NeuroImage 84 (2014) 299–306

Contents lists available at ScienceDirect

NeuroImage

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

Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects

Hugo G. Schnack *[,1], Mireille Nieuwenhuis [1], Neeltje E.M. van Haren, Lucija Abramovic, Thomas W. Scheewe, Rachel M. Brouwer, Hilleke E. Hulshoff Pol, René S. Kahn

Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Accepted 26 August 2013
Available online 1 September 2013

Keywords:
Schizophrenia
Bipolar disorder
MRI
Machine learning
Classification

ABSTRACT

Although structural magnetic resonance imaging (MRI) has revealed partly non-overlapping brain abnormalities in schizophrenia and bipolar disorder, it is unknown whether structural MRI scans can be used to separate individuals with schizophrenia from those with bipolar disorder. An algorithm capable of discriminating between these two disorders could become a diagnostic aid for psychiatrists. Here, we scanned 66 schizophrenia patients, 66 patients with bipolar disorder and 66 healthy subjects on a 1.5 T MRI scanner. Three support vector machines were trained to separate patients with schizophrenia from healthy subjects, patients with schizophrenia from those with bipolar disorder, and patients with bipolar disorder from healthy subjects, respectively, based on their gray matter density images. The predictive power of the models was tested using cross-validation and in an independent validation set of 46 schizophrenia patients, 47 patients with bipolar disorder and 43 healthy subjects scanned on a 3 T MRI scanner. Schizophrenia patients could be separated from healthy subjects with an average accuracy of 90%. Additionally, schizophrenia patients and patients with bipolar disorder could be distinguished with an average accuracy of 88%.The model delineating bipolar patients from healthy subjects was less accurate, correctly classifying 67% of the healthy subjects and only 53% of the patients with bipolar disorder. In the latter group, lithium and antipsychotics use had no influence on the classification results. Application of the 1.5 T models on the 3 T validation set yielded average classification accuracies of 76% (healthy *vs* schizophrenia), 66% (bipolar *vs* schizophrenia) and 61% (healthy *vs* bipolar). In conclusion, the accurate separation of schizophrenia from bipolar patients on the basis of structural MRI scans, as demonstrated here, could be of added value in the differential diagnosis of these two disorders. The results also suggest that gray matter pathology in schizophrenia and bipolar disorder differs to such an extent that they can be reliably differentiated using machine learning paradigms.

Introduction

Currently, the diagnosis of psychiatric disorders such as schizophrenia and bipolar disorder is based predominantly on their clinical manifestations. While psychiatrists can establish the presence of illness (as distinct from its absence) with relative ease, discrimination between several possible diagnoses is far more complicated, especially in the early phase of schizophrenia and bipolar disorder. The availability of additional (objective) measures would assist psychiatrists in the process of diagnosis, with obvious benefits to efficiency of treatment and improved outcome. Magnetic resonance imaging (MRI) has proven to be an effective technique to detect structural brain abnormalities at group-level in schizophrenia patients (meta-analyses: Haijma et al., 2012; Olabi et al., 2011) and those with bipolar disorder (meta-analyses: Kempton et al., 2008; McDonald et al., 2004). Unfortunately, statistical group differences do not translate to discovering deviations from normal on an individual basis and therefore are not sufficient as a diagnostic aid.

Using machine learning techniques, promising results have been obtained for the classification of schizophrenia patients and healthy subjects based on MRI scans. Pioneering work was done by Davatzikos et al. (2005), followed by numerous other investigations. The support vector machine (SVM; Fan et al., 2008; Ingalhalikar et al., 2010; Koutsouleris et al., 2009; Pohl and Sabuncu, 2009; Vapnik, 1999) and the Discriminant Function Analysis (Karageorgiou et al., 2011; Kasparek et al., 2011; Leonard et al., 1999; Liu et al., 2004; Nakamura et al., 2004; Takayanagi et al., 2011) are the most frequently used methods (for an overview of schizophrenia classification studies using

* Corresponding author at: Department of Psychiatry, A01.126, University Medical Center Utrecht, Heidelberglaan 100, NL-3584 CX, Utrecht, The Netherlands. Fax: +31 88 75 55443.
E-mail address: h.schnack@umcutrecht.nl (H.G. Schnack).
[1] Authors contributed equally.

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.08.053

structural MRI, see Nieuwenhuis et al. (2012)). We recently demonstrated in two large independent samples that a classification model built from one data set can be used to classify new subjects as schizophrenia patients or healthy subjects with 71% accuracy. To the best of our knowledge, no studies have been published investigating the use of MRI to separate bipolar patients from healthy subjects or schizophrenia patients [although one study combined structural MRI brain measures and neuropsychological test scores for this purpose (Pardo et al., 2006)]. Given the brain abnormalities found in bipolar disorder and schizophrenia and the differences between these abnormalities (Arnone et al., 2009; Ellison-Wright and Bullmore, 2010; Hulshoff Pol et al., 2012; Koo et al., 2008; McDonald et al., 2005; Qiu et al., 2008; Rimol et al., 2010, 2012), it may be fruitful to apply these classification models to help separate these two disorders. We train three SVM models to separate patients with schizophrenia from those with bipolar disorder, healthy subjects from patients with schizophrenia, and healthy subjects from patients with bipolar disorder. Although it could be of theoretical interest to build a three-group classifier that separates the three groups in a single step, our approach addresses the clinical relevant issue of separating the two disorders using MRI. Furthermore, it provides brain patterns that discriminate between the respective groups, which can be analyzed to indicate which features are unique to the discrimination between schizophrenia and bipolar disorder. We test the predictive power of the models both in the dataset they were built on and in an independent dataset.

Materials and methods

General

In this study we used two datasets. The first set, called discovery sample, was used to build classification models for the separation of healthy subjects and patients with schizophrenia and bipolar disorder. The models were tested on this set too. On the second set, called validation sample, no models were built; this independent sample was used to test the generalizability of the models built on the first set.

Subjects — discovery sample

Schizophrenia patients (SZ), patients with bipolar disorder (BP) and healthy subjects (HC) were selected from our database. Since the quality of machine learning models strongly benefits from large and balanced training data sets, we extracted the largest possible groups of subjects that were same-sized and matched on gender (exactly) and age. This resulted in three groups of each 66 subjects (24 males), aged 37 ± 11 years. The subjects overlap to a large extent with the sample described in Hulshoff Pol et al. (2012); additional SZ patients and HC subjects are part of the study described by Hulshoff Pol et al. (2001). The sample included singletons and twins. To ensure independency between the three groups, only one twin from discordant twin pairs was included. The presence or absence of psychopathological abnormality was established using the Comprehensive Assessment of Symptoms and History (Andreasen et al., 1992) and Schedule for Affective Disorders and Schizophrenia Lifetime Version (Endicott and Spitzer, 1978) assessed by at least one independent rater who was trained to assess this interview. All healthy subjects met Research Diagnostic Criteria (Spitzer et al., 1978) of "never [being] mentally ill." Patients in the schizophrenia group met *DSM-IV* criteria for schizophrenia. Patients in the bipolar group met *DSM-IV* criteria for bipolar I (N = 50), II (N = 14) or NOS (N = 2). Subjects were matched for age, sex and socioeconomic status of their parents expressed as the highest completed level of education by one of their parents.

