Psychological Medicine

cambridge.org/psm

Prediction of estimated risk for bipolar disorder using machine learning and structural MRI features

Pavol Mikolas[1], Michael Marxen[1], Philipp Riedel[1], Kyra Bröckel[1], Julia Martini[1], Fabian Huth[1], Christina Berndt[1], Christoph Vogelbacher[2,3,4], Andreas Jansen[2,3,4], Tilo Kircher[2,3,4], Irina Falkenberg[2,3,4], Martin Lambert[5], Vivien Kraft[5], Gregor Leicht[5], Christoph Mulert[4,5,6], Andreas J. Fallgatter[7], Thomas Ethofer[7], Anne Rau[7], Karolina Leopold[8], Andreas Bechdolf[8], Andreas Reif[9], Silke Matura[9], Felix Bermpohl[10], Jana Fiebig[10], Thomas Stamm[10,11], Christoph U. Correll[12,13,14], Georg Juckel[15], Vera Flasbeck[15], Philipp Ritter[1], Michael Bauer[1] and Andrea Pfennig[1]

[1]Department of Psychiatry and Psychotherapy, Carl Gustav Carus University Hospital, Technische Universität Dresden, Dresden, Germany; [2]Core-Facility Brainimaging, Faculty of Medicine, University of Marburg, Marburg, Germany; [3]Department of Psychiatry, University of Marburg, Marburg, Germany; [4]Center for Mind, Brain and Behavior (CMBB), University of Marburg and Justus Liebig University Giessen, Germany; [5]Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [6]Centre for Psychiatry, Justus-Liebig University Giessen, Giessen, Germany; [7]Department of Psychiatry, Tuebingen Center for Mental Health, University of Tuebingen, Tuebingen, Germany; [8]Department of Psychiatry, Psychotherapy and Psychosomatic Medicine, Vivantes Hospital Am Urban and Vivantes Hospital Im Friedrichshain, Charité-Universitätsmedizin Berlin, Berlin, Germany; [9]Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital Frankfurt – Goethe University, Frankfurt am Main, Germany; [10]Department of Psychiatry and Psychotherapy, Charité Campus Mitte, Charité University Medicine, Berlin, Germany; [11]Department of Clinical Psychiatry and Psychotherapy, Brandenburg Medical School Theodor Fontane, Neuruppin, Germany; [12]Department of Child and Adolescent Psychiatry, Charité Universitätsmedizin Berlin, Berlin, Germany; [13]Department of Psychiatry, Northwell Health, The Zucker Hillside Hospital, Glen Oaks, NY, USA; [14]Department of Psychiatry and Molecular Medicine, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA and [15]Department of Psychiatry, Psychotherapy and Preventive Medicine, LWL University Hospital, Ruhr-University, Bochum, Germany

Original Article

Cite this article: Mikolas P *et al* (2024). Prediction of estimated risk for bipolar disorder using machine learning and structural MRI features. *Psychological Medicine* **54**, 278–288. https://doi.org/10.1017/S0033291723001319

Received: 26 July 2022
Revised: 20 January 2023
Accepted: 18 April 2023
First published online: 22 May 2023

Keywords:
Diagnostic classification; machine learning; risk of bipolar disorder; structural MRI

Corresponding author:
Pavol Mikolas;
Email: pavol.mikolas@uniklinikum-dresden.de

Abstract

Background. Individuals with bipolar disorder are commonly correctly diagnosed a decade after symptom onset. Machine learning techniques may aid in early recognition and reduce the disease burden. As both individuals at risk and those with a manifest disease display structural brain markers, structural magnetic resonance imaging may provide relevant classification features.

Methods. Following a pre-registered protocol, we trained linear support vector machine (SVM) to classify individuals according to their estimated risk for bipolar disorder using regional cortical thickness of help-seeking individuals from seven study sites ($N = 276$). We estimated the risk using three state-of-the-art assessment instruments (BPSS-P, BARS, EPI*bipolar*).

Results. For BPSS-P, SVM achieved a fair performance of Cohen's κ of 0.235 (95% CI 0.11–0.361) and a balanced accuracy of 63.1% (95% CI 55.9–70.3) in the 10-fold cross-validation. In the leave-one-site-out cross-validation, the model performed with a Cohen's κ of 0.128 (95% CI −0.069 to 0.325) and a balanced accuracy of 56.2% (95% CI 44.6–67.8). BARS and EPI*bipolar* could not be predicted. In post hoc analyses, regional surface area, subcortical volumes as well as hyperparameter optimization did not improve the performance.

Conclusions. Individuals at risk for bipolar disorder, as assessed by BPSS-P, display brain structural alterations that can be detected using machine learning. The achieved performance is comparable to previous studies which attempted to classify patients with manifest disease and healthy controls. Unlike previous studies of bipolar risk, our multicenter design permitted a leave-one-site-out cross-validation. Whole-brain cortical thickness seems to be superior to other structural brain features.

Introduction

Early detection of mental disorders has become a growing field with remarkable progress. Validated techniques for the individualized prediction of transition to diagnosed disorder

are becoming increasingly available (Fusar-Poli et al., 2017, 2019; Koutsouleris et al., 2021). In the case of bipolar disorder, early detection plays a special role, since the correct diagnosis using current diagnostic approaches occurs in average 8.7–12.4 years after the appearance of first symptoms (Kessler et al., 2005; Lambert et al., 2013; Merikangas et al., 2011; Pfennig et al., 2011). This goes along with risks of incorrect treatment, such as antidepressant-induced (or unrecognized) mania (Lambert et al., 2013; Pfennig, Bschor, Falkai, & Bauer, 2013).

Aggregation of big data from multiple centers and machine learning has enabled individualized predictions for diagnostics, prognosis, and therapy response (Dwyer, Falkai, & Koutsouleris, 2018). In the field of early recognition, psychosis risk has received the largest attention (Kambeitz-Ilankovic et al., 2015; Koutsouleris et al., 2015, 2016, 2018). Prediction of transition to psychosis in high-risk subjects can be substantially improved using machine learning, achieving up to 85.5% accuracy when combined with clinicians' judgments (Koutsouleris et al., 2021). Disproportionately fewer machine learning studies have focused on the early recognition of bipolar disorder (Claude, Houenou, Duchesnay, & Favre, 2020).

Among neuroimaging data, structural magnetic resonance imaging (MRI) is especially suitable for diagnostic and prognostic analyses using machine learning techniques. Most psychiatric disorders have been associated with brain structural markers or alterations. Recent large-scale multicentric studies of major psychiatric disorders within the ENIGMA consortium showed that along with schizophrenia, bipolar disorder ranks highest in cortical thinning among major conditions beginning in early- to mid-adulthood (Abé et al., 2020; Ching et al., 2020). Unlike major depression, attention-deficit hyperactivity disorder (ADHD), obsessive-compulsive disorder, or autism, both disorders seem to be associated with similar patterns of large-scale cortical thinning in frontal, temporal, and parietal regions with relatively high effect sizes. From a practical point of view, structural MRI (sMRI) requires relatively short scanning sequences, modest compliance, and displays high test–retest reliability (Hedges et al., 2022). Unlike genetic predisposition, which is a major risk for bipolar disorder with transition rates of 4.2–22.4% by first-degree relatives (Hafeman et al., 2017; Kerner, 2014; Post et al., 2018), using sMRI in assessment of risk for bipolar disorder has been rarely investigated.

Individuals at risk for bipolar disorders have been studied using two major approaches – family cohorts, i.e. first-degree relatives (Hajek et al., 2013), and help-seeking populations (Pfennig et al., 2020). The latter approach enables for studying a broader range of risk factors including specific subsyndromal manic or depressive symptoms, mood swings, changes in sleep and circadian rhythm, anxiety, ADHD, specific character traits, stressful life events, or substance use (Faedda et al., 2019; Leopold et al., 2012). For this purpose, and in order to facilitate the risk recognition in help-seeking cohorts, several risk assessment tools have been developed, including (extended) bipolar-at-risk criteria [BAR(S)] (Bechdolf et al., 2014; Fusar-Poli et al., 2018), Bipolar Prodrome Symptom Interview and Scale (BPSS-P) (Correll et al., 2014), and the EPI*bipolar* interview (Leopold et al., 2012). It is a strength of our study that all of these three scores are available for our cohort and were investigated as the dependent variable.

