Published in final edited form as:

Biol Psychiatry Cogn Neurosci Neuroimaging. 2016 March 1; 1(2): 186–194. doi:10.1016/j.bpsc.2016.01.001.

Individualized Prediction and Clinical Staging of Bipolar Disorders using Neuroanatomical Biomarkers

Benson Mwangi, Ph.D.¹, Mon-Ju Wu, Ph.D.¹, Bo Cao, Ph.D.¹, Ives C. Passos, M.D., Ph.D.¹, Luca Lavagnino, M.D., Ph.D.¹, Zafer Keser, M.D.², Giovana B. Zunta-Soares, M.D.¹, Khader M. Hasan, Ph.D.³, Flavio Kapczinski, M.D., Ph.D.^{4,5}, and Jair C. Soares, M.D., Ph.D.¹

¹UT Center of Excellence on Mood Disorders, The University of Texas Health Science Center Houston, Houston, TX, USA

²Department of Physical Medicine and Rehabilitation and TIRR Memorial Hermann Neuro-Recovery Research Center, The University of Texas Health Science Center Houston, Houston, TX, USA

³Department of Diagnostic and Interventional Radiology, The University of Texas Health Science Center Houston, Houston, TX, USA

⁴Bipolar Disorder Program, Laboratory of Molecular Psychiatry, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, RS, Brazil

⁵Graduation Program in Psychiatry and Department of Psychiatry, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

Abstract

Background—Neuroanatomical abnormalities in Bipolar disorder (BD) have previously been reported. However, the utility of these abnormalities in distinguishing individual BD patients from Healthy controls and stratify patients based on overall illness burden has not been investigated in a large cohort.

Methods—In this study, we examined whether structural neuroimaging scans coupled with a machine learning algorithm are able to distinguish individual BD patients from Healthy controls in a large cohort of 256 subjects. Additionally, we investigated the relationship between machine learning predicted probability scores and subjects' clinical characteristics such as illness duration and clinical stages. Neuroimaging scans were acquired from 128 BD patients and 128 Healthy controls. Gray and white matter density maps were obtained and used to 'train' a relevance vector

Corresponding Author: Benson Mwangi, PhD, Department of Psychiatry & Behavioral Sciences, The University of Texas Health Science Center, 1941 East Road, Houston, TX, 77054, USA, Phone: +17134862686, Fax: +17134862553, benson.irungu@uth.tmc.edu.

FINANCIAL DISCLOSURES

JCS has participated in research funded by Forest, Merck, BMS, GSK and has been a speaker for Pfizer and Abbott. All other authors report no biomedical financial interests or potential conflicts of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

machine (RVM) learning algorithm which was used to distinguish individual patients from Healthy controls.

Results—The RVM algorithm distinguished patients from Healthy controls with 70.3 % accuracy (74.2 % specificity, 66.4 % sensitivity, chi-square $p < 0.005$) using white matter density data and 64.9 % accuracy (71.1 % specificity, 58.6 % sensitivity, chi-square $p < 0.005$) with gray matter density. Multiple brain regions – largely covering the fronto – limbic system were identified as ‘most relevant’ in distinguishing both groups. Patients identified by the algorithm with high certainty (a high probability score) – belonged to a subgroup with more than ten total lifetime manic episodes including hospitalizations (late stage).

Conclusions—These results indicate the presence of widespread structural brain abnormalities in BD which are associated with higher illness burden – which points to neuroprogression.

Keywords

Bipolar Disorders; Manic Episodes; Clinical Staging; Neuroimaging; Machine Learning. Big Data

INTRODUCTION

Bipolar disorder (BD) is one of the most debilitating illnesses with an approximate lifetime prevalence of 4–5% in the general population (1). In the last two decades, neuroimaging studies have extensively reported volumetric abnormalities in patients with BD as compared to healthy controls. Specifically, gray matter density reductions have been reported in the orbitofrontal cortex, superior temporal gyrus, and Insula (2–4). In addition, white matter density reductions have also been reported in the corpus callosum, cingulate gyrus and prefrontal white matter (5, 6). Noticeably, a high number of total lifetime manic episodes have been associated with greater gray matter density reductions in the prefrontal brain regions, cerebellum and the ventricular systems (7–10). However, while these studies have undoubtedly offered significant insights into neuroanatomical abnormalities of BD – subsequent findings have not been translated into *objective* and clinically useful biomarkers. Markedly, a significant first step in realizing this goal is the ability to use neuroimaging scans and associated clinical measurements to objectively distinguish individual BD patients from healthy controls – as well as discern clinically relevant biological pathways and brain circuitries as hypothesized elsewhere (11–14). Consequently, machine learning algorithms are ideal computational solutions able to contribute to the search of much needed objective biomarkers for the following three reasons. First, machine learning algorithms allow predictions at an *individual* subject level and therefore able to facilitate *individualized* clinical decisions (12, 14). Second, these algorithms are largely ‘multivariate’ and therefore able to analyze multiple biological measurements simultaneously – as opposed to traditional ‘univariate’ statistical methods which are only able to analyze single measurements at a time (15). Third, machine learning algorithms utilize robust ‘cross-validation’ methods to establish generalizability of results by ‘testing’ the algorithm using previously ‘unseen’ observations (16–19). A detailed overview of machine learning in psychiatric neuroimaging is given elsewhere (15–17). Other fields of psychiatric research that have benefitted from machine learning include; prediction of suicide attempts (20), prediction of treatment

response (21) – among others as summarized elsewhere (22). However, although promising these techniques are still not available for clinical use.