All SZ patients had received antipsychotic medication in the past and all but one patient received antipsychotic medication at the time of the MRI scan. Medication included typical (N = 36) and atypical (N = 23) antipsychotic agents. Forty-five BP patients were using lithium at the time of the scan and 13 patients were using antipsychotics (of whom 5 were using both lithium and antipsychotics). See Table 1 for demographic information.

Subjects — validation sample

Forty-six SZ patients, 47 BP patients and 43 HC subjects were drawn from our 3 Tesla MRI database. The SZ patients and part of the HC subjects were part of an earlier study (Scheewe et al., 2012). When composing the validation data set, we had to make a trade-off between the size of the sample and the matching on sex and age with the discovery set. A fair test of the validity of the models requires distributions of subjects with respect to variables such as age and sex that match those of the discovery sample as close as possible, but not at the cost of excluding too many subjects, since this would reduce the power of the validation test. The resulting sample included about equal numbers of subjects per group, and matched the discovery sample on age, but not on gender (significantly more males). Patients in the SZ group met *DSM-IV* criteria for schizophrenia (N = 35) or schizoaffective disorder (N = 11). Patients in the bipolar group all met *DSM-IV* criteria for bipolar I. Medication of SZ patients included typical (N = 4) and atypical (N = 38) antipsychotic agents. Thirty-one BP patients were using lithium at the time of the scan and 36 patients were using antipsychotics (of whom 22 were using both). See Table 1 for demographic information.

All participants gave written informed consent to participate in the study. The study was approved by the Medical Ethical Research Committee for human research (METC) from the University Medical Centre Utrecht and was carried out under the directives of the Declaration of Helsinki (Amendment South Africa 2000).

Imaging and preprocessing

All scans from the discovery sample were acquired on a 1.5 Tesla Philips NT scanner (Philips, Best, The Netherlands). Three-dimensional T1-weighted, fast field echo scans with 160 to 180 contiguous coronal slices (echo time [TE], 4.6 ms; repetition time [TR], 30 ms; flip angle, 30°; field of view [FOV], 256 mm; 1 × 1 × 1.2 mm^3 voxels) were made of all subjects. All scans from the validation sample were acquired on a 3 Tesla Philips Achieva scanner. Three-dimensional T1-weighted, fast field echo scans with 180 contiguous sagittal slices (TE, 4.6 ms; TR, 10 ms; flip angle, 90°; FOV, 240 mm; 0.75 × 0.75 × 0.80 mm^3 voxels) were made of all subjects.

The scans were processed with our standard image processing pipeline (Brouwer et al., 2010; Hulshoff Pol et al., 2001) on the computer network of the Department of Psychiatry at the University Medical Center Utrecht. The features we used were extracted from the processed T1-weighted images. The images were transformed into Talairach orientation (no scaling), after which they were corrected for scanner RF-field nonuniformity. Using a partial volume segmentation technique

Table 1
Demographics.

	Discovery sample			Validation sample		
	SZ	BP	HC	SZ	BP	HC
N	66[a]	66[b]	66[c]	46	47	43
Male/female	24/42	24/42	24/42	33/13	22/25	21/22
Age (year), mean (SD)	36.5 (11.0)	37.7 (11.0)	38.2 (10.8)	31.0 (7.5)	41.6 (10.0)	33.8 (9.4)
Range	18–57	18–60	18–62	19–48	22–60	19–60
Duration of illness, mean (SD) (year)	15.4[d] (11.0)	12.8[e] (9.7)	– –	7.2 (6.5)	20.5[f] (7.4)	– –
Medication:						
Antipsychotics (yes/no)	59/1[g]	12/54	–	46/0	36/11	–
Lithium (yes/no)	0/66	45/21	–	0/46	31/15[h]	–

[a]24 twins (no complete pairs) [b]32 twins and 13 complete concordant twin pairs; [c]23 twins and 18 complete pairs. Information missing in [d]17, [e]8 [f]5, [g]6, [h]1 patients.

(Brouwer et al., 2010) the brain was segmented into gray matter, white matter and cerebrospinal fluid. The gray matter segments were blurred using a three-dimensional Gaussian kernel (full-width half-maximum (FWHM) = 8 mm). The voxel values of these blurred segments reflect the local presence, or concentration, of gray matter and will be referred to as gray matter 'densities' (GMDs). In order to compare GMDs at the same anatomical location between all subjects, the GMD images were transformed into a standardized coordinate system using a two-step process. First, the T1-weighted images were linearly transformed to a model brain (Hulshoff Pol et al., 2001). In this linear step joint entropy mutual information metric was optimized. In the second step nonlinear (elastic) transformations were calculated to register the linearly transformed images to the model brain up to a scale of 4 mm (FWHM), thus removing global shape differences between the brains, but retaining local differences (ANIMAL; Collins et al., 1995). The GMD maps were then transformed to the model space by applying the concatenated linear and nonlinear transformations. The GMD maps were not modulated, thus representing relative amounts of gray matter. We made this choice (i) to keep the preprocessing steps and interpretation the same as for our VBM studies on these data and (ii) because of recent evidence that unmodulated analyses outperform modulated analyses, both in VBM (Radua et al., in press) and analyses aiming to discriminate between groups of patients and controls (Dashjamts et al., 2012). Since the density maps have been blurred to an effective resolution of 8 mm, it is not necessary to keep this information at the 1-mm level. Therefore, the maps were resampled to voxels of size $2 \times 2 \times 2.4$ mm^3, i.e., doubling the original voxel sizes. The use of smoothed, resampled, GMD maps removes noise originating from imperfect segmentation and warping (due to image noise and different brain topology). This is especially important when combining 1.5 T and 3 T images, since the latter are much more detailed. For all voxels, GMD was regressed on age and sex for all subjects in the sample together. The resulting b-maps were used to correct the GMD maps for the effects of these factors and calculate GMD residuals, which were used as features for the support vector machine model.