Several studies have explored the use of machine learning in classifying diagnosed bipolar disorder (Hajek et al., 2015; Nunes et al., 2020) and individuals with high genetic risk for bipolar disorder (i.e. first-degree relatives). A review by Claude et al. (2020) identified five studies that aimed to classify persons with genetic risk using different modalities, achieving accuracies from 59.7% up to 83.21%. Among those, two studies used regional cortical volumes (Hajek et al., 2015; Lin et al., 2018) and two used functional MRI (Frangou, 2019; Mourão-Miranda et al., 2012; Roberts et al., 2017). To the best of our knowledge, no multicenter machine learning study has yet been conducted to classify risk scores for bipolar disorder while including, but not being limited to the genetic risk. Based on the data from the Early-BipoLife study (Pfennig et al., 2020), we aimed to train a machine learning classifier using 10-fold cross-validation to stratify help-seeking subjects by estimated risk using sMRI. In contrast to single-center studies, we also used the multicenter design to validate it on test data from an 'unseen' study site through a leave-one-site-out cross-validation. Our results may provide a proof-of-concept for the utility of sMRI data for individualized risk prediction in subjects seeking help.

Methods

Pre-registration

We pre-registered our analyses at the Open Science Framework (https://osf.io/c4hfn).

Sample

The data were collected within the multicenter Early-BipoLife study (Pfennig et al., 2020; Ritter et al., 2016). Early-BipoLife is a multicenter, naturalistic, prospective-longitudinal observational cohort study of adolescents and young adults (age 15–35) at risk for bipolar disorder. From 10 participating German university and teaching hospitals with early detection centers/facilities for bipolar disorder, seven centers (Berlin, Bochum, Frankfurt, Hamburg, Dresden, Marburg, Tübingen) acquired MRI data. For this study, we accessed the baseline clinical and MRI data. For a detailed description of data collection procedures, see Pfennig et al. (2020). Briefly, of the total $N = 1229$ recruited adolescents and young adults at risk, $N = 313$ opted to receive MRI. In order to include all proposed risk factors for bipolar disorder, we recruited the participants in three recruitment pathways: $N = 123$ were consulting early detection centers/facilities and were screened positive for $\geqslant 1$ proposed risk factor for bipolar disorder (see online Supplementary note 1), $N = 146$ were young in- and outpatients with a depressive syndrome, and $N = 44$ had an established diagnosis of ADHD. In order to include older individuals who might have an unrecognized bipolar disorder (e.g. due to presence of exclusively depressive episodes, but no full-blown mania or hypomania yet), we extended the age inclusion criterion beyond the typical age of onset based on available studies on time to diagnosis. For more details on inclusion/exclusion criteria, see online Supplementary note 1. The study was approved by the Ethics Committee of the Medical Faculty of the Technische Universität Dresden (No: EK290082014), as well as local ethics committees at each study site. We obtained a written informed consent after comprehensive information about study aims and procedures. Additionally, parents of adolescents gave their informed consent about their children's participation.

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press

MRI acquisition, preprocessing and quality assessment

We acquired high-resolution structural T1-weighted images using Siemens Magnetom MR scanners at 6 sites (Trio, Skyra, Prisma) and a Philips Achieva scanner at 1 site. We standardized the pulse sequence parameters across all sites to the extent permitted by each platform. For a detailed description of the scanning protocol including the detail of MRI scanners, specific hardware configurations, and pulse sequence parameters, see Vogelbacher et al. (2021).

Prior to preprocessing, we performed the data acquisition and quality assessment according to the BipoLife study protocol (Vogelbacher et al., 2021). Briefly, we analyzed the MRI images using the MRIQC tool (Esteban et al., 2017). Two authors visually inspected the obtained reports of several metrics including a movement plot and a plot of the background noise. In this way, 23 subjects were excluded from further analysis due to strong movement ($N = 18$), ghosting ($N = 1$), or fold-over artifacts ($N = 4$).

We preprocessed the T1-weighted sMRI using Freesurfer 6.0 software integrated in our processing pipeline NICePype (Müller, Küttner, & Hannig, 2015). We obtained regional cortical thicknesses and surface area values for 68 cortical brain areas (34 left/34 right) defined by the Desikan–Killiany atlas (Desikan et al., 2006) and 14 subcortical volumes (7 left/7 right) (Fischl et al., 2004).

We performed a standardized quality control of the cortical and subcortical segmentations and parcellations according to the established protocols of the ENIGMA working group (http://enigma.ini.usc.edu/protocols/imaging-protocols). This included a visual inspection of the segmented regions using the internal and external surface methods, as well as statistical outlier detection. The outliers were subjected for further visual inspection. Three subjects did not pass the quality control or displayed major segmentation errors and were discarded.

Risk assessment instruments

We assessed the risk for the development of bipolar disorder using three state-of-the-art assessment instruments – the Bipolar At-Risk (BAR) criteria (Bechdolf et al., 2014) and the extended BAR criteria (BARS; Fusar-Poli et al., 2018), the Bipolar Prodrome Symptom Scale (BPSS-P; Correll et al., 2014), and the Early Phase Inventory for bipolar disorders (EPI*bipolar*; Leopold et al., 2012).

BAR(S) criteria comprise a set of subthreshold clinical and behavioral symptoms as well as genetic risk. A person is assessed as having high risk if one or more risk syndromes are fulfilled: sub-threshold mania, sub-threshold depression, sub-threshold depression with genetic risk, mixed symptoms, or mood swings. BARS criteria showed an adequate prognostic accuracy of conversion to bipolar disorder (conversion rate 18.5% in $N = 27$ participants) in a longitudinal cohort (Fusar-Poli et al., 2018). BPSS-P and EPI*bipolar* are semi-structured interviews. BPSS-P was developed based on the DSM-IV criteria for bipolar disorder and major depression and established rating scales for these conditions. BPSS-P combines all these criteria to a mania symptom index, depression symptom index, and general symptom index. It implies two at-risk states: attenuated mania symptom syndrome (AMSS) and genetic mania risk and deterioration syndrome (GMRDS). BPSS-P has good internal consistency, convergent validity, and inter-rater reliability (Correll et al., 2014). EPI*bipolar* contains elements from BPSS-P and additionally captures risk factors that have been identified through a systematic literature review, such as subsyndromal manic or depressive symptoms, mood swings, changes in sleep and circadian rhythm, anxiety, ADHD, specific character traits, stressful life events, or changing patterns of substance use (Leopold et al., 2012). It defines three risk categories: no-risk, low-risk, and high-risk. For the purpose of this analysis, we pooled subjects from the low-risk and high-risk groups assessed by EPI*bipolar*, as these participants, unlike those from the no-risk group, displayed several clinically relevant risk factors or symptoms and are intended for targeted interventions in early recognition services. The term 'no-risk' group in EPI*bipolar* was originally established to describe the lack of need for a specialized clinical intervention in the participants with only minor risk factors (Leopold et al., 2012). Of note, all recruited participants, even those who did not fulfill the criteria of any risk syndrome/group on any of the three risk instruments, displayed at least one known risk factor for bipolar disorder (see online Supplementary note 1). In research settings, this label might be misleading, as participants in the no-risk group might also display minor risk factors and are not to be confused with healthy controls. The final binary outcomes were as follows: any symptom syndrome/no symptom syndrome for BPSS-P; any risk group/no risk group for BARS; high-risk + low-risk groups/no-risk group for EPI*bipolar* (see also Table 1 for demographics). As we discarded subjects with missing data on corresponding assessment tools, the sample sizes for each of the three risk assessment tools varied ($N_{BARS} = 264$, $N_{BPSS-P} = 276$, $N_{EPIbipolar} = 273$). For details on the risk assessment tools, see online Supplementary Table S1 and Pfennig et al. (2020). All three instruments/criteria sets were obtained from the respective authors and can be administered after appropriate training. The administration of the complete risk assessment battery takes 2–3h.