Several studies have recently enlisted machine learning algorithms to distinguish BD patients from Healthy controls and reported above chance (>50%) prediction accuracy. Schnack and colleagues (23) recently reported 61% prediction accuracy in discriminating BD patients from Healthy controls using structural neuroimaging scans from a total of 132 subjects. In a two cohort replication study of 80 subjects, Rocha-rego and colleagues (24) reported prediction accuracies ranging (72 – 73%). Most recently, a multicenter study reported prediction accuracies of 62% and 87.6% respectively using neuroimaging scans acquired from two centers – each with a total of 58 subjects (25). However, although these studies have undoubtedly made significant contributions in establishing the predictive validity of neuroimaging scans in BD, multiple research questions still remain unanswered. First, predictive results (specificity and sensitivity) have not been established using large samples. Second, the utility of these computational algorithms in supporting clinically relevant applications (e.g. patient stratification or clinical staging) has not been fully explored.

Therefore, the three main objectives of the current study were; 1) Establish the utility of structural T_1 -weighted scan data together with a machine learning algorithm in distinguishing BD patients from Healthy controls. 2) Elucidate gray and white matter neuroanatomical characteristics ‘most relevant’ in distinguishing BD patients from Healthy controls. 3) As a *post-hoc* test, investigate the relationship between machine learning algorithm’s prediction outputs or probability scores and individual subjects’ clinical stages. Subjects were assigned multiple clinical stages (Healthy control, BD-I Early stage, BD-I Intermediate, BD-I Late stage and BD type II) based on the International Society for Bipolar Disorders (ISBD) task report on staging systems (26). We hypothesized that BD patients predicted by the machine learning algorithm with high certainty (high probability scores) will belong to the BD-I Late stage subgroup.

METHOD

Subjects

This study was approved by the local Institutional review board (IRB) and written informed consent obtained from all subjects. Study participants included 128 BD patients with Diagnostic and Statistical Manual, 4th edition (DSM-IV) diagnosis and 128 demographically matched Healthy controls as summarized in Table 1. A diagnosis of BD was established by a research psychiatrist using the Structural Criteria for Diagnosis (SCID) interview (27). Demographically matched Healthy controls were also recruited and did not have first degree relatives with axis I DSM-IV psychiatric disorders. Participants with a history of head trauma, neurological disorders and current medical condition such as active liver disease or kidney problems were excluded from the study. Patients’ current mood status was evaluated using the Young Mania Rating Scale (YMRS) (28) and the 21-item Hamilton Depression Rating Scale (HAMD) (29) (see Table 1).

Image acquisition and preprocessing

All structural neuroimaging scans were acquired using a 1.5 T magnetic resonance imaging (MRI) Philips scanner (Philips Medical System, Andover, MA) using an axial three-dimensional T_1 -weighted fast field echo sequence with the following parameters. Repetition time (TR) = 24 ms, echo time (TE) = 5 ms, flip angle = 40°, field of view (FOV) = 256 mm, Slice thickness = 1 mm, matrix size = 256x256 and 150 slices. All scans were visually inspected to rule out any gross artefacts and spatially normalized into a template to allow inter-subject statistical comparisons (30). The spatial normalization step involved creating a study-specific template as implemented in the statistical parametric mapping (SPM8) toolbox (30). Specifically, this step was implemented as follows. First, all T_1 -weighted scans were segmented into three different tissue types (gray matter, white matter and cerebrospinal fluid). Second, a study specific template was created by averaging subjects' tissue probability maps using the diffeomorphic anatomical registration through exponential lie algebra (DARTEL) algorithm (30). Third, tissue segmentations attained in the first step were aligned onto the study specific template (30). The DARTEL algorithm is commonly used in voxel-based neuroimaging studies – including those using machine learning to generate a study-specific template (12, 31, 32). Resulting spatially normalized (unmodulated) gray and white matter scans - which we refer to as gray or white matter density were resampled into a $3 \times 3 \times 3$ voxel size. The 'modulation' step (30) – which entails scaling the spatially normalized scans using a 'jacobian matrix' obtained from the registration step was avoided as recent studies have reported empirical evidence pointing to loss of sensitivity following the 'modulation' operation (33, 34). Lastly, gray matter density and white matter density data were used as predictors in the ensuing machine learning analyses with 59,598 and 37,436 features (predictor variables) respectively.

Machine learning

We assumed a diagnostic classification problem with N number of observations (subjects) – which we represent as $\{x_i, t_i\}_{i=1}^N$, where x_i represents individual subjects' scan data (e.g. gray matter density) and t_i stands for categorical target labels (0 – Healthy control, 1 – BD). The main objective here was to 'train' a machine learning algorithm able to estimate $p(t|X)$ – which is a *probability of a subject belonging to either BD or Healthy control groups given previously 'unseen' subjects' scan data*. In the current study, the relevance vector machine (RVM) algorithm (35) was used to estimate $p(t|X)$ as detailed in the supplementary materials and elsewhere (12, 35, 36). RVM was chosen as it offers several benefits compared to equivalent machine learning algorithms (e.g. support vector machines). First, RVM estimates 'nuisance' parameters automatically and no requirement for 'user defined' parameters when used with a linear kernel mapping function (35, 37). Second, RVM solutions result into probabilistic scores – which may be used in patient stratification as opposed to categorical or binary outputs (12). This is in line with recent psychiatric studies calling for development and validation of staging systems of BD that place individual subjects on a probabilistic continuum of increasing disease severity (38, 39). In the current study, the RVM algorithm was implemented using a MATLAB (The Mathworks, Inc) toolbox (35) and in house custom routines as detailed elsewhere (12, 40).