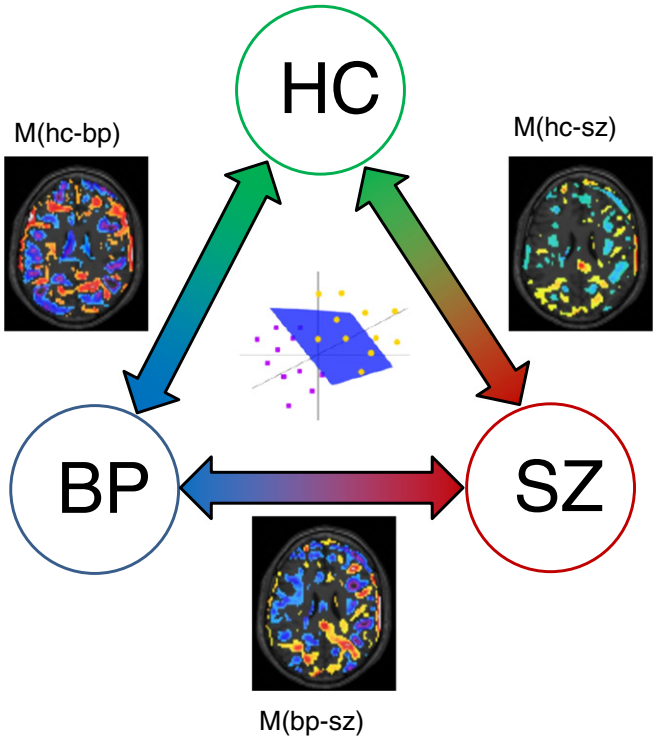

Fig. 1. Classification scheme. The three groups (healthy subjects (HC), patients with bipolar disorder (BP) and schizophrenia patients (SZ)) are depicted by circles. The three models that are trained to perform pair-wise separations of the groups are indicated by arrows, labeled with the model's name (M) and a symbolic picture of its discriminative brain pattern (w-map). In the center a schematic picture of the support vector machine (SVM): an optimal separation plane (OSH; blue) separates the two classes of subjects based on their positions in a high-dimensional feature space (yellow and purple dots).

Support vector machine models

The support vector machine (SVM) is a high-dimensional, pattern recognition, supervised learning algorithm (Vapnik, 1999) used to solve classification problems. In our case this problem consists of separating three groups of subjects. Therefore, we build three models: M(sz-hc) to separate SZ from HC; M(hc-bp) to separate HC from BP; and M(bp-sz) to separate BP from SZ (Fig. 1). The SVM model is trained to classify subjects based on their features, in our case gray matter densities. We integrated LIBSVM (Chang, 2011) with our software to carry out the classification.

Subjects are represented by features congregated into a vector x_i per subject. These vectors exist in a high dimensional feature space, in which a flat decision surface is constructed to separate the subjects from different classes (shown schematically in the center of Fig. 1). This is accomplished by the introduction of a decision function $y(x_i)$:

$$y(x_i) = w^T \cdot x_i - b, \quad (1)$$

that vanishes at the decision surface. The weight vector w is a normal vector to this surface; b is an offset. In the training phase each subject has a label t_i (e.g., patients 1; healthy control -1), and the function is optimized by requiring $y(x_i) < 0$ if $t_i = -1$, and $y(x_i) > 0$ if $t_i = +1$. When applying the model this decision function is used to classify the test subjects according to the sign of $y(x_i)$. The weight-vector not only contains information on feature importance, but also on whether it is either an increase or decrease of a particular feature's value that contributes to being classified as a patient.

There can be several surfaces that exactly separate the classes. The SVM chooses the so called *optimal* separating hyperplane (OSH) such that the space between the two classes, which is called the margin, is made as large as possible. This is a necessary condition for generalization of the model to new subjects. There is a free parameter C in SVM that influences the narrowness of the margin. It was shown earlier (Franke et al., 2010) that tuning C can increase the model's performance. We used C as optimized by Nieuwenhuis et al. (2012).

Limiting and avoiding the influences of medication

The size of the striatum is known to be affected by (typical) antipsychotic medication (Smieskova et al., 2009). To exclude this possible confounding effect, we created a model where the striatum was masked out. The striatum was segmented manually from the model brain image and, using mathematical morphology operations, enlarged, to ensure the entire striatum was excluded for all subjects. We applied two-sample t-tests to test for possible differences in model performance between BP patients on antipsychotic medication and those who were not. Lithium has been shown to influence gray matter (density) (Kempton et al., 2008). Two-sample t-tests were used to test for differences in model performance (ie, accuracy) between BP patients using/not-using lithium. To avoid separating lithium users from non-users, instead of BP patients from other subjects, we also built SVM models including only subjects not on lithium (these models are referred to as MŁ). Since only 25 of the BP patients did not use lithium, models using these subjects would include only 50 subjects in total. To solve this problem of small training sets, we built 100 of such models, each time using random selections (but accounting for twin dependency — see 'Quality measures' section) of SZ and HC subjects. For comparison, we also built 100 models MŁ(hc-sz) including random selections of 25 SZ patients and 25 healthy subjects.

Table 2
Classification results (discovery sample).

Model	BP-SZ	HC-BP	HC-SZ
	M/MŁ	M/MŁ	M/MŁ
Group:			
HC	–	66.7/63.4	87.9/87.1
SZ	89.4/82.2	–	92.4/84.8
BP all	86.4	53.0	–
L	90.2	51.2	–
Ł	80.0/84.0	56.0/54.8	–

Percentages correctly classified subjects for the three models (HC-BP, HC-SZ, BP-SZ), built on all subjects (M) and lithium non-users (MŁ). BP percentages have been split into percentages for lithium users (L) and non-users (Ł). Leave-4-out cross-validation was used for the M models, and leave-2-out for the MŁ models.

Quality measures I: tests on the discovery sample

The quality of a model M(g1–g2) is assessed by the percentage correctly classified subjects belonging to group g1 and the percentage correctly classified subjects belonging to group g2. We tested the accuracy of the models by using cross validation, which gives an estimate of how well the model will generalize to a new data set. In a leave-*k*-out cross validation setup, each time *k* subjects are left out of the training set, which are subsequently used to test the model. Often $k = 1$ is chosen, but to keep equal numbers of subjects in the two groups, $k = 2$ would be more appropriate here. We chose $k = 4$, which gave us the opportunity to leave complete twin pairs out of the training set simultaneously. This avoids a bias in the prediction of a twin's class if his/her cotwin (in the same class) is in the training set. The procedure is thus as follows: First a model is trained on all subjects but four, which are then used to test this model. This is repeated with four different subjects left out, until all subjects have been left out once. In our case each model M is thus trained 33 times. For the MŁ models, twin bias was avoided by not including the cotwin when a twin from a complete pair was included in the selection. This gave us the opportunity to apply leave-2-out cross-validation, leaving the training set as large as possible.