Machine learning classification

In accordance with a previous study of subjects with diagnosed bipolar disorder by the ENIGMA consortium (Nunes et al., 2020) and to increase reproducibility, we used a linear support vector machine (SVM) classifier with the hyperparameter $C = 1$ for the primary analysis. We performed independent binary classifications for each risk instrument (BPSS-P, BARS, and EPI*bipolar*). Using Scikit-learn 1.0 package for Python 3.8.3 (Pedregosa et al., 2011), we utilized two cross-validation methods: 10-fold and leave-one-site-out (i.e. data from one study center was taken to be the test-data, while the training dataset included the data from all other centers). In each fold, we standardized features in the training and testing sets separately by removing the mean and scaling to unit variance using standard scaler (Scikit-learn 1.0 package, see above). We took the following measures to manage the imbalanced class distribution within the data: (A) we used a stratified cross-validation to ensure, that the class ratio in all folds stays approximately the same, (B) we used random oversampling of the minority class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) in the training set, so that the class ratios in each fold was balanced. For the primary analysis, we used the 68 regional cortical thickness values as features and we performed both cross-validation methods (10-fold and leave-one-site-out), i.e. we trained six models altogether. As the class ratios for all three risk instruments were imbalanced, we used following two performance measures which are commonly used for imbalanced classification problems: *Cohen's* κ (i.e. the

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press

Table 1. Socio-demographic characteristics

Risk assessment instrument	BPSS-P ($N = 276$)			BARS ($N = 264$)			EPI*bipolar* ($N = 273$)		
Risk criterion fulfilled	No	Yes	Test	No	Yes	Test	No	Yes	Test
N (%)	220 (79.7)	56 (20.3)	n/a	77 (29.2)	187 (70.8)	n/a	32 (11.7)	241 (88.3)	n/a
Female (%)	97 (44.1)	34 (60.7)	$\chi^2 = 4.947$, $p = 0.026$*	35 (45.5)	92 (49.2)	$\chi^2 = 0.306$, $p = 0.58$	10 (31.3)	120 (49.8)	$\chi^2 = 3.894$, $p = 0.048$*
Age (s.d.)	24.91 (4.2)	24.57 (4.98)	$t = 0.523$, df = 274, $p = 0.62$	24.35 (3.7)	25.1 (4.6)	$t = 1.242$, df = 262, $p = 0.215$	24.13 (3.08)	24.95 (4.52)	$t = -1.001$, df = 271, $p = 0.318$
Education high school (%)	177 (80.5)	39 (69.6)	$\chi^2 = 9.516$, $p = 0.147$	65 (84.4)	141 (75.4)	$\chi^2 = 13.789$, $p = 0.032$*	26 (81.3)	188 (78.0)	$\chi^2 = 2.437$, $p = 0.875$
Recruitment pathway									
Early recognition (%)	97 (44.1)	19 (33.9)	$\chi^2 = 1.915$, $p = 0.384$	36 (46.8)	77 (41.2)	$\chi^2 = 23.149$, $p \leq 0.001$***	15 (46.9)	101 (41.9)	$\chi^2 = 23.149$, $p \leq 0.001$***
Depression (%)	95 (43.2)	29 (51.8)		24 (31.2)	91 (48.7)		5 (15.7)	116 (48.1)	
ADHD (%)	28 (12.7)	8 (14.3)		17 (22.1)	19 (10.2)		12 (37.5)	24 (9.9)	
Psychiatric medication									
Yes (%)	118 (53.6)	36 (64.3)	$\chi^2 = 2.053$, $p = 0.152$	35 (45.5)	112 (59.9)	$\chi^2 = 4.608$, $p = .032$*	11 (34.4)	142 (58.9)	$\chi^2 = 6.909$, $p = 0.009$**
Substance use									
Smoking status									
Never smoked (%)	102 (46.4)	21 (37.5)	$\chi^2 = 2.376$, $p = 0.305$	44 (57.1)	76 (40.6)	$\chi^2 = 6.008$, $p = 0.05$*	18 (56.3)	104 (43.2)	$\chi^2 = 6.036$, $p = 0.049$*
Current smoker (%)	102 (46.4)	28 (0.5)		28 (36.4)	93 (49.7)		9 (28.1)	119 (49.4)	
Past smoker (%)	16 (7.3)	7 (12.5)		5 (6.5)	18 (9.6)		5 (15.6)	18 (7.5)	
Cannabis present									
No use (%)	155 (70.5)	44 (78.6)	$p = 0.078$[a]	62 (80.5)	129 (67.0)	$p = 0.506$[a]	26 (81.3)	172 (71.4)	$p = 0.545$[a]
<1x/month (%)	21 (9.5)	3 (5.4)		5 (6.5)	16 (8.6)		1 (3.1)	22 (9.1)	
~1x/month (%)	12 (5.5)	2 (3.6)		3 (3.9)	11 (5.9)		0 (0)	14 (5.8)	
2–9x/month (%)	17 (7.7)	0 (0)		3 (3.9)	13 (7.0)		2 (6.3)	14 (5.8)	
⩾10x/month (%)	15 (6.8)	7 (12.5)		4 (5.2)	18 (9.6)		3 (9.4)	19 (7.9)	
Cannabis lifetime									
No use (%)	86 (39.1)	23 (41.7)	$p = 0.980$[a]	39 (50.6)	67 (35.8)	$p = 0.253$[a]	15 (46.9)	94 (39.0)	$p = 0.762$[a]
<1x/month (%)	48 (21.8)	10 (17.9)		16 (20.8)	39 (20.9)		8 (25.0)	49 (20.3)	
~1x/month (%)	9 (4.1)	2 (3.6)		2 (2.6)	9 (4.8)		0 (0)	11 (4.6)	
2–9x/month (%)	25 (11.4)	6 (10.7)		7 (9.1)	24 (12.8)		3 (9.4)	28 (11.6)	
⩾10x/month (%)	50 (22.7)	14 (25.0)		13 (16.9)	46 (24.6)		6 (18.8)	56 (23.2)	
Genetic risk									
FDR (%)	17 (7.7)	4 (7.1)	$\chi^2 = 0.022$, $p = 0.883$	9 (11.7)	11 (5.9)	$\chi^2 = 2.626$, $p = 0.105$	0 (0)	21 (8.7)	$p = 0.148$[a]

*$p \leqslant 0.05$; **$p \leqslant 0.01$; ***$p \leqslant 0.001$. FDR, first-degree relatives of BD patients.
aFisher–Freeman–Halton's exact test was used for variables with ⩾1 expected cell counts <5.

measure of agreement between the classifier and a random classifier relative to the frequency of classes, <0 no agreement, 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1 almost perfect agreement) (Landis & Koch, 1977), and *balanced accuracy* [balanced accuracy = (sensitivity + specificity)/2]. Additionally, we report sensitivity and specificity. We do not report other common measures such as accuracy and area under receiver operating characteristic curve, as these are not suitable for imbalanced data (He & Ma, 2013). As this was a population-based, observational study, the samples in each site were not balanced regarding participants at risk, some were even smaller than the recommended size $N > 20$ for a test set (Flint et al., 2021; as well as see online Supplementary Table S2). For this reason, we report the performance in both 10-fold as well as leave-one-site-out cross-validations.

For risk assessment instruments achieving an above chance prediction (i.e. Cohen's $\kappa > 0$ and lower confidence interval > 0), we assessed the possible effects of confounds using post-hoc statistical tests comparing the correctly and incorrectly classified subjects. This is a more valid approach to account for possible confounders than regressing out covariates prior to analysis, which would disrupt the train/test separation (Pereira, Mitchell, & Botvinick, 2009). We also report the post-hoc tests for the leave-one-site-out cross-validation by BPSS-P, where the lower confidence interval slightly crossed the zero boundary. Given the above-mentioned limitations of the leave-one-site-out cross-validation (low sample size of some sites, imbalanced classes), we consider both measures relevant. We accounted for following confounds: age, sex, medication (yes/no), recruitment pathway (early recognition services/depression/ADHD), smoking status (never smoked/current smoker/past smoker), present cannabis use (no use/<1 per month/~1 per month/2–9 per month/⩾10 per month), lifetime cannabis use (no use/<1x month/~1x month/2–9x month/⩾10x month), site and scanner type (for the list of sites and scanner types see above).

