APSIPA Transactions on Signal and Information Processing, 2022, 11, e14

Original Paper

Maximum Credibility Voting (MCV) – An Integrative Approach for Accurate Diagnosis of Major Depressive Disorder from Clinically Readily Available Data

Yu Shimizu[1], Junichiro Yoshimoto[2], Masahiro Takamura[3], Go Okada[3], Tomoya Matsumoto[3], Manabu Fuchikami[3], Satoshi Okada[3], Shigeru Morinobu[3], Yasumasa Okamoto[3], Shigeto Yamawaki[3] and Kenji Doya[1]*

[1] *Okinawa Institute of Science and Technology, Japan*

[2] *Nara Institute of Science and Technology, Graduate School of Information Science, Japan*

[3] *Hiroshima University, Japan*

ABSTRACT

Diagnosis of Major Depressive Disorder (MDD) is currently a lengthy procedure due to the low diagnostic accuracy of clinically readily available biomarkers. We integrate predictions from multiple datasets based on a credibility parameter defined on the probabilistic distributions of the respective models. We demonstrate by means of structural and resting-state functional magnetic resonance imaging and blood markers obtained from 62 treatment naive MDD patients (age 40.63 ± 9.28, 36 female, HRSD 20.03 ± 4.94) and 66 controls without mental disease history (age 35.52 ± 12.91, 30 female), that our method called Maximum Credibility Voting (MCV) significantly increases diagnostic accuracy from about 65% average classification accuracy of individual biomarker models) to 80% (accuracy after integration of the models). Classification results from different combinations of the available datasets validate the method's stability with respect to redundant or contradictory predictions. By

*Corresponding author: Yu Shimizu, yu.shimizu@aikomi.co.jp.

Received 31 July 2021; Revised 02 November 2021
ISSN 2048-7703; DOI 10.1561/116.00000042

definition, MCV is applicable to any desired data and compatible with missing values, ensuring continued improvement of diagnostic accuracy and patient comfort as new data acquisition methods and markers emerge.

Keywords: Biomarkers, major depression, machine learning, multimodal

1 Introduction

Due to a lack of profound knowledge of functional and physiological characteristics, the diagnosis of major depressive disorder (MDD) is currently based on lengthy and tedious evaluations of behavioral symptoms [9]. The complexity in expression, as well as progression of these symptoms, further impede diagnostic procedures. However, reliable diagnosis and connection of the symptoms to physiology is a prerequisite for effective psychological and pharmacological interventions. Hence, with the number of MDD patients surging worldwide, identification of accurate and distinctive fingerprints of the disease is becoming increasingly urgent [36]. Advances in neuroimaging and data analysis techniques have triggered an intensive search for MDD-relevant biomarkers, continuously revealing statistically significant differences between depression patients and healthy controls [36] and encouraging diagnosis of MDD through machine learning algorithms [25]. Despite significant progress, however, accurate clinically translatable biomarkers for MDD are yet to be defined [13, 34, 36], one pervasive limitation being the heterogeneity inherent in MDD studies. As a result, individual biomarkers generally exhibit low specificity and sensitivity and are prone to confounding factors.

Integrative evaluation methods using multiple biomarkers have thus been suggested to improve prediction. Hahn *et al.* [15] trained Gaussian process (GP) classifiers of brain activity during three separate tasks involving emotional and reward processing and integrated their predictions using a decision tree algorithm. This method resulted in a classification accuracy of 83% (sensitivity, 80%; specificity, 87%), an improvement in accuracy of 11% compared with the single best of all GP classifiers, suggesting that several neurological pathways contribute to a more robust classification. A similar effect was demonstrated with multiplex protein assays, where over 90% prediction accuracy was achieved [1, 24] by mathematically integrating nine blood protein markers as well as physical measures into a single MDD score, again suggesting the consideration of multiple biological pathways as robust source of tissue-based MDD biomarkers with trait and state characteristics [36]. Considering the multiple facets of MDD symptoms and development, this seems a reasonable if not necessary approach. While both of the aforementioned studies are very promising, an obstacle for clinical application is that they are both highly specific with respect to the acquired biomarkers, as well as data evaluation.

We propose a novel, more flexible method, which allows for integration of any biomarker acquired by any modality. The idea is to simplify the process of deciding which of the available biomarkers delivers the most reliable diagnosis for a certain subject. In the studies described above, this is done by creating a decision tree and mathematical formula, respectively, both of which are tailored specifically for the available biomarkers. In a previous study [28] we reported that probabilistic classification algorithms can provide an estimation of the prediction reliability based on the overlap of the odds ratio distributions for target and control group. Here, we make this idea explicit and introduce a prediction credibility measure $C(z)$, defined on these odds ratio distributions, which expresses the chance that a certain odds ratio predicts the correct diagnosis. Essentially, this measure can be seen as local prediction accuracy. This credibility measure can be estimated for any probabilistic model, enabling ranking of multiple predictions according to their diagnostic reliability. The most reliable prediction is considered diagnostically valid. Commonly, it is believed that the higher the absolute value of the odds ratio, the higher the chance of correct prediction. However, this is only true if the distributions are to a certain extent well separated and even then the problem of prediction reliability for small odds ratios, where the distributions overlap, still remains. We demonstrate that in contrast to other simple integration methods applied to the odds ratios (majority of votes, sum, maximum, and mean of odds ratio), our method called *Maximum Credibility Voting (MCV)*, consistently improves prediction accuracy. It is further superior to common approaches such as support vector machine (SVM) and regression tree with respect to its formulation. It is formulated such that the addition of biomarkers remains optional and its application to data with missing values (i.e., measurements) is possible without modification of the algorithm. The only requirement is the estimation of the credibility measure for the model of the new data.

We demonstrate the aforementioned benefits of MCV for clinical MDD diagnosis by means of anatomical magnetic resonance imaging (MRI) data, resting state functional MRI (rsfMRI) data, and the methylation rate at several CpG islands in the promotor region of the brain derived neurotrophic factor (BDNF) gene. We further demonstrate how the results of MCV can be used to stratify subjects based on which data yields the most reliable diagnosis and which MDD symptoms they exhibit. In contrast to task-based functional MRI data acquisition, structural MRI and rsfMRI scans can be rapidly acquired and processed. Instructions are simple and prognostic procedures can be kept short and uncomplicated, making them suitable for a wide range of subjects (i.e., subjects with difficulties adhering to task paradigms). Further, identified discriminative brain regions can be directly related to their physiological cause, namely, substance loss or increase in brain tissue in the case of structural data and loss or increase of spontaneous neural activity in the case of rsfMRI. The same is true for the methylation rate, which inhibits gene transcription and in turn inhibits neurogenesis in cortex [35]. Unlike protein identification, determination of the methylation rate at several sites can be accomplished in a single measurement.

The only requirement for classification models of the data we wish to integrate is that they are of probabilistic nature. In this study, we chose Elastic Net and sparse logistic regression with a least absolute shrinkage and selection operator (LASSO). These models limit the number of effective variables by penalizing the sum of the absolute weights or the sum of squared weights and setting small weights to zero depending on a given threshold. As a result, they return a model based on features with the most discriminative relevance only (compare with SVM, which is inherently based on all input features, so that there is a risk of noise or nothing at all). In addition, we have previously shown that these algorithms can successfully handle a number of features that are several times larger than the number of training instances [28]. For this reason, we opted for whole brain analysis rather than targeted brain area analysis, with the intent to construct an unbiased classification model and to reveal new brain areas as MDD indicators. The same holds for the application of these algorithms to the methylation data.

2 Methods

2.1 Subjects

Sixty-two MDD patients (age 40.63 ± 9.28, 30 female), free of substance-related disorders other than alcohol and any co-morbidity were recruited by the Psychiatry Department of Hiroshima University. They were diagnosed by senior psychiatrists according to DSM-IV [9] criteria, interviews and information from medical records. Diagnosis was reconfirmed by experienced psychiatrists and psychologists at the time of participation in the study, using the Japanese version of the Mini-International Neuropsychiatric Interview (M.I.N.I [27]), which has been shown to have good to excellent interrater and test-retest reliability [23]. All patients had been treated less than 14 days at the time of participation in the study. Beck Depression Inventory (BDI) scores for this group ranged from 11 to 53 (average 30.52 ± 9.08) and Patient Health Questionnaire (PHQ9) scores from 6 to 26 (average 17.71 ± 4.5). About half of the patients had experienced a previous depression period. Age of depression onset was 38.22 ± 1.07. The length of the episode at time of the study varied considerably (160.98 ± 209.54 days). Depression severity was evaluated using the Hamilton Rating Scale for Depression (HRSD17 20.03 ± 4.94). In addition the scores for following self-reported measures were recorded:

- Snaith-Hamilton Pleasure Scale (SHAPS): A 14 items questionnaire assessing four domains of pleasure response and hedonic experience within the four pleasure domains interest and pastimes, social interaction, sensory experience, and food and drink. Subjects can answer with: Strongly disagree, Disagree, Agree or Strongly agree. Either of the Disagree responses receive a score of 1 and either of the Agree responses receives a score of 0. The SHAPS is scored as the sum of the 14 items so that total scores ranged from 0 to 14. A higher total SHAPS score indicated higher levels of anhedonia. A cut-off score of 2 provides the best discrimination between "normal" and "abnormal" level of hedonic tone.

- State Trait Anxiety Inventory (STAI): Also based on a 4-point scale, the STAI consists of 40 questions on a self-report basis. The STAI measures two types of anxiety: state anxiety, or anxiety about an event, and trait anxiety, or anxiety level as a personal characteristic. Higher scores are positively correlated with higher levels of anxiety.