The significance of a model M(g1–g2) is tested by means of bootstrapping, testing the null-hypothesis that the observed separation of the two groups is by chance and not due to true group differences. All subject-labels (g1, g2) are randomly permuted prior to the model's training and testing phase. This process is repeated a thousand times to estimate the null-distribution of separation accuracy percentages and calculate the probability of finding by chance an accuracy at least as extreme as observed (p-value).

To test whether being a twin biases a subject's chance to be correctly classified, a Fisher's exact test was applied to the frequencies of correctly and wrongly classified singletons and twins.

In addition, a multi-class SVM [max-wins-voting SVM, MWV_SVM (Knerr et al., 1990)] model was trained and tested using the same cross-validation set-up.

Quality measures II: tests on the validation sample

The generalizability of the models built on the discovery sample was tested by applying them to the data of the validation sample. Large systematic differences in scan quality between the two samples (1.5 T *vs* 3 T), which will be reflected in the GMD values and thus in the feature vectors *x*, were expected to lead to shifts in the output values *y* of the classifiers (see Eq. (1)), and thus to a misbalance between false positives and false negatives in the classification of subjects scanned at 3 T. These possible shifts were accounted for in two steps. Based on a 1.5 T–3 T calibration study (Brouwer et al., unpublished; see Appendix A) a reliability mask was applied to the features, thus only including voxels with GMD values comparable between 1.5 T and 3 T into the model. The 1.5 T models were thus rebuilt on a restricted feature set (76% of the voxels). A receiver operating characteristic (ROC) curve analysis was used to remove the remaining misbalance between false positives and false negatives. In a *post hoc* analysis the accuracy percentages on the validation set were recalculated, weighing the male and female percentages according to the discovery sample's gender distribution.

Within the SZ group, a Fisher's exact test was used to test whether diagnosis (schizophrenia vs schizoaffective) biases a subject's chance to be correctly classified.

To test the significance of the results obtained in the validation sample the same 1000 label permutations of the discovery set were used. The models built from these 1000 sets (without leaving out any subjects this time) were applied to the validation set's subjects to estimate the null-distribution of separation accuracy percentages.The validation accuracies of the true models were compared with these null-distributions to calculate the probability that these accuracies were found by chance.

Analysis of the discriminative patterns

If the different models are built from the same set of features (as is the case in our approach), their weight vectors lie in the same space and can be directly compared. The gray matter pattern that discriminates between schizophrenia and bipolar disorder may share part of it with the pattern that discriminates between schizophrenia and health. Mathematically, the vectors w in Eq. (1) of the different models may not be orthogonal. By projection, we can decompose w(bp-sz) into a part that coincides with w(hc-sz) and a part perpendicular to it:

$$w(\text{bp}-\text{sz}) = w_{//\text{hc-sz}}(\text{bp-sz}) + w_{\perp\text{hc-sz}}(\text{bp-sz}), \tag{2}$$

which can be rewritten as

$$w(\text{bp-sz}) = \cos(\varphi) w(\text{hc-sz}) + w_{\perp\text{hc-sz}}(\text{bp-sz}), \tag{3}$$

with $\cos(\varphi) = w^{T}(\text{bp-sz}) \cdot w(\text{hc-sz})/(||w(\text{bp-sz})||\ ||w(\text{hc-sz})||)$, ie, the cosine of the angle φ between the weight vectors, determining the size of the parallel, or shared part. The perpendicular part is obtained by subtraction: $w_{\perp} = w - w_{//}$. The offset b in Eq. (1) is divided accordingly: $b = b_{//} + b_{\perp}$. The separation of SZ and BP is thus built from a part that employs the same pattern as is used for the HC-SZ separation, and a part that is unique for the discrimination between SZ and BP. The latter tells us which brain pattern drives the discrimination between the two disorders. We will refer to the vectors w as w-maps. The two components of w-map(bp-sz) will also be applied to the data.

Results

The classification results of the models are given in Table 2. In the discovery sample, the models involving SZ patients performed best, with classification accuracies of 86% or higher for the models including all subjects (M). SZ patients can thus be separated well from HC subjects *and* from BP patients ($p < 0.001$). The separation of BP patients and HC subjects, on the other hand, turned out to be less accurate: 67% of the HC subjects were correctly classified ($p = 0.02$) and only 53% of the BP patients ($p = 0.4$). There was no significant difference in classification accuracy

Table 3
Contingency table (multi-class SVM).

	Classification:		
	HC	SZ	BP
Group:			
HC	59	11	30
SZ	6	86	6
BP	41	8	50

Class predictions for subjects from the three groups in %. In the SZ and BP groups one subject could not be assigned to a class because there was a tie.

Table 4
Classification results (validation).

	Validation set			Discovery set	
Model	Raw	ROC[a]	ROC-reweighted	Calib-masked	Original
BP/SZ	78.7/47.8	66.0/65.2	64.6/71.0	86.4/89.4	86.4/89.4
HC/BP	69.8/48.9	62.8/55.3	63.0/55.5	66.7/54.5	66.7/53.0
HC/SZ	90.7/54.3	76.7/73.9	79.2/79.2	86.4/90.9	87.9/92.4

Classification accuracies in %. Raw: application of discovery-model to validation set; ROC: application results after ROC curve analysis; ROC-reweighted: the ROC results reweighted to match gender distribution of discovery sample; calib-masked: model performance on the discovery sample (using L4O) using the reduced feature set, for direct comparison with the validation set results; the 'original' results are taken from Table 2, for comparison.
[a] Corresponding shifts in b (see Eq. (1)): −0.16 (BP/SZ), −0.08 (HC/BP), and −0.20 (HC/SZ).

between BP patients on antipsychotic medication and those who were not (M(sz-bp): 83% and 87%, respectively, $t(14.7) = 0.32$, $p = 0.63$; M(hc-bp): 50% and 54%, $t(15.7) = 0.23$, $p = 0.41$). There was no significant difference in classification accuracy between BP lithium users and non-users (M(sz-bp): 90% and 80%, $t(39.6) = 1.11$, $p = 0.14$; M(hc-bp): 51% and 56%, $t(50.6) = 0.38$, $p = 0.71$). The models including only non-users (MŁ) produced comparable outcomes. These accuracies were on average a few percentage points lower. There was no significant difference in classification accuracy between twins and singletons. Contingency Table 3 presents the multi-class results, which indicate a comparable classification of SZ subjects (86% correct) and less accurate separation of HC and BP subjects, although with accuracy percentages of 59 and 50 well above chance (33%).