We estimated the magnitude of contribution of brain regions to the SVM classification using SVM coefficients. Coefficients of a linear classifier can be interpreted as relative measure of feature importance (Pereira et al., 2009) for the classification process. Note that this is not to say that a highly weighted feature contains necessarily a lot of information about the target class (Haufe et al., 2014). We used the freesurfer_statsurf_display library (https://chrisadamsonmcri.github.io/freesurfer_statsurf_ display) to visualize the results.

Secondary analyses

To investigate whether a lower feature/sample size ratio might improve classification performance, we selected 20 features based on the available literature from other relevant large-scale multicenter studies and included these in our pre-registration. We chose 20 features in order to approach the similar ratio of features as the prior study on bipolar disorder using SVM by Nunes et al. (2020), which reported having 20 times more participants, than features. We selected those features from another large-scale ENIGMA study of bipolar disorder and healthy controls by Hibar et al. (2018) which identified a pattern of significant reductions of cortical thickness in frontal, temporal, and parietal regions in a sample of 6503 participants and bipolar patients. We selected the 20 features displaying the highest effect sizes in that study (see online Supplementary note 2 for the list of features).

In order to better compare the performance of the SVM on our sample of help-seeking individuals at risk and patients with established disease published by Nunes et al. (2020), we also performed the classification using the same feature set of 150 features including 68 regional cortical thickness and 68 surface area values as well as volumes of 14 subcortical features plus the estimated total intracranial volume.

Lastly, we investigated whether hyperparameter optimization using a nested cross-validation would improve the results. In each fold, we divided the train set into train and test subsets once more and ran multiple nested SVM classifications with different SVM regularization parameters C (1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2}, 1×10^{-1}, 1, 10, and 100) (grid search method). We selected the best possible model according to the achieved balanced accuracy. Finally, we tested the selected model on the unseen test data from the primary loop.

Results

Demographics

For detailed demographics, see Table 1. The participants who fulfilled any risk syndrome according to BPSS-P did not differ from those not fulfilling any risk syndrome in any of the measured variables. The participants who fulfilled any risk syndrome according to BARS were more likely to take medication ($\chi^2 = 4.608$, $p = 0.032$), to smoke ($\chi^2 = 6.008$, $p = 0.05$), and suffer from diagnosed depression, but less likely to suffer from ADHD ($\chi^2 = 23.149$, $p \leq 0.001$) and were more likely to have attended high-school ($\chi^2 = 13.789$, $p = 0.032$) than those not fulfilling any risk syndrome. The participants who fulfilled any risk syndrome according to EPI*bipolar* were more likely to be female ($\chi^2 = 3.894$, $p = 0.048$), to take medication ($\chi^2 = 6.909$, $p = 0.009$), to smoke ($\chi^2 = 6.036$, $p = 0.049$), and suffer from diagnosed depression, but less likely to suffer from ADHD ($\chi^2 = 23.149$, $p \leq 0.001$), than those not fulfilling any risk syndrome. The participants removed due to movement during the scan and quality control did not differ from those in the final dataset in the proportion of any of the risk syndromes: BPSS-P (df = 1, $\chi^2 = 0.004$, $p = 0.949$), BARS (df = 1, $\chi^2 = 0.412$, $p = 0.521$), and EPI*bipolar* (df = 2, $\chi^2 = 1.092$, $p = 0.579$).

Primary analysis

Performance measures of the classification for all three risk instruments (BPSS-P, BARS, and EPI*bipolar*) using all regional cortical thickness values as features are given in Table 2. Only for BPSS-P, both performance measures reached levels above chance for the 10-fold CV approach with following performance: Cohen's κ 0.235 (95% CI 0.11–0.361), balanced accuracy 63.1% (95% CI 55.9–70.3), sensitivity 48% (95% CI 36–60), and specificity 78.2% (95% CI 72.5–83.9). The correctly and incorrectly classified subjects did not differ in age (df = 274, $t = 0.987$, $p = 0.114$), sex (df = 1, $\chi^2 = 0.152$, $p = 0.698$), medication (df = 1, $\chi^2 = 0.068$, $p = 0.795$), recruitment pathway (df = 2, $\chi^2 = 0.673$, $p = 0.714$), first-degree relatives (df = 1, $\chi^2 = 0.334$, $p = 0.563$), smoking status (df = 2, $\chi^2 = 2.254$, $p = 0.324$), cannabis use lifetime (Fisher–Freeman–Halton's exact test $p = 0.28$), site (Fisher–Freeman–Halton's exact test $p = 0.119$), and scanner type (Fisher–Freeman–Halton's exact test $p = 0.225$). There was a significant difference in the present cannabis use (Fisher–Freeman–Halton's exact test $p = 0.043$), however, using residuals of cortical

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press

Table 2. Detailed performance metrics of the SVM classification using all regional cortical thickness values (68 features) and all three risk assessment tools

	Cohen's κ (%)			Balanced accuracy (%)			Sensitivity (%)			Specificity (%)		
		95% CI			95% CI			95% CI			95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper		Lower	Upper
BPSS-P												
10-fold	0.235	0.11	0.361	63.1	55.9	70.3	48.0	36.0	60.0	78.2	72.5	83.9
Leave-one-site-out	0.128	−0.069	0.325	56.2	44.6	67.8	33.0	12.4	53.7	79.4	72.3	86.4
BARS												
10-fold	−0.049	−0.189	0.092	47.3	38.9	55.7	51.2	42.4	60.0	43.4	30.1	56.7
Leave-one-site-out	−0.005	−0.195	0.184	50.8	39.5	62.1	57.9	43.4	72.3	43.7	16.0	71.5
EP*Ibipolar*												
10-fold	0.02	- 0.156	0.197	50.2	39.4	61.1	80.5	73.3	87.7	20.0	3.3	36.7
Leave-one-site-out	0.007	−0.115	0.129	52.6	43.8	61.4	75.7	66.9	84.4	29.6	9.3	50.0

features after regressing out present cannabis use resulted in a comparable performance Cohen's κ 0.240 (95% CI 0.102–0.379), balanced accuracy 62.6% (95% CI 55.3–69.8).

In the leave-one-site-out cross-validation, the classifier based on BPSS-P achieved Cohen's κ 0.128 (95% CI −0.069 to 0.325), balanced accuracy 56.2% (95% CI 44.6–67.8), sensitivity 33% (95% CI 12.4–53.7), and specificity 79.4% (95% CI 72.3–86.4). The correctly and incorrectly classified subjects did not differ in age ($df = 274$, $t = 0.523$, $p = 0.601$), sex ($df = 1$, $\chi^2 < 0.001$, $p = 0.994$), medication ($df = 1$, $\chi^2 = 2.268$, $p = 0.132$), recruitment pathway ($df = 2$, $\chi^2 = 2.951$, $p = 0.229$), first-degree relatives ($df = 1$, $\chi^2 = 2.125$, $p = 0.145$), smoking status ($df = 2$, $\chi^2 = 3.595$, $p = 0.166$), cannabis use present (Fisher–Freeman–Halton's exact test $p = 0.281$), cannabis use lifetime (Fisher–Freeman–Halton's exact test $p = 0.518$), site (Fisher–Freeman–Halton's exact test $p = 0.905$), and scanner type (Fisher–Freeman–Halton's exact test $p = 0.694$).

See Table 2 for the summary of performance measures.