- Child Abuse Trauma Scale (CATS): 38-item measure designed to assess subjective memories and perspectives of adolescents and adults with respect to child abuse and maltreatment. Respondents report on their experiences with both parents combined when they were a child or teenager, responding on a 5-point scale from 0 (never) to 4 (always).

- Life Event Scale (LES): a 23-item scale that assesses whether or not a stressful event happened over the past year and how stressful the event was (stressful (0) to very stressful (3)).

As a control group, 66 persons (age 35.52 ± 12.91, 36 female) free of mental or neurological disease history were recruited by advertisements in local newspapers. All healthy controls (HC) underwent the same self-assessments and examinations administered to the MDD group (except for MDD-specific HRSD). BDI scores were between 0 and 24 (average 6.74 ± 5.88) and PHQ9 ranged from 0 to 18 (average 3.36 ± 3.76). 61 MDD subjects (out of 62) had values ≥ 14 (standard cutoff for MDD at Hiroshima University), while 57 controls (out of 66) had scores under 14. For PHQ9 scores with a cutoff value of 10, 60 patients had scores ≥ 10 and 61 controls under 10. Subjects of both groups completed the Japanese version of the National Adult Reading Test [21] for an estimate of their IQ (108.37 ± 9.81 for the MDD group and 113.32 ± 8.03 for controls, Table 1).

Written informed consent was obtained from all participants (approved by the Research Ethics Committee of the Okinawa Institute of Science and Technology and the Research Ethics Committee of Hiroshima University).

Table 1: Demographic and clinical characteristics of all subjects included in the study.

	MDD	Control	p-value
Number of subjects	62	66	–
Sex (male/female)	32/30	30/36	0.49
Age (years)	40.63 ± 9.28	35.52 ± 12.91	0.04^{*a}
IQ	108.37 ± 9.81	113.32 ± 8.03	0.6
Alcohol dependent subjects	5	0	0.02^{*}
BDI II[b]	30.52 ± 9.08	6.74 ± 5.88	2.03e-09***
PHQ 9	17.71 ± 4.50	3.36 ± 3.76	9e-14***
SHAPS	37.26 ± 5.46	23.62 ± 6.13	2.34e-09***
STAI	56.48 ± 7.76	40.5 ± 8.82	1.45e-05***
CATS	34.75 ± 23.20	24.89 ± 14.30	0.56
LES	−6.57 ± 6.39	−0.71 ± 3.90	0.006^{**}
HRSD	20.03 ± 4.94	–	–
Age of depression onset (years)	38.22 ± 1.07	–	–
Number of previous episodes	0.61 ± 0.94	–	–
Length of current episode (days)	160.98 ± 209.54	–	–
Lexapro single agent	52	–	–
Lexapro combination	2	–	–
Other single agent	2	–	–
No treatment	6	–	–

Note: [a]Asterisks denote significant differences: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
[b]Nomenclature: BDI II = Beck's Depression Inventory II, PHQ9 = 9 Question Patient Health Questionnaire, SHAPS = Snaith-Hamilton Pleasure Scale, STAI = State Trait Anxiety Inventory, CATS = Child Abuse and Trauma Scale, LES = Life Event Stress, HRSD17 = 17 Question Hamilton Rating Scale of Depression.

2.2 Data

Anatomical and resting state functional MRI data were collected from all participants. Blood samples could only be obtained for a subgroup.

2.2.1 MRI Data

Anatomical T1 images were acquired on a 3T GE Signa HDx scanner (IRP FSPGR, TR = 6.32 ms, FA = 20, voxel size 1 × 1 × 1 mm, matrix size

256 × 256 × 180) and processed using VBM8 (Christian Gaser, University of Jena, Department of Psychiatry), yielding voxel wise white matter (WM) and gray matter (GM) density maps.

For acquisition of resting state functional MRI (rsfMRI) data, subjects were asked to close their eyes and relax. Images were obtained over 5 minutes, resulting in 145 images (2D EP/GR, TR = 2000 ms, no gaps, interleaved, matrix size 64 × 64 × 32, voxel size 4 × 4 × 4 mm). Debriefing routinely conducted after the scans revealed that two subjects had fallen asleep during the measurement. Their data were thus excluded from analysis. Measurements during which patients had moved more than 3 mm or 3 degrees translationally or rotationally, respectively, were also excluded (exclusion of the whole time series). The difference in motion between MDD and HC subjects of which data was used to estimate classification models was not significant ($p = 0.68$, average framewise displacement for controls 0.08 ± 0.10 mm and 0.07 ± 0.05 mm for MDD subjects, see appendix).

Images were realigned, normalized and smoothed (FWHM = 8 mm) using SPM8 (Wellcome Trust Centre for Neuroimaging, UCL, London). Motion was regressed out using the standard six-head motion parameters. Time series were band pass filtered (0.009–0.1 Hz) and de-trended using the rsfMRI Data Analysis Toolkit (REST [31]). Using the band pass filtered rsfMRI measurements, functional connectivity between regions of interest (ROIs) was evaluated as correlation coefficients between the average time series of BOLD fMRI signals of ROIs. The ROIs consist of 90 brain regions across 14 intrinsic connectivity networks that were derived by means of Independent Component Analysis [29]. These networks comprise (number of ROIs in parentheses): Anterior Salience (7), Auditory (3), Basal Ganglia (5), Dorsal Default Mode (9), Language (7), Left Executive Control (6), Precuneus (4), Posterior Salience (12), Right Executive Control (6), Ventral Default Mode (10), Visuospatial (11), Primary Visual (2), Higher Visual (2), and Sensorimotor (6) network. Nifti templates of the ROIs are publicly available [22]. We discarded correlations if the probability that there is an actual relationship between the time series, was small ($p > 0.01$) for all healthy subjects. This ensures that connections passed to the classification algorithms are functionally meaningful in terms of synchronization. The threshold of 0.01 was used in order to retain connections for which there is substantial evidence against the null hypothesis (i.e., that the found correlation is coincidence), while leaving a margin for weaker correlations that might be of importance to establish group differences. This resulted in 108 connections.

While functional connectivity identifies spatial patterns of synchronous low-frequency oscillations on a network level, it does not reveal information on localized dysfunctions of specific brain regions, which ultimately contribute to network abnormalities. Such local brain activity can be assessed by evaluating low-frequency oscillations themselves. We have done this by calculating the Amplitude of Low frequency fluctuation (ALFF), which accounts for the summed amplitude in the low frequency range (0.009–0.1 Hz), and Regional Homogeneity (ReHo), which accounts for signal homogeneity between neighbouring voxels. We also assessed fractional ALFF (fALFF), which evaluates the ratio of a low frequency amplitude (0.009–0.1 Hz) with respect to the amplitude of the whole frequency spectrum. As a results we have an absolute measure of low frequency fluctuations (ALFF) and a relative measure (fALFF). While fALFF is robuster to physiological noise, ALFF shows higher test-retest reliability in gray matter regions and thus more sensitive for differences between groups [38]. However, we decided to evaluate both measures due to their different characteristics. All three local parameters were assessed using REST [31].

2.2.2 Bloodmarkers

Genomic DNA was extracted from the acquired blood samples and the methylation rate at 32 CpG islands at promoters of the BDNF exon1 gene was assessed using a MassArrayH system (SEQUENOM). The majority of these sites have previously been shown to be related to depression [12] (see the appendix for details).

For each acquired diagnostic feature, data were age- and sex-matched with respect to patient and control group. The number of subjects in each group was matched in order to avoid sample size bias during model estimation. This resulted in 60 subjects per group for the anatomical data, 42 subjects for per group for the resting state data and 33 subjects per group for the BDNF methylation data (see the appendix for demographic and clinical comparison of the groups used for each feature).

2.3 Classification

2.3.1 Datawise Classification

The sole requirement for MCV is for the applied models to be probabilistic, i.e., the models provide odds ratio distributions for control and target (i.e., MDD patient) group. As previously mentioned, we chose algorithms that are effective in handling a large number of features relative to the number of subjects [28]. In this way, we can make use of all information in the datasets and achieve classification without bias introduced by prior assumptions.

We applied Elastic net and logistic regression with Least Absolute Shrinkage and Selection Operator (sLASSO) regularization to the brain area-wise mean of WM, GM, rsfMRI ALFF, and ReHo maps. We also applied both methods to FC and BDNF methylation. The whole volume data of WM, GM, ALFF, and ReHo were also subjected to group LASSO (gLASSO) regression, where the sum of weights for voxels located in the same brain area (defined in the anatomical labeling atlas AAL [33]) is constrained, resulting in brain area-wise reduction of discriminative voxels [28]. These sparse algorithms return only features with the greatest diagnostic relevance, clearly identifying MDD correlates.

Validity of models and their regularization parameters were assessed using 10-fold nested cross-validation, repeated 10 times, each time shuffling training and test data.

Model evaluation was based on standard parameter accuracy (percentage of correctly diagnosed subjects), specificity (the percentage of healthy controls, correctly identified as such) and sensitivity (percentage of patients correctly identified as such).

2.3.2 Integrated Classification – MCV

Probabilistic classifiers yield negative and positive log odds ratios $z = \ln(p/(1-p))$, where p is the predictive probability that the subject belongs to the target group (here, the MDD group). These log odds ratios (usually) assume different distributions for control and target groups. The overlap of these two distributions gives information on the reliability of a prediction with a certain odds ratio. We fit the normalized log odds ratio distributions for HCs and patients in the training data using the Weibull distribution function (Figure 1):

$$W_{\lambda k}(z) = \frac{k}{\lambda}(\frac{z+1}{\lambda})^{k-1} e^{-(\frac{z+1}{\lambda})^k}, \tag{1}$$

$z > 0$, where $k > 0$ allows for a skewed shape of the distribution and $\lambda > 0$ determines the width of the distribution. These parameters were fitted using maximum likelihood estimation. Through these parameters the Weilbull function can assume the properties of a whole family of distributions including the normal distribution. While we do not expect the distributions of the odds ratios to be eg exponentially distributed, they can very well be skewed normal distributions. The usage of the Weibull distribution thus frees us from assumptions on the odds ratio distribution. A minor drawback is that its accuracy is dependent on the availability of enough data. Since the Weibull function exists strictly only for positive values, log odds ratios were shifted by 1 before fitting and were shifted back thereafter.