To establish generalization ability (sensitivity and specificity) of the algorithm in predicting ‘previously unseen’ subjects – training and testing data were separated using a *leave-one-out* cross-validation (LOOCV) approach (see Figure S1 of supplementary materials). Briefly, the LOOCV process entailed training RVM with all subjects’ data except one – while the ‘left-out’ subject was used in testing the algorithm. This procedure was repeated until all subjects were left out for algorithm ‘testing’ at-least once as detailed elsewhere (16). A recent study comparing cross-validation methods (e.g. *K*-fold) reported that the LOOCV approach becomes stable (low variance) as the sample size increases – particularly in studies with large samples such as the current study (41). The algorithm performance in distinguishing BD patients from Healthy controls was evaluated using commonly used parameters such as accuracy, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV) and receiver operating characteristic (ROC) curves as detailed elsewhere (16, 42).

Feature Selection

Noticeably, the number of predictor variables or features (e.g. gray matter density) greatly exceeded the number of observations (subjects) – leading to the ‘*small-n-large-p*’ or ‘*curse-of-dimensionality*’ problem (17). Therefore, to mitigate this problem a feature selection calculation was implemented to remove redundant predictor variables as suggested elsewhere (40, 43–45). Specifically, a univariate analysis of variance (ANOVA) feature selection filter was performed within a 10-fold cross-validation to evaluate predictors with maximal difference among patients and Healthy control groups and ANOVA *p*-value threshold with the best prediction accuracy evaluated using the RVM algorithm. The feature selection process was implemented using a 10-fold cross-validation – using training data only to avoid circularity also known as ‘double dipping’ as explored in details elsewhere (12, 17). Univariate feature selection filters have previously been used in neuroimaging machine learning (12, 43, 44, 46, 47). A detailed discussion of feature selection in psychiatric neuroimaging is given elsewhere (17, 48, 49).

Relationship between Predicted Probability scores and Clinical stages and illness duration

Individual subjects were assigned a clinical stage label (Healthy controls, BD-I Early, BD-I Intermediate, BD-I Late and BD-II) – based on reported total lifetime manic episodes and hospitalizations as described by the ISBD task force report on staging systems in BD (26). Specifically, BD-I patients reporting < 3 total lifetime manic episodes without hospitalization were assigned to the BD-I Early stage subgroup. In contrast, patients reporting >10 total lifetime manic episodes including hospitalizations were assigned to the BD-I Late stage subgroup. All other BD-I (Bipolar disorder type I) patients were assigned to the BD-I intermediate stage subgroup. Importantly, BD type II patients (BD-II) were not stratified into early, intermediate and late stages - as these patients do not report full manic episodes. Therefore, all subjects were assigned into four categories namely; Healthy controls (N=128), BD-I Early stage (N=17), BD-I Intermediate stage (N=33), BD-I Late stage (N=21) and BD-II (N=32). A total of 25 patients had missing data on total lifetime manic episodes and were not included in the *post-hoc* analyses. A *post-hoc* statistical test was performed to evaluate the null hypothesis of no difference in predicted subjects’ probability scores between clinical stages (e.g. BD-I Late stage) and Healthy controls. We hypothesized that probability scores between BD-I Late stage patients and Healthy controls will differ

significantly as compared to the contrast between BD-I Early stage patients and Healthy controls. Lastly, a partial correlation test was performed to assess the relationship between predicted probability scores and total illness duration while controlling for both age and gender.

Machine learning model with both gray and white matter maps

To establish whether a combination of gray and white matter density maps would improve prediction accuracy, both gray and white matter density maps were integrated into a single feature vector through a 'side by side concatenation'. The resulting predictor variables were entered into the RVM algorithm which was implemented through similar LOOCV and feature selection approaches, as above and prediction accuracy, specificity and sensitivity calculated.

RESULTS

The patient and Healthy control groups did not differ significantly by age, gender and ethnicity as summarized in Table 1. The RVM algorithm 'trained' with white matter density data distinguished BD patients from Healthy controls with accuracy = 70.4%, specificity = 74.2%, sensitivity = 66.4%, PPV = 71.4.0%, NPV = 68.5% and area under receiver operating characteristic curve (AUROC) = 0.72 as shown in Figure 1a. These predictions were significant at $p < 0.005$ through a chi-square calculation between actual and RVM predicted labels. Individual subjects' predicted probabilities are shown in Table S1 of supplementary materials. In contrast, the RVM algorithm 'trained' using gray matter density distinguished patients from Healthy controls with accuracy = 64.9%, specificity = 71.1%, sensitivity = 58.6 %, PPV = 67.1%, NPV = 63.4%, AUROC = 0.7 and chi-square $p < 0.005$ as shown in Figure 1b. Incorporating a 5% population prevalence rate of BD (1) by following Altman and Bland's formula (50) (see supplementary materials) resulted in a PPV and NPV rate of 12 % and 98 % respectively. White matter brain regions most relevant in distinguishing BD patients from Healthy controls and identified during the feature selection step included; Cerebellum, brainstem, cingulate gyrus and corpus callosum as summarized in Table 2 and shown in Figure 2a. On the other hand, gray matter brain regions most relevant in distinguishing both groups and also identified through the feature selection step included; medial orbitofrontal cortex, superior frontal gyrus, temporal lobes and the midbrain as summarized in Table 2 and shown in Figure 2b.

The RVM trained using combined features from both gray and white matter density maps distinguished individual BD patients from Healthy controls with accuracy = 64 %, specificity = 70 % and sensitivity = 59 %.

RVM Predicted probability scores between Healthy controls and BD-I Early stage patients did not differ significantly ($p > 0.05$) – indicating both groups were largely indistinguishable. In contrast, predicted probability scores between Healthy controls and BD-I Intermediate as well as BD-I Late stage patients differed significantly ($p = 0.001$ and $p < 0.001$) respectively as shown in Figure 3a. This is an indication that the RVM algorithm was able to distinguish Healthy controls from BD-I Intermediate and BD-I Late stage patients whilst on the contrary the algorithm was unable to distinguish Healthy controls from BD-I Early stage patients.