Secondary analyses

Both literature-derived feature selection of 20 regional cortical thickness features, as well as an extended feature set including whole-brain regional surface area and volumes of subcortical regions did not yield significantly higher accuracies, as the confidence intervals overlapped with those from the primary analysis (see Table 3 for the summary of classification metrics). The lower difference in performance measures between the 10-fold and the leave-one-site-out cross-validation using the 20 regional cortical features rather than all regional cortical values by BPSS-P suggests a non-significant trend toward better model validity when using the 20 cortical features.

Hyperparameter optimization

Using hyperparameter optimization, we achieved Cohen's κ 0.212 (95% CI 0.123–0.302), balanced accuracy 62.3% (95% CI 56.7–68.0), sensitivity 48% (95% CI 37.7–59.0), and specificity 76.4% (95% CI 72.4–80.3) in a 10-fold cross-validation and Cohen's κ 0.136 (95% CI −0.075 to 0.346), balanced accuracy 57.1% (95% CI 43.6–70.6), sensitivity 33.7% (95% CI 10.2–57.3), and specificity 80.4% (95% CI 72.4–88.4) in the leave-one-site-out cross-validation. The mostly selected C parameter was 100 (7 out of 10 and 4 out of 7) for 10-fold and leave-one-site-out, respectively.

SVM coefficients

The mean over folds of the absolute values of the SVM coefficients by feature (brain region) for the BPSS-P, 10-fold cross-validation, and whole-brain regional cortical thickness features are depicted in Fig. 1. For the values of all coefficients, see online Supplementary Table S3.

Discussion

The linear SVM classifier detected individuals with increased estimated risk for bipolar disorder as defined by the BPSS-P interview with a Cohen's κ of 0.227/0.141 and balanced accuracy of 63.1/56.2% (based on pooled sample and leave-one-site-out cross-validations, respectively). Precuneus, inferior frontal gyrus, and posterior cingulate cortex ranked among the highest contributing features according to SVM coefficients. SVM could not detect

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press

Table 3. Results of the secondary classification for all three risk instruments using two different feature selection methods: CT 20–20 selected cortical features, CT, all regional cortical values; SC, all subcortical volumes; SA, all regional surface area values

	CT 20						CT + SC + SA					
	Cohen's κ (%)			Balanced accuracy (%)			Cohen's κ (%)			Balanced accuracy (%)		
		95% CI			95% CI			95% CI			95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper		Lower	Upper
BPSS-P												
10-fold	0.227	0.051	0.402	65.6	53.9	77.3	0.226	0.088	0.365	63.0	54.4	71.7
Leave-one-site-out	0.141	0.006	0.277	62.2	50.1	74.4	0.028	−0.209	0.265	50.3	36.6	64.0
BARS												
10-fold	−0.141	−0.251	−0.03	41.8	35.0	48.5	<0.001	−0.103	0.105	50.2	44.5	55.9
Leave-one-site-out	<0.001	−0.142	0.142	51.8	42.5	61.2	0.021	0.218	0.26	52.7	38.8	66.6
EPI*bipolar*												
10-fold	−0.03	−0.114	0.05	47.0	39.2	54.7	−0.03	−0.167	0.107	47.3	39.2	55.5
Leave-one-site-out	0.018	−0.044	0.08	55.2	45.9	64.4	−0.085	0.179	0.01	51.0	35.1	66.9

participants with increased risk for bipolar disorder based on EPI*bipolar* or BARS criteria. Whole-brain cortical thickness yielded the highest accuracy, whereas reducing the features based on literature, or expanding the features by surface area or subcortical volumes did not change the performance significantly given the large confidence intervals. However, there might be a trend toward better model validity using fewer cortical thickness features.

Our results suggest that young participants at risk of bipolar disorder according to the BPSS-P display distinct structural brain features that permit better-than-chance classification. Importantly, using both the pooled sample (10-fold cross-validation), as well as leave-one-site out cross-validation, we achieved accuracies comparable to the previous multicenter study by Nunes et al. (2020) that differentiated patients with *manifest* bipolar disorder from healthy controls with balanced accuracies of 65.23% (95% CI 63.47–67.00) and 58.67% (95% CI 56.70–60.63), respectively (Nunes et al., 2020). Compared to their study, the 95% confidence intervals in our study were considerably wider, which was to be expected given that our sample was more than 10 times smaller (276 *v.* 3020 participants). Larger sample sizes tend to yield more stable performance (Nieuwenhuis et al., 2012). Post-hoc tests suggested an effect of present cannabis use on the classification using 10-fold cross-validation; however, regressing out present cannabis use did not impair the performance. Moreover, there was no such effect in the leave-one-site-out validation. Other demographic variables did not show an effect on the classification (see also online Supplementary note 3). As such, this would be consistent with the notion that differences in brain structure in bipolar disorder are not a result of the disorder but are a pre-morbid risk factor potentially related to genetics. On the other hand, as the age of participants in our sample was higher than the typical age of onset of bipolar disorder, we might have included older participants with a yet undiagnosed bipolar disorder. Those participants would have possibly displayed more structural differences than participants before the age of onset, which might in turn have led to higher classification accuracies.

Unlike previous attempts to detect participants with genetic risk within family cohorts (Hajek et al., 2015), we estimated the individual risk state using state-of-the-art screening instruments, which better address the clinical realities of early recognition centers. Given the variable estimated transition rates of 4.2–22.4% by known genetic risk (Hafeman et al., 2017; Kerner, 2014; Post et al., 2018), there is a need for more differentiated risk assessment including state markers in order to provide targeted interventions. Moreover, most people seeking for early recognition services do not have genetic risk (12.9% or 15 out of 116 recruited via the early recognition pathway). BPSS-P provides a conservative risk assessment, selecting persons displaying an AMSS or a GMRDS. In total, 20.3% of the participants screened positive on one of these syndromes.

Surprisingly, SVM could not detect participants at risk estimated using EPI*bipolar*, although we detected significant differences in cortical thickness between the high-risk and no-risk individuals in our previous study (Mikolas et al., 2021). Given similar sample size (previous study $N = 263$), we pooled the individuals in the high-risk and low-risk groups in order to allow for a binary classification. As a result, the no-risk group had only 32 participants, which might have had a negative influence on the learning phase. In a post hoc analysis (see online Supplementary note 4), a three-category classification using all three risk groups did not

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press

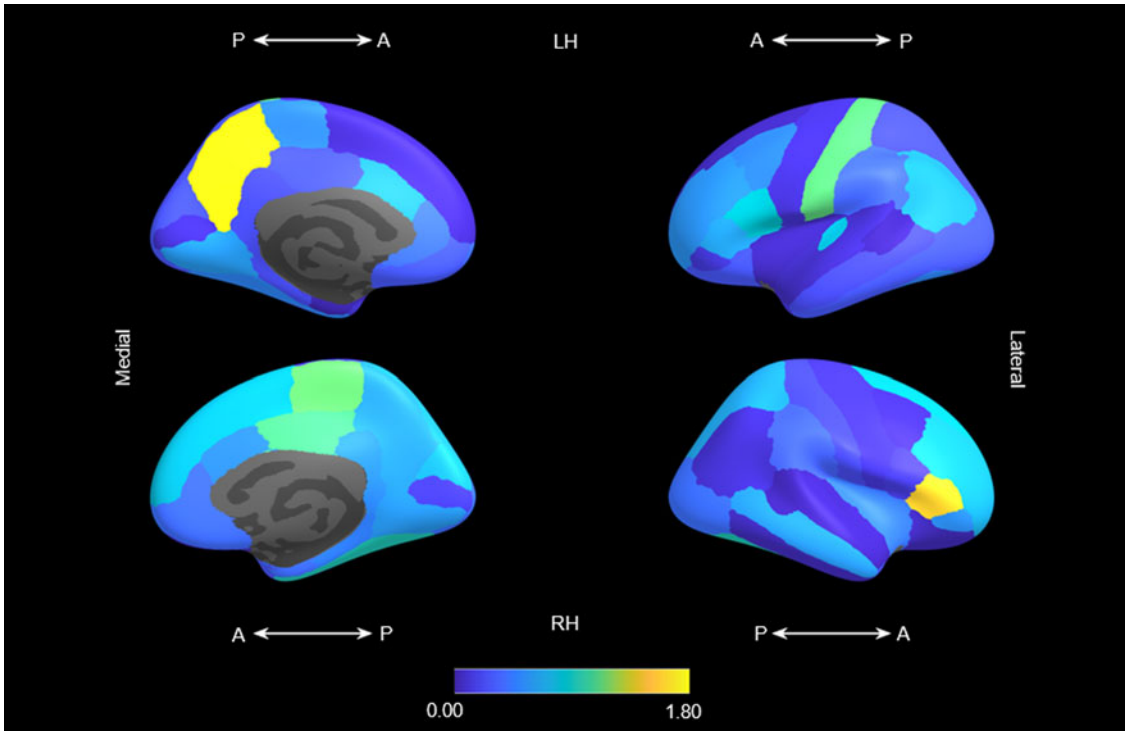

Figure 1. Magnitude of contribution of brain regions to SVM classification. The coefficients of a linear classifier can be interpreted as relative measure of feature importance. The color represents the mean over folds of the absolute value of the SVM coefficients for each region. The classification was based on BPSS-P risk instrument using regional cortical thickness values as features in a 10-fold cross-validation.