Application of the discovery sample's models to the validation sample yielded unbalanced specificity/sensitivity percentages: 91%/54% (HC/SZ); 79%/48% (BP/SZ); 70%/49% (HC/BP). Adjustment of the classification thresholds after ROC curve analyses led to accuracies of 77%/74% (HC/SZ), 66%/65% (BP/SZ), 63%/55% (HC/BP). The performance of the HC/SZ and BP/SZ models was significantly better than chance ($p < 0.005$ and $p < 0.05$); the HC/BP model did not perform significantly better ($p = 0.2$). Reweighting the male/female accuracy percentages to match the discovery sample's gender distribution lead to specificity/sensitivity percentages of 79%/79% (HC/SZ), 65%/71% (BP/SZ), 63%/55% (HC/BP). There was no significant difference in classification accuracy between the diagnoses schizophrenia and schizoaffective disorder. The results are presented in Table 4, together with those from the discovery sample, for comparison.

Fig. 2 shows the w-maps overlaid on the brain and the decomposition of w-map(bp-sz) into the part shared with w-map(hc-sz), ie, $w_{//\text{hc-sz}}$(bp-sz) and the part perpendicular to it, $w_{\perp\text{hc-sz}}$(bp-sz). The cosine of the angle between the two weight vectors was $\cos(\varphi) = 0.65$ (Eq. (3)). Schizophrenia patients are separated from other subjects by a pattern that includes decreases in (pre- and orbito)frontal and (superior)temporal gray matter. The pattern that drives the discrimination between patients with bipolar disorder and schizophrenia patients includes decreases (in SZ relative to BP) in superior frontal and parietal gray matter; deeper in the frontal, parietal and occipital cortices, paired regions of increases and decreases are found (Fig. 2, third column). Fig. 3 displays the y-values (Eq. (1)) after application of the different w-maps to the individual brain data, ie, the scattering around the optimal separating hyperplanes (OSHs), here depicted as separation lines. The figure illustrates the good separability of SZ patients from both BP patients and HC subjects. $M_{//\text{hc-sz}}$(bp-sz), the part shared by HC and BP, already produces a good separation of SZ patients and the other subjects; $w_{\perp\text{hc-sz}}$(bp-sz), the part unique for BP-SZ separation, further increases the distance of the BP subjects to the OSH (all subjects' $y < 0$ in the training phase), while it does not displace the HC subjects from the OSH (subjects' mean y close to 0). In the test phase (leave-4-out), the discriminating effect of $w_{\perp\text{hc-sz}}$(bp-sz) was smaller, but significant (mean difference in y between the two groups = 0.11, $t = 2.06$, $p < 0.041$).

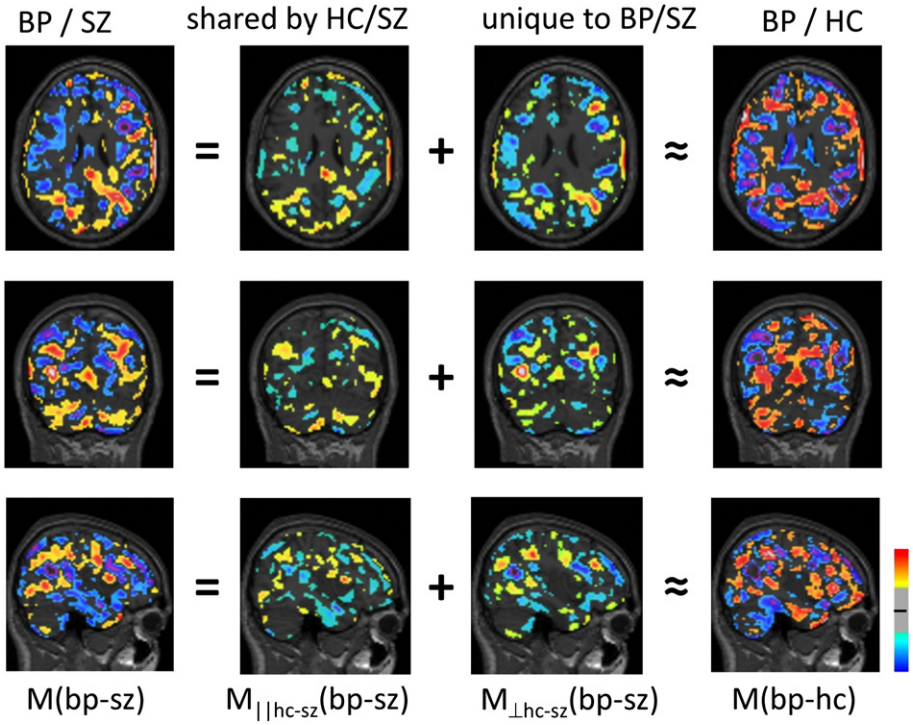

Fig. 2. Axial, coronal and sagittal slices (from top to bottom) of the template brain with w-maps projected on it. Separation pattern (w-map) of BP and SZ (M(bp-sz); first column), which has been decomposed into the part shared with the separation of HC and SZ ($M_{//\text{hc-sz}}$(bp-sz) = 0.65 * M(hc-sz); second column) and the remaining part, orthogonal to the former part, being unique to the separation of BP and SZ ($M_{\perp\text{hc-sz}}$(bp-sz); third column). The latter pattern can be compared to M(bp-hc) (=−M(hc-bp; fourth column). The pattern in the third column is characterized by neighboring pairs of regions with hot and cool colors. For all models M(g1–g2), hot colors (clamped between 0.0005 and 0.0025) refer to relative increase of GM density in g2 subjects with respect to g1 subjects, and vice versa for cool colors (clamped between −0.0005 and −0.0025).