We chose the Weibull distribution over the normal distribution, because it is very flexible. Through its parameters, it can assume the properties of several other distributions. While we do not expect the distributions of the odds ratios to be eg exponentially distributed, they can very well be skewed normal distributions. The usage of the Weibull distribution frees us from assumptions on the distribution. A drawback is that its accuracy is dependent on the availability of enough data. We have added a comment in the method section near the definition of the Weibull function.

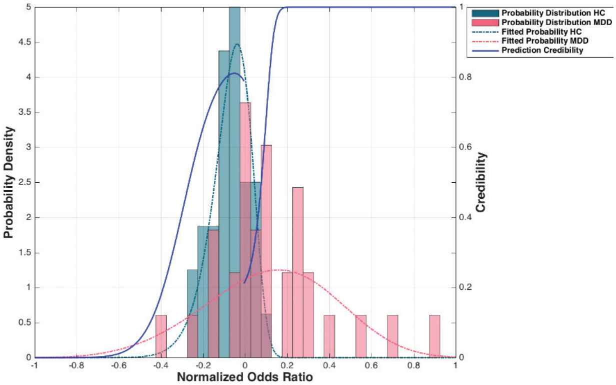

Figure 1: The *credibility* $C(z)$ (Eq. 2) of each log odds ratio is defined through the ratio of the Weibull distributions fitted to the log odds ratio distributions of healthy controls (W_+) and MDD patients (W_-), respectively (Eq. 1). It reflects the portion of true predictions among all predictions with certain odds ratio. Here, the low credibility of small negative odd ratios compared to the high credibility for large positive log odds ratios, is a result of false negatives toward the end of the spectrum.

In the following, we denote the Weibull distribution fitted for log odds ratios of HCs and MDD patients as W_- and W_+, respectively. The values of W_- and W_+ at a certain log odds ratio allow estimation of how many false positive or negatives in comparison to true positives or negatives we can expect. In other words, they give an estimate on how high the chance is for a prediction with a specific log odds ratio to be correct. We define the credibility function C as:

$$C(z) = \begin{cases} W_-/(W_- + W_+) & \text{for } z < 0 \\ W_+/(W_- + W_+) & \text{for } z > 0, \end{cases} \tag{2}$$

the ratio of true (negative or positive) predictions within all (negative or positive) predicted outcomes (Figure 1). For the sake of completion, we define the $C(z) = C(1)$ for $z > 1$ and $C(z) = C(-1)$ for $z < -1$ (i.e., predictions with log odds ratios outside the normalised range are assigned the respective credibility at the far ends of the distributions).

We use this credibility measure to estimate diagnosis reliability for predictions obtained from each dataset. The prediction with the highest credibility is chosen as the final diagnosis. We refer to this method as MCV:

$$MCV(z_1, \ldots, z_N) = I(z_j > 0), \tag{3}$$

with $j = \text{argmax}_{i=1:N}\, C(z_i)$, z_i the normalized log odds ratios obtained from N models yielded by different data obtained from the same subject and z_j the odds ratio with the highest credibility value ($C(z_i)$ the credibility function as

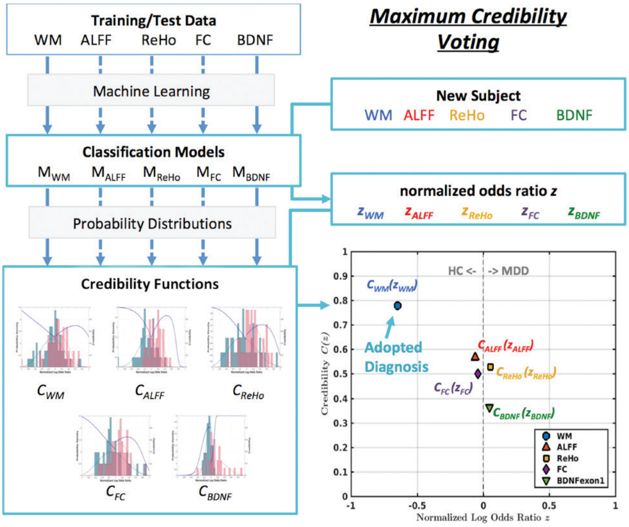

Figure 2: *MCV*. Diagnosis is based on the log odds ratios z_i of multiple classification models derived from different types of data (here, WM, ALFF, ReHo, FC, and BDNF related data) and determined by the prediction (odds ratio) z_j with the highest credibility value $C_j(z_j)$ (Eq. 2). The credibility functions C_j are derived from the odds ratio distributions of the respective classification models. In the depicted case, WM decides with nearly 80% credibility, that the subject is a healthy control.

defined in Eq. 2). For $MCV(z_1, \ldots, z_N) = 1$, the subject's diagnosis is MDD and for $MCV(z_1, \ldots, z_N) = 0$ healthy. Intuitively, this procedure is straight forward. Out of several predictions, we pick the one we can trust the most (Figure 2). We remark that MCV itself does not require training data per se, but a method that can be applied as is. However, the more training data are available for each of the underlying classification models, the more accurate their credibility functions (due to the increased number of data points (i.e., odds ratios) outlining the probability distributions, Figure 2, left column). As a result, MCV is more effective.

We validated MCV using 10-fold cross validation, where prediction credibility functions for each model were estimated based on the log odds ratio distributions obtained in each cross validation of the training data. In this way, test and training data are kept independent throughout the model estimation procedure, as well as in the following MCV procedure.

We compared MCV to other model integration methods, where the number of negative or positive predictions (most votes), sum, maximum and mean of log odds ratio, respectively, determine diagnosis. Further, comparisons to the performance of SVM and classification tree are made. As opposed to the arithmetic approaches, SVM and classification tree cannot be applied to data with missing values.

3 Results

3.1 Diagnostic Accuracy for Each Diagnostic Feature

For clarity, we only consider one model per data modality and restrict the MCV results to the models that achieved the highest classification accuracy in each data modality. Models with accuracy lower than 60% are disregarded. Further, only features selected in more than 80% of all cross validated models are presented and considered of diagnostic relevance. A performance summary for all models and their diagnostic features can be found in the appendix.

3.1.1 Anatomical MRI

For the white matter and gray matter density volumes, only white matter classification using group LASSO yielded an accuracy over 60% ($63 \pm 2\%$ accuracy, $58 \pm 3\%$ specificity, $68 \pm 4\%$ sensitivity). Left and right post central cortex, left frontal superior cortex and right middle temporal cortex were assigned negative weights, indicating that these areas are denser in white matter in healthy controls than in depression patients. Left middle temporal cortex was assigned positive weight.

3.1.2 Resting State fMRI

Mean brain area ALFF subjected to Elastic Net, showed the best classification among the resting state data with an accuracy of $68 \pm 1\%$ (specificity $67 \pm 2\%$, sensitivity $69 \pm 1\%$). Left and right posterior cingulate cortex (PCC) and left thalamus thereby showed negative weights indicating that the amplitude in these areas is lower in MDD patients than healthy controls.

L1 and Elastic Net applied to the mean ReHo values in brain areas showed similar performance of 66% accuracy, 65% specificity and 67% sensitivity, but with slightly smaller variance for the model estimated with Elastic Net. Both models assigned negative weights to left PCC, medial orbitofrontal cortex (OFC). Elastic Net also showed negative weight for the left amygdala. Positive weight was assigned to the left cerebellum pars8 in both models and additionally to the right cerebellum pars 8 in the Elastic Net.

For the FC data L1 LASSO yielded the best classification with $65 \pm 4\%$ accuracy, $65 \pm 5\%$ specificity and $66 \pm 6\%$ sensitivity. Two connections with negative weights were selected as diagnostic: the connection between right parahippocampus and right retrosplenial cortex including a part of the posterior cingulate (network 13, areas 08 and 05) and the connection between PCC/Precuneus and medial prefrontal/anterior cingulate/orbitofrontal area (network 3, areas 04 and 01). While the first connection is part of the ventral default mode network (DMN), the other is part of the dorsal DMN.

3.1.3 BDNF exon1 methylation

Elastic Net yielded $78 \pm 3\%$ accuracy ($84 \pm 5\%$ specificity, $73 \pm 2\%$ sensitivity) and assigned negative weights to 12 (out of 32) sites, positive weights to 11 sites. CpG1, CpG18, CpG24, CpG52, CpG61, CpG63, CpG77 with negative weights were in agreement with the results given in Fuchikami *et al.* [12]. The contributing islands, CpG19.20.21, CpG28, CpG32 showed reversed relation to the results in that study; however, the difference in CpG28, Cpg32 methylation between the two investigated subject groups was not significant. CpG25.26.27, CpG29.30.31 methylation, both without significant group differences and CpG33.34 methylation with significant group difference were not measured in Fuchikami *et al.* Indication of a role in MDD diagnosis opposite to the one found in their study was also true for CpG5, CpG15, CpG36, CpG37, CpG48 and CpG78, which were assigned positive weights in our model, but were more highly methylated in healthy controls than in MDD subjects in their study. In our study, methylation differences in these sites exist, but were not significant. CpG8.9, CpG14 also countered the relation given in their study, but in both their and our study, group differences were not significant. CpG17, CpG50.51, CpG74.75, CpG22 agreed with Fuchikami *et al.*, but here again, significant group differences could not be found.