yield an above chance classification. However, after removing the low-risk group, we obtained a balanced accuracy of 60.9/55.5% (10-fold/leave-one-site-out). This suggests that whereas in a hypothesis-driven region-of-interest analysis, EPI*bipolar* selected participants displaying significantly thinner cortex in the left pars opercularis (Mikolas et al., 2021), BPSS-P selected participants displaying widespread structural alterations enabling for more accurate, single subject classification using machine learning. Interestingly, in our above-mentioned previous study (Mikolas et al., 2021), the pars opercularis was not significantly thinner in participants scoring positive in BPSS-P; however, the low p value might have suggested a non-significant trend. Additionally, among the participants scoring positive on any risk criterion in both EPI*bipolar* and BARS, those with depression were more represented comparing to BPSS-P. As a result, more participants with unipolar depression might have been selected by EPI*bipolar* and BARS, which might have impeded the classification. Indeed, the cortical thickness differences in major depression seem to be less prominent than in bipolar disorder (Ching et al., 2020; Schmaal et al., 2017). Finally, unlike in BPSS-P, the participants who fulfilled any risk syndrome according to BARS or EPI*bipolar* differed from those not fulfilling any risk syndrome in several other demographic variables which might have confounding effects on cortical thickness, such as medication or smoking.

The regions with highest contribution toward the classification (i.e. with the highest values of SVM coefficients) partially overlapped with those contributing to classification of patients with manifest bipolar disorder and healthy controls obtained by Nunes et al. (2020) in their study. Of 33 cortical thickness weights reported by Nunes et al., 69.7% hat the same sign as in our study. Notably, the inferior frontal gyrus is a region structurally and functionally associated with the genetic risk for bipolar disorder (Hajek et al., 2013; Roberts et al., 2013, 2017). This suggests a consistent structural pattern of individuals at risk estimated by BPSS-P and patients with manifest disease or genetic risk. A direct comparison, however, of feature weights between Nunes et al. and our study is to be viewed with caution because of the complex covariance structure within the feature set (Haufe et al., 2014), the difference in the number and type of features, and the limited number of training samples in our study. While multivariate machine learning techniques have the potential to optimize prediction accuracies, univariate, between-group comparisons are more straight forward to interpret in terms of relative feature importance, as we have done in our previous work (Mikolas et al., 2021).

The achieved accuracy is not sufficient to suggest sMRI as a single risk assessment method. Even using the best performing model, among the subjects, who did not fulfill any risk criterion, 21.8% were classified as positive (type I error). Among the subjects at risk, 51.8% were classified as negative (type II error). Even feature selection approaches or hyperparameter optimization did not achieve a more accurate classification. Earlier machine learning neuroimaging studies reported accuracies well beyond the 80% boundary that roughly demarks the clinical utility (Nunes et al., 2020; Radua & Carvalho, 2021). However, many earlier studies did not comply with recently established criteria (Dwyer et al., 2018), for example, by using insufficiently small samples [i.e. $N < 130$ (Nieuwenhuis et al., 2012)] or omitting validation samples (Radua & Carvalho, 2021). Studies that used validation samples generally reported lower accuracies.

Differentiating between healthy, non-help-seeking persons and help-seeking persons with higher risk for bipolar disorder might lead to higher accuracies. For a potential clinical application, however, this might be misleading, as clinicians are required to make

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press

predictions by persons who already display symptoms and therefore seek for help. Thus, our population-based sample of help-seeking individuals reflects the real clinical setting better than using a healthy-control group. The very fact that we chose a conservative approach by comparing only help-seeking individuals and yet are able to obtain a clear above-chance prediction of the score in an established risk instrument with mere structural neuroimaging data demonstrates the potential of sMRI in risk stratification. A major advantage over functional MRI is that structural T1 images are part of any standard clinical exam and would not invoke additional costs for specialized scanning protocols.

Overall, our results suggest that in order to achieve clinically meaningful predictions, future approaches using brain imaging should aim at integrating multimodal data such as clinical data, such as body mass index (McWhinney et al., 2021) or genetics, rather than focusing on brain structure only. An 'augmentation' of clinical judgments of trained professional by a machine learning-based algorithm might be a realistic scenario. Koutsouleris et al. (2021) showed in individuals with psychosis risk, that in a multimodal application, sMRI might contribute to the overall prediction by several percent. As our study suggests, sMRI, especially cortical thickness, might contribute to the diagnostic performance of such algorithms aimed at estimating the risk for bipolar disorder.

An important limitation that needs to be addressed in future studies was the use of the estimated risk as outcome. The concept of high risk for bipolar disorder is still in development (Keramatian, Chakrabarty, Saraf, & Yatham, 2021). Participants scoring positive on those risk criteria might benefit from a more intensive diagnostic and prevention process. However, in order to further individualize the risk prediction, larger, longitudinal studies with sufficient number of participants who develop a first manic episode should be performed in the future.

Lower occurrence of ADHD in the high-risk group was due to the distribution of risk factors within the three different recruitment pathways. Although ADHD as a risk factor enabled the participants to enter the study through all three recruitment pathways, the risk factor ADHD was 'enriched' in the overall sample due to in- and outpatients entering the study through the 'ADHD' recruitment pathway. However, these participants displayed fewer additional risk factors, so that most did not fulfill the criteria for the higher risk groups.

An interesting research objective for future studies would be to include participants with borderline personality disorder, as these might be hard to clinically distinguish from bipolar disorder in its initial or at-risk state. Especially the question whether people that transition to different disorders also differ in brain structure would be highly relevant.

In summary, we show that machine learning techniques can detect brain structural alterations in young individuals at risk for bipolar disorder with a performance comparable to previous studies of patients with manifest disease and healthy controls. Whole-brain cortical thickness might be superior to other structural brain features in predicting the risk to develop bipolar disorder. Future studies should aim to improve the performance of predictive models for individuals at risk by using larger cohorts and multimodal data. Even more sophisticated machine learning methods or methods of feature extraction may contribute to clinically meaningful predictions. Our own study may contribute to this effort in the future (Böhle, Eitel, Weygandt, & Ritter, 2019).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291723001319

Author contributions. A. P., M. B., P. Ritter, and A. J. designed the study. K. B., C. B., J. M., A. J. F., T. E., A. Rau, T. K., I. F., M. L., G. L., C. M., V. K., K. L., A. B., A. Reif, S. M., T. S., F. B., J. F., G. J., V. F., and C. U. C. participated in the patient recruitment. P. M., M. M., P. Riedel, and C. V. performed the MRI data analyses and statistics. P. M., M. M., P. Riedel, F. H., and A. P. wrote the article. M. M., K. B., C. B., J. M., A. J. F., T. E., A. Rau, T. K., I. F., M. L., G. L., C. M.,V. K., K. L., A. B., A. Reif, S. M., T. S., F. B., J. F., G. J., V. F., C. U. C., M. B., and P. Ritter revised it critically for important intellectual content. All of the authors reviewed and approved the manuscript for publication.