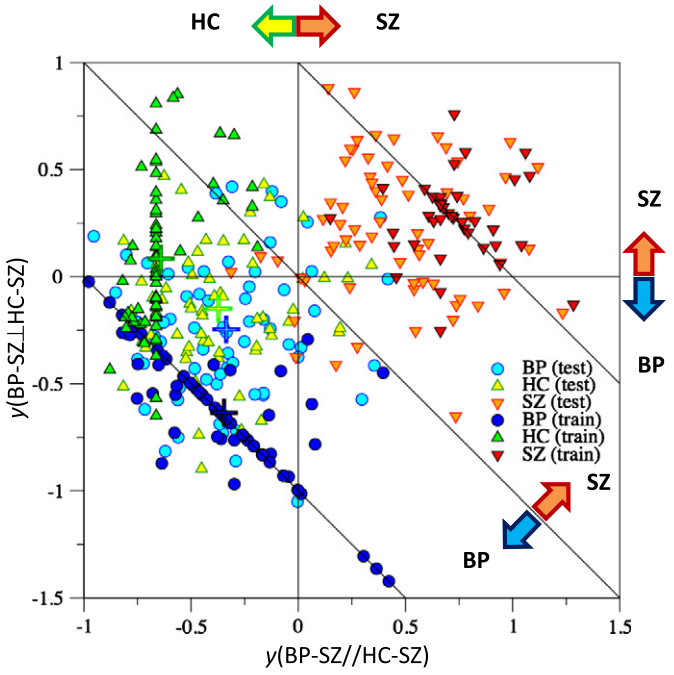

Fig. 3. Application of M(bp-sz) to all subjects. The horizontal axis shows the effect (y-value, Eq. (1)) of the part shared with M(hc-sz), ie, $0.65(w^T(\text{hc-sz}) \cdot x_i - b)$, with $\cos(\varphi) = 0.65$ (Eq. (3)). A y-value smaller than 0 means not-SZ, larger than 0 means SZ. The vertical axis shows the effect of the remaining part of the model (unique for BP-SZ), ie, the part perpendicular to M(hc-sz). A y-value smaller than 0 means classification as BP, larger than 0 means classification as SZ. The sum of the two parts gives the y-value of the M(bp-sz) model itself, with subjects on the lower-left of the diagonal being classified as BP (or not-SZ), and on the upper-right of the diagonal as SZ. Darker symbols (red, blue, green) refer to the effects using the models trained on all subjects; brighter symbols (orange, cyan, yellow) to effects on subjects that were left out of the model (test). The majority of training subjects is placed at distance ± 1 (or $\pm \cos(\varphi)$) by application of the model they are part of: this is reflected by the lines of blue (BP) and green (HC) symbols; since SZ is part of all models, the majority of SZ subjects is placed at $(\cos(\varphi), 1 - \cos(\varphi)) = (0.65, 0.35)$. The plus symbols are placed at the means of the distributions.

Discussion

The purpose of this study was to classify patients with schizophrenia, bipolar disorder, and healthy controls on the basis of their structural MRI scans. We used a support vector machine (SVM) to create three models from gray matter density images, each separating two of the three groups. We confirmed that it is possible to separate schizophrenia patients from healthy subjects with an average accuracy of 90%. Additionally, we demonstrated that schizophrenia patients can be separated from those with bipolar disorder with a nearly identical degree of precision (88%). The delineation of patients with bipolar disorder and healthy subjects reached much lower accuracy: 67% of the healthy subjects and only 53% of the patients with bipolar disorder were classified correctly. In an independent data set we replicated these results, yielding average classification accuracies of 75% and 66% for the separation of schizophrenia patients from healthy subjects and patients with bipolar disorder, and 59% for the separation of bipolar patients from healthy subjects. The accuracy of the HC-SZ and BP-SZ models in this set is lower than in the discovery set, but remains significantly better than chance.

Since lithium is known to affect gray matter density (Kempton et al., 2008; Moore et al., 2000; van der Schot et al., 2010), patients with bipolar disease were split into those who were on lithium and those who were not; separate models were created including only subjects not on lithium. Lithium turned out not to influence classification accuracy. Models including only non-users performed with slightly lower precision, but this was a systematic effect also present in the SZ-HC model with reduced number of subjects, and thus likely to be a decrease in robustness due to fewer training subjects (Nieuwenhuis et al., 2012). It is well known that (typical) antipsychotic medication affects striatal volume, which is why we excluded features from the striatum in the models. Although some studies also suggest that antipsychotic medication affects cortical gray matter (Ho et al., 2011), the reported effects are inconsistent (Shepherd et al., 2012) and so it is not clear which brain regions should be left out of the model. However, when we compared the patients with bipolar disorder who were on antipsychotic medication with those who were not, we did not find an effect of antipsychotic medication on the classification of these patients suggesting that, for classification purposes at least, the effect of antipsychotics on cortical gray matter is limited. Nevertheless, effects of medication on classification can only be ruled out by studies in medication naive patients.

Since the discriminative brain pattern is a description of the cumulative contributions of all features, the interpretation of the effects of single brain regions on the separation of the groups is complicated. Some marked contributions are, however, present. Schizophrenia patients are separated from other subjects by a pattern that includes decreases in (pre- and orbito)frontal and (superior)temporal gray matter, consistent with reported structural brain abnormalities in schizophrenia (Fornito et al., 2009; Haijma et al., 2012). The pattern that separates bipolar patients from healthy subjects shows widespread contributions, lacking any marked regions, "consistent" with recent literature that concludes that many brain abnormalities in BP may be obscured by (clinical) heterogeneity (Kempton et al., 2008; Selvaraj et al., 2012). The pattern that drives the discrimination between patients with bipolar disorder and schizophrenia is diffuse, but includes decreases (in SZ relative to BP) in superior frontal and parietal gray matter. Deeper in the frontal, parietal and occipital cortices, neighboring regions with opposite contributions to the discrimination are found: decreases are flanked by increases (Fig. 2; third column). These paired regions may reflect local shape changes in which some structure is larger in one disorder relative to the other, at the cost of a neighboring structure. The large extent of these changes over the (right) hemisphere may reflect a brain network problem in SZ (van den Heuvel et al., 2013), reflecting symptoms that could be attributable to problems with the proper integration of information.

The classification accuracies in the validation sample are all the more remarkable given the differences in scanner field strength and demographics between discovery and validation samples. While a 1.5 T–3 T calibration allowed us to exclude the least reliable features, the 24% reduction in number of features to build the model from could have reduced the model's generalizability. Correcting for the difference in gender distribution improved the performance in the validation sample. Other factors that may have affected the performance reproducibility are the differences in duration of illness and the number of SZ patients on typical antipsychotic medication, although we tried to reduce the impact of the latter as much as possible by excluding the striatum from the model (Nieuwenhuis et al., 2012). The classification accuracy of patients with schizoaffective disorder was not significantly different from those with schizophrenia. The relatively low performance of the model separating BP from HC may be attributable to the less marked brain abnormalities found in BP (meta-analyses: Kempton et al., 2008; McDonald et al., 2004) as compared to those found in SZ (meta-analyses: Haijma et al., 2012; Olabi et al., 2011).

This study has two major implications. First, the accurate separation of individuals suffering from schizophrenia and bipolar disorder on the basis of their structural MRI scans suggests that gray matter pathology in schizophrenia and bipolar disorder differs to such an extent that they can be reliably differentiated using machine learning paradigms. The second, and more practically useful, implication is that structural MRI could aid at separating schizophrenia from bipolar disorder in the differential diagnostic process.