There was a significant difference in cerebellar white matter density between Healthy controls and BD-I Intermediate stage ($p < 0.05$) and BD-I Late stage ($p < 0.001$) and not BD-I Early stage as shown in Figure 3b. Demographic details of patients stratified in different BD stages are summarized in Table 3.

There was no significant correlation between patients' predicted probability scores and illness duration. This calculation was performed using partial correlations analysis whilst accounting for age and gender ($r = -0.141$, $p = 0.173$). All BD patients except 39 reported a previous history of taking psychotropic medications. There was no significant relationship between RVM misclassified subjects and history of taking psychotropic medications (White matter $\chi^2 = 0.1$, $p = 0.75$, Gray matter $\chi^2 = 0.1$, $p = 0.75$). Similarly, there was no significant association between BD patients' age and predicted probabilities (White matter $r = 0.1$, $p = 0.1$, Gray matter $r = 0.1$, $p = 0.2$) as well as healthy controls' predicted probabilities and age (White matter $r = 0.08$, $p = 0.4$, Gray matter $r = 0.1$, $p = 0.1$).

The relationship between RVM predicted probability scores and current mood status was also examined. Predicted probabilities and subjects' current mood status (euthymic, depressed manic, hypomanic and mixed) did not differ significantly ($F = 1.4$, $p = 0.23$) also detailed in Table S2 of supplementary materials.

DISCUSSION

We report a machine learning algorithm able to distinguish BD patients from Healthy controls with 70.3% and 64.9% accuracy using white matter and gray matter density respectively. These predictions were above chance level (chi-square p -value < 0.005) and established using a large sample. In addition, we report prediction results comparable to previous psychiatric neuroimaging machine learning studies in BD – which have reported accuracy results ranging 61% – 87% (23–25). Furthermore, we report a significant relationship between subjects' predicted probability scores and clinically stratified stages of BD (Early, Intermediate and Late stages). Specifically, patients predicted by RVM with high certainty (high probability scores) belonged to the BD-I Late stage category and noticeably with a significantly lower cerebellar white matter density as compared to Healthy controls. To the best of our knowledge, this is the largest neuroimaging machine learning study in BD – as well as the first to report associations of objectively derived probability scores and a clinically relevant stratification or staging system in BD. The algorithm's ability to distinguish early, intermediate and late stage patients is noteworthy, as it may allow us to elucidate the timing of neuroprogression as well as guide early interventions and neuroprotection studies.

The feature selection process discerned a circuitry of abnormalities in BD largely covering the fronto-limbic brain circuitry which has also been associated with BD in previous studies. Importantly, feature selection was performed through a 'nested' 10-fold cross-validation step using training data only (without the 'left-out' testing subject) to avoid circularity or *double dipping* a detailed elsewhere (12, 17). Several studies have reported gray matter density reductions in the temporal lobes (51, 52), Inferior frontal gyrus (52, 53), paracentral lobule (54) and superior frontal gyrus (52) – as also detailed in the ensuing meta-analytic studies

(3, 4). Noticeably, our findings on gray matter density reductions in the midbrain agree with previous findings as the midbrain is known to host dopaminergic and serotonergic neuronal systems through the substantia nigra and dorsal raphe nucleus and alterations in this system have been associated with BD (55–57). Indeed, these observations have also been supported by positron emission tomography studies which have reported lower midbrain serotonin transporter binding in patients with BD as compared to Healthy controls (58). The white matter model identified widespread white matter density reductions as ‘most relevant’ in distinguishing BD patients from Healthy controls. These ‘most relevant’ regions were selected through a feature subset selection process which entails selecting features that maximize prediction accuracy using algorithm ‘training data’ only as detailed elsewhere (17). These findings align with previously reported white matter abnormalities in BD in the cerebellum (24), brainstem (59) and corpus callosum (6, 24, 60). All together, we observed widespread gray and white matter density reductions – largely covering the fronto-limbic system which is known for emotion regulation and widely reported in neuroanatomical studies of BD (57). The algorithm’s high performance in identifying BD-II patients from healthy controls highly suggests that there may be substantial overlap of neuroanatomical abnormalities among BD-I and BD-II subtypes.

An association between RVM predicted probability scores and clinically stratified BD stages was observed. Precisely, BD patients predicted with a higher certainty (high probability scores) reported more total lifetime manic episodes and prior hospitalizations (BD-I Late stage). On the contrary, early stage (BD-I Early) patients were predicted with less certainty (low probability scores) and largely indistinguishable from Healthy controls. It is noteworthy that, the machine learning algorithm discerned this pattern without being exposed to any clinical characteristics other than subjects’ diagnostic labels (BD -1, Healthy controls - 0). Indeed, this is a strong indication that predicted probability scores may qualify as a composite bio-signature able to place individual subjects on a continuum of increasing disease severity as well as encapsulate varied patterns of neuroanatomical abnormalities as also hypothesized elsewhere (38). Illness duration was not associated with probability scores which may suggest that the brain signature presented here may represent lifetime severity of illness as opposed to cumulative illness effects.

We argue that the difference in probability scores as well as average cerebellar white matter density between Healthy controls, BD-I Intermediate and BD-I Late stage patients may be an indication of neuroprogression through the patient group as previously postulated (26, 61, 62). For example, DelBello and colleagues (10) observed a similar pattern in the cerebellum when comparing Healthy controls, first episode BD patients and patients with multiple episodes – with the latter group showing reduced cerebellar volume. A potential mechanistic interpretation of the observed progressive brain changes is increased microglia activation which leads to the release of inflammatory markers potentially translating into tissue loss which is reflected in the neuroimaging scan (63).