Financial support. Early-BipoLife is funded by the Federal Ministry of Education and Research (BMBF, grant numbers: 01EE1404A, 01EE1404E, and 01EE1404F). M. M. was supported by the Deutsche Forschungsgemeinschaft (DFG grant Nos. 178833530 [SFB 940] and 402170461 [TRR 265]).

Conflict of interest. K. Leopold has been a consultant and/or advisor to or has received honoraria from: Janssen/J&J, Lundbeck, Otsuka, Recordati, and ROVI. She has received grant support from Janssen/J&J and Otsuka. All other authors declared no conflict of interest.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References

Abé, C., Liberg, B., Song, J., Bergen, S. E., Petrovic, P., Ekman, C. J., … Landén, M. (2020). Longitudinal cortical thickness changes in bipolar disorder and the relationship to genetic risk, mania, and lithium use. *Biological Psychiatry*, *87*(3), 271–281. Retrieved from https://doi.org/10.1016/j.biopsych.2019.08.015

Bechdolf, A., Ratheesh, A., Cotton, S. M., Nelson, B., Chanen, A. M., Betts, J., … McGorry, P. D. (2014). The predictive validity of bipolar at-risk (prodromal) criteria in help-seeking adolescents and young adults: A prospective study. *Bipolar Disorders*, *16*(5), 493–504. Retrieved from https://doi.org/10.1111/bdi.12205

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in Aging Neuroscience*, *11*, 194. Retrieved from https://doi.org/10.3389/fnagi.2019.00194

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. Retrieved from https://doi.org/10.1613/jair.953

Ching, C. R. K., Hibar, D. P., Gurholt, T. P., Nunes, A., Thomopoulos, S. I., Abé, C., … ENIGMA Bipolar Disorder Working Group. (2020). What we learn about bipolar disorder from large-scale neuroimaging: Findings and future directions from the ENIGMA bipolar disorder working group. *Human Brain Mapping*, ***43***(1), hbm.25098. Retrieved from https://doi.org/10.1002/hbm.25098

Claude, L., Houenou, J., Duchesnay, E., & Favre, P. (2020). Will machine learning applied to neuroimaging in bipolar disorder help the clinician? A critical review and methodological suggestions. *Bipolar Disorders*, *22*(4), 334–355. Retrieved from https://doi.org/10.1111/bdi.12895

Correll, C. U., Olvet, D. M., Auther, A. M., Hauser, M., Kishimoto, T., Carrión, R. E., … Cornblatt, B. A. (2014). The bipolar prodrome symptom interview and scale-prospective (BPSS-P): Description and validation in a psychiatric sample and healthy controls. *Bipolar Disorders*, *16*(5), 505–522. Retrieved from https://doi.org/10.1111/bdi.12209

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.

NeuroImage, *31*(3), 968–980. Retrieved from https://doi.org/DOI: 10.1016/j.neuroimage.2006.01.021

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. Retrieved from https://doi.org/10.1146/annurev-clinpsy-032816-045037

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE*, *12*(9), e0184661. Retrieved from https://doi.org/10.1371/journal.pone.0184661

Faedda, G. L., Baldessarini, R. J., Marangoni, C., Bechdolf, A., Berk, M., Birmaher, B., … Correll, C. U. (2019). An international society of bipolar disorders task force report: Precursors and prodromes of bipolar disorder. *Bipolar Disorders*, *21*(8), 720–740. Retrieved from https://doi.org/10.1111/bdi.12831

Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23*(Suppl 1), S69–S84. Retrieved from https://doi.org/10.1016/j.neuroimage.2004.07.016

Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D. M. A., Emden, D., … Hahn, T. (2021). Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*, *46*(8), 1510–1517. Retrieved from https://doi.org/10.1038/s41386-021-01020-7

Frangou, S. (2019). Neuroimaging markers of risk, disease expression, and resilience to bipolar disorder. *Current Psychiatry Reports*, *21*(7), 52. Retrieved from https://doi.org/10.1007/s11920-019-1039-7

Fusar-Poli, P., De Micheli, A., Rocchetti, M., Cappucciati, M., Ramella-Cravaro, V., Rutigliano, G., … Falkenberg, I. (2018). Semistructured interview for bipolar at risk states (SIBARS). *Psychiatry Research*, *264*, 302–309. Retrieved from https://doi.org/10.1016/j.psychres.2018.03.074

Fusar-Poli, P., Rutigliano, G., Stahl, D., Davies, C., Bonoldi, I., Reilly, T., & McGuire, P. (2017). Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. *JAMA Psychiatry*, *74*(5), 493–500. Retrieved from https://doi.org/10.1001/jamapsychiatry.2017.0284

Fusar-Poli, P., Werbeloff, N., Rutigliano, G., Oliver, D., Davies, C., Stahl, D., … Osborn, D. (2019). Transdiagnostic risk calculator for the automatic detection of individuals at risk and the prediction of psychosis: Second replication in an independent national health service trust. *Schizophrenia Bulletin*, *45*(3), 562–570. Retrieved from https://doi.org/10.1093/schbul/sby070

Hafeman, D. M., Merranko, J., Goldstein, T. R., Axelson, D., Goldstein, B. I., Monk, K., … Birmaher, B. (2017). Assessment of a person-level risk calculator to predict new-onset bipolar spectrum disorder in youth at familial risk. *JAMA Psychiatry*, *74*(8), 841. Retrieved from https://doi.org/10.1001/jamapsychiatry.2017.1763

Hajek, T., Cooke, C., Kopecek, M., Novak, T., Hoschl, C., & Alda, M. (2015). Using structural MRI to identify individuals at genetic risk for bipolar disorders: A 2-cohort, machine learning study. *Journal of Psychiatry & Neuroscience: JPN*, *40*(5), 316–324. Retrieved from https://doi.org/10.1503/jpn.140142

Hajek, T., Cullis, J., Novak, T., Kopecek, M., Blagdon, R., Propper, L., … Alda, M. (2013). Brain structural signature of familial predisposition for bipolar disorder: Replicable evidence for involvement of the right Inferior frontal gyrus. *Biological Psychiatry*, *73*(2), 144–152. Retrieved from https://doi.org/10.1016/j.biopsych.2012.06.015

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. Retrieved from https://doi.org/10.1016/j.neuroimage.2013.10.067

He, H., & Ma, Y. (Eds.) (2013). *Imbalanced learning: Foundations, algorithms, and applications*. Hoboken, NJ: IEEE Press, Wiley.

Hedges, E. P., Dimitrov, M., Zahid, U., Brito Vega, B., Si, S., Dickson, H., … Kempton, M. J. (2022). Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream. *NeuroImage*, *246*, 118751. Retrieved from https://doi.org/10.1016/j.neuroimage.2021.118751

Hibar, D. P., Westlye, L. T., Doan, N. T., Jahanshad, N., Cheung, J. W., Ching, C. R. K., … Andreassen, O. A. (2018). Cortical abnormalities in bipolar disorder: An MRI analysis of 6503 individuals from the ENIGMA bipolar disorder working group. *Molecular Psychiatry*, *23*(4), 932–942. Retrieved from https://doi.org/10.1038/mp.2017.73

Kambeitz-Ilankovic, L., Meisenzahl, E. M., Cabral, C., von Saldern, S., Kambeitz, J., Falkai, P., … Koutsouleris, N. (2015). Prediction of outcome in the psychosis prodrome using neuroanatomical pattern classification. *Schizophrenia Research*, ***173***(3), Retrieved from https://doi.org/10.1016/j.schres.2015.03.005

Keramatian, K., Chakrabarty, T., Saraf, G., & Yatham, L. (2021). Transitioning to bipolar disorder: A systematic review of prospective high-risk studies. *Current Opinion in Psychiatry*, Publish Ahead of Print. Retrieved from https://doi.org/10.1097/YCO.0000000000000762

Kerner, B. (2014). Genetics of bipolar disorder. *The Application of Clinical Genetics*, *7*, 33–42. Retrieved from https://doi.org/10.2147/TACG.S39297

