). We used the following convention: class 1 was the BD group, with labels +1, and class 2 was the control group, with labels −1. In the discrimination map, positive coefficients indicate...
voxels with a predictive value for BD (class 1; visualized in red color scale) and negative coefficients indicate voxels with a predictive value for controls (class 2; visualized in blue color scale); by definition, a voxel with negative predictive weight for patients has the same positive predictive weight for controls. There is significant overlap between the cohorts. In terms of GM, regions within the frontopolar and ventral prefrontal cortex, the parietal lobules, the middle/superior temporal gyri, the lingual gyrus and cuneus and within the thalamus and cerebellum emerge as being implicated most consistently in the diagnosis of BD. A similar conclusion can be drawn for WM tracts traversing ventral prefrontal regions, parietal and postcentral regions, the middle occipital gyrus and the cuneus, the cingulum and genu of the corpus callosum. Discriminative regions were more extensive in cohort 2, which consisted of younger and less medicated patients than cohort 1, suggesting that the results of the GPCs are not driven by medication or age.

**VBM**

At the threshold of \( p<0.0001 \) uncorrected, GM and WM volumetric differences between patients and controls were noted in multiple brain regions in both cohorts. Details of the regional maxima are provided in Supplementary Tables S1–S4. As discussed, the output of the GPC and VBM analyses are not directly comparable as the former reflects the predictive value of voxels in discriminating between patients and controls whereas the latter represents the mean differences between patients and controls. Nevertheless, there is significant overlap between the two outputs and between cohorts. An effect of medication was identified only in cohort 1, with cumulative lithium exposure being positively associated with the right anterior cingulate GM volume (\( x=5.2, y=41.1, z=4 \), cluster size=85). No correlations with medication dose were noted in cohort 2.

**Discussion**

To our knowledge, this is the first study to evaluate the feasibility of using pattern recognition algorithms for the automatic classification of sMRI data of patients with BD and healthy controls. We found that GPCs applied to GM reliably achieved above chance discriminative power solely on the basis of anatomical

Table 4. Cohort 2: gray matter (GM) regions discriminating between individuals with bipolar disorder (BD) and controls

<table>
<thead>
<tr>
<th>Region</th>
<th>Laterality</th>
<th>BA</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>Cluster size(^a)</th>
<th>( W_1 )^(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frontal lobe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle frontal gyrus</td>
<td>R</td>
<td>10</td>
<td>28.0</td>
<td>41.0</td>
<td>17.0</td>
<td>650</td>
<td>14.369</td>
</tr>
<tr>
<td><strong>Parietal lobe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferior parietal lobe</td>
<td>L</td>
<td>7</td>
<td>−42.0</td>
<td>−69.0</td>
<td>47.0</td>
<td>49</td>
<td>−11.18</td>
</tr>
<tr>
<td>Temporal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>L</td>
<td>39</td>
<td>−44.0</td>
<td>−65.0</td>
<td>17.0</td>
<td>29</td>
<td>−73.482</td>
</tr>
<tr>
<td><strong>Occipital lobe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle occipital gyrus</td>
<td>L</td>
<td>18</td>
<td>−30.0</td>
<td>−83.0</td>
<td>−7.0</td>
<td>245</td>
<td>64.639</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>18</td>
<td>24.0</td>
<td>−89.0</td>
<td>11.0</td>
<td>17</td>
<td>−10.218</td>
</tr>
<tr>
<td>Inferior occipital gyrus</td>
<td>R</td>
<td>18</td>
<td>26.0</td>
<td>−91.0</td>
<td>−7.0</td>
<td>184</td>
<td>34.588</td>
</tr>
<tr>
<td>Lingual gyrus</td>
<td>L</td>
<td>17</td>
<td>−14.0</td>
<td>−97.0</td>
<td>−17.0</td>
<td>193</td>
<td>−11.478</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>17</td>
<td>10.0</td>
<td>−95.0</td>
<td>−17.0</td>
<td>25</td>
<td>−19.075</td>
</tr>
<tr>
<td>Cuneus</td>
<td>R</td>
<td>23</td>
<td>12.0</td>
<td>−71.0</td>
<td>9.0</td>
<td>90</td>
<td>−21.117</td>
</tr>
<tr>
<td><strong>Limbic lobe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anterior cingulate gyrus</td>
<td>L</td>
<td>32</td>
<td>−3.0</td>
<td>42.0</td>
<td>7.0</td>
<td>650</td>
<td>12.45</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>32</td>
<td>5.0</td>
<td>40.0</td>
<td>6.0</td>
<td>650</td>
<td>12.19</td>
</tr>
<tr>
<td>Thalamus</td>
<td>L</td>
<td></td>
<td>−6.0</td>
<td>−11.0</td>
<td>5.0</td>
<td>69</td>
<td>21.513</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>R</td>
<td></td>
<td>28.0</td>
<td>−87.0</td>
<td>−39.0</td>
<td>9</td>
<td>−0.753</td>
</tr>
</tbody>
</table>