Although the current MRI classification results are promising, they are limited by the small number of patients with bipolar disorder not using lithium ($N = 25$). Indeed, larger datasets and independent

replications are needed before this model can be put to clinical use. Also, the model was built, and tested on, chronically ill patients, while its clinical use will be mostly in the more recently ill subjects. Thus, longitudinal studies including recent onset psychotic patients with diagnostic follow-up (of many years) are needed to test the clinical usefulness of this model. Although the psychiatrist's diagnosis serves as gold standard to train our SV machine, his/her reliability is not 100%; in fact, for SCID-I, inter-rater reliabilities for bipolar disorder are in the range of 80% and for schizophrenia are 94% (Skre et al., 1991), while agreement between clinical diagnoses and SCID-I is poor (Steiner et al., 1995). Machines cannot be expected to approach 100% accuracy, and, consequently, an MRI-SVM can only become a tool aiding the diagnostic process. However, this tool comes with two virtues: it is based on information not available to the psychiatrist, and it is objective.

In conclusion, we have shown that structural MRI can be used to separate schizophrenia patients from both healthy subjects and bipolar patients with good accuracy, despite using discovery and validation samples acquired on scanners with different field strengths and varying in gender distribution. This suggests not only that gray matter pathology is different in schizophrenia and bipolar disorder, but that, eventually, MRI may prove to be a useful instrument in the diagnostic process in psychiatry.

Acknowledgments

We thank Dorret Boomsma for providing us with healthy twin subjects from the Netherlands Twin Registry. We thank and Astrid van der Schot for assistance with data acquisition and processing.

Appendix A. 1.5 Tesla–3 Tesla calibration

Fifteen healthy volunteers (12 female; mean (SD) age 26.1 (8.1) year) were scanned on both our 1.5 T and 3 T MRI scanner with the acquisition protocols as described in the Methods section. Twelve of them were scanned twice on the same day and the others less than 4 months apart. Scans were preprocessed using our standard image processing pipeline (including putting scans in Talairach orientation, non-uniformity correction, and partial volume segmentation of gray/white matter and CSF, as described in the Methods section). Like the images from this study, the calibration scans were non-linearly transformed to a model brain. These transformations were then applied to the partial volume gray matter maps. The transformed segments were blurred (full-width-half-max, 8 mm) and then resampled to a $2 \times 2 \times 2.4$ mm^3 resolution to reduce noise and increase statistical power as would be done in a voxel-based analysis. To assess systematic offsets and reliability between the scanners, mean (SD) gray matter density per scanner and intraclass correlation (Shrout and Fleiss, 1979) between scanners was computed in each voxel. A between-scanner reliability mask was created by selecting voxels with ICC > 0.7 or SDs < 0.05.

References

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (CASH). An instrument for assessing diagnosis and psychopathology. Arch. Gen. Psychiatry 49, 615–623.
Arnone, D., Cavanagh, J., Gerber, D., Lawrie, S.M., Ebmeier, K.P., McIntosh, A.M., 2009. Magnetic resonance imaging studies in bipolar disorder and schizophrenia: meta-analysis. Br. J. Psychiatry 195, 194–201.
Brouwer, R.M., Hulshoff Pol, H.E., Schnack, H.G., 2010. Segmentation of MRI brain scans using non-uniform partial volume densities. NeuroImage 49, 467–477.
Chang, C.-C., 2011. A library for support vector machines, In: Lin, C.-J. (Ed.), ACM Transactions on Intelligent Systems and Technology, 2nd ed. pp. 1–27.
Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C., 1995. Automatic 3-d model-based neuroanatomical segmentation. Hum. Brain Mapp. 3, 190–208.
Dashjamts, T., Yoshiura, T., Hiwatashi, A., Togao, O., Yamashita, K., Ohyagi, Y., Monji, A., Kamano, H., Kawashima, T., Kira, J., Honda, H., 2012. Alzheimer's disease: diagnosis by different methods of voxel-based morphometry. Fukuoka Igaku Zasshi 103, 59–69.
Davatzikos, C., Shen, D., Gur, R.C., Wu, X., Liu, D., Fan, Y., et al., 2005. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. Arch. Gen. Psychiatry 62, 1218–1227.
Ellison-Wright, I., Bullmore, E., 2010. Anatomy of bipolar disorder and schizophrenia: a meta-analysis. Schizophr. Res. 117, 1–12.
Endicott, J., Spitzer, R.L., 1978. A diagnostic interview: the schedule for affective disorders and schizophrenia. Arch. Gen. Psychiatry 35, 837–844.
Fan, Y., Gur, R.E., Gur, R.C., Wu, X., Shen, D., Calkins, M.E., Davatzikos, C., 2008. Unaffected family members and schizophrenia patients share brain structure patterns: a high-dimensional pattern classification study. Biol. Psychiatry 63, 118–124.
Fornito, A., Yücel, M., Patti, J., Wood, S.J., Pantelis, C., 2009. Mapping grey matter reductions in schizophrenia: an anatomical likelihood estimation analysis of voxel-based morphometry studies. Schizophr. Res. 108, 104–113.
Franke, K., Ziegler, G., Kloppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. NeuroImage 50, 883–892.
Haijma, S.V., Van Haren, N., Cahn, W., Koolschijn, P.C., Hulshoff Pol, H.E., Kahn, R.S., 2012. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. Schizophr. Bull. 39, 1129–1138.
Ho, B.C., Andreasen, N.C., Ziebell, S., Pierson, R., Magnotta, V., 2011. Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. Arch. Gen. Psychiatry 68, 128–137.
Hulshoff Pol, H.E., Schnack, H.G., Mandl, R.C., van Haren, N.E., Koning, H., Collins, D.L., et al., 2001. Focal gray matter density changes in schizophrenia. Arch. Gen. Psychiatry 58, 1118–1125.
Hulshoff Pol, H.E., van Baal, C.M., Schnack, H.G., Brans, R., van der Schot, A.C., Brouwer, R.M., et al., 2012. Overlapping and segregating structural brain abnormalities in twins with schizophrenia or bipolar disorder. Arch. Gen. Psychiatry 69, 349–359.
Ingalhalikar, M., Kanterakis, S., Gur, R., Roberts, T.P., Verma, R., 2010. DTI based diagnostic prediction of a disease via pattern classification. Med. Image Comput. Comput. Assist. Interv. 13, 558–565.
Karageorgiou, E., Schulz, S.C., Gollub, R.L., Andreasen, N.C., Ho, B.C., Lauriello, J., et al., 2011. Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. Neuroinformatics 9, 321–333.
Kasparek, T., Thomaz, C.E., Sato, J.R., Schwarz, D., Janousova, E., Marecek, R., et al., 2011. Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. Psychiatry Res. 191, 174–181.
Kempton, M.J., Geddes, J.R., Ettinger, U., Williams, S.C., Grasby, P.M., 2008. Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. Arch. Gen. Psychiatry 65, 1017–1032.
Knerr, S., Personnaz, L., Dreyfus, G., 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Fogelman-Soulie, Herault (Ed.), Neurocomputing: Algorithms, Architectures and Applications, NATO ASI. Springer.
Koo, M.S., Levitt, J.J., Salisbury, D.F., Nakamura, M., Shenton, M.E., McCarley, R.W., 2008. A cross-sectional and longitudinal magnetic resonance imaging study of cingulate gyrus gray matter volume abnormalities in first-episode schizophrenia and first-episode affective psychosis. Arch. Gen. Psychiatry 65, 746–760.
Koutsouleris, N., Meisenzahl, E.M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., et al., 2009. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch. Gen. Psychiatry 66, 700–712.
Leonard, C.M., Kuldau, J.M., Breier, J.I., Zuffante, P.A., Gautier, E.R., Heron, D.C., et al., 1999. Cumulative effect of anatomical risk factors for schizophrenia: an MRI study. Biol. Psychiatry 46, 374–382.
Liu, Y., Teverovskiy, L., Carmichael, O., Kikinis, R., Shenton, M., Carter, C.S., et al., 2004. Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification. Lect. Notes Comput. Sci. 3216, 393–401.
McDonald, C., Zanelli, J., Rabe-Hesketh, S., Ellison-Wright, I., Sham, P., Kalidindi, S., et al., 2004. Meta-analysis of magnetic resonance imaging brain morphometry studies in bipolar disorder. Biol. Psychiatry 56, 411–417.
McDonald, C., Bullmore, E., Sham, P., Chitnis, X., Suckling, J., MacCabe, J., et al., 2005. Regional volume deviations of brain structure in schizophrenia and psychotic bipolar disorder: computational morphometry study. Br. J. Psychiatry 186, 369–377.
Moore, G.J., Bebchuk, J.M., Wilds, I.B., Chen, G., Manji, H.K., 2000. Lithium-induced increase in human brain grey matter. Lancet 356, 1241–1242.
Nakamura, K., Kawasaki, Y., Suzuki, M., Hagino, H., Kurokawa, K., Takahashi, T., et al., 2004. Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects. Schizophr. Bull. 30, 393–404.
Nieuwenhuis, M., van Haren, N.E., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. NeuroImage 61, 606–612.
Olabi, B., Ellison-Wright, I., McIntosh, A.M., Wood, S.J., Bullmore, E., Lawrie, S.M., 2011. Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. Biol. Psychiatry 70, 88–96.
Pardo, P.J., Georgopoulos, A.P., Kenny, J.T., Stuve, T.A., Findling, R.L., Schulz, S.C., 2006. Classification of adolescent psychotic disorders using linear discriminant analysis. Schizophr. Res. 87, 297–306.
Pohl, K.M., Sabuncu, M.R., 2009. A unified framework for MR based disease classification. Inf. Process. Med. Imaging 21, 300–313.
Qiu, A., Vaillant, M., Barta, P., Ratnanather, J.T., Miller, M.I., 2008. Region-of-interest-based analysis with application of cortical thickness variation of left planum temporale in schizophrenia and psychotic bipolar disorder. Hum. Brain Mapp. 29, 973–985.