To summarize, none of the methylation sites that were assigned positive weights showed significant group differences. All sites with significant methylation differences between healthy and MDD subjects were negatively weighted and selected in over 98% of all cross-validated models. Hence, the contribution of these sites with positive weights, to the diagnostic power lie in the combination with the other selected methylation sites. Mean methylation rates were generally lower in our study than in Fuchikami *et al.*, which can be attributed to the greater number of subjects (HC/MDD = 33/33 to 18/20).

3.2 MCV

For demonstration of MCV, we chose the highest performing model for each diagnostic data: the group LASSO model for white matter, the Elastic Net model applied to mean brain areas values of ALFF and ReHo, L1 for FC and the Elastic Net model for the BDNF methylation data.

To evaluate MCV for different data combinations, appropriate subsets of the available data were used. Differences between results were considered significant at $p < 0.05$. We also give the F-scores in order to account for the imbalances in cohort sizes in the test data.

All individual models were evaluated repeating ten 10-fold nested cross validations, each time shuffling the subjects. As a result, we obtain ten odds ratios for each subject and dataset available for the subject. Equally, we have ten different odds ratio distributions for controls and MDD subjects, in which the test subjects are not included. We use these to construct the credibility functions and evaluate MCV prediction accuracy.

3.2.1 MCV(All), Application to Dataset with no Missing Data

Within all subjects, structural MRI, resting state MRI and methylation data were available for 23 healthy controls (age 41.22 ± 11.94, 7 female) and 20 MDD patients (age 35.1 ± 6.03, 10 female). For these subjects, the average BDI2 and PHQ9 scores were 8.78 ± 7.00 and 4.30 ± 4.07 for the HC group and 29.60 ± 10.7 and 18.70 ± 4.53 for the MDD group. Detailed demographic and clinical characteristics are given in Table 2. Figure 3 shows the data sets acquired for each subject.

The average accuracy when diagnosing this group of subjects based on each biological data modality alone was $66.22 \pm 6.53\%$. Application of MCV significantly improved accuracy to 80 ± 3% (F-score 77 ± 4%), with a specificity of 87 ± 3% and sensitivity of 73 ± 7%. This is an average increase in accuracy of 14% compared to the accuracies delivered by each individual data model alone (Table 3, MCV (WM, rsfMRI, BDNFexon1)).

Note that MCV performance relies solely on the accuracy of the credibility functions, which are constructed based on training data provided for the different classification models, i.e., the number of subjects to which MCV is applied does not influence MCV performance. For the same reason, overfitting can per definition not occur.

Comparing MCV results with those of SVM, Decision Tree, maximum number of votes, sum, mean and maximum absolute value of the odds ratios, we find significant superiority of MCV over all other approaches ($p < 0.001$ for all comparisons, Table 3 and Figure 4). For voting arithmetic methods (Most Votes, Sum, Mean and Maximum Absolute Value of the odds ratios), specificity was higher than that of MCV, but could not achieve comparable accuracy due to low sensitivity.

Adding the predictions from the different models to MCV one by one (Table 4) shows the stepwise increase in accuracy. Naturally, the accuracy significantly increases with addition of models that have higher accuracy to begin with. However, the importance lies in the fact that per construction, the accuracy should never significantly decrease when adding new predictions. We can see this in the decrease in accuracy when adding FC, which is not significant ($p = 0.81$), but shows the same credibility as in the previous step. This confirms that MCV is robust to unreliable and redundant predictions, so that we always end up with a diagnosis about which we can be more confident, even if the accuracy has not improved. We corroborate this fact by evaluating the results for different combinations of data, namely MCV (rsfMRI), MCV (rsfMRI, WM), MCV (WM, BDNF), see the appendix.

Table 2: Demographic and clinical characteristics of subjects for which all data sets were available.

	MDD	Control	p-value
Number of subjects	20	23	
Sex (male/female)	10/10	16/7	0.20
Age (years)	35.1 ± 6.03	41.22 ± 11.94	0.04^{*a}
IQ	109.47 ± 9.45	111.00 ± 9.23	0.59
Alcohol dependent subjects	0	0	1
BDI II[b]	29.6 ± 10.70	8.78 ± 6.99	$2.03e\text{-}09^{***}$
PHQ 9	18.7 ± 4.53	4.30 ± 4.07	$9.00e\text{-}14^{***}$
SHAPS	36.65 ± 4.16	24.04 ± 6.31	$2.34e\text{-}09^{***}$
STAI	55.8 ± 6.04	43.35 ± 9.81	$1.45e\text{-}05^{***}$
CATS	35.25 ± 21.33	31.70 ± 18.71	0.56
LES	-5.5 ± 7.08	−0.52 ± 3.93	0.006^{**}
HRSD17	19.5 ± 4.49	–	–
Age of depression onset (years)	30.5 ± 7.17	–	–
Number of previous episodes	0.75 ± 0.44	–	–
Length of current episode (days)	118.65 ± 85.74	–	–
Lexapro single agent	16	–	–
Lexapro combination	0	–	–
Other single agent	2	–	–
No treatment	2	–	–

Note: [a]Asterisks denote significant differences: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
[b]Nomenclature: BDI II = Beck's Depression Inventory II, PHQ9 = 9 Question Patient Health Questionnaire, SHAPS = Snaith-Hamilton Pleasure Scale, STAI = State Trait Anxiety Inventory, CATS = Child Abuse and Trauma Scale, LES = Life Event Stress, HRSD17 = 17 Question Hamilton Rating Scale of Depression.

The number of true negative (TN) and true positive (TP) subjects in this dataset was 20 ± 1 and 15 ± 1, respectively. In approximately 50% of the cases, their final diagnosis was determined by BDNF methylation (53 ± 7% for TN and 49 ± 11% for TP subjects). ALFF was the second most frequent determining factor, comprising 27±4% of true negatives. For the true positives, this proportion was with 47 ± 14% nearly as high as the BDNF proportion. WM was determining for 13 ± 6% of the true negatives, but only for 2 ± 3% for true positives. For neither group was FC the diagnostic factor.

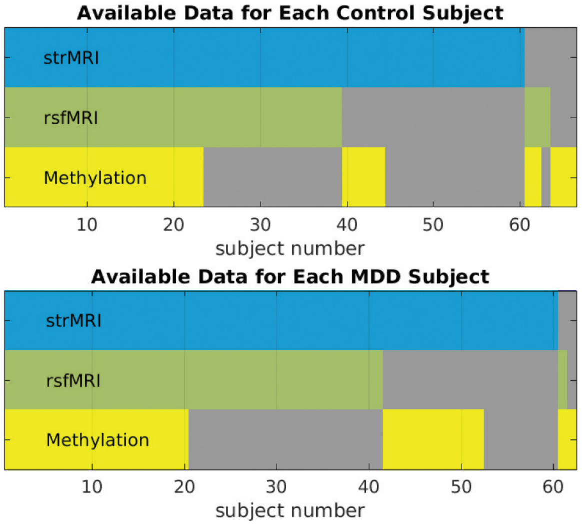

Figure 3: Available data for each subject: structural MRI data (strMRI), resting state fMRI data (rsfMRI) and BDNFexon1 methylation data were evaluated for 60, 42 and 33 subjects, respectively, for each experimental group. For 23 controls and 20 MDD patients strMRI, rsfMRI and methylation data were available. For all the others, only one or two of the biomarkers could be acquired.

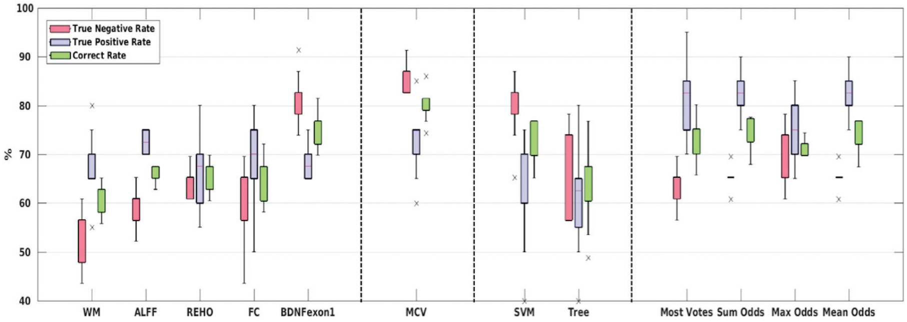

Figure 4: MCV classification accuracy: MCV significantly outperforms single data models, but also the more intuitive integration methods SVM, Decision Tree, most votes, sum of odds, maximum of odds and mean of odds (see also Table 3).

Nineteen subjects were diagnosed as true negative and 15 as true positive in over half the cross validations. For controls in this group, the most frequent diagnostic dataset was BDNF methylation (10 subjects), followed by ALFF (5 subjects), WM (3 subjects) and finally ReHo (1 subjects). Depression subjects in this group were exclusively diagnosed by BDNF methylation (8 subjects) and ALFF (7 subjects).

Table 3: MCV over all available diagnostic datasets outperforms single data classifiers as well as the integration methods SVM, Classification Tree, Most Votes, sum, max and mean of odds ratio. p-values are given with respect to MCV accuracy.