\( \text{BA, Brodmann area; R, right; L, left.} \)

\( \text{Coordinates are shown in Montreal Neurological Institute (MNI) standard space; x=sagittal, y=coronal, z=axial.} \)

\( \text{\(^a\) Number of voxels.} \)

\( \text{\(^b\) Highest weights within individual clusters.} \)
data, with classification accuracy ranging between 69% and 78%.

**Prediction accuracy of GPCs applied to sMRI in BD**

Neuroanatomical studies using conventional analyses have established the presence of morphological changes in BD (Kempton et al. 2008; Arnone et al. 2009; Vita et al. 2009; Bora et al. 2010; Ellison-Wright & Bullmore, 2010; Kempton et al. 2011; Selvaraj et al. 2012). However, these findings have had limited translational application primarily for three reasons: (a) there is considerable between-group overlap in brain morphological variables derived from group-level neuroimaging analyses (Kempton et al. 2008, 2009, 2011), (b) voxel-based analysis methods are significantly biased toward detecting group differences that are highly localized in space but are limited in detecting group differences that are spatially distributed and subtle (Davatzikos, 2004), and (c) voxel-based analyses do not lend themselves to making predictions at the level of individual subjects.

The data presented here demonstrate that these limitations may be surmounted with the aid of multivariate pattern recognition techniques. The application of GPC analysis to anatomical scans in BD provided diagnostic accuracy in the range 69–78%. As with any new test, the accuracy of the GPC classification for BD was determined against ‘gold standard’ diagnostic assessments. In this study ‘true positive cases’ (i.e. patients with BD) were identified using the SCID-I, conducted by clinicians with expertise in mood disorders. The SCID-I is designed to elicit the presence or absence of the operational criteria that define the syndrome of BD itself and is therefore expected to have the highest diagnostic accuracy (Williams et al. 1992; Fennig et al. 1994; Segal et al. 1994). A more appropriate comparison would be with behavior-based case-finding instruments, whose sensitivity and specificity are about 70%, such as the Mood Disorder Questionnaire (MDQ; Hirschfeld et al. 2000; Hirschfeld, 2010; Zimmerman et al. 2011). However, even more important is the comparison of our results to ‘real world’ clinical assessments where BD is either missed or misdiagnosed resulting in nearly a third of patients having to wait for approximately 10 years before they receive an accurate diagnosis (Lish et al. 1994; Hirschfeld et al. 2003; Berk et al. 2007). This is because of the substantial overlap between clinical symptoms of BD and those of other disorders, particularly major depressive disorder (MDD) because depressive symptoms are commonly present at onset (Perugi et al. 2000) and often dominate the clinical picture thereafter (Judd et al. 2002). Additionally, the presence of psychosis during manic or depressive
episodes often leads to difficulties in distinguishing BD from schizophrenia and schizo-affective disorder (Schimmelmann et al. 2005). Further diagnostic challenges arise from the high level of co-morbidity of BD with other disorders, particularly substance abuse and anxiety disorders (McElroy et al. 2001; Merikangas et al. 2011). In this context, a classifier that is trained to identify true positives BD cases might have an important role in assisting clinicians when used in combination with other clinical measures.