Limitations of the current study should be noted. All patients’ except 39 had a history of taking multiple types of psychotropic medications but there was no association between predicted labels and a history of psychotropic mediations. Only six subjects were currently on lithium. A majority of BD patients were females (72%) and future studies should confirm

this hypothesis using a sample with equal males and females. However, gender was not considered a confounding factor in the machine learning analyses as male and female subjects were balanced across both groups. Probability scores among BD stages and healthy controls in male subjects only resulted in a similar pattern of findings (see supplementary materials Table S3 and Figure S1). Although cross-validation was used in the current study to examine results generalizability an independent cohort was not included but work is ongoing in our group to acquire another cohort and replicate findings. While we report a high prediction accuracy using white matter density (70.3%) and subsequently observe overall white matter density reductions in multiple regions (e.g. cerebellum) – findings should be interpreted with caution as the biological properties of T_1 -weighted scan derived white matter density are unclear. For example it may not be clear whether reduced white matter density represents altered axonal density packing or demyelination. However, future studies examining individualized prediction and clinical staging in BD will utilize diffusion weighted or myelin water fraction imaging – which are more suitable in investigating white matter related pathological anomalies. Nonetheless, neuroanatomical regions identified in the current study may be used in future as ‘landmarks’ in white matter connectivity studies. Ongoing work in our group has been designed to examine these hypotheses. Integrating both gray and white matter density maps did not lead to improved prediction accuracy (64%) as compared to white matter alone (70.3%) an outcome which was also observed in a recent BD vs Healthy controls machine learning study (25). Notably though, whilst the current study reports prediction accuracy ranging (64.9–70.3%), work is ongoing in our group to integrate other biological data such as neurocognition, diffusion tensor imaging measurements which may potentially improve the prediction accuracy. Following Altman and Bland’s (50) work PPV value reduced to 12% whilst the NPV value increased to 98%. This is in line with their conclusion that the rarer the disease the more certainty that a negative test points towards no disease and therefore less certainty that a positive result points to presence of disease.

Whilst it is also possible to correlate RVM predicted probability scores with total number of manic episodes, the latter was highly skewed (skewness = 0.48) and therefore grouping patients into (early, intermediate and late stages) was adopted. A recent study in major depression observed a similar trend and adopted a similar categorization approach (64). The current study included BD patients with comorbidities similar to other studies as well as in clinical practice and may affect interpretation of results. However, we did not find any relationship between BD stages and comorbidities. A large sample size acquired from a single scanning center is a major strength of the current study. Lastly, current findings on neuroprogression in BD patients need to be confirmed using a longitudinal design which may give more insights as compared to a cross-section design as presented here. Indeed, a very recent longitudinal study with a relatively small sample (31 BD patients) reported prefrontal cortex gray matter volume reductions in BD patients experiencing multiple manic episodes as compared to those without manic episodes over a period of six years and these findings corroborate our results (65). There are two potential applications of the method presented in this study 1) Individualized prediction of patients likely to follow a neuroprogressive course – and therefore elicit timely therapeutic interventions. 2)

Individualized identification of patients likely to benefit from functional remediation therapies.

In summary, we report accurate predictions in distinguishing individual BD patients from Healthy controls using structural neuroimaging scan data. We also report associations between subjects' predicted probability scores and clinically relevant stratifications or illness stages. Neuroimaging scans coupled with advanced machine learning as reported in this study – may be relevant in stratifying psychiatric patients into clinically relevant subgroups (e.g. illness stages) – eventually translating into individual patient tailored therapeutic interventions and the ultimate realization of personalized medicine.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported in part by NIMH grant R01 085667, The Dunn Foundation and the Pat Rutherford, Jr. Endowed Chair in Psychiatry to JCS.

References

1. Merikangas KR, Akiskal HS, Angst J, Greenberg PE, Hirschfeld RM, Petukhova M, et al. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of general psychiatry*. 2007; 64:543–552. [PubMed: 17485606]
2. Arnone D, Cavanagh J, Gerber D, Lawrie S, Ebmeier K, McIntosh A. Magnetic resonance imaging studies in bipolar disorder and schizophrenia: meta-analysis. *The British Journal of Psychiatry*. 2009; 195:194–201. [PubMed: 19721106]
3. Selvaraj S, Arnone D, Job D, Stanfield A, Farrow TF, Nugent AC, et al. Grey matter differences in bipolar disorder: a meta-analysis of voxel-based morphometry studies. *Bipolar disorders*. 2012; 14:135–145. [PubMed: 22420589]
4. Bora E, Fornito A, Yücel M, Pantelis C. Voxelwise meta-analysis of gray matter abnormalities in bipolar disorder. *Biological psychiatry*. 2010; 67:1097–1105. [PubMed: 20303066]
5. Emsell L, Langan C, Van Hecke W, Barker GJ, Leemans A, Sunaert S, et al. White matter differences in euthymic bipolar I disorder: a combined magnetic resonance imaging and diffusion tensor imaging voxel-based study. *Bipolar disorders*. 2013; 15:365–376. [PubMed: 23621705]
6. Brambilla P, Nicoletti MA, Sassi RB, Mallinger AG, Frank E, Kupfer DJ, et al. Magnetic resonance imaging study of corpus callosum abnormalities in patients with bipolar disorder. *Biological psychiatry*. 2003; 54:1294–1297. [PubMed: 14643097]
7. Ekman C, Lind J, Ryden E, Ingvar M, Landén M. Manic episodes are associated with grey matter volume reduction—a voxel-based morphometry brain analysis. *Acta Psychiatrica Scandinavica*. 2010; 122:507–515. [PubMed: 20712826]
8. Lyoo IK, Kim MJ, Stoll AL, Demopulos CM, Parow AM, Dager SR, et al. Frontal lobe gray matter density decreases in bipolar I disorder. *Biological psychiatry*. 2004; 55:648–651. [PubMed: 15013835]
9. Strakowski SM, DelBello MP, Zimmerman ME, Getz GE, Mills NP, Ret J, et al. Ventricular and periventricular structural volumes in first-versus multiple-episode bipolar disorder. *American Journal of Psychiatry*. 2002; 159:1841–1847. [PubMed: 12411217]
10. DelBello MP, Strakowski SM, Zimmerman ME, Hawkins JM, Sax KW. MRI analysis of the cerebellum in bipolar disorder: a pilot study. *Neuropsychopharmacology*. 1999; 21:63–68. [PubMed: 10379520]