	HC/MDD = 23/20	Specificity	Sensitivity	Accuracy (F-score)	p-value
Single feature Classification	ALFF	60 ± 4	73 ± 2	66 ± 2 (62 ± 4)	2.5e-10***
	FC	67 ± 8	61 ± 4	64 ± 4 (64 ± 5)	3.0e-08***
	WM	53 ± 6	69 ± 7	60 ± 3 (64 ± 6)	1.1e-10***
	BDNFexon1	80 ± 5	68 ± 4	75 ± 3 (71 ± 3)	0.0017**
Integrating methods	SVM	79 ± 6	64 ± 11	72 ± 4 (68 ± 8)	8.5e-06***
	Classification Tree	65 ± 9	61 ± 11	63 ± 8 (60 ± 9)	2.1e-04***
	Most Votes	64 ± 4	81 ± 7	73 ± 4 (73 ± 5)	6.5e-04***
	Sum of odds ratio	65 ± 3	82 ± 5	74 ± 3 (74 ± 3)	2.8e-04***
	Max of odds ratio	69 ± 6	75 ± 6	72 ± 2 (71 ± 3)	9.1e-07***
	Mean of odds ratio	65 ± 3	82 ± 5	73 ± 3 (74 ± 3)	9.9e-05***
	MCV	87 ± 3	73 ± 7	80 ± 3 (77 ± 4)	–

3.2.2 Missing Data Compatibility

MCV can also to be applied to data with missing values. The credibility for missing values is simply set 0; therefore, it does not interfere with the Voting process. For 47 healthy controls and 53 depression patients, at least two different, but not all measurements were available. MCV significantly outperformed all other odds ratio integration methods (here, most votes, sum, max and mean of odds ratio) with at least 7% higher correct rate (see the appendix for details). SVM and classification tree are inherently incompatible with datasets that comprise missing values.

4 Discussion

The results show that MCV allows easy integration of predictions from different datasets and significantly improves classification accuracy. The diagnostic accuracy of MCV when integrating over all data given in this study (81.56 ± 2.58%) is on average 15.33 ± 6.53% more higher than the classification accuracy yielded by the individual diagnostic datasets and 6.40 ± 3.76% higher than the other integration methods investigated (SVM, decision tree, maximum number as votes, sum, maximum and mean odds ratio, Table 3).

In detail, MDD diagnosis is significantly improved through application of MCV to the three characteristics, ALFF, ReHO and FC obtained from rsfMRI data (see the appendix). It could be further improved by integrating white matter density or information on BDNFexon1 methylation at certain sites. The addition of BDNFexon1 methylation to rsfMRI data significantly improved the specificity, while no significant improvement could be observed

Table 4: Stepwise MCV for data without missing values: p-values are given in comparison to the accuracy in the previous step.

HC/MDD = 23/20	Specificity	Sensitivity	Accuracy (F-score)	p-value	Credibility
WM	53 ± 6	69 ± 7	60 ± 3 (62 ± 4)	–	61 ± 2
MCV (WM, ALFF)	60 ± 4	73 ± 4	66 ± 4 (66 ± 3)	0.004***	73 ± 11
MCV (WM, ALFF, ReHo)	63 ± 3	74 ± 3	68 ± 3 (68 ± 3)	0.20	75 ± 10
MCV (WM, ALFF, ReHo, FC)	62 ± 3	74 ± 3	67 ± 3 (68 ± 3)	0.85	75 ± 10
MCV (WM, rsfMRI, BDNFexon1)	87 ± 3	73 ± 7	80 ± 3 (77 ± 4)	1.6e-08***	84 ± 2

for the sensitivity. Looking at the reverse operation, adding rsfMRI to already obtained diagnosis from methylation data, does not improve specificity, but significantly improves sensitivity. For the integration of methylation and white matter density data, only the specificity is significantly improved. However, the sensitivity did not significantly decrease, either. In general, all examined cases showed improved or comparable specificity and sensitivity, confirming that MCV is robust with respect to redundant or little reliable predictions.

With respect to clinical application, the results suggest the following: MDD diagnosis with reasonable accuracy is provided by sLASSO regression of BDNF methylation markers (84 ± 5% specificity, 73 ± 2% sensitivity). Integration of rsfMRI and anatomical data can further increase the accuracy by approximately 6%, lowering the risk of false negatives by nearly 10% and the risk of false positives by 2%. If a blood test is not available, acquisition and MCV integration of rsfMRI data is advisable (The often used PHQ9, which is fast to acquire and to evaluate, has a specificity as well as sensitivity of 88% [18], however, as a self-administered questionnaire the differences in diagnostic accuracy in comparison to the accuracy acquired from biomarkers is difficult to interpret.). Here, the procedure is based on structural MRI, rsfMRI, and BDNF methylation rate data, but the extension of MCV to an arbitrary number of appropriate data is straightforward. Moreover, in contrast to SVM and decision tree, MCV is not restricted by lack of available data for a subject. Credibility for missing data is simply set to zero.

MCV differs from the introduced integration methods [24] in that it does not try to combine all predictions into a new model. It merely decides which of the predictions is the most reliable. Intuitively, that would be the prediction derived from the model with the highest accuracy, but this accuracy is based on global evaluation of the model, ie. the sum of all positive and negative predictions, regardless of their odds ratio, are weighed against the sum of the actual positive and negative labels. In contrast, MCV considers the local accuracy of these models. It quantifies the accuracy of the model for different odds ratios, expressing it as credibility function. Despite low general diagnostic power, for example, the WM model is quite accurate for certain odds-ratios, even more accurate than the other models, thus helping to boost accuracy when using MCV. Finding the data with the highest credibility thus equals identifying the data with the most pronounced MDD or HC characteristics for the subject in question. If we consider all data as part of a single mechanism, MCV points to the weakest or strongest link in the mechanism of a specific subject, respectively. Multiple subjects with the same weakest link suggest an MDD subtype. In our data, for example, ALFF and BDNFexon1 methylation were the most pronounced determining factors for true positives. In an experiment with a bigger cohort, these groups could point to two depression types that might need different treatment. If the demographic and clinical characteristics of these groups are the same, these depression types would indicate physiological subtypes of depression. If the characteristics are different, these depression types would indicate subtypes whose demographic and clinical characteristics are linked to different physiological phenomena. In the latter case, further investigation would be needed to establish if regulation of the physiological symptoms is possible and if it can be used to remedy MDD symptoms.

We exploit the fact that MCV can be used with any type of probabilistic models, by using sparse classification algorithms, which ensure that the effective variables (here, brain regions and methylation islands) are explicit, a crucial aspect for development of effective medication. Knowing details about the contributing factors allows for insight into their possible relation to depression as well as into the relation between the different data modalities, and the underlying mechanism of MDD. The discriminative white matter brain areas, for example, are responsible for somatosensory information processing (post central gyrus), cognitive (frontal superior cortex), and language related functions (middle temporal lobe); functions, which are hypofunctional in MDD [8, 19, 32]. ALFF data identified the PCC and left thalamus (left hemisphere), as MDD discriminative, both of them exhibiting lower activation in depressed patients. Investigation why the spontaneous activity in these areas is suppressed becomes necessary. Low ALFF could be, for example, an indicator of the volumetric changes in PCC seen in first episode MDD subjects [20], but which were not big enough yet in our study to be captured by the anatomical data. Such investigations reveal if and maybe also how ALFF can be altered. If subtypes like indicated above exist, subtype specific treatment and prevention methods could be developed.

Similar implications can be drawn for subjects for which other data modalities are the predominant decisive factor: Left PCC and amygdala showed lower ReHO in MDD, i.e., spontaneous activation is locally badly synchronized, which may affect not the only regional functioning of this brain area, but also the connectivity with other brain areas. For both PCC, the central node of the default mode network, and left amygdala, which has significant functional connectivity with the ventral striatum, this may have a significant impact on behaviour. Increased ReHo was found in the cerebellum, which during the last years has been shown to participate in emotion regulation, inhibition of impulsive decision making, attention, and working memory [3]. This increase might be a result of the cerebellar cortical connections known to be disrupted in depression subjects [26]. However, in line with the above, whether these ReHO alterations are entirely functional or cause to subtle anatomical changes, such as mentioned above, remains to be investigated.

For FC, the merit of sparse discriminative feature selection becomes especially apparent. So far, predictions are mainly based on SVM procedures [25], yielding models based on more than a hundred relevant connections [37], despite preselection procedures. Here, two connections were selected as discriminative, one that is located within the ventral DMN (vDMN) involving right PCC/RSC and parahippocampus, and one in the dorsal DMN (dDMN), concerning the PCC/Precuneus and medial prefrontal cortex(mPFC)/anterior cingulate (ACC)/orbitofrontal cortex (OFC), respectively (Appendix Table A5). Both connections were weaker in MDD subjects. Both connections involve part of the PCC, which has been shown in our data to exhibit altered ReHO and ALFF characteristics in MDD subjects and is thus subject to further investigation on whether and how these phenomena are related. If strongly correlated, treatment of one deficiency could be sufficient to correct several of them.

The importance of BDNF in MDD is evidenced not only by its supportive role in serotonin signaling [5, 16] and in the dopaminergic system [14, 17], but also by its impact on neurogenesis, neural differentiation and cell survival, and thus on formation, stabilization and continuity of long-term memory [30]. Its effects on white matter density are obvious. Indeed, Choi *et al.* [7] have shown an inverse association of BDNF DNA methylation and reduced white matter integrity in the anterior corona radiata in major depression. Here again, these WM changes might have been too subtle to detect in the MRI images of our cohort, but might already be reflected in the DNA methylation characteristics, the major mechanism for neural plasticity [2]. This is supported by the fact that for controls, WM (which is denser in controls) was more often involved in the affirmation of the diagnosis than for MDD subjects.

Finally, examining all biomarkers identified in this study (Appendix Table A6), we can outline their relations to each other. The cerebellum indicated as distinctive in the ReHO data, targets prefrontal as well as temporal cortices through the thalamus. This is also reflected in the functional connectivity of the cerebellum with these regions [4]. While prefrontal and temporal areas were discriminative in the WM, the thalamus was distinctive in the ALFF data. Further, the middle OFC, also selective in the ReHO data, has strong connections with the hippocampus and associated areas of the cingulate and RSC (identified in the FC data), as well as with the anterior thalamus (ALFF) [11].