The results presented here for the predictive value of sMRI data in BD compare favorably with classification accuracies of approximately 80% reported for Alzheimer’s disease and schizophrenia, even though the magnitude of neuroanatomical deviance is greater for these disorders (reviewed by Klöppel et al. 2012).

Brain regions discriminating patients with BD from controls

The GM and WM discriminative maps generated by the GPCs show that clusters contributing to the distinction between BD patients and healthy controls are spatially distributed within cortical and subcortical regions (Tables 2–5). GM discriminative clusters consistently associated with BD in both cohorts were localized primarily within the frontopolar and ventral prefrontal cortex, the inferior parietal lobule, the medial and lateral temporal cortex, the cingulate cortex, occipital regions in the lingual gyrus and cuneus, the thalamus and cerebellum. This is in keeping with previous morphometric studies that have repeatedly shown an association between volumetric changes in these regions and disease expression for BD.
et al (Kempton et al. 2008, 2009, 2011; Scherk et al. 2008; Arnone et al. 2009; Yu et al. 2010; Hallahan et al. 2011). Previous research on global (Scherk et al. 2008; Vita et al. 2009) and regional ( McIntosh et al. 2005; Stanfield et al. 2009) WM volume changes in BD yielded variable results. However, there is increasing consensus for an association between disease expression for BD and WM pathology within the cingulum (Vederine et al. 2011) and the genu of the corpus callosum (Bellani et al. 2009; Walterfang et al. 2009a,b; Bearden et al. 2011). Our findings in both cohorts suggest that WM regions of predictive value for BD patients are widespread but consistently include the cingulum and genu. Although the results of the VBM and GPC analyses cannot be compared directly, they showed significant overlap in terms of the spatial distribution of regions influenced by the diagnosis of BD.

Methodological considerations and future directions

In cohort 1, the majority of BD patients were medicated with antipsychotic or mood-stabilizing medication or both. Treatment with lithium has been associated with volumetric changes in BD, and specifically increases in global and regional volumes (Bearden et al. 2007; Kempton et al. 2008, 2009; Phillips et al. 2008; Germana et al. 2010; van Erp et al. 2012). In line with this, we observed a positive association between lithium dose and GM volume in the anterior cingulate in cohort 1. However, medication effects related to lithium or antipsychotics cannot fully explain the results because patients in cohort 2 were not on these medications.

The GPCs were trained to segregate healthy controls from patients with BD. This represents the necessary first step in developing pattern recognition approaches for use as neurodiagnostic tools. Future studies are required to replicate these findings in larger samples and across different sites. Another important task is to evaluate the performance of pattern recognition classifiers for the identification of biologically meaningful subtypes of BD and for the differential diagnosis of BD from disorders with overlapping clinical phenotypes.

In summary, our results demonstrate that GPC-based neuroanatomical pattern recognition techniques may prove clinically useful in improving the timely diagnosis of BD, which currently relies entirely on clinical symptoms.

Supplementary material

For supplementary material accompanying this paper visit http://dx.doi.org/10.1017/S0033291713001013.

Acknowledgments

This work was partially supported by the European College of Neuropsychopharmacology, Networks Initiative, Neuroimaging Network, which had no further role in the study design, in the collection, analysis and interpretation of data, in the writing of the manuscript, and in the decision to submit it for publication. A.M. gratefully acknowledges support from the King’s College London Centre of Excellence in Medical Engineering, funded by the Wellcome Trust and the Engineering and Physical Sciences Research Council (EPSRC) under grant no. WT088641/Z/09/Z. J.M.M. was funded by a Wellcome Trust Career Development Fellowship under grant no. WT086565/Z/08/Z. V.R.R. and A.S. were supported by the National Institute for Health Research (NIHR) Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, King’s College London.

Declaration of Interest

None.

References