11. Oquendo M, Baca-Garcia E, Artes-Rodriguez A, Perez-Cruz F, Galfalvy H, Blasco-Fontecilla H, et al. Machine learning and data mining: strategies for hypothesis generation. *Molecular psychiatry*. 2012; 17:956–959. [PubMed: 22230882]
12. Mwangi B, Ebmeier KP, Matthews K, Steele JD. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain*. 2012; 135:1508–1521. [PubMed: 22544901]
13. Krystal JH, State MW. Psychiatric disorders: diagnosis to therapy. *Cell*. 2014; 157:201–214. [PubMed: 24679536]
14. Linden DE. The challenges and promise of neuroimaging in psychiatry. *Neuron*. 2012; 73:8–22. [PubMed: 22243743]
15. Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*. 2012; 36:1140–1152. [PubMed: 22305994]
16. Johnston B, Mwangi B, Matthews K, Coghill D, Steele J. Predictive classification of individual magnetic resonance imaging scans from children and adolescents. *European child & adolescent psychiatry*. 2013; 22:733–744. [PubMed: 22930323]
17. Mwangi B, Tian TS, Soares JC. A Review of Feature Reduction Techniques in Neuroimaging. *Neuroinformatics*. 2013:1–16. [PubMed: 23224666]
18. Mwangi B, Matthews K, Steele JD. Prediction of illness severity in patients with major depression using structural MR brain scans. *Journal of Magnetic Resonance Imaging*. 2012; 35:64–71. [PubMed: 21959677]
19. Reiss PT. Cross-validation and hypothesis testing in neuroimaging: an irenic comment on the exchange between Friston and Lindquist et al. *NeuroImage*. 2015
20. Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA psychiatry*. 2015; 72:49–57. [PubMed: 25390793]
21. Hahn T, Kircher T, Straube B, Wittchen H-U, Konrad C, Ströhle A, et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. *JAMA psychiatry*. 2015; 72:68–74. [PubMed: 25409415]
22. Wolfers T, Buitelaar JK, Beckmann C, Franke B, Marquand AF. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews*. 2015
23. Schnack HG, Nieuwenhuis M, van Haren NE, Abramovic L, Scheewe TW, Brouwer RM, et al. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *Neuroimage*. 2014; 84:299–306. [PubMed: 24004694]
24. Rocha-Rego V, Jogia J, Marquand A, Mourao-Miranda J, Simmons A, Frangou S. Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach. *Psychological medicine*. 2014; 44:519–532. [PubMed: 23734914]
25. Redlich R, Almeida JR, Grotegerd D, Opel N, Kugel H, Heindel W, et al. Brain Morphometric Biomarkers Distinguishing Unipolar and Bipolar Depression: A Voxel-Based Morphometry–Pattern Classification Approach. *JAMA psychiatry*. 2014; 71:1222–1230. [PubMed: 25188810]
26. Kapczinski F, Magalhães P, Balanzá-Martínez V, Dias V, Frangou S, Gama C, et al. Staging systems in bipolar disorder: an International society for bipolar disorders task force report. *Acta Psychiatrica Scandinavica*. 2014; 130:354–363. [PubMed: 24961757]
27. First M, Spitzer R, Gibbon M, Williams J. Structured Clinical Interview for DSM-IV Axis I Disorders, Patient Edition (SCID-P). New York, New York State Psychiatric Institute. Biometrics Research. 1996
28. Young R, Biggs J, Ziegler V, Meyer D. A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry*. 1978; 133:429–435. [PubMed: 728692]
29. Hamilton M. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*. 1960; 23:56.

30. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage*. 2007; 38:95–113. [PubMed: 17761438]
31. Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, et al. Prognostic prediction of therapeutic response in depression using high-field MR imaging. *Neuroimage*. 2011; 55:1497–1503. [PubMed: 21134472]
32. Johnston BA, Mwangi B, Matthews K, Coghill D, Konrad K, Steele JD. Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification. *Human brain mapping*. 2014; 35:5179–5189. [PubMed: 24819333]
33. Klauser P, Fornito A, Lorenzetti V, Davey CG, Dwyer DB, Allen NB, et al. Cortico-limbic network abnormalities in individuals with current and past major depressive disorder. *Journal of affective disorders*. 2015; 173:45–52. [PubMed: 25462395]
34. Radua J, Canales-Rodríguez EJ, Pomarol-Clotet E, Salvador R. Validity of modulation and optimal settings for advanced voxel-based morphometry. *Neuroimage*. 2014; 86:81–90. [PubMed: 23933042]
35. Tipping ME. Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*. 2001; 1:211–244.
36. Phillips CL, Bruno M-A, Maquet P, Boly M, Noirhomme Q, Schnakers C, et al. Relevance vector machine consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. *Neuroimage*. 2011; 56:797–808. [PubMed: 20570741]
37. Bishop, CM. *Pattern recognition and machine learning*. Springer; New York: 2006.
38. Berk M, Berk L, Dodd S, Cotton S, Macneil C, Daglas R, et al. Stage managing bipolar disorder. *Bipolar disorders*. 2014; 16:471–477. [PubMed: 23782499]
39. Berk M, Conus P, Lucas N, Hallam K, Malhi GS, Dodd S, et al. Setting the stage: from prodrome to treatment resistance in bipolar disorder. *Bipolar disorders*. 2007; 9:671–678. [PubMed: 17988356]
40. Mwangi B, Hasan KM, Soares JC. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach. *Neuroimage*. 2013; 75:58–67. [PubMed: 23501046]
41. Combrisson E, Jerbi K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*. 2015
42. Savitz J, Rauch SL, Drevets W. Clinical application of brain imaging for the diagnosis of mood disorders: the current state of play. *Molecular psychiatry*. 2013; 18:528–539. [PubMed: 23546169]
43. Chu C, Hsu A-L, Chou K-H, Bandettini P, Lin C. Initiative AsDN. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*. 2012; 60:59–70. [PubMed: 22166797]
44. De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*. 2008; 43:44–58. [PubMed: 18672070]
45. Kerr WT, Douglas PK, Anderson A, Cohen MS. The utility of data-driven feature selection: Re: Chu et al. 2012. *NeuroImage*. 2014; 84:1107–1110. [PubMed: 23891886]
46. Craddock RC, Holtzheimer PE, Hu XP, Mayberg HS. Disease state prediction from resting state functional connectivity. *Magnetic resonance in Medicine*. 2009; 62:1619–1628. [PubMed: 19859933]
47. Wu M-J, Wu HE, Mwangi B, Sanches M, Selvaraj S, Zunta-Soares GB, et al. Prediction of pediatric unipolar depression using multiple neuromorphometric measurements: A pattern classification approach. *Journal of psychiatric research*. 2015; 62:84–91. [PubMed: 25687738]
48. Guyon I, Elisseeff A. An introduction to variable and feature selection. *The Journal of Machine Learning Research*. 2003; 3:1157–1182.
49. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *bioinformatics*. 2007; 23:2507–2517. [PubMed: 17720704]
50. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 2: predictive values. *Bmj*. 1994; 309:102. [PubMed: 8038641]

51. Chen X, Wen W, Malhi GS, Ivanovski B, Sachdev PS. Regional gray matter changes in bipolar disorder: a voxel-based morphometric study. *Australian and New Zealand journal of psychiatry*. 2007; 41:327–336. [PubMed: 17464719]
52. Ha TH, Ha K, Kim JH, Choi JE. Regional brain gray matter abnormalities in patients with bipolar II disorder: a comparison study with bipolar I patients and healthy controls. *Neuroscience letters*. 2009; 456:44–48. [PubMed: 19429131]
53. Stanfield AC, Moorhead TWJ, Job DE, McKirdy J, Sussmann JE, Hall J, et al. Structural abnormalities of ventrolateral and orbitofrontal cortex in patients with familial bipolar disorder. *Bipolar disorders*. 2009; 11:135–144. [PubMed: 19267696]
54. Adler CM, DelBello MP, Jarvis K, Levine A, Adams J, Strakowski SM. Voxel-based study of structural changes in first-episode patients with bipolar disorder. *Biological psychiatry*. 2007; 61:776–781. [PubMed: 17027928]
55. Matthews PR, Harrison PJ. A morphometric, immunohistochemical, and in situ hybridization study of the dorsal raphe nucleus in major depression, bipolar disorder, schizophrenia, and suicide. *Journal of affective disorders*. 2012; 137:125–134. [PubMed: 22129767]
56. Chou Y-H, Lin C-L, Wang S-J, Lirng J-F, Yang K-C, Chang AC, et al. Aggression in bipolar II disorder and its relation to the serotonin transporter. *Journal of affective disorders*. 2013; 147:59–63. [PubMed: 23123132]
57. Soares, JC.; Young, AH. *Bipolar disorders: basic mechanisms and therapeutic implications*. CRC Press; 2007.
58. Oquendo MA, Hastings RS, Huang Y-y, Simpson N, Ogden RT, Hu X-z, et al. Brain serotonin transporter binding in depressed patients with bipolar disorder using positron emission tomography. *Archives of general psychiatry*. 2007; 64:201–208. [PubMed: 17283287]
59. Brambilla P, Harenski K, Nicoletti M, Mallinger AG, Frank E, Kupfer DJ, et al. MRI study of posterior fossa structures and brain ventricles in bipolar patients. *Journal of psychiatric research*. 2001; 35:313–322. [PubMed: 11684138]
60. Arnone D, McIntosh A, Chandra P, Ebmeier K. Meta-analysis of magnetic resonance imaging studies of the corpus callosum in bipolar disorder. *Acta Psychiatrica Scandinavica*. 2008; 118:357–362. [PubMed: 18644004]
61. Lin A, Reniers RL, Wood SJ. Clinical staging in severe mental disorder: evidence from neurocognition and neuroimaging. *The British Journal of Psychiatry*. 2013; 202:s11–s17. [PubMed: 23288495]
62. Reinares M, Papachristou E, Harvey P, Bonnín CM, Sánchez-Moreno J, Torrent C, et al. Towards a clinical staging for bipolar disorder: defining patient subtypes based on functional outcome. *Journal of affective disorders*. 2013; 144:65–71. [PubMed: 22862890]
63. Berk M, Kapczynski F, Andreazza A, Dean O, Giorlando F, Maes M, et al. Pathways underlying neuroprogression in bipolar disorder: focus on inflammation, oxidative stress and neurotrophic factors. *Neuroscience & biobehavioral reviews*. 2011; 35:804–817. [PubMed: 20934453]
64. Treadway MT, Waskom ML, Dillon DG, Holmes AJ, Park MTM, Chakravarty MM, et al. Illness progression, recent stress, and morphometry of hippocampal subfields and medial prefrontal cortex in major depression. *Biological psychiatry*. 2015; 77:285–294. [PubMed: 25109665]
65. Abé C, Ekman C-J, Sellgren C, Petrovic P, Ingvar M, Landén M. Manic episodes are related to changes in frontal cortex: a longitudinal neuroimaging study of bipolar disorder I. *Brain*. 2015;awv266.