Radua, J., Canales-Rodríguez, E.J., Pomarol-Clotet, E., Salvador, R., 2013. Validity of modulation and optimal settings for advanced voxel-based morphometry. NeuroImage. http://dx.doi.org/10.1016/j.neuroimage.2013.07.084 (in press).

Rimol, L.M., Hartberg, C.B., Nesvåg, R., Fennema-Notestine, C., Hagler Jr., D.J., Pung, C.J., et al., 2010. Cortical thickness and subcortical volumes in schizophrenia and bipolar disorder. Biol. Psychiatry 68, 41–50.

Rimol, L.M., Nesvåg, R., Hagler Jr., D.J., Bergmann, O., Fennema-Notestine, C., Hartberg, C.B., et al., 2012. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. Biol. Psychiatry 71, 552–560.

Scheewe, T.W., van Haren, N.E., Sarkisyan, G., Schnack, H.G., Brouwer, R.M., de Glint, M., et al., 2012. Exercise therapy, cardiorespiratory fitness and their effect on brain volumes: a randomised controlled trial in patients with schizophrenia and healthy controls. Eur. Neuropsychopharmacol. 23, 675–685.

Selvaraj, S., Arnone, D., Job, D., Stanfield, A., Farrow, T.F., Nugent, A.C., Scherk, H., Gruber, O., Chen, X., Sachdev, P.S., Dickstein, D.P., Malhi, G.S., Ha, T.H., Ha, K., Phillips, M.L., McIntosh, A.M., 2012. Grey matter differences in bipolar disorder: a meta-analysis of voxel-based morphometry studies. Bipolar Disord. 14, 135–145.

Shepherd, A.M., Laurens, K.R., Matheson, S.L., Carr, J.V., Green, M.J., 2012. Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. Neurosci. Biobehav. Rev. 36, 1342–1356.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 2, 420–428.

Skre, I., Onstad, S., Torgersen, S., Kringlen, E., 1991. High interrater reliability for the Structured Clinical Interview for DSM-III-R Axis I (SCID-I). Acta Psychiatr. Scand. 84, 167–173.

Smieskova, R., Fusar-Poli, P., Allen, P., Bendfeldt, K., Stieglitz, R.D., Drewe, J., et al., 2009. The effects of antipsychotics on the brain: what have we learnt from structural imaging of schizophrenia? — a systematic review. Curr. Pharm. Des. 15, 2535–2549.

Spitzer, R.L., Endicott, J., Robins, E., 1978. Research diagnostic criteria: rationale and reliability. Arch. Gen. Psychiatry 35, 773–782.

Steiner, J.L., Tebes, J.K., Sledge, W.H., Walker, M.L., 1995. A comparison of the structured clinical interview for DSM-III-R and clinical diagnoses. J. Nerv. Ment. Dis. 183, 365–369.

Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., et al., 2011. Classification of first-episode schizophrenia patients and healthy subjects by automated MRI measures of regional brain volume and cortical thickness. PLoS One 6, e21047.

van den Heuvel, M.P., Sporns, O., Collin, G., Scheewe, T., Mandl, R.C., Cahn, W., Goñi, J., Hulshoff Pol, H.E., Kahn, R.S., 2013. Abnormal rich club organization and functional brain dynamics in schizophrenia. JAMA Psychiatry 70, 783–792.

van der Schot, A.C., Vonk, R., Brouwer, R.M., van Baal, G.C., Brans, R.G., van Haren, N.E., et al., 2010. Genetic and environmental influences on focal brain density in bipolar disorder. Brain 133, 3080–3092.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10, 988–999.