Considering the above, the aspects found in ALFF, ReHO, FC and WM density each seem to depict potential deficiencies in MDD within a common functional and anatomical network that connects the limbic system and cortical areas. This hypothesis is supported by Chen *et al.* [6], who found that the frontal-striatal-thalamic pathways are affected in MDD. A more recent study found clusters of functional connections in the above mentioned areas to lie at the core of four anxiety- and anhedonia-related subtypes of depression [10]. The well known symptom variability in MDD can be assumed to result from these equally variable network deficiencies. While they cannot always be picked up by the data modality one might expect (e.g., FC), they might be reflected in mechanism-related characteristics of other data modalities (here, ALFF, ReHo, WM or BDNF methylation).

We remark that this study is limited by the number of subjects that were available for each classification model. The fitted distribution functions may therefore not reflect the true distribution of positive and negative odds ratio. The higher the number of subjects, the more accurate the true to all predictions ratio, and thus credibility of the newly calculated diagnosis. When more accurate models are used, the biomarker found decisive as for the final diagnosis for each subject in this study, might therefore be a different one.

Note, that if a posterior probability for the decision could be obtained, tuning of the decision threshold would be possible and therefore a risk assessment for MDD. Unfortunately, we cannot benefit from this fact due to the use of the indicator function in MCV. However, per construction of MCV, the risk of a false MDD prediction, positive and negative, is lower than that of any of the predictions obtained from the single models.

5 Conclusion

We reiterate that in MCV, identifying the most reliable diagnostic dataset for a subject equates to finding the dataset with the most pronounced MDD or HC characteristics for a given subject. In other words, MCV pinpoints which of the biological factors in the variable network is dysfunctional (in a typical way). At the same time, it is just that consideration of variability, i.e., consideration of potential MDD subtypes, that results in higher prediction accuracy. The derived effective variables are limited by the small number of subjects and need to be reconfirmed on a lager data sample. However, the proposed MCV method itself is by construction robust and flexible, so that we are confident that it will allow simple to use and accurate MDD diagnosis in clinical settings. Its transparency with respect to the “weakest link” aids the identification of MDD subtypes and consequently the development of adequate medical treatment. As cheaper and more accessible neural and physiological markers become available, this method will naturally become an increasingly useful clinical tool. Finally, we wish to underscore the fact that MCV is not restricted to MDD diagnosis alone, but is widely applicable to situations where decisions have to be made based on multiple predictions from models with low accuracy.

Financial Support

This research was supported by the Strategic Research Program for Brain Sciences, Japan Agency for Medical Research and Development (AMED), JP15dm0107012, JP18dm0107093, JP18dm0107096, JP19dm020700, JP21dm0307009 and the Japan society for Promotion of Science (JSPS) KAKENHI JP16H06563. Thanks to Dr. Steven D. Aird, technical editor of the Okinawa Institute of Science and Technology, for thorough proofreading.

Ethical Standards

This study was approved by Research Ethics Committee of the Okinawa Institute of Science and Technology and the Research Ethics Committee of Hiroshima University (permission nr.172). All methods were performed in accordance with the relevant guidelines and regulations.

Due to potentially identifying information, data for this study are restricted by the Ethical Committee for Epidemiology of Hiroshima University, Japan. Interested, qualified researchers may request the data by contacting Dr. Shoji Karatsu (kasumi-kenkyu@office.hiroshimau.ac.jp).

Biographies

Yu Shimizu received a Ph.D. degree in applied mathematics from the University of Vienna in 2003. After working as a researcher at Kyoto University, Advanced Telecommunications Research International and finally at the Okinawa Institute of Science and Technology (OIST), she is now product director at Aikomi Ltd (https://aikomi.co.jp) and visiting researcher at OIST. Her main interests are machine learning, brain imaging and data analysis in mental disease.

Junichiro Yoshimoto received a Ph.D. in 2002 from the Nara Institute of Science and Technology (NAIST). After working as a postdoctoral fellow at the Japan Science and Technology Agency, he was a researcher at the Okinawa Institute of Science and Technology in 2004 and promoted to a group leader in 2010. Since 2015, he has been an associated professor at NAIST. His current research interests are machine learning, neuroinformatics, and biomedical data science.

Kenji Doya is a Professor at the Neural Computation Unit, Okinawa Institute of Science and Technology Graduate University. He took Ph.D. in 1991 at U. Tokyo. After postdoctoral training in U. C. San Diego and Salk Institute, he joined Advanced Telecommunications Research International (ATR) in 1994 and became the head of Computational Neurobiology Department in 2003. In 2004, he was appointed as the Principal Investigator of Neural Computation Unit, Okinawa Institute of Science and Technology (OIST) and served as the Vice Provost for Research as OIST established itself as a graduate university in 2011. He serves as the Co-Editor in Chief of Neural Networks since 2008 and received Donald O. Hebb Award in 2018. Contact: doya@oist.jp.

Appendix

CpGs used for classification: CpG1, CpG4, CpG5, CpG7, CpG8.9, CpG14, CpG15, CpG17, CpG18, CpG19.20.21, CpG22, CpG23, CpG24, CpG25.26.27, CpG28, CpG29.30.31, CpG32, CpG33.34, CpG36, CpG37, CpG47, CpG48, CpG50.51, CpG52, CpG59, CpG61, CpG63, CpG72.73, CpG74.75, CpG77, CpG78, CpG79.

Table A1: Average framewise translational and rotational displacement during rsfMRI scan in mm and degrees, respectively.

	HC	MDD	p-value
$\triangle trans_x$	0.0069 ± 0.0028	0.0061 ± 0.0032	0.21
$\triangle trans_y$	0.0381 ± 0.0466	0.0423 ± 0.0417	0.67
$\triangle trans_z$	0.0364 ± 0.0560	0.0255 ± 0.0161	0.22
$\triangle rot_x$	3.63e-04 ± 2.48e-04	3.34e-04 ± 2.13e-04	0.56
$\triangle rot_y$	1.86e-04 ± 1.47e-04	1.65e-04 ± 8.29e-05	0.41
$\triangle rot_z$	1.42e-04 ± 5.45e-05	1.36e-04 ± 6.05e-05	0.62
Framewise displacement	0.08 ± 0.10	0.07 ± 0.05	0.68

Table A2: Demographic and clinical characteristics of subjects included in WM and GM model.

	MDD	Control	p-value
Number of subjects	60	60	–
Sex (male/female)	29/31	32/28	0.5738
Age (years)	40.57 ± 8.99	36.95 ± 12.6657	0.0738
IQ	108.52 ± 9.7	1113.34 ± 7.4	$0.0028^{**\mathrm{a}}$
Alcohol Dependent Subjects	5	0	0.0224^{*}
BDI2	30.15 ± 8.79	6.63 ± 6.03	1.08e-33***
PHQ9	17.65 ± 4.41	3.13 ± 3.80	2.38e-38***
SHAPS	37.12 ± 5.48	23.70 ± 6.21	1.92e-23***
STAI	56.33 ± 7.74	40.6 ± 8.88	2.97e-18***
CATS	34.69 ± 23.41	24.47 ± 14.86	0.0060^{**}
LES	-6.64 ± 6.43	0.42 ± 3.37	2.48e-11***
HRSD17	19.95 ± 4.96	–	–
Age of Depression Onset (years)	38.06 ± 10.83	–	–
Number of Previous Episodes	0.51 ± 0.6	–	–
Length of Current Episode (days)	164.68 ± 212.01	–	–
Lexapro single agent	50	–	–
Lexapro combination	2	–	–
Other single agent	2	–	–
No Treatment	6	–	–

Note: ^{a}asterisks denote significant group differences, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table A3: Demographic and clinical characteristics of subjects included in ALFF, REHO and FC model.

	MDD	Control	p-value
Number of subjects	42	42	–
Sex (male/female)	23/19	25/17	0.6592
Age (years)	36.24 ± 6.04	38.36 ± 13.22	0.3476
IQ	108.65 ± 10.84	112.39 ± 7.40	0.0686
Alcohol Dependent Subjects	4	0	0.0404*
BDI 2	30.17 ± 9.25	7.64 ± 6.35	1.21e-21***
PHQ9	17.88 ± 4.40	3.83 ± 4.08	1.60e-25***
SHAPS	37.71 ± 5.21	25.05 ± 6.11	2.73e-16***
STAI	56.02 ± 7.93	41.26 ± 8.69	3.84e-12***
CATS	37.23 ± 25.32	27.48 ± 15.90	0.0390*
LES	-6.275 ± 6.71	0.02 ± 3.58	8.44e-07***
HRSD17	19.55 ± 5.076	–	–
Age of Depression Onset (years)	32.83 ± 8.36	–	–
Number of Previous Episodes	0.56 ± 0.63	–	–
Length of Current Episode (days)	148.90 ± 195.61	–	–
Lexapro single agent	35	–	–
Lexapro combination	2	–	–
Other single agent	2	–	–
No Treatment	3	–	–

Table A4: Demographic and clinical characteristics of subjects included in the methylation model.

	MDD	Control	p-value
Number of subjects	33	33	–
Sex (male/female)	15/18	21/12	0.1380
Age (years)	40.36 ± 10.26	36.42 ± 12.76	0.1716
IQ	108.15 ± 9.46	111.42 ± 9.79	0.1721
Alcohol Dependent Subjects	1	0	0.3136
BDI	30.76 ± 9.76	8.18 ± 6.26	1.05e-16***
PHQ9	18.30 ± 4.86	4.18 ± 3.73	5.28e-20***
SHAPS	36.97 ± 4.44	23.33 ± 6.22	4.01e-15***
STAI	56.54 ± 6.60	42.61 ± 9.14	1.22e-09***
CATS	34.17 ± 19.56	29.15 ± 16.69	0.2767
LES	-6.07 ± 6.50	0.303 ± 4.61	2.92e-05***
HRSD17	20.36 ± 4.97	–	
Age of Depression Onset (years)	37.29 ± 11.96	–	
Number of Previous Episodes	0.52 ± 0.51	–	
Length of Current Episode (days)	162.45 ± 196.34	–	
Lexapro single agent	27	–	–
Lexapro combination	0	–	–
Other single agent	2	–	–
No Treatment	4	–	–

Table A5: Classification Accuracy for each individual data modality.