Figure 1.
 ‘Confusion matrix’ and ROC curves summarizing RVM prediction results from A) white matter density with 70.3% accuracy and B) Gray matter density with 64.9 % accuracy.

Figure 2.

A) White matter brain regions identified during feature selection as most relevant in distinguishing BD patients from Healthy controls. B) Gray matter regions most relevant in identifying patients from Healthy controls. The colors Red – Blue indicate anatomical regions with reduced gray/white matter density in BD patients as compared to Healthy controls. The cerebellum was used for *post hoc* tests as it reported the largest cluster size (391 voxels). White matter brain regions include; CG – Cingulate gyrus, Corpus callosum, MB –midbrain and CR – Cerebellum. On the other hand, Gray matter brain regions include; MOFC – Medial orbitofrontal cortex and thalamus.

Figure 3.

A) Comparisons of average predicted probability scores between Healthy controls, BD Early stage, BD intermediate stage, BD late stage and BD type II patients. Analysis of variance (ANOVA) statistical test was performed among groups ($F=13.54$, $p<0.005$). *Post hoc* tests with Bonferroni multiple comparison corrections were also performed to examine differences among groups. BD I Early patients' probability scores did not differ significantly from those of Healthy controls. On the other hand BD intermediate and BD late stage probability scores were significantly different from Healthy controls (significantly higher probability scores). Probability scores were normalized by subtracting the 'chance level' score (0.5) for visualization purposes only. B) Comparisons of average cerebellar white matter density between Healthy controls and clinical stages. Similar to probability scores – white matter density between Healthy controls and BD early stage patients did not differ significantly. The progressive white matter density reductions in BD intermediate and BD late stages which were not observed in BD early as compared to Healthy controls may be an indication of neuroprogression. Analysis of Variance (ANOVA) statistical tests were performed in SPSS Version 20, IBM Inc and corrected for multiple comparisons using the Bonferroni method. White matter density values were normalized by subtracting the average value in the Healthy control group for visualization purposes only.

Table 1

Demographic and clinical details

	Healthy controls mean (SD)	BD ^c mean (SD)	P-value
Age (years)	36.33 (12.25)	37.56 (11.6)	0.4086 ^a
Female/total	84 (128)	92 (128)	0.2807 ^b
Bipolar Type	-	BD-I 96/128 BD-II 32/128	-
HAMD	0.72 (1.07)	12.92 (7.84)	P<0.0001 ^a
YMRS	0.31 (0.84)	6.9 (6.8)	P<0.0001 ^a
GAF	91.46 (6.12)	62.52 (11.63)	P<0.0001 ^a
MADRAS	0.38 (1.01)	17.18 (11.1)	P<0.0001 ^a
Currently or previously taken any psychotropic medication	-	120	-
Current mood	-	Euthymic- 34 Depressed- 64 Manic-6 Hypomanic- 7 Mixed- 14 Undetermined- 3	-

^a student t-test,^b chi-square test,

^c Comorbidities included (*n* =): 22 generalized anxiety disorder, 11 obsessive-compulsive disorder, 36 panic disorder, 25 posttraumatic stress disorder, 12 specific/simple phobia, 24 social phobia, 26 agoraphobia, 6 alcohol abuse, 1 anorexia, 4 anxiety disorder, 3 binge eating disorder, 2 bulimia.

BD - bipolar depression, HAMD- Hamilton depression rating scale, YMRS- Young mania rating scale, GAF- Global assessment of functioning, MADRAS - Montgomery asberg depression rating scale.

Table 2

Gray and white matter regions most relevant in distinguishing individual BD patients from Healthy controls.

Tissue type (Gray/white matter)	Anatomical region	Brodman area	MNI Coordinates (x, y, z)	Cluster size	RVM weight *
Gray Matter	Temporal lobe (middle temporal gyrus)	21	-63 -15 -18	500	-23.0
	Postcentral gyrus	3	-9 -39 69	496	-10.8
	Medial orbitofrontal cortex	11	-18 36 -24	407	-13.2
	Temporal lobe (middle temporal gyrus)	21	66 -30 0	263	-17.2
	Midbrain (Thalamus, medial dorsal nucleus)	-	9 -15 9	158	-15.0
White Matter	Superior frontal gyrus	8	6 66 -12	114	-10.6
	Cerebellum	-	3 -57 -30	391	-38.0
	Brainstem	-	6 -12 -12	141	-35.8
	Cingulate gyrus	23	6 -15 27	67	-27.3
	Corpus callosum	-	-3 15 21	50	-18.7

* RVM weighting factor – Negative weights indicate gray/white matter density reductions in BD

Table 3

Demographic details of different BD stages.

	BD-I Early	BD-I Intermediate	BD-I Late	BD-II	p-value
N (Total = 128)	17	33	21	32	-
Age (years)	40.1 (12.1)	36.3 (11.8)	41.1 (11.4)	37.0 (11.0)	0.39 ^a
Females/Total	13	23	16	23	0.78 ^b

^a Analysis of variance (ANOVA),

^b chi-square test.