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*(6), 593. Retrieved from https://doi.org/10.1001/archpsyc.62.6.593

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., … PRONIA Consortium. (2021). Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry*, *78*(2), 195–209. Retrieved from https://doi.org/10.1001/jamapsychiatry.2020.3604

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., … Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *The Lancet Psychiatry*, *3*(10), 935–946. Retrieved from https://doi.org/10.1016/S2215-0366(16)30171-7

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., … Borgwardt, S. (2018). Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression. *JAMA Psychiatry*, *75*(11), 1156–1172. Retrieved from https://doi.org/10.1001/jamapsychiatry.2018.2165

Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., … Borgwardt, S. (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin*, *41*(2), 471–482. Retrieved from https://doi.org/10.1093/schbul/sbu078

Lambert, M., Bock, T., Naber, D., Löwe, B., Schulte-Markwort, M., Schäfer, I., … Karow, A. (2013). Die psychische gesundheit von kindern, jugendlichen und jungen erwachsenen – teil 1: Häufigkeit, störungspersistenz, belastungsfaktoren, service-inanspruchnahme und behandlungsverzögerung mit konsequenzen. *Fortschritte der Neurologie Psychiatrie*, *81*(11), 614–627. Retrieved from https://doi.org/10.1055/s-0033-1355843

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Leopold, K., Ritter, P., Correll, C. U., Marx, C., Özgürdal, S., Juckel, G., … Pfennig, A. (2012). Risk constellations prior to the development of bipolar disorders: Rationale of a new risk assessment tool. *Journal of Affective Disorders*, *136*(3), 1000–1010. Retrieved from https://doi.org/10.1016/j.jad.2011.06.043

Lin, K., Shao, R., Geng, X., Chen, K., Lu, R., Gao, Y., … So, K.-F. (2018). Illness, at-risk and resilience neural markers of early-stage bipolar disorder. *Journal of Affective Disorders*, *238*, 16–23. Retrieved from https://doi.org/10.1016/j.jad.2018.05.017

McWhinney, S. R., Abé, C., Alda, M., Benedetti, F., Bøen, E., Mar Bonnin, C., … the ENIGMA Bipolar Disorders Working Group. (2021). Diagnosis of bipolar disorders and body mass index predict clustering based on similarities in cortical thickness – ENIGMA study in 2436 individuals. *Bipolar Disorders*, ***24***(5), bdi.13172. Retrieved from https://doi.org/10.1111/bdi.13172

Merikangas, K. R., Jin, R., He, J.-P., Kessler, R. C., Lee, S., Sampson, N. A., … Zarkov, Z. (2011). Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of General Psychiatry*, *68*(3), 241–251. Retrieved from https://doi.org/10.1001/archgenpsychiatry.2011.12

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press

Mikolas, P., Bröckel, K., Vogelbacher, C., Müller, D. K., Marxen, M., Berndt, C., … Pfennig, A. (2021). Individuals at increased risk for development of bipolar disorder display structural alterations similar to people with manifest disease. *Translational Psychiatry*, *11*(1), 485. Retrieved from https://doi.org/10.1038/s41398-021-01598-y

Mourão-Miranda, J., Almeida, J. R. C., Hassel, S., de Oliveira, L., Versace, A., Marquand, A. F., … Phillips, M. L. (2012). Pattern recognition analyses of brain activation elicited by happy and neutral faces in unipolar and bipolar depression. *Bipolar Disorders*, *14*(4), 451–460. Retrieved from https://doi.org/10.1111/j.1399-5618.2012.01019.x

Müller, D. K., Küttner, R., & Hannig, R. (2015). NICePype: A Web-based pipeline manager for processing neuroimaging data based on Nipype. *Proceedings of the International Society for Magnetic Resonance*, *23*, 3743. Retrieved from https://cds.ismrm.org/protected/15MProceedings/PDFfiles/3743.pdf

Nieuwenhuis, M., van Haren, N. E. M., 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*(3), 606–612. Retrieved from https://doi.org/10.1016/j.neuroimage.2012.03.079

Nunes, Nunes, A., Schnack, H. G., Ching, C. R. K., Agartz, I., Akudjedu, T. N., … Hajek, T. (2020). Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Molecular Psychiatry*, *25*(9), 2130–2143. Retrieved from https://doi.org/10.1038/s41380-018-0228-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1 Suppl), S199–S209. Retrieved from https://doi.org/10.1016/j.neuroimage.2008.11.007

Pfennig, A., Bschor, T., Falkai, P., & Bauer, M. (2013). The diagnosis and treatment of bipolar disorder. *Deutsches Aerzteblatt Online*, *110*(6), Retrieved 21 June 2021 from https://doi.org/10.3238/arztebl.2013.0092

Pfennig, A., Jabs, B., Pfeiffer, S., Weikert, B., Leopold, K., & Bauer, M. (2011). Health care service experiences of bipolar patients in Germany survey prior to the introduction of the S3 Guideline for diagnostics and treatment of bipolar disorders. *Nervenheilkunde*, *30*(05), 333–340. Retrieved from https://doi.org/10.1055/s-0038-1627819

Pfennig, A., Leopold, K., Martini, J., Boehme, A., Lambert, M., Stamm, T., … Bauer, M. (2020). Improving early recognition and intervention in people at increased risk for the development of bipolar disorder: Study protocol of a prospective-longitudinal, naturalistic cohort study (Early-BipoLife). *International Journal of Bipolar Disorders*, *8*(1), 22. Retrieved from https://doi.org/10.1186/s40345-020-00183-4

Post, R. M., Altshuler, L. L., Kupka, R., McElroy, S. L., Frye, M. A., Rowe, M., … Nolen, W. A. (2018). Multigenerational transmission of liability to psychiatric illness in offspring of parents with bipolar disorder. *Bipolar Disorders*, *20*(5), 432–440. Retrieved from https://doi.org/10.1111/bdi.12668

Radua, J., & Carvalho, A. F. (2021). Route map for machine learning in psychiatry: Absence of bias, reproducibility, and utility. *European Neuropsychopharmacology*, *50*, 115–117. Retrieved from https://doi.org/10.1016/j.euroneuro.2021.05.006

Ritter, P. S., Bermpohl, F., Gruber, O., Hautzinger, M., Jansen, A., Juckel, G., … Bauer, M. (2016). Aims and structure of the German Research Consortium BipoLife for the study of bipolar disorder. *International Journal of Bipolar Disorders*, *4*(1), 26. Retrieved from https://doi.org/10.1186/s40345-016-0066-0

Roberts, G., Green, M. J., Breakspear, M., McCormack, C., Frankland, A., Wright, A., Levy, F., Lenroot, R., Chan, H. N., ... Mitchell, P. B. (2013) Reduced inferior frontal gyrus activation during response inhibition to emotional stimuli in youth at high risk of bipolar disorder. *Biological Psychiatry*, *74*, 55–61.

Roberts, G., Lord, A., Frankland, A., Wright, A., Lau, P., Levy, F., … Breakspear, M. (2017). Functional dysconnection of the inferior frontal gyrus in young people with bipolar disorder or at genetic high risk. *Biological Psychiatry*, *81*(8), 718–727. Retrieved from https://doi.org/10.1016/j.biopsych.2016.08.018

Schmaal, L., Hibar, D. P., Sämann, P. G., Hall, G. B., Baune, B. T., Jahanshad, N., … Veltman, D. J. (2017). Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA major depressive disorder working group. *Molecular Psychiatry*, *22*(6), 900–909. Retrieved from https://doi.org/10.1038/mp.2016.60

Vogelbacher, C., Sommer, J., Schuster, V., Bopp, M. H. A., Falkenberg, I., Ritter, P. S., … Jansen, A. (2021). *The German research consortium for the study of bipolar disorder* (*BipoLife*): *A magnetic resonance imaging study protocol* (preprint). Retrieved 18 June 2021 from In Review: 10.21203/rs.3.rs-339978/v1.

https://doi.org/10.1017/S0033291723001319 Published online by Cambridge University Press