Data	HC/MDD	Algorithm	Specificity	Sensitivity	Accuracy
WM	66/66	gLASSO	58 ± 2	68 ± 4	63 ± 2
		sLASSO	<60	<60	<60
		Elastic Net	<60	<60	<60
GM	66/66	gLASSO	<60	<60	<60
		sLASSO	<60	<60	<60
		Elastic Net	<60	<60	<60
ALFF	42/42	gLASSO	64 ± 4	66 ± 4	65 ± 2
		sLASSO	66 ± 3	69 ± 2	68 ± 2
		Elastic Net	67 ± 2	69 ± 1	68 ± 1
fALFF	42/42	gLASSO	<60	<60	<60
		sLASSO	<60	<60	<60
		Elastic Net	<60	<60	<60
ReHo	42/42	gLASSO	<60	<60	<60
		sLASSO	65 ± 4	67 ± 6	66 ± 3
		Elastic net	65 ± 2	67 ± 6	66 ± 2
FC	42/42	sLASSO	65 ± 5	66 ± 6	65 ± 4
		Elastic Net	<60	<60	<60
BDNFexon1	33/33	sLASSO	81 ± 2	74 ± 4	77 ± 2
		Elastic Net	84 ± 5	73 ± 3	78 ± 3

Note: Each MRI dataset was subjected to group LASSO, L1 LASSO and Elastic Net. All models were evaluated based on ten times repeated 10-fold cross validation.

Table A6: Diagnostic features, individual datasets: Brain areas and methylation sites, which had discriminative capability in more than 80% of the crossvalidated classification models (Areas as defined in [29]). Negatively weighted features exhibit larger values in healthy controls than depression subjects. Positively weighted features show larger values in depression subjects.

Data (HC/MDD)	Accuracy (Algorithm)	Negative weights	Positive weights
WM (66/66)	63 ± 2 (gLASSO)	Post Central Gyrus L/R Frontal Superior Cortex L Middle Temporal Cortex R	Middle Temporal Cortex L
ALFF (42/42)	68 ± 1 (Elastic Net)	Posterior Cingulum L/R Thalamus L	
ReHo (42/42)	66 ± 2 (Elastic Net)	Posterior Cingulum L Middle Frontal Orbitalis L Amygdala L	Cerebelum8 L/R
FC (42/42)	65 ± 4 (sLASSO)	Ventral DMN 8 - Ventral DMN 5[a] Dorsal DMN 4 - Dorsal DMN 1[b]	
BDNFexon1 (33/33)	78 ± 3 (Elastic Net)	CpG33.34 CpG24 CpG1 CpG63 CpG77 CpG52 CpG61 CpG19.20.21 CpG18 CpG25.26.27 CpG29.30.31 CpG32	CpG8.9 CpG14 CpG5 CpG37 CpG48 CpG78 CpG36 CpG22 CpG74.75 CpG17 CpG15

Note: [a] Ventral DMN:
Area 5: Right Retrosplenial Cortex, Posterior Cingulate Cortex (BA 20, 23),
Area 8: Right Parahippocampal Gyrus (BA 37,30)
[b] Dorsal DMN:
Area 1: Medial Prefrontal Cortex, Anterior Cingulate Cortex, Orbitofrontal Cortex (BA 9,10,24,32,11)
Area 4 : Posterior Cingulate, Precuneus (BA 23, 30)

Table A7: MCV for rsfMRI data: Prediction Rate based on individual datasets and for MCV combinations. If not indicated otherwise p-values indicate performance difference with respect to MCV(rsfMRI).

HC/MDD = 42/42	Specificity	Sensitivity	Accuracy	*p*-value	Credibility
ALFF	67 ± 2	69 ± 1	68 ± 1	3.6e-05***	71 ± 2
ReHo	65 ± 2	67 ± 5	66 ± 2	3.1e-06***	61 ± 4
FC	65 ± 5	66 ± 6	65 ± 4	6.3e-05***	55 ± 3
MCV (ALFF, ReHo)	65 ± 3	77 ± 2	71 ± 2	2.7e-04*** wrt ALFF 1.3e-05*** wrt ReHo 0.716	73 ± 2
MCV (ALFF, FC)	67 ± 3	77 ± 3	72 ± 2	1.2e-04*** wrt ALFF 1.3e-04*** wrt FC 0.6	72 ± 2
MCV (ReHo, FC)	64 ± 3	73 ± 7	68 ± 4	0.104 wrt ReHo 0.096 wrt FC 0.007**	64 ± 3
MCV (rsfMRI)	66 ± 4	79 ± 3	73 ± 2	–	74 ± 2

Table A8: MCV for rsfMRI and WM: Prediction Rate based on individual datasets and for MCV combinations. If not indicated otherwise p-values indicate performance difference with respect to MCV(rsfMRI,WM).

HC/MDD = 39/41	Specificity	Sensitivity	Accuracy (F-score)	p-value	Credibility
ALFF	67 ± 1	64 ± 2	70 ± 2 (69 ± 1)	1.5e-07***	70 ± 2
ReHo	63 ± 3	66 ± 5	64 ± 2 (66 ± 3)	2.4e-06***	61 ± 4
FC	65 ± 6	66 ± 6	66 ± 4 (67 ± 5)	0.002**	55 ± 3
WM	52 ± 3	70 ± 6	61 ± 3 (65 ± 4)	5.7e-09***	62 ± 2
MCV(rsfMRI)	64 ± 4	79 ± 3	71 ± 2 (74 ± 2)	0.033*	73 ± 2
MCV (rsfMRI, WM)	69 ± 4	78 ± 3	74 ± 2 (75 ± 0)	–	75 ± 2

Table A9: MCV for rsfMRI and BDNF methylation: Prediction Rate based on individual datasets and for MCV combinations. If not indicated otherwise p-values rate performance difference with respect to MCV(rsfMRI, BDNFexon1)

HC/MDD = 25/21	Specificity	Sensitivity	Accuracy (F-score)	p-values	Credibility
ALFF	63 ± 4	66 ± 2 (66 ± 2)	5.1e-07***	71 ± 2	
ReHo	67 ± 3	68 ± 7	67 ± 3 (65 ± 4)	3.2e-06***	61 ± 4
FC	60 ± 7	65 ± 10	62 ± 4 (61 ± 6)	1.6e-07***	54 ± 3
BDNFexon1	82 ± 5	65 ± 3	70 ± 4 (70 ± 3)	1.4e-08***	74 ± 3
MCV (rsfMRI)	59 ± 5	76 ± 3	67 ± 3 (68 ± 2)	5.0e-06***	74 ± 2
MCV (rsfMRI, BDNFexon1)	83 ± 4	70 ± 7	77 ± 4 (74 ± 5)	–	82 ± 2

Table A10: MCV for WM and BDNF methylation: Prediction Rate for individual datasets and for MCV combinations. If not indicated otherwise p-values indicate performance difference with respect to MCV.

HC/MDD = 28/31	Specificity	Sensitivity	Accuracy (F-score)	p-values	Credibility
WM	53 ± 4	70 ± 6	62 ± 3 (66 ± 4)	1.6e-12***	61 ± 2
BDNFexon1	81 ± 6	75 ± 2	78 ± 3 (78 ± 3)	0.005**	72 ± 4
MCV(WM, BDNFexon1)	89 ± 5	73 ± 5	80 ± 4 (79 ± 4)	–	78 ± 3

Table A11: MCV vs straight forward integrated prediction methods for data with missing values. For all subjects at least two different, but not all measurements were available. p-values indicate performance difference with respect to MCV.

HC/MDD = 47/53	Specificity	Sensitivity	Accuracy (F-score)	p-value
Most Votes	61 ± 2	67 ± 4	64 ± 3 (78 ± 3)	7.1e-10***
Sum of odds ratio	69 ± 3	76 ± 4	72 ± 3 (75 ± 3)	1.9e-05***
Max of oxdds ratio	69 ± 2	72 ± 5	70 ± 2 (72 ± 3)	2.5e-07***
Mean of odds ratio	69 ± 2	76 ± 4	73 ± 3 (75 ± 3)	2.8e-05***
MCV	84 ± 4	77 ± 5	80 ± 3 (81 ± 3)	–

References

[1] J. A. Bilello, L. M. Thurmond, K. M. Smith, B. Pi, R. Rubin, S. M. Wright, F. Taub, M. E. Henry, R. C. Shelton, and G. I. Papakostas, "MDDScore: Confirmation of a Blood Test to Aid in the Diagnosis of Major Depressive Disorder," *The Journal of Clinical Psychiatry*, 76(2), 2015, 99–206.

[2] E. Borrelli, E. J. Nestler, C. D. Allis, and P. Sassone-Corsi, "Decoding the Epigenetic Language of Neuronal Plasticity," *Neuron*, 60(6), 2008, 961–74.

[3] R. L. Buckner, "The Cerebellum and Cognitive Function: 25 Years of Insight from Anatomy and Neuroimaging," *Neuron*, 80(3), 2013.

[4] R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. T. Yeo, "The Organization of the Human Cerebellum Estimated by Intrinsic Functional Connectivity," *Journal of Neurophysiology*, 106(5), 2011, 2322–45.

[5] F. Calabrese, G. Guidotti, A. Middelman, G. Racagni, J. Homberg, and M. Riva, "Lack of Serotonin Transporter Alters BDNF Expression in the Rat Brain During Early Postnatal Development," *Molecular Neurobiology*, 48(1), 2013, 244–56.

[6] G. Chen, X. Hu, L. Li, X. Huang, S. Lui, and W. Kuang, "Disorganization of White Matter Architecture in Major Depressive Disorder: A Meta-analysis of Diffusion Tensor Imaging with Tract-based Spatial Statistics," *Scientific Reports*, 6, 2016 Feb, 21825.

[7] S. Choi, K. M. Han, E. Won, B. J. Yoon, M. S. Lee, and B. J. Ham, "Association of Brain-derived Neurotrophic Factor DNA Methylation and Reduced White Matter Integrity in the Anterior Corona Radiata In Major Depression," *Journal of Affective Disorders*, 172, 2015, 74–80.

[8] T. W. Chow and J. L. Cummings, *The Human Frontal Lobes: Functions and Disorders*, Vol. Chapter 3, The Guilford Press, 2007.

[9] "Diagnostic and Statistical Manual of Mental Disorders (DSM)," *American Psychiatric Publishing*, 2013.

[10] A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin, A. F. Schatzberg, K. Sudheimer, J. Keller, H. S. Mayberg, F. M. Gunning, G. S. Alexopoulos, M. D. Fox, A. Pascual-Leone, H. U. Voss, B. J. Casey, M. J. Dubin, and C. Liston, "Resting-State Connectivity Biomarkers Define Neurophysiological Subtypes of Depression," *Nature Medicine*, 23(1), 2017, 28–38.

[11] R. Elliott, R. Dolan, and C. D. Frith, "Dissociable Functions in the Medial and Lateral Orbitofrontal Cortex: Evidence from Human Neuroimaging Studies," *Cerebral Cortex*, 10(3), 2015, 308–17.

[12] M. Fuchikami, S. Morinobu, M. Segawa, Y. Okamoto, S. Yamawaki, N. Ozaki, T. Inoue, I. Kusumi, T. Koyama, K. Tsuchiyama, and T. Terao, “DNA Methylation Profiles of the Brain-Derived Neurotrophic Factor (BDNF) Gene as a Potent Diagnostic Biomarker in Major Depression,” *PLoS ONE*, 6(8), 2011, e23881.

[13] S. Gao, V. D. Calhoun, and J. Sui, “Machine Learning in Major Depression: From Classification to Treatment Outcome Prediction,” *CNS Neuroscience & Therapeutics*, 24(11), 2018, 1037–52.

[14] O. Guillin, J. Diaz, P. Carroll, N. Griffon, J. Schwartz, and P. Sokoloff, “BDNF Controls Dopamine D3 Receptor Expression and Triggers Behavioural Sensitization,” *Nature*, 411, 2001, 86–9.

[15] T. Hahn, A. Marquand, A. Ehlis, T. Dresler, S. Kittel-Schneider, T. A. Jarczok, K.-P. Lesch, P. M. Jakob, J. Mourão-Miranda, M. Brammer, and A. J. Fallgatter, “Integrating Neurobiological Markers of Depression,” *Archives of General Psychiatry*, 68(4), 2011, 361–8.

[16] J. Homberg, R. Molteni, F. Calabrese, and M. Riva, “The Serotonin-BDNF Duo: Developmental Implications for the Vulnerability to Psychopathology,” *Neuroscience & Biobehavioral Reviews*, 43, 2014, 35–47.

[17] C. Hyman, M. Hofer, Y. A. Barde, M. Juhasz, G. D. Yancopoulos, S. P. Squinto, and R. M. Lindsay, “BDNF is a Neurotrophic Factor for Dopaminergic Neurons of the Substantia Nigra,” *Nature*, 350(6315), 1991, 230–2.

[18] K. Kroenke, R. L. Spitzer, and J. B. Williams, “The PHQ-9: Validity of a Brief Depression Severity Measure,” *Journal of General Internal Medicine*, 16(9), 2001, 606–13.

[19] S. Kurita, Y. Takei, Y. Maki, S. Hattori, T. Uehara, M. Fukuda, and M. Mikuni, “Magnetoencephalography Study of the Effect of Attention Modulation on Somatosensory Processing in Patients with Major Depressive Disorder,” *Psychiatry and Clinical Neurosciences*, 70(2), 2016, 116–25.

[20] Y. Lu, H. Liang, D. Han, Y. Mo, Z. Li, Y. Cheng, X. Xu, Z. Shen, C. Tan, W. Zhao, Y. Zhu, and X. Sun, “The Volumetric and Shape Changes of the Putamen and Thalamus in First Episode, Untreated Major Depressive Disorder,” *Neuroimage Clinical*, 11, 2016, 658–66.

[21] K. Matsuoka, M. Uno, K. Kasai, K. Koyama, and Y. Kim, “Estimation of Premorbid IQ in Individuals with Alzheimer’s Disease using Japanese Ideographic Script (Kanji) Compound Words: Japanese version of National Adult Reading Test,” *Psychiatry and Clinical Neurosciences*, 60(3), 2006, 332–9.

[22] “Neuropsychiatric Disorders Lab at Stanford University,” http://findlab.stanford.edu/functional_ROIs.html.

[23] T. Otsubo, K. Tanaka, R. Koda, J. Shinoda, N. Sano, S. Tanaka, H. Aoyama, M. Mimura, and K. Kamijima, "Reliability and validity of Japanese version of the Mini-International Neuropsychiatric Interview," *Psychiatry and Clinical Neurosciences*, 59(5), 2005, 517–26.

[24] G. I. Papakostas, R. C. Shelton, G. Kinrys, M. E. Henry, B. R. Bakow, S. H. Lipkin, B. Pi, L. Thurmond, and J. A. Bilello, "Assessment of a Multi-assay, Serum-based Biological Diagnostic Test for Major Depressive Disorder: A Pilot and Replication Study," *Molecular Psychiatry*, 18(3), 2013, 332–9.

[25] M. J. Patel, A. Khalaf, and H. J. Aizenstein, "Studying Depression Using Imaging and Machine Learning Methods," *NeuroImage: Clinical*, 10, 2016, 115–23.

[26] J. R. Phillips, D. H. Hewedi, A. M. Eissa, and A. A. Moustafa, "The Cerebellum and Psychiatric Disorders," *Front Public Health*, 3(66), 2015.

[27] D. Sheehan, Y. Lecrubier, and K. Sheehan, "The Mini-International Neuropsychiatric Interview (MINI): The Development and Validation of a Structured Diagnostic Psychiatric Interview for DSM-IV and ICD-10," *The Journal of Clinical Psychiatry*, 50(suppl 20), 1998, 22–33.

[28] Y. Shimizu, J. Yoshimoto, S. Toki, M. Takamura, S. Yoshimura, Y. Okamoto, S. Yamawaki, and K. Doya, "Toward Probabilistic Diagnosis and Understanding of Depression Based on Functional MRI Data Analysis with Logistic Group LASSO," *PLoS ONE*, 10(5), 2015, e0123524.

[29] W. Shirer, S. Ryali, E. Rykhlevskaiam, V. Menon, and M. Greicius, "Decoding Subject-driven Cognitive States with Whole-brain Connectivity Patterns," *Cerebral Cortex*, 22(1), 2011, 158–65.

[30] K. K. Singh, K. J. Park, E. J. Hong, B. M. Kramer, M. E. Greenberg, D. R. Kaplan, and F. D. Miller, "Developmental Axon Pruning Mediated by BDNF-p75NTR-dependent Axon Degeneration," *Nature Neuroscience*, 11(6), 2008, 649–58.

[31] X. W. Song, Z. Y. Dong, X. Y. Long, S. F. Li, X. N. Zuo, C.-Z. Zhu, Y. He, C.-G. Yan, and Y.-F. Zang, "REST: A Toolkit for Resting-State Functional Magnetic Resonance Imaging Data Processing," *PLoS ONE*, 6(9), 2011, e25031.

[32] A. U. Turken and N. F. Dronkers, "The Neural Architecture of the Language Comprehension Network: Converging Evidence from Lesion and Connectivity Analyses," *Frontiers in Systems Neuroscience*, 5, 2011.

[33] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain," *NeuroImage*, 15(1), 2002, 273–89.

[34] G. Venkatasubramanian and M. S. Keshavan, "Biomarkers in Psychiatry – A Critique," 23(3–5), 2016.

[35] E. Won, S. Choi, J. Kang, A. Kim, K.-M. Han, H. S. Chang, W. S. Tae, K. R. Son, S.-H. Joe, M.-S. Lee, and B.-J. Ham, "Association between Reduced White Matter Integrity in the Corpus Callosum and Serotonin Transporter Gene DNA Methylation in Medication-naive Patients with Major Depressive Disorder," *Translational Psychiatry*, 6(8), 2016, e866.

[36] J. J. Young, T. Silber, D. Bruno, I. R. Galatzer-Levy, N. Pomara, and C. R. Marmar, "Is there Progress? An Overview of Selecting Biomarker Candidates for Major Depressive Disorder," *Frontiers in Psychiatry*, 7(72), 2016.

[37] X. Zhu, X. Wang, J. Xiao, J. Liao, M. Zhong, W. Wang, and S. Yao, "Evidence of a Dissociation Pattern in Resting-State Default Mode Network Connectivity in First-Episode, Treatment-Naive Major Depression Patients," *Biological Psychiatry*, 71(7), 2012, 611–7.

[38] X.-N. Zuo, A. Di Martino, C. Kelly, Z. E. Shehzad, D. G. Gee, D. F. Klein, F. X. Castellanos, B. B. Biswal, and M. P. Milham, "The Oscillating Brain: Complex and Reliable," *NeuroImage*, 49(2), 2010, 1432